Transfer learning for abusive language detection

Tulika Bose

UNIVERSITÉ DE LORRAINE

Thèse de Doctorat

Tulika Bose

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Lorraine
Mention Informatique

École doctorale : IAEM

Unité de recherche : Laboratoire Lorrain de Recherche en Informatique et ses Applications UMR 7503

Soutenue le 30 Janvier 2023
Thèse N°:

Transfer Learning for Abusive Language Detection

JURY

Directrice de thèse :	**Irina ILLINA**, HDR, Maître de conférence, Université de Lorraine, LORIA-INRIA, France
Co-directeur de thèse :	**Dominique FOHR**, Chargé de Recherche, CNRS, LORIA-INRIA, France
Rapporteur :	**Björn GAMBÄCK**, Professeur, Université Norvégienne des Sciences et Technologies, Norvège
Rapporteuse :	**Veronique HOSTE**, Professeur, Université de Ghent, Belgique
Examinateur : (et président du jury)	**Benjamin LECOUTEUX**, Professeur, Université Grenoble Alpes, France
Examinateur :	**Benoit FAVRE**, Professeur, Aix-Marseille Université, France
Examinatrice :	**Claire GARDENT**, Directrice de Recherche, CNRS, LORIA-INRIA, France
Examinatrice :	**Angeliki MONNIER**, Professeur, Université de Lorraine, CREM, France

Résumé

La prolifération des médias sociaux, malgré ses nombreux avantages, a entraîné une augmentation des propos injurieux. Ces propos, qui sont généralement blessants, toxiques ou empreints de préjugés à l'encontre d'individus ou de groupes, doivent être détectés et modérés rapidement par les plateformes en ligne. Les modèles d'apprentissage profond pour la détection de propos abusifs ont montré des niveaux de performance élevé quand ils sont évalués sur des données similaires à celles qui ont servi à entraîner les modèles, mais sont nettement moins performants s'ils sont évalués sur des données dont la distribution est différente. En outre, ils nécessitent une quantité considérable de données étiquetées coûteuses pour l'apprentissage. C'est pour cela qu'il est intéressant d'étudier le transfert efficace de connaissances à partir de corpus annotés existants de propos abusifs. Cette thèse étudie le problème de l'apprentissage par transfert pour la détection de propos abusifs et explore diverses solutions pour améliorer le transfert de connaissances dans des scénarios inter corpus.

Tout d'abord, nous analysons la généralisabilité inter-corpus des modules de détection de propos abusifs sans accéder à des données cibles pendant le processus d'apprentissage. Nous examinons si la combinaison des représentations issues du thème (topic) avec des représentations contextuelles peut améliorer la généralisabilité. Nous montrons que l'association de commentaires du corpus cible avec des thèmes du corpus d'entraînement peut fournir des informations complémentaires pour un meilleur transfert inter-corpus.

Ensuite, nous explorons l'adaptation au domaine non supervisée (UDA, Unsupervised Domain Adaptation), un type d'apprentissage par transfert transductif, avec accès au corpus cible non étiqueté. Nous explorons certaines approches UDA populaires dans la classification des sentiments pour la détection de propos abusifs dans le cadre de corpus croisés. Nous adaptons ensuite une variante du modèle BERT au corpus cible non étiqueté en utilisant la technique du modèle de langue avec masques (MLM, Masked Language Model). Alors que cette dernière améliore les performances inter-corpus, les autres approches UDA ont des performances sous-optimales. Notre analyse révèle leurs limites et souligne le besoin de méthodes d'adaptation efficaces pour cette tâche.

Comme troisième contribution, nous proposons deux approches d'adaptation au domaine utilisant les attributions de caractéristiques (feature attributions), qui sont des explications a posteriori du modèle. En particulier, nous étudions le problème des corrélations erronées (spurious correlations) spécifiques à un corpus qui limitent la généralisation pour la détection des discours de haine, un sous-ensemble des propos abusifs. Alors que les approches de la littérature reposent sur une liste de termes établie manuellement, nous extrayons et pénalisons automatiquement les termes qui causent des corrélations erronées. Nos approches dynamiques améliorent les performances dans le cas de corpus

croisés par rapport aux travaux précédents, à la fois indépendamment et en combinaison avec des dictionnaires prédéfinis.

Enfin, nous considérons le transfert de connaissances d'un domaine source avec beaucoup de données étiquetées vers un domaine cible, où peu d'instances étiquetées sont disponibles. Nous proposons une nouvelle stratégie d'apprentissage, qui permet une modélisation flexible de la proximité relative des voisins récupérés dans le corpus source pour apprendre la quantité de transfert utile. Nous incorporons les informations de voisinage avec une méthode de transport optimal (Optimal Transport) qui exploite la géométrie de l'espace de représentation (embedding space). En alignant les distributions conjointes de l'embedding et des étiquettes du voisinage, nous montrons des améliorations substantielles dans des corpus de discours haineux de taille réduite.

Abstract

The proliferation of social media, despite its multitude of benefits, has led to the increased spread of abusive language. Such language, being typically hurtful, toxic, or prejudiced against individuals or groups, requires timely detection and moderation by online platforms. Deep learning models for detecting abusive language have displayed great levels of in-corpus performance but underperform substantially outside the training distribution. Moreover, they require a considerable amount of expensive labeled data for training. This strongly encourages the effective transfer of knowledge from the existing annotated abusive language resources that may have different distributions to low-resource corpora. This thesis studies the problem of transfer learning for abusive language detection and explores various solutions to improve knowledge transfer in cross-corpus scenarios.

First, we analyze the cross-corpus generalizability of abusive language detection models without accessing the target during training. We investigate if combining topic model representations with contextual representations can improve generalizability. The association of unseen target comments with abusive language topics in the training corpus is shown to provide complementary information for a better cross-corpus transfer.

Secondly, we explore Unsupervised Domain Adaptation (UDA), a type of transductive transfer learning, with access to the unlabeled target corpus. Some popular UDA approaches from sentiment classification are analyzed for cross-corpus abusive language detection. We further adapt a BERT model variant to the unlabeled target using the Masked Language Model (MLM) objective. While the latter improves the cross-corpus performance, the other UDA methods perform sub-optimally. Our analysis reveals their limitations and emphasizes the need for effective adaptation methods suited to this task.

As our third contribution, we propose two DA approaches using feature attributions, which are post-hoc model explanations. Particularly, the problem of spurious corpus-specific correlations is studied that restricts the generalizability of classifiers for detecting hate speech, a sub-category of abusive language. While the prior approaches rely on a manually curated list of terms, we automatically extract and penalize the terms causing spurious correlations. Our dynamic approaches improve the cross-corpus performance over previous works both independently and in combination with pre-defined dictionaries.

Finally, we consider transferring knowledge from a resource-rich source to a low-resource target with fewer labeled instances, across different online platforms. A novel training strategy is proposed, which allows flexible modeling of the relative proximity of neighbors retrieved from the resource-rich corpus to learn the amount of transfer. We incorporate neighborhood information with Optimal Transport that permits exploiting the embedding space geometry. By aligning the joint embedding and label distributions of neighbors, substantial improvements are obtained in low-resource hate speech corpora.

Acknowledgements

My doctoral journey has been the most transformative experience of my life in more ways than one. During these years, I was blessed to know some of the most amazing, supportive, and compassionate people. Everyone I encountered contributed to this journey in their particular ways. I take this opportunity to extend my sincere appreciation to everyone who became an integral part of my doctoral adventure.

First and foremost, I express my heartfelt gratitude to my supervisors – Irina Illina and Dominique Fohr. It would not have been possible for me to achieve this thesis without the overwhelming trust you placed in me and the freedom you gave me to pursue my research ideas. I could not have grown as a researcher without that amount of confidence and freedom. I am perpetually thankful for your guidance, encouragement, and the enormous time you offered me throughout these years from your already busy schedules, whenever I needed. I would also like to extend my earnest gratitude to Nikolaos Aletras for making my research visit to the University of Sheffield comfortable and gratifying during the days of the pandemic. I am immensely grateful for the guidance, advice, and support that you offered me not only during my stay in Sheffield but also after I returned.

I extend my gratitude to Björn Gambäck and Veronique Hoste for investing their valuable time in reviewing my thesis and providing their detailed and enriching feedback. I would like to thank Benjamin Lecouteux, Benoit Favre, Claire Gardent, and Angeliki Monnier for agreeing to be a part of my jury. I truly enjoyed the insightful questions and discussions during my defense that broadened my research perspectives. A special thanks to Christophe Cerisara, who was one of my thesis monitoring committee members, for extending his constant support and encouragement throughout my Ph.D.

I consider myself fortunate to be a part of the MULTISPEECH team of LORIA-INRIA and for being surrounded by extremely talented, friendly, and supportive teammates. Indeed, your support and warmth made the roller coaster ride of the doctoral journey comfortable and worthwhile. I shall always treasure the memories we created together in the lab. A big thanks to all the former and present members of the team – including but not limited to – Denis Jouvet, Slim Ouni, Emmanuel Vincent, Mostafa Sadeghi, Antoine Deleforge, Romain Serizel, Paul Magron, Vincent Colotte, Ashwin D'sa, Michel Olvera, Sandipana Dowerah, Marina Krémé, Marie-Anne Lacroix, Seyed Hosseini, Vinicius Ribeiro, Ali Golmakani, Can Cui, Colleen Beaumard, Emre Canbazer, Nicolas Zampieri, Louis Abel, Louis Delebecque, Joris Cosentino, Ajinkya Kulkarni, Md Sahidullah, Sunit Sivasankaran, Brij Srivastava, Shakeel Ahmad Sheikh, Sewade Ogun, Prerak Srivastava, Sofiane Azzouz, Nasser-Eddine Monir. A special thanks goes to George Chrysostomou from the University of Sheffield for all those fruitful technical conversations during coffee breaks that enriched my stay at Sheffield.

During these years of my doctoral journey, some days were more challenging than others. I express my deepest gratitude to Imran Sheikh for being a true friend and a mentor right from the time I stepped into Nancy and through the toughest times of my Ph.D, for all the long technical discussions that enabled me to develop my scientific rigor and those philosophical discussions that positively influenced my thought process. Thank you for always selflessly being there as my support system at both professional and personal levels.

Alongside my lab mates, I was fortunate enough to earn some more beautiful friendships outside the lab who turned this new country into a comfortable home for me. My heartfelt thanks to Krupali Donda, Ashmita Bhattacharya, Rituraj Kaushik, Priyanka Tyagi, Himanshu Maheshwari, Neelam Rout, Anupama Chingacham, Asma Hasil, Kasturi Raraone, Anil C M for all those wonderful memories that I shall cherish forever.

I could invest the maximum amount of time into my research because all the administrative formalities were timely and efficiently handled on my behalf. I take this opportunity to thank Hélène Cavallini, Delphine Hubert, Souad Boutaguermouchet, Anne-Marie Messaoudi, Aurore Tranchina, Sabrina Ferry-Tritz, and Ferdally Jacobs for making every administrative procedure a smooth ride for me.

I would like to thank the financial support that I received from the french PIA project "Lorraine Université d'Excellence", reference ANR-15-IDEX-04-LUE. A special thanks goes to the OLKi (LUE IMPACT Open Language and Knowledge for Citizens) project and Aurore Coince for all the support that I received. The experiments presented in my thesis were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER, and several Universities as well as other organizations (see https://www.grid5000.fr). My heartfelt gratitude goes to the support staff of the Grid'5000 testbed for their efficient maintenance, without which I could not have finished my experiments on time.

Last but never least, I am incredibly grateful to my family who have constantly been there for me through my uncertainties and celebrations, and who never let me bother about anything else throughout these years other than my doctoral work. I can never thank my parents – K K Bose and Rita Bose – enough for everything they have done for me, and for giving me all the resources, motivation, and courage that made me the person I am today. I couldn't have achieved anything without their unconditional love, support, and faith in me. I am truly grateful to my parents-in-law – Mrityunjay Saha and Gita Saha – for their understanding, love, and support. I am profoundly thankful to my brother Rajat Bose for the encouragement that he always gave me. Finally, I wholeheartedly thank my spouse Mayukh Saha for being an extremely supportive partner and taking care of everything back home while I was away from the country pursuing my thesis.

Warning

This thesis contains content that may be offensive and distressing. It is only intended for a better scientific analysis of the machine learning models used in the experiments for research purposes.

Contents

List of figures

List of tables

List of acronyms

AAD adversarial adaptation with distillation
AAE african-american english
ADASYN adaptive synthetic
ADDA adversarial discriminative domain adaptation
BERT bidirectional encoder representations from transformers
Bi-LSTM bidirectional long short-term memory
BoW bag of words
BPE byte pair encoding
BTM biterm topic model
CBOW continuous bag of words
C-GRU convolutional-gated recurrent unit
CNN convolutional neural network
DAN deep averaging network
DNN deep neural network
DMM dirichlet multinomial mixture
D-Ref dynamic refinement
DeepLIFT/DL deep learning important features
ED embedding distance
ELMo embeddings from language models
FEDA frustatingly easy domain adaptation
FN false negative
FP false positive
GloVe global vectors for word representation
GPT generative pre-trained transformer
GPU graphics processing unit
GRL gradient reversal layer
GRU gated recurrent unit
HATN hierarchical attention transfer network
HBERT HateBERT
IG integrated gradients
JDOT joint distribution optimal transport
KD knowledge distillation

LB-SOINN load-balancing self-organizing incremental neural network
LC label consistency
LDA latent dirichlet allocation
LF-DMM latent feature dirichlet multinomial mixture
LIME local interpretable model-agnostic explanations
LR Logistic regression
LRP layer-wise relevance propagation
LSTM long short-term memory
MAML model-agnostic meta-learning
MLM masked language model
MLP multi-layer perceptron
MTL multi-task learning
NER named entity recognition
NLP natural language processing
NVDM neural variational document model
OOV out-of-vocabulary
OT optimal transport
OT$_u$ unbalanced optimal transport
OT$_u^{NN}$ neighborhood-aware (unbalanced) optimal transport
PBLM pivot based language modeling
PCA principal component analysis
PERL pivot-based encoder representation of language
POS part-of-speech
PTM pseudo-document-based topic modeling
ReLU rectified linear unit
RoBERTa robustly optimized BERT approach
RNN recurrent neural network
ROS random oversampling
SAE standard american english
SATM self-aggregation-based topic modeling
SBERT sentence-BERT
SCL structural correspondence learning
SDA stacked denoising autoencoder
SFA spectral feature alignment
SGNS skip-gram negative-sampling
SMOTE synthetic minority over-sampling technique
SVM support vector machine
TCAV testing concept activation vector

TCT topical coherence transfer
TDLM topically driven neural language model
TF-IDF term frequency-inverse document frequency
TN true negative
TP true positive
UDA unsupervised domain adaptation
URL uniform resource locators
USE universal sentence encoder
U-TDLM universal sentence encoder-based topically driven neural language model
VRL variational representation learning
VDCNN very deep convolutional neural network
WNTM word network topic model

1 Introduction

The phenomenal growth of the Internet in the last decades has radically transformed almost every aspect of our daily lives. According to a recent report[1] on digital usage around the globe, over 5 billion people use the internet worldwide now, up from only 413 million[2] in 2000, which comprise 63% of the world population. The Internet has facilitated easy access to information sources, along with social, cultural, economic, and personal benefits (Van Deursen and Helsper, 2018). In particular, the advent of 'social media' that include blogs, microblogs, wikis, social networking sites, video-sharing sites, and other platforms (Kaplan and Haenlein, 2010; Kane et al., 2014) has brought persuasive changes in the mode of communication as it holds the core promise of largely increased connectivity to the wider world. At present, the number of active social media users is around 4.7 billion comprising 59% of the global population, as per the same report. Social media have provided platforms for self-expression in public and for potentially reaching a large audience by publishing a variety of content. Users may easily connect with others who share their interests based on their own personal profiles and join vibrant communities of interest. Indeed, social media encourage prolonged and productive engagement in many civic spheres and offer a wealth of advantages.

However, the lived experience on these platforms often devalues their potential (Papacharissi, 2004; Jurgens et al., 2019). Particularly, social media have emerged as a fertile ground for abusive language, which is a form of antisocial, hurtful, and aggressive communication from certain sections of people. This has prevented these platforms from providing a safe online environment to their users. Recent polls suggest that around 40% of online users have been victims of online abuse at some point (Duggan, 2017; Jurgens et al., 2019). Abusive language causes comment threads to become poisonous and unproductive as insults exchanged in a vicious loop deter participants who would be ready to contribute positively to the conversation. Hate speech, a kind of abusive language, particularly targets minority communities (Herring et al., 2002; DeAngelis, 2009; Waseem and Hovy, 2016; Jurgens et al., 2019) and spreads prejudiced opinions and stereotypes against them, thus further marginalizing the underrepresented groups. They can even culminate in severe implications like long-lasting trauma (Vidgen et al., 2021a; Hinduja and Patchin, 2019) for the victims and lead to actual incidents of violence (Burnap and Williams, 2014; Alnazzawi, 2022). For instance, the Christchurch mosque shooting in 2019 and other incidents of mass shootings have been linked to the '8chan' website, which is infamous for allowing and promoting racism and other kinds of hateful

[1] https://datareportal.com/reports/digital-2022-july-global-statshot
[2] https://ourworldindata.org/internet

content[3]. Another example is the 'COVID-19' outbreak when a surge in online hate against Asian people led to incidents of physical assaults (Yu et al., 2020).

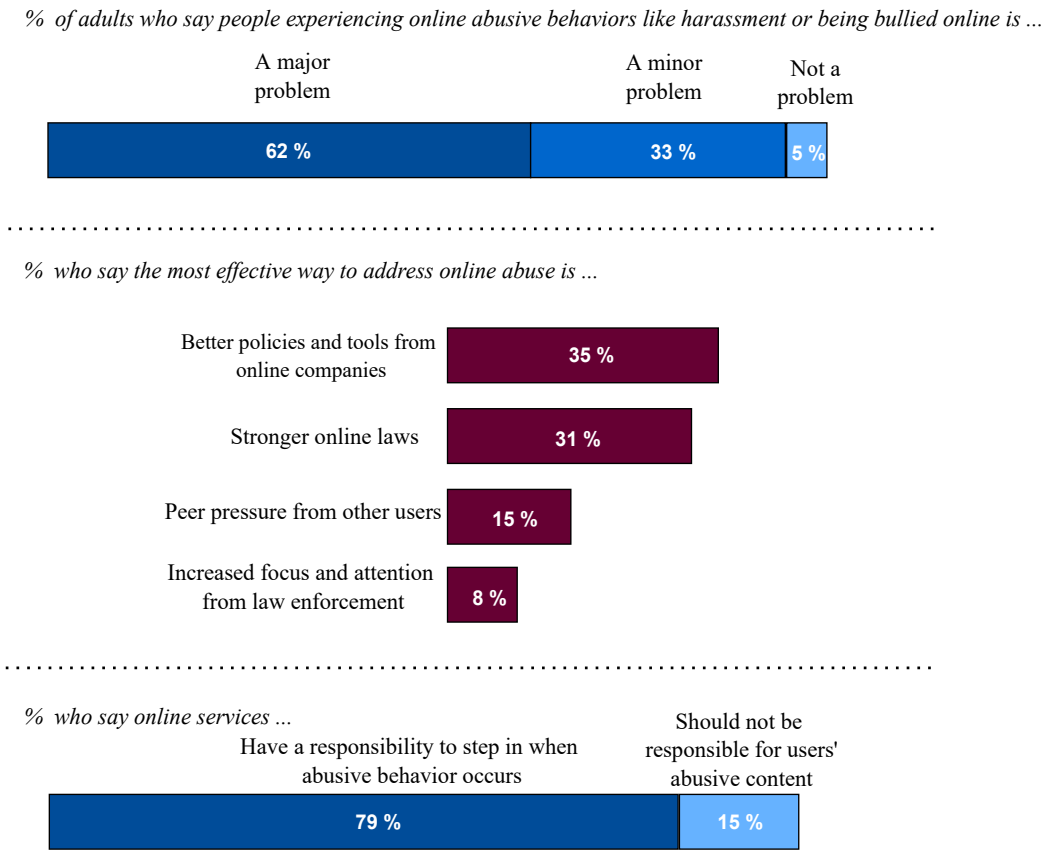


Figure 1.1: Survey on Americans for their perspectives on the responsibility of online companies in addressing abuse on their platforms (Survey conducted Jan. 9-23, 2017). This figure is adapted from Duggan (2017).

There has been an increased focus on the regulation of abusive language on social media platforms in recent years, owing to law enforcement in different countries. As an example, the European Union Commission has directed Twitter, Facebook, YouTube, and Microsoft to sign an online "code of conduct" for fighting hate speech that calls for immediate action of reviewing the content reported by their users, typically within 24 hours[4]. Some platforms have been accused for not doing enough to combat hate speech by removing them in a reasonable time (Kottasová, 2017). This can negatively affect the

[3] https://hackinghate.eu/news/when-online-hate-speech-goes-extreme-the-case-of-hate-crimes/

[4] https://www.theguardian.com/technology/2016/may/31/facebook-youtube-twitter-microsoft-eu-hate-speech-code

quality of service (Oboler and Connelly, 2014) and the reputation of these companies and they may risk high penalties (Chong, 2018). Figure 1.1 presents the report of a survey conducted about public opinion on the accountability of online platforms for regulation of abusive language (Duggan, 2017).

Machine learning systems that are trained on the data from online platforms worsen the issue. These systems themselves become propagators of abusive language if such content is not filtered adequately, thus creating a vicious cycle (Sarwar and Murdock, 2022). For instance, it has been discovered that YouTube encourages abusive language through its recommendation algorithm by simply learning from the abusive interactions amongst online users (Tufekci, 2019). Microsoft created 'Tay', a conversational agent in 2016 that learned from user-generated communications on Twitter. It had to be removed soon afterward because it began generating racist language within a few hours of its release (Schwartz, 2016). However, moderating the huge flow of online content generated every day is a tremendous task. The manual review and removal of abusive comments are time-consuming, expensive, and also have detrimental psychological effects on moderators (Hern, 2019), especially faced with the urgency of quick reviews and the toxicity of such data. This creates a strong motivation for the automatic detection of abusive language. Automatic detection systems can scan the enormous volume of text and report the flagged content to the appropriate authorities, making the process of moderation much faster. Such systems can be developed using Natural Language Processing (NLP) techniques.

The detection of abusive language can typically be considered a classification task. The traditional machine learning-based classification methods for detecting abusive language relied on bag-of-words (BoW), handcrafted dictionaries of abusive words, and other lexical features (Razavi et al., 2010; Kwok and Wang, 2013). Further studies involved finding grammatical relations among words in sentences using Part-of-Speech (POS) tags and dependency parsers (Xu and Zhu, 2010; Chen et al., 2012b; Gitari et al., 2015). Researchers incorporated the information about the count of abusive words, hashtags, mentions, punctuations, tokens in comments, and other features such as character and token n-grams (Nobata et al., 2016; Davidson et al., 2017) with traditional classifiers like Logistic Regression, Naïve Bayes, Support Vector Machines (SVMs) (Cortes and Vapnik, 1995), Decision Trees, etc. (Davidson et al., 2017; Fauzi and Yuniarti, 2018).

Later, the use of Deep Neural Networks (DNN)-based models gained popularity in this task because of their representative power. Several comparative studies (Badjatiya et al., 2017; Gambäck and Sikdar, 2017; Zhang et al., 2018) using models like Convolutional Neural Networks (CNNs) (Kim, 2014), Long Short-Term Memory Networks (LSTMs) (Hochreiter and Schmidhuber, 1997), Gated Recurrent Unit (GRU) (Cho et al., 2014a,b) and others have indicated that DNN-based models usually perform substantially better than traditional models. Pre-trained word-embeddings that capture semantic information of words have been extensively used (Badjatiya et al., 2017; Gambäck and Sikdar, 2017; van Aken et al., 2018) as inputs to these models that enable them to begin the training with better prior knowledge. With the advent of Transformers (Vaswani et al., 2017), models like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and other transformer-based models that are pre-trained on mas-

sive amounts of unlabeled data have become the state-of-the-art methods in the abusive language detection task (Mozafari et al., 2019; Bodapati et al., 2019; Banerjee et al., 2020; D'Sa et al., 2020). In light of these recent advancements, **we have adopted the pre-trained transformer-based models throughout our thesis**.

1.1 Transferring Knowledge in Abusive Language Detection

Most machine learning models are built on the basic assumption that the data used for training and evaluation are sampled from the same underlying probability distribution. However, this assumption does not hold true in many real-world applications when the test data is issued from a different generating process compared to the training data (Daumé III and Marcu, 2006). In particular, the nature of conversation over social media is subject to changes over the course of time across many different dimensions. This often leads to a considerable shift in the distribution of data sampled in different manners or at different time-periods (Yin and Zubiaga, 2021). The spontaneity of these online discourses frequently results in neologism, i.e. coining of new terms (Grieve et al., 2018; Würschinger, 2021). For example, the word *finfluencer*, which is used for influencers on social media who create content on financial topics, is a very recently emerged term on the internet. Besides, Eisenstein et al. (2010, 2014) discuss that the written forms of language used in social media vary depending on factors like social and geographic context. Another dimension is the rapid evolution of topics over time caused by the ever-changing socio-political events across the globe (Florio et al., 2020; Saha and Sindhwani, 2012). Abusive language, in particular, is highly sensitive to these dynamics as users often respond quickly to breaking news and other occurrences that trigger sudden outbursts of abusive content. For similar reasons, abusive language is targeted against different communities or individuals at different points in time. Therefore, most abusive language corpora sampled in a certain time frame have a limited vocabulary with regard to the diversity in language usage, topics, targets, etc. For example, a corpus having abusive language that involves only anti-woman terms may not be sufficient for detecting xenophobic content.

Moreover, abusive language corpora themselves incorporate different biases primarily due to the sampling strategy used for their creation (Wiegand et al., 2019). Some corpus-specific biases occur typically due to the use of certain topic-related keywords as seeds to sample abusive content, which results in the disproportionate representation of certain terms across the annotated classes in the corpus. For example, Wiegand et al. (2018a) reported that such a sampling strategy resulted in having topic and user style-specific biases in a popular corpus provided by Waseem and Hovy (2016). Classifiers trained on such corpora are likely to capture these biases rather than learn the generalizable concept of abuse.

In NLP, the term *domain* usually refers to some coherent type of corpus, with regards to the topic, genre, style, linguistic register, etc. (Plank, 2011; Ramponi and Plank, 2020). Since different abusive language corpora are sampled from different underlying

distributions, the term 'corpus' and 'domain' can be used interchangeably. Hence, these factors result in *domain shift* across these corpora. In order to assess how abusive language classifiers generalize to new data, it is recommended to analyze their cross-corpus or cross-domain performance, i.e. trained on one corpus and tested on another, as a more practical evaluation setting (Wiegand et al., 2018a; Karan and Šnajder, 2018).

However, due to the aforementioned variations in abusive language, it is challenging to develop a corpus that is robust to new instances of abuse. Even within the same language or culture, instances that represent certain distributions of topics, targets, and language patterns of abuse may not represent others sufficiently well. Hence, abusive language detection systems trained on older sets of corpora may not be adequate to monitor new comments, undermining their practical value. Recently, many prior works have shown that abusive language classifiers perform well on their respective test sets but degrade substantially when evaluated on data from a different corpus (Karan and Šnajder, 2018; Swamy et al., 2019; Wiegand et al., 2019; Yin and Zubiaga, 2021). Furthermore, annotating data in the new domain of abusive language requires a lot of effort, is expensive and time-consuming, and also has a negative psychological effect on the annotators (Schmidt and Wiegand, 2017; Poletto et al., 2019). Therefore, it is often desirable to transfer knowledge from the existing labeled resources in the best possible way when building a model. The central intent of this thesis is **to learn strategies that can effectively transfer knowledge to minimize the adverse effect of domain shift as well as reduce the annotation effort required for achieving decent levels of performance in a new domain.**

Transfer learning is a concept used in machine learning where previously acquired knowledge from one domain or task is applied to solve a problem in a different but related domain or task (Pan and Yang, 2009; Ruder et al., 2019; Mozafari et al., 2020). In our application, since the task remains the same but the domain changes, we either seek to build generalizable models without exposure to the new data in a different domain or adapt our models to a new domain when some data is accessible. The latter scenario, called *domain adaptation*, is a particular category of transfer learning, called *transductive transfer learning* (Pan and Yang, 2009). Unlike inductive transfer learning, where the source and target tasks differ, in transductive transfer learning the source and target tasks remain unchanged, while the source and target domains are different in terms of their underlying probability distributions. In this thesis, we explore different ways to effectively transfer knowledge from one domain to another by answering the following research questions:

1 How do the topic-memberships of target corpus comments with respect to the topics in the source corpus contribute to the generalizability of a classifier?

2 How do the domain adaptation methods from other tasks like sentiment classification fare for cross-corpus abusive language detection?

3 How to reduce the corpus-specific bias for improving cross-corpus performance?

4 How to effectively transfer knowledge from a resource-rich source corpus to a low-resource target corpus across varied online platforms?

1.2 Contribution of the Thesis

We describe below our contributions in this thesis that attempt to answer the above research questions through different approaches. Table 1.1 specifies whether each of the contributions addresses abusive language detection in general, which includes all kinds of abuse, or particularly targets hate speech detection, a sub-category of abusive language. Moreover, it mentions the assumptions made about the amount of information available from the target corpus.

Contribution	**Detection of Abuse/Hate**	**Information available from the Target Corpus**
Topic Models for Analyzing Generalizability	Abuse	No exposure to the target
Unsupervised Domain Adaptation	Abuse	Unlabeled train set
Model Explanations for Penalizing Spurious Correlations	Hate	Unlabeled train set and small-sized labeled validation set
Neighborhood-aware Optimal Transport	Hate	Small-sized labeled train set

Table 1.1: Task and the availability of information from the target corpus.

1.2.1 Topic Models for Analyzing Generalizability

Topic models are typically unsupervised mechanisms that are capable of discovering the latent topics in a corpus. In an abusive language corpus, every comment can be represented as a probability distribution of topics, such that every topic is a distribution of words in the corpus. Assuming zero exposure to the target corpus during training, we apply unsupervised topic modeling on the source corpus and obtain the topic memberships of the unseen target corpus comments during inference. In Chapter 4, the topic distributions of comments as used as additional information along with the representations obtained from a contextual model to address our first research question: "*How do the topic-memberships of target corpus comments with respect to the source domain contribute to the generalizability of a classifier?*". In particular, we adopt a neural topic model called the Topically Driven Neural Language Model (TDLM) (Lau et al., 2017) and modify it to use pre-trained sentence embeddings for obtaining topic representations. These topic representations are combined with contextualized representations obtained from a pre-trained model called HateBERT (Caselli et al., 2021) to train a classifier on the source corpus. We study how the associations of unseen target comments with abusive language topics in the source corpus impact the cross-corpus performance of classifiers (Bose et al., 2021a).

1.2.2 Unsupervised Domain Adaptation

While considerable efforts are required to annotate a new corpus for abusive language, it is relatively easy to obtain unlabeled data from the target domain. With the availability of this unlabeled data, one solution to improve cross-corpus performance would be to adapt a model trained on the labeled source corpus to the unlabeled target corpus using Unsupervised Domain Adaptation (UDA) methods. UDA approaches have been applied to many NLP tasks but one of the most related tasks to abusive language detection, where these methods have been extensively explored, is sentiment classification. Therefore, in Chapter 5, we analyze the applicability of different widely used pivot-based and adversarial UDA methods from cross-domain sentiment classification in our task to investigate the second research question: "*How do the domain adaptation methods from other tasks like sentiment classification fare for cross-corpus abusive language detection?*". In addition, we train the HateBERT model on the unlabeled target corpus using the Masked Language Model (MLM) objective as another adaptation mechanism before fine-tuning it on the labeled source corpus. Its performance is compared with the other UDA approaches and empirical analysis is performed. Our findings reveal the shortcomings of the UDA approaches from sentiment classification when applied to our task and highlight the necessity to build domain adaptation approaches suited to the requirements of the abusive language detection task (Bose et al., 2021b).

1.2.3 Model Explanations for Penalizing Spurious Correlations

As discussed previously, the disproportionate presence of certain terms across classes results in corpus-specific biases, which manifest in the form of spurious or undesirable correlations. The extent to which these correlations affect a deep learning model can be identified and analyzed by attempting to explain the model predictions. In Chapter 6, we use a class of post-hoc model explanations, called *feature attribution methods* (Lundberg and Lee, 2017), to address the third research question: "*How to reduce the corpus-specific bias for improving cross-corpus performance?*". Feature attribution methods assign scores to the terms in input instances according to their contributions to the predictions. In this chapter, we study the abusive language sub-category of hate speech as it is usually more concerning compared to the other forms of abuse (Davidson et al., 2017). Two domain adaptation approaches are proposed that reduce the effect of spurious correlations caused by corpus-specific terms and improve the generalizability of models on the target corpus. Unlike prior works that rely on pre-defined lists of terms acquired through domain knowledge, the proposed approaches automatically extract and penalize the terms causing spurious correlations in a dynamic manner while training a classifier on the source corpus. This is done through the use of some well-known attribution methods. The first approach uses a small amount of labeled data from the target corpus to extract the terms while the second approach leverages the unlabeled content in the target corpus for the extraction. We demonstrate substantial improvements in cross-corpus performance (Bose et al., 2022a,b) compared to prior works.

1.2.4 Neighborhood-aware Optimal Transport

The effort and cost involved in annotating hate speech comments can be minimized by transferring knowledge from a corpus that has a higher amount of annotated data to a corpus with only a few labeled instances. In addition, the transfer mechanism should be able to effectively address the domain shift across corpora while transferring knowledge to the target. Optimal Transport (OT) (Peyré and Cuturi, 2019; Villani, 2009; Kantorovich, 2006), which is a mathematical theory for comparing two probability distributions, can find correspondences between the source and target domain instances in a geometrically sound manner. Therefore, it is a viable solution for domain transfer. Moreover, it has been found that modeling the relation between the source and target instances using a neighborhood framework is effective for transferring knowledge across corpora. Therefore we propose the incorporation of neighborhood information with *joint distribution OT* (Courty et al., 2017a; Damodaran et al., 2018; Fatras et al., 2021) as a solution to our last research question: "*How to effectively transfer knowledge from a resource-rich source corpus to a low-resource target corpus across varied online platforms?*'. The joint framework models both the embedding and the label distributions of the two domains simultaneously. Our neighborhood-aware joint distribution framework enables learning the amount of transfer between target corpus instances and their neighbors in the source corpus based on their proximity in a sentence embedding space as well as their corresponding labels. Since the available resource-rich source corpus can be collected from a different platform compared to the target corpus in practical scenarios, we perform experiments in cross-platform settings. In these settings, the source and target corpora are sampled from different online platforms. Our approach obtains considerable improvements in cross-corpus performance (Bose et al., 2022c) compared to different transfer learning strategies existing in the literature.

1.2.5 Publications

Parts of the thesis have been published in the following articles:

1 **Tulika Bose**, Irina Illina, Dominique Fohr. Generalisability of Topic Models in Cross-corpora Abusive Language Detection. In *Proceedings of the Fourth NLP4IF Workshop on NLP for Internet Freedom, NAACL*, 2021.

2 **Tulika Bose**, Irina Illina, and Dominique Fohr. Unsupervised Domain Adaptation in Cross-corpora Abusive Language Detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, NAACL*, 2021.

3 **Tulika Bose**, Nikolaos Aletras, Irina Illina, and Dominique Fohr. Dynamically Refined Regularization for Improving Cross-corpora Hate Speech Detection. In *Findings of the Association for Computational Linguistics (ACL)*, 2022.

4 **Tulika Bose**, Nikolaos Aletras, Irina Illina, and Dominique Fohr. Domain Classification-based Source-specific Term Penalization for Domain Adaptation in Hate-speech Detection. In *Proceedings of the Twenty-ninth International Conference on Computational Linguistics, (COLING)*, 2022.

5 **Tulika Bose**, Irina Illina, and Dominique Fohr. Transferring Knowledge via Neighborhood-aware Optimal Transport for Low-resource Hate Speech Detection. In *Proceedings of the Second Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the Twelfth International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, 2022.

1.3 Organization of the Thesis

The remaining part of the thesis is organized as follows.

Chapter 2 describes the panorama of the existing literature on abusive language detection. It begins with presenting the definitions of abusive language and hate speech, provided by different entities, followed by a discussion on the evolution of automatic detection systems, various challenges confronted in the task, and the state-of-the-art solutions to address them. It ends with an overview of existing studies for transfer learning in abusive language detection.

Chapter 3 describes the deep-neural network models, corpora, text pre-processing steps, and the evaluation metric that we have used throughout the thesis.

Chapter 4 deals with our approach of using topic models for exploring the generalizability of abusive language detection systems.

Chapter 5 concerns the analysis of UDA approaches from sentiment classification for our task.

Chapter 6 deals with our proposed approaches of automatic extraction and penalization of corpus-specific spurious correlations using feature attribution methods for improved domain adaptation.

Chapter 7 deals with our proposed approach of neighborhood-aware optimal transport for transferring knowledge in low-resource scenarios.

Chapter 8 concludes the thesis by summarizing the results. We further provide our perspectives and directions for future research.

2 State of the Art

2.1 Introduction

In the past few years, research on the automatic detection of abusive language and hate speech using different NLP techniques has continuously grown with a surge in interest. There has been extensive research aimed to address different challenges in this task. However, owing to its complexity, building a reliable and generalizable classifier for the task remains a challenge. There are numerous problems that contribute to this complexity. Although the objective of this thesis is not to provide solutions to all these problems, studying the wide landscape of the challenges and the existing literature in this field from a computer science standpoint is indeed important. This can provide guidelines and research directions for other future studies that may be conducted in this area. In this chapter, we first study the existing definitions of abusive language and hate speech provided by various sources in Section 2.2. We then discuss the evolution of automatic abusive language classifiers from traditional approaches to state-of-the-art methods in Section 2.3. Thereafter, in Section 2.4, the different challenges present in this task are discussed, along with the studies that have attempted to address them. Finally, in Section 2.5, we discuss the prior work on transfer learning in abusive language detection.

2.2 Defining Abusive Language and Hate Speech

Determining what constitutes 'abusive language' and 'hate speech' is one of the first and major challenges in its automatic detection. There is no precise and universally accepted definition for these terms (Davidson et al., 2017; Schmidt and Wiegand, 2017; Fortuna et al., 2020). A specific definition could possibly make the annotation procedure easier and more reliable (Ross et al., 2016). However, because of the fine line between such phenomena and the right to free expression, it is difficult to agree on a single precise definition. The definitions provided by law are restricted to certain jurisdictions and may not capture all kinds of abusive language and hate speech (Ross et al., 2016; Matsuda, 2018). In practice, social media platforms usually provide their own definitions. Based on the recommendations from online platforms and prior literature, the authors of different abusive and hate speech corpora also present their own annotation guidelines and adjust the existing definitions to obtain more reliable labeling of the sampled data (Davidson et al., 2017; Waseem and Hovy, 2016).

Nonetheless, there exist some commonalities across these definitions. As pointed out by Kumar et al. (2018), it is crucial to reach at least some level of common understanding in order to make progress toward addressing a complex phenomenon like this. Abusive

language has been used as an umbrella term that includes different forms of harmful speech (Nobata et al., 2016; Waseem et al., 2017; Fortuna and Nunes, 2018; Caselli et al., 2020; Stanković et al., 2020; Yin and Zubiaga, 2021). Caselli et al. (2020) defined abusive language as '*hurtful language that a speaker uses to insult or offend another individual or a group of individuals based on their personal qualities, appearance, social status, opinions, statements, or actions. This might include hate speech, derogatory language, profanity, toxic comments, racist and sexist statements.*' Nobata et al. (2016) considered abusive language as a superset of hate speech, derogatory language, and profanity. Thus, indeed hate speech has been agreed upon as a sub-category of abusive language. To understand what constitutes hate speech in particular, we present the definitions provided by different sources[1], such as governing bodies like United Nations[2], EU code of conduct[3], social media platforms like Twitter[4], YouTube[5], Facebook[6], and those adopted by researchers for annotating hate speech resources in Table 2.1.

Table 2.1: Definitions of hate speech

Source	Definition
Code of Conduct between European Union (EU) and companies on illegal online hate speech	"All conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin"
United Nations	"Any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor."
Twitter	"You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories."

[1] online sources accessed in July 2022
[2] https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf
[3] https://ec.europa.eu/commission/presscorner/detail/en/IP_16_1937, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM:l33178
[4] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
[5] https://support.google.com/youtube/answer/2801939?hl$=$en
[6] https://transparency.fb.com/policies/community-standards/hate-speech/

YouTube	"Hate speech is not allowed on YouTube. We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: age, caste, disability, ethnicity gender identity and expression, nationality, race immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin, veteran status."
Facebook	"Direct attack against people – rather than concepts or institutions – on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes, which we define as dehumanizing comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence."
Nockleby (2000)	"Any communication that disparages a person or a group on the basis of some characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic."
Nobata et al. (2016)	"Language which attacks or demeans a group based on race, ethnic origin, religion, disability, gender, age, disability, or sexual orientation/gender identity."
Davidson et al. (2017)	"Language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group"
Fortuna and Nunes (2018)	"Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used."

The definitions in Table 2.1 range from extreme cases of inciting violence, in the definitions provided by the EU Code of Conduct, Twitter, YouTube, Facebook, etc., to including subtle forms of hate, such as the use of humor to demean a group of people in the one provided by Fortuna and Nunes (2018). Despite these nuances in the definitions, all of them have a similar theme: hate speech is targeted toward a social group based on certain attributes in a way that is potentially harmful to them or humiliates them. It is different from the use of general profanity (Malmasi and Zampieri, 2018) or personal

attacks (Wulczyn et al., 2017) that are not targeted toward members of a social group on the basis of the characteristics of that group. As discussed before, following Caselli et al. (2020) and Nobata et al. (2016), *since abusive language subsumes hate speech, in the remaining of the thesis while using the terms 'abuse/abusive', we also include hate speech. However, when using the term 'hate', we refer only to hate speech in particular that follows the aforementioned common notion of wrongfully targeting the members of a social group.* For example, the first comment present below comprises abusive language but not hate speech, whereas the second abusive comment is considered hate speech as it targets the social group 'blacks'.

- Keep your f*cking opinion to yourself. No one gives a damn about an assh*le like you.
- Blacks just don't having the same mental capacity as the whites.

The automatic detection of abusive language can be considered a text classification problem from the NLP perspective: detecting if a message/comment is abusive or not. To perform this task accurately, machine learning models should be trained on data that are annotated reliably (Ross et al., 2016; Fortuna et al., 2020). Typically such annotations are done manually by human annotators. Ross et al. (2016) found low agreement between annotators because of unclear definitions and their own subjective judgments. They further emphasized the need to provide clear and well-formulated guidelines to annotators. Waseem and Hovy (2016), Davidson et al. (2017), Basile et al. (2019) and other authors who later contributed to abusive language resources provided a more precise set of guidelines to the annotators to enhance the annotation quality. These guidelines often included a series of rules and sample comments to increase the agreement among the annotators and thus the reliability of the annotations. As a result, they obtained higher levels of inter-annotation agreement scores (see Chapter 3).

2.3 Evolution of Automatic Abusive Language Detection

In this section, we present an overview of the evolution of abusive language detection models from traditional classifiers in the pre-deep learning era to more sophisticated deep learning-based approaches.

2.3.1 Traditional Pre-deep Learning Methods

The traditional machine learning-based methods for abusive language detection require a careful selection of hand-crafted features provided to classifiers. Therefore, the classification performance depends on how well these features can capture the complexity of the task and the selection of appropriate classification methods along with the features. The advantage of using these traditional methods is that the model decisions are easily interpretable.

Razavi et al. (2010) created an abusive language dictionary by gathering different words, phrases, and expressions and assigning different weights to them based on their

potential impact. They performed multilevel classification, initially using the Complement Naïve Bayes classifier (Witten and Frank, 2002) in the first level that extracted the most discriminative features. Subsequently, they used the Multinomial Updatable Naïve Bayes classifier (Witten and Frank, 2002) in the second level to obtain a new set of features derived from the first and then ran a rule-based classifier called Decision Table/Naïve Bayes hybrid classifier (Hall and Frank, 2008) on the derived features to make the final decision. Kwok and Wang (2013) also used the Naïve Bayes classifier with bag-of-words (BoW) to detect racist tweets. They concluded that using BoW as features is not sufficient and recommended the incorporation of knowledge from more refined methods like performing sentiment analysis, word sense disambiguation, etc. to achieve robust classification.

Xu and Zhu (2010) introduced a sentence-level filtering approach that can semantically remove abusive content from sentences. They used a dictionary of abusive words to find other affected words in sentences that should be removed along with the identified abusive words to maintain the readability of the sentences. For this, they identified grammatical relations among words using Part-of-Speech (POS) tags and dependency relations, and then performed rule-based filtering of abusive content. Chen et al. (2012b) also constructed an abusive word dictionary and then adjusted the offensiveness scores associated with these words based on their context using dependency parsing. These scores were then combined to get the offensiveness value of a sentence to finally detect abusive content. Gitari et al. (2015) used a rule-based method to create a lexicon for detecting abusive language using semantic and subjectivity features inclined toward abuse. They used a POS tagger to detect subjective sentences using sentiment lexicons. Using these subjective sentences, the authors derived three different sets of features: words with negative polarity, hate words, and theme-based grammatical patterns to perform rule-based classification.

Nobata et al. (2016) employed different features such as character and token n-grams, linguistic features like number of tokens in comments, number of punctuations, number of politeness words, number of abusive words, etc., syntactic features like POS tags, dependency relations derived using dependency parsing, and embedding-derived features. The n-gram features gave better performance compared to the other features and combining all the features yielded the best performance in the abusive language detection task. Davidson et al. (2017) used different features like POS tags, n-grams, sentiment scores for every comment using a sentiment lexicon, binary and count indicators for hashtags, mentions, retweets, etc., number of characters, words, and syllables, and other features. The authors experimented with different models like Logistic Regression, Naïve Bayes, Decision Trees, Random Forests, and linear Support Vector Machines (SVMs) (Cortes and Vapnik, 1995) and found that Logistic Regression and SVM outperformed the other models. They chose to employ a Logistic Regression with L2 regularization in the final model as it provides predicted probabilities of class memberships.

Fauzi and Yuniarti (2018) used BoW features with Term Frequency - Inverse Document Frequency (TF-IDF) (Salton and Buckley, 1988) weighting and also experimented with different classifiers like Naïve Bayes, K-Nearest Neighbours, Maximum Entropy,

Random Forests, and SVM. They finally combined these classifiers with ensemble approaches of hard voting, i.e. majority voting, and soft voting, i.e. taking average class probabilities as the voting score, and showed that such ensemble methods yield robust and improved performance. Van Hee et al. (2018) used multiple features like word n-grams, character n-grams, indicators for proper names, 'allness' indicators (e.g. 'always', 'everybody'), diminishers (e.g. 'slightly', 'relatively'), intensifiers (e.g. 'absolutely', 'amazingly'), negation words, profane words, topic model features, etc. Linear SVMs were used along with optimization for the features and hyperparameter combinations.

2.3.2 Deep Learning-based Methods

Unlike the traditional classifiers that require a proper selection of appropriate features along with domain-specific knowledge, Deep Neural Network (DNN)-based approaches can learn useful representations of the input text through the automatic selection of features. They use a multi-layer structure to learn new abstract representations from the input. Since DNN models require a large amount of data for learning useful representations, pre-trained representations have been used extensively in different downstream NLP tasks. The pre-trained word embeddings like word2vec (Mikolov et al., 2013a, 2018) are learned in such a way that two semantically similar words occur close in the high dimensional representation space. Currently, deep learning-based methods have emerged as state-of-the-art in the task of abusive language detection.

Badjatiya et al. (2017) were one of the first to investigate the impact of deep learning in the abusive language detection task and they found that deep learning-based approaches significantly outperformed the traditional machine learning methods. The authors experimented with different traditional methods by using features like character n-grams, TF-IDF, BoW Vectors over Global Vectors for Word Representation (GLoVe) (Pennington et al., 2014) with classifiers like SVMs, logistic regression, Random Forest, and Gradient Boosted Decision Trees (GBDTs). They compared these approaches with randomly initialized embeddings and pre-trained embeddings from GLoVe provided as input to different deep learning architectures using Convolutional Neural Networks (CNNs) (Kim, 2014), Long Short-Term Memory Networks (LSTMs) (Hochreiter and Schmidhuber, 1997), etc. Finally, they found that these deep learning-based approaches yielded the best performance. A similar comparative study was performed by Lee et al. (2018). CNN-based classifiers were also explored for abusive language detection by Gambäck and Sikdar (2017). They found that a CNN model with pre-trained word vectors gave the best overall performance compared to Logistic Regression with character n-grams and having random vectors or character n-grams as inputs to the CNN model. Zhang et al. (2018) proposed a combination of CNN and Gated Recurrent Unit (GRU) (Cho et al., 2014a,b), where the input was passed sequentially through the CNN and GRU layers. They compared this model to SVM with different surface-based, linguistics, and sentiment features and with the CNN model, and found that their proposed combined architecture gave the best classification performance.

Transformer-based models (Vaswani et al., 2017) like Bidirectional Encoder Repre-

sentations from Transformers (BERT) (Devlin et al., 2019), Robustly optimized BERT approach (RoBERTa) (Liu et al., 2019b), XLNeT (Yang et al., 2019) are trained on massive unlabeled corpora and can be simply fine-tuned to different tasks using the relevant task-specific corpora. While performing fine-tuning for a specific task, additional task-specific layers are added to the model and these task-specific, as well as the pre-trained parameters in many cases, are updated. Mozafari et al. (2019) adopted different strategies to fine-tune BERT for the abusive language detection task: fine-tuning the model without changes, inserting additional non-linear layers that take the output embeddings for the `[CLS]` token, inserting a Bidirectional LSTM (BiLSTM) layer that takes the outputs for all the tokens from the last layer of BERT and applying convolution operation on the output from all the layers. The best results were obtained from fine-tuning along with the convolution operation. Similarly, Liu et al. (2019a) fine-tuned BERT and observed that it outperformed an LSTM-based classifier.

Banerjee et al. (2020) adopted an ensemble of different pre-trained Transformer-based models like BERT, XLNeT, RoBERTa. Bodapati et al. (2019) compared different models like fastText (Joulin et al., 2017), subword-based model, joint word and character embedding model (Kim, 2014; Peters et al., 2018) that use character CNN along with pre-trained word embeddings, Very Deep Convolutional Neural Network (VDCNN) (Conneau et al., 2017) that works at the character level, Byte Pair Encoding (BPE)-based model (Sennrich et al., 2016), and the pre-trained BERT model. They observed that fine-tuning BERT achieved the best performance in most cases. Similarly, d'Sa et al. (2020) compared the performance of providing fastText and pre-trained BERT embeddings as features to DNN classifiers like CNN, and BiLSTM, with BERT fine-tuning. They found that simply fine-tuning BERT obtained significantly improved performance.

2.4 Challenges in Automatic Abusive Language Detection

In this section, we present a discussion on some of the major challenges that exist in the task of abusive language detection and the attempts made in the literature to address them.

2.4.1 Explicit versus Implicit

A prominent distinction made in the abusive language literature is whether the abuse expressed is explicit or implicit (Waseem et al., 2017). While the first one is expressed generally unambiguously through profanity, derogatory terms, or threats, the second involves subtle expressions of abuse, eg. using sarcasm, irony, negative stereotypes (Sap et al., 2020), humor, metaphors (Mohammad et al., 2016; Mishra et al., 2019), etc. In many cases, the latter requires additional context, complex reasoning, or external world knowledge about the cultural and social dimensions involved to be identified as abusive (van Aken et al., 2018; Espinosa Anke et al., 2019). This makes both the annotation and eventually the detection of implicit abuse challenging (Justo et al., 2014; Botelho et al., 2021; Wiegand et al., 2021; Pal et al., 2022). Kumar et al. (2018) introduced a corpus

sampled from Facebook that was annotated by linguistic experts as 'overt' and 'covert' aggression, which are respectively similar to explicit and implicit abuse. Some examples of explicit and implicit abuse are given below:

Explicit Abuse:

- So go ahead, go nuts, go apesh*t, f*cking n*gs. (Vidgen et al., 2021c)
- Lol, chop her head off and rape the b*tch. (Basile et al., 2019)
- This guy is the biggest f*ggot omfg. (Davidson et al., 2017)

Implicit Abuse:

- Nothing. I like women in pink. It's a tipoff that they don't angrily reject femininity like...ahem, well, you know. (Waseem and Hovy, 2016)
- I can distinguish a Muslim woman from a letterbox, but it does take me some time! (Vidgen et al., 2021c)
- Is this one of hard working immigrants just looking for a better life in the US that we hear about from the left all the time? #BuildTheWall #SendThemBack. (Basile et al., 2019)

Implicit abuse is no less harmful than explicit abuse in its intention to marginalize its targets (Alorainy et al., 2019; Caselli et al., 2020). For instance, the first example of implicit abuse is a sexist statement in connection with feminism. The second example is a sarcastic statement that insults Muslim women without using expletives. Because of their subtle nature, implicitly abusive comments are often misclassified as non-abusive (false negatives cases) (Qian et al., 2018; Zhang and Luo, 2019; Basile et al., 2019; Mozafari et al., 2019). van Aken et al. (2018) found that many of the false negatives were caused by abusive comments without swear words, comments with rhetorical questions, metaphors, and sarcasm. Furthermore, Wiegand et al. (2019) showed that biased sampling strategies (e.g. sampling using specific keywords or topics that may coincide with abusive language) caused abusive language corpora to have a high degree of implicit abuse. Models trained on such corpora gave a worse performance in cross-domain evaluations, i.e. evaluated on samples from a different corpus, compared to training on corpora with explicit abuse. They concluded that classifiers could not learn implicit abuse well.

Several studies have attempted to address the detection of implicit abuse. The work of Magu et al. (2017) specifically targeted the problem in which online users substitute the terms identifying social groups with code words in abusive comments to evade their automatic detection, e.g in the comment '*Gas the skypes*', 'skypes' is a code word for 'jews'. They collected and annotated tweets with such phenomena, and trained an SVM classifier over tweets represented by BoW features. Caselli et al. (2020) presented annotation guidelines to indicate the degree of explicitness. They re-annotated the 'offensive' class in the corpus provided by Zampieri et al. (2019a,b) into the classes of 'implicit abuse', 'explicit abuse', and 'no abuse'. The authors demonstrated that the BERT model had better classification performance than dictionary-based approaches for distinguishing abuse versus non-abuse. Furthermore, the authors performed a multi-class classification

with BERT for the three aforementioned classes. They showed that even for BERT, detecting implicitly abusive samples is much more challenging compared to the explicitly abusive ones. Sap et al. (2020) designed a considerably different task and corpus to learn implied stereotypes in comments. The authors asked annotators to specifically write simple statements on what stereotype or characteristic of the targeted group is implied by the intentionally hurtful comments. They subsequently defined a sequence-to-sequence task of generating the implied statements along with the classification task of identifying different categorical variables in the comment, such as offensiveness, intent to offend, etc.

There are only a few studies that designed models to specifically address implicit abuse. Gao et al. (2017) built a weakly-supervised model having two components with the goal of capturing both explicit and implicit abusive language. The model's 'slur learner' path aimed to capture explicit cases using an initial slur matching and the LSTM path attempted to detect the implicit content. However, as discussed by Yin and Zubiaga (2021) since the LSTM path was also trained on abusive language cases obtained through the slur-learner, it is questionable if it actually learns to recognize implicit abuse. Wang et al. (2020) employed representations from a sarcasm detection model (Ghosh and Veale, 2016) as extra input for improving the detection of implicit abuse. However, this methodology may not necessarily be suited to other forms of implicit abuse.

Waseem et al. (2017) observed that researchers may find it challenging to develop models that can aid in the detection of implicit abuse in the absence of good-quality labeled data to learn these representations. Moreover, the authors suggested using features other than those obtained from the comments, such as user profiling (Dadvar et al., 2013) and other extra-textual features. Since in many cases, implicitly abusive comments can only be understood within their specific context, Gao and Huang (2017) introduced context information such as the original news article along with the comments in a news corpus. de Gibert et al. (2018) provided context information to the annotators for labeling the comments. They used a 'relation' label in their corpus to distinguish comments that can only be understood properly in the context of their neighbors.

Alatawi et al. (2021) collected a corpus for detecting white supremacist hate speech and initially tried to obtain annotations for the categories of 'explicit white supremacy', 'implicit white supremacy', 'other hate speech', and 'neutral'. However, they observed very low inter-annotator agreement for these labels with a Cohen kappa score Cohen (1960) of 0.07. The highest disagreement observed while differentiating the neutral and implicit categories. Therefore, they had to collapse the four labels into the categories of 'white supremacy' and 'non-white supremacy', which had an improved but still poor inter-annotator score of 0.11. Their work pointed out the difficulty of the annotation process in this task, especially annotating implicit cases, due to its subjective nature.

ElSherief et al. (2021) provided a taxonomy for characterizing different forms of implicit hate speech and presented a corpus with fine-grained annotations for these different forms. The authors hired and trained expert annotators for this process. They demonstrated that the BERT classifier outperformed SVMs, but the fine-grained classification performance of implicit abuse was comparatively poor. Wiegand et al. (2021) also provided a list of sub-types of implicitly abusive language. They recommended the creation

of corpora for the individual sub-types of implicit abuse rather than having corpora that cover many different forms of abuse. They argued that by doing so, classifiers are less prone to getting biased to other unintended artifacts present in the data and can learn the particular sub-type of implicit abuse in a better way. Indeed, the detection of implicit abuse still remains a challenging open research problem. We do not specifically address this problem in the thesis.

2.4.2 Role of Contextual Information

One of the key issues raised by Prabhakaran et al. (2020) is that 'abuse is contextual', while the authors reported and analyzed the discussion content of a panel comprising NLP researchers and human rights experts at the Human Rights Conference RightsCon 2020. The panelists also discussed how temporal and geographical contexts as well as the power held by the perpetrator might be crucial in determining whether a message is abusive or not. In the position paper by Jurgens et al. (2019), the authors stated that understanding community norms and context is important for detecting abuse. Vidgen et al. (2019) discussed that what is considered abusive in one context or to one user may be considered non-abusive to a different user or in another context. This is also observed in the analysis of the inter-annotator agreement of annotators by Salminen et al. (2018), who showed that even though the annotators were experienced, they often had different perceptions of what is abusive. This perception can vary based on different factors, e.g. the gender of the annotator (Binns et al., 2017).

Moreover, factors such as the topic of a comment, its discourse context, accompanying media objects, such as images, videos, and audio, its time of posting, global events, the author's identity, and the identity of the intended audience also have an effect based on which a message is considered abusive (Schmidt and Wiegand, 2017). For instance, the comment '*barryswallows: Merkel would never say NO*' (Gao and Huang, 2017) may not be considered abusive unless the context of the comment is provided. This comment was posted for the news titled "German lawmakers approve 'no means no' rape law after Cologne assaults" and was an insult directed against a female politician.

In most of the available abusive language corpora (Waseem, 2016; Waseem and Hovy, 2016; Basile et al., 2019; Davidson et al., 2017; Wulczyn et al., 2017), the previous comments in a conversation are not included. As a result, conversational context is not taken into account by algorithms trained on these corpora. As discussed in Section 2.4.1 on implicit abuse, de Gibert et al. (2018) and Gao and Huang (2017) included context information in the annotation process and the corpus. However, Pavlopoulos et al. (2020) mentioned that the context provided by Gao and Huang (2017) is limited to the title of the news article. The preceding comments included in this corpus cannot be linked with the corresponding target comments in the absence of sufficient information. Menini et al. (2021) re-annotated the Twitter corpus provided by Founta et al. (2018) with and without preceding context. They showed that when the context information was provided, many previously labeled abusive tweets were labeled as non-abusive by the annotators. They also found that the performance of classifiers dropped significantly

when context information was provided as additional input, compared to providing only the actual tweet as input. This suggested that classification using context-aware corpora may be more challenging but reflects the real application scenario whereas classifiers trained without context may be overly optimistic. Fehn Unsvåg and Gambäck (2018) investigated the effect of incorporating different user features from Twitter into an abusive language classifier. The experiments showed that the network-related user features (i.e., the number of followers and friends) gave the most consistent improvements, while other user features resulted in inconsistent effects on the performance. Pavlopoulos et al. (2020) created a corpus with context information from Wikipedia conversations and found that using context had a significant effect on the annotation. However, they found no evidence that including context actually improves classification performance.

There are a few works that use *multi-modal features* beyond just textual features as additional context for detecting abusive content. Hosseinmardi et al. (2015) created a corpus containing images from Instagram, their corresponding comments, and metadata such as the number of followings, followers, shared media, etc. for cyberbullying detection. Moreover, annotators were asked to categorize the images based on their content and train classifiers that incorporate all these multi-modal features and demonstrated largely improved classification performance. Zhong et al. (2016) also collected images and comments from Instagram along with the image captions and other metadata for the same task. They used the features derived from this information along with representations from a pre-trained CNN applied to image pixels to train their classifiers. They found that the topic proportions in image captions serve as crucial features for indicating future bullying. Gomez et al. (2020) discussed that many multi-modal messages are abusive only in the combination of texts with their associated images. Therefore they collected samples from Twitter having both textual and image information. However, they observed that their multi-modal models could not outperform the textual models. Rezvani et al. (2020) used the Instagram corpus provided by Hosseinmardi et al. (2015) and a Twitter corpus (Sui, 2015) for cyberbullying detection and designed a model that incorporates different contextual information apart from text such as user metadata and image features. They demonstrated that incorporating this contextual information improved the performance of classifiers compared to using only the textual comments. Kiela et al. (2020) presented a 'hateful memes challenge' that involved challenging memes whose hateful content could not be easily identified with textual models.

Incorporating additional context and multi-modal features in classifiers is an emerging area in abusive language classification. It nevertheless remains a challenge given the few context-based corpora and studies in this area. In this thesis, we have used the abusive language corpora that comprise only the individual comments without additional context information to maintain uniformity of the available information across corpora.

2.4.3 Non-standard and Ungrammatical Language

The language used in social media is colloquial with irregular capitalization, punctuation, spelling, lexicon, and syntax (Eisenstein, 2013), e.g. using 'whaaaat' instead of 'what',

'grt' instead of 'great'. Hosseini et al. (2017) designed experiments to provide adversarial examples to Google's `Perspective`[7] API, which is a project to detect abusive language. They misspelled some abusive words or added unusual punctuations and could deceive the system into reducing the toxicity score of comments. For instance, a comment '*They are liberal idiots who are uneducated*' with a toxicity score of 90% got reduced to 15% with a minor spelling variation to '*They are liberal i.diots who are un.educated*' at the time of their experiments[8]. Spelling variations result in misclassifications as they create 'out-of-vocabulary' (OOV) words (Serrà et al., 2017).

Gröndahl et al. (2018) used more challenging adversarial inputs such as using typos and leetspeak (Perea et al., 2008) ('idiot' to 'id10t'), inserting or removing whitespace to change word boundaries (e.g. 'feminismisawful'), adding unrelated positive words (e.g. 'love') to abusive comments, without changing the perceived meaning of the comments for humans. They showed that different state-of-the-art abusive language detection models perform well only when evaluated on the same kind of data as they were trained on but suffer a significant drop in performance while encountering adversarial inputs with different forms of spelling and word variations. They showed that especially removing word boundaries and adding words like 'love' to abusive comments had a drastic effect on a classifier's performance. Similar adversarial attacks have been analyzed by Eger et al. (2019).

Some authors have performed spelling checks by using or building dictionaries or available spell-check tools as a part of text preprocessing (Pratiwi et al., 2019; Moh et al., 2020). Moh et al. (2020) proposed different defense mechanisms to shield against the adversarial attacks of text morphing as mentioned before and also introduced the attack of adding random letters to word edges (e.g. 'wall' to 'xxxwallxxx'). These mechanisms include word segmentation, i.e. separating combined words without whitespaces into separate words, a rule-based correction of morphed letters, identifying words that break grammatical rules, finding the longest word based on vowel positional searching, and retraining on the morphed text. However, these schemes may not specifically address obfuscations such as repetition of letters (e.g. 'woman' to 'woooomaaann') or characters in words separated by spaces ('gay' to 'g a y') and other noises. Rather than using spell check tools, Hu et al. (2020) represented the text as phonemes (Trager and Bloch, 1941) because phonemes of the actual, morphed words with spelling variations and repetitions stay the same. With phoneme representations, they obtained improvements in classification performance. Eger et al. (2019) employed various strategies to defend against morphed texts: *adversarial training*, i.e. training on the morphed text, *visual image-based character embeddings* to shield against perturbed characters that can be correctly interpreted by humans visually (e.g. 's' to '$'), combining the two approaches, and *rule-based recovery* that replaced every non-standard character in the input with its nearest standard neighbor in the character embedding space.

Wulczyn et al. (2017) showed that the model using character n-grams outperform word n-grams, indicating higher robustness of character n-grams to spelling variations.

[7] https://www.perspectiveapi.com/

[8] Currently the API is robust to this example.

Similarly, Nobata et al. (2016) used character n-grams for addressing noisy data. Besides, the rule-based and dictionary-based approaches mentioned above may not be able to handle misspelled words that have different meanings in different contexts. For example, the word 'cnt' in the sentences '*I cnt understand this!*' and '*You feminist cnt!*' (Mishra et al., 2018b) have different meanings: its means '*cannot*' in the first example and a slur '*cunt*' in the second example. Mishra et al. (2018b) obtained context-aware character representations by averaging the forward and backward states of a bi-LSTM network and learned to predict the embeddings of unseen words to address such situations.

Mou et al. (2020) incorporated subword information to make models resilient to noisy text. Likewise, Bodapati et al. (2019) analyzed the effectiveness of character, subword, or Byte Pair Encoding (BPE) (Gage, 1994; Sennrich et al., 2016) models in the presence of noisy text. Subword or BPE tokenization basically concatenates the frequently occurring character n-grams within a word boundary to address the OOV problem. For example, if the words in a training corpus are given by {'*greatest*', '*fastest*', '*older*', '*old*'}, they are first split into constituent characters. Then the characters that frequently occur together in the corpus, such as {'e', 's', 't'} and {'o', 'l', 'd'} would get merged together into 'est' and 'old'. This would help split rare OOV words into meaningful subwords. Bodapati et al. (2019) showed that character-based models are less effective compared to subword-based models as the latter retain important word-boundary information. They further demonstrated that the pre-trained WordPiece tokenization of the BERT model had an overall better performance and substantial gains compared to character, subword, or custom BPE models.

2.4.4 Diverse Labeling Choices

A wide range of labeling schemes is used in the annotation of the abusive language corpora, which results in variations across corpora and additional difficulties in the classification of abusive language, especially in cross-corpus settings. Wulczyn et al. (2017) studied personal attacks on Wikipedia and used the labels of *toxic speech* and *non-toxic speech*. This was later extended and used with the fine-grained multi-labels (a single comment having more than one label) of *toxic*, *severe toxic*, *obscene*, *threat*, *insult*, and *identity hate* in the Kaggle's toxic comment challenge[9]. However, while toxic speech is different from hate speech, it may be considered a category of abusive language as per the definition of abusive language provided by Caselli et al. (2020) and presented in Section 2.2.

Both hate speech and other forms of abusive language may use profane language, but general profanity can also be used in informal expressions without intending harm to an individual or a community. Indeed, filtering all instances of profanity may not be desirable (Malmasi and Zampieri, 2018). Therefore, the corpus provided by Davidson et al. (2017) comprises three classes: *hate speech*, *offensive* but not hate speech, and *neither*. Founta et al. (2018) followed a more principled strategy by performing a controlled statistical study of the labels selected by the annotators in an iterative manner. They initially

[9] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

started with the label set of {*offensive, abusive, hateful, aggressive, cyberbullying, spam, normal*}. From the first two rounds of annotation, they realized that cyberbullying was rarely used by the annotators as they are meant to be a repetitive form of abuse in nature, which could not get reflected in individual comments. Further, they observed that the labels *abusive*, *offensive*, and *aggressive* were highly correlated and *abusive* was the most popular amongst the three. On the other hand *hateful* was not significantly correlated with any other label. Therefore, their final set of labels was {*abusive, hateful, spam, normal*}, by eliminating the label *cyberbullying* and merging *offensive*, *aggressive*, and *abusive* into a single label *abusive*.

Different from these schemes, Zampieri et al. (2019a,b) used a hierarchical labeling strategy with the first labeling level of *offensive* and *non-offensive*, the second level where offensive comments were further categorized into *targeted insult* and *untargeted*, and the third level of categorizing targeted offensive comments into *individual*, *group*, and *other*. Caselli et al. (2020) further re-annotated the offensive comments in this corpus, directed to either individuals or groups, into the labels of *explicit abuse*, *implicit abuse*, and *none*. Cercas Curry et al. (2021) took a step further and used a rating scale to measure the severity of abuse, ranging from +1 (*friendly/non-abusive*) to -3 (*strongly abusive*) for studying the user conversations with three conversational AI systems. Similar to the previously mentioned labeling schemes, in addition to binary labels of *abusive* and *non-abusive* and the rating scale, they used a 3-way labeling for the target of abuse, i.e. *group*, *individual–system*, or *individual–3rd party* and binary labeling for the directedness of abuse, i.e. *explicit* or *implicit*. Finally, they categorized the targets into more finely-grained labels of *general*, *sexist*, *sexual harassment*, *homophobic*, *racist*, *transphobic*, *ableist*, or *intellectual*.

The corpus provided by Waseem and Hovy (2016) focused more on fine-grained categories within hate speech itself. They thus used the labels *racism*, *sexism*, and *neither*. Basile et al. (2019) used a multi-label scheme where every comment has a binary label of *hateful* and *non-hateful*, and then every hateful comment comprises another kind of binary label indicating whether the target is a specific *individual* or a *generic* group. Finally, they used another set of labels for the hateful comments to specify whether the author of the comment is *aggressive* or *non-aggressive*. Some authors focused on specific types of hate speech only, for instance, sexism or misogyny (Jha and Mamidi, 2017; Fersini et al., 2018a; Samory et al., 2021), Islamophobia (Chung et al., 2019), or East Asian prejudice (Vidgen et al., 2020), etc. The corpus provided by Vidgen et al. (2021c) is more diverse through a different labeling strategy. This includes a binary label of *hate* and *non-hate*, labels for the types of hate, i.e. *derogation*, *animosity*, *threatening language*, *support for hateful entities*, and *dehumanization*, and labels for many different targets of hate. Similarly, Mollas et al. (2022) also used a diverse set of targets for their labeling scheme and used a multi-label annotation strategy. The labels involve both the binary labels of the presence or absence of hate speech (*isHate*) and multi-labeled categorization in terms of *violence*, *directed_vs_generalised*, *gender*, *race*, *national_origin*, *disability*, *sexual_orientation*, and *religion*.

The heterogeneity in the labeling choices of different corpora raises difficulties in the

adaptivity or generalizability of abusive language detection models, i.e. when they are trained on one corpus and evaluated on different corpora (Gröndahl et al., 2018; Pamungkas and Patti, 2019; Fortuna et al., 2020, 2021). Pamungkas and Patti (2019) demonstrated that training on a corpus with more general and broader coverage of abusive language generalizes better when evaluate on a corpus with specific types of abuse (e.g. racism, sexism). Fortuna et al. (2020, 2021) discussed the need for more coherent categorization and standardization of labels, i.e. assigning the same labels to equivalent categories in different corpora, for better generalizability.

2.4.5 Class Imbalance

Founta et al. (2018) estimated that the proportion of online abusive content is very low (0.1% -3% on Twitter) in practice. Therefore, many corpus creators use the keyword-based sampling strategy, using terms that are more frequently present in abusive discussions, to increase the abusive content in the data. Still, many of the widely used corpora exhibit class imbalance (Davidson et al., 2017; Waseem and Hovy, 2016; Waseem, 2016). For example, in the corpus provided by Davidson et al. (2017), the proportions for the classes 'offensive speech', 'hate speech', and 'neither' are 77.4%, 5.6%, and 16.9%, respectively. The class of 'offensive speech' has an unusually high number of instances as the corpus has been sampled using slurs and other abusive keywords. In the corpus introduced by Waseem and Hovy (2016), 20% of the comments were annotated sexist, 11.7% racist, and 68.3% neither. When classifiers are dominated by the majority class, they tend to prefer predicting the test instances as belonging to the majority class. This results in sub-optimal classification performance (Chawla et al., 2004).

Agrawal and Awekar (2018) performed Random Oversampling (ROS), i.e. oversampling of the minority class by replicating them randomly to address the class imbalance. However, they oversampled the entire training data before splitting the data into train-test partitions. Arango et al. (2019) argued that such sampling should only be done in the training partition of the data as otherwise, the same data instances can appear both in the train and test partitions. Rathpisey and Adji (2019) also used ROS along with Random Undersampling (RUS), Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002), and Adaptive Synthetic (ADASYN) (He et al., 2008). They showed that ROS outperformed the other methods to handle unequal class distributions. Rizos et al. (2019) used different text-based data-augmentation strategies that involved (i) substituting words with their synonyms from the word-embedding space, (ii) shifting, and (iii) warping word tokens in a sentence within the boundaries of a padded sequence, and generation of new sentences using class-conditional, recurrent neural language models. They also performed undersampling of the majority class instances at every training epoch such that the model is trained on equal proportions of instances from both classes while being able to learn from all the training instances. The authors demonstrated significant improvements in the classification performance. Prasad et al. (2021) demonstrated that random oversampling of the minority class followed by undersampling (randomly deleting instances) of the majority class gave a better performance than using only oversampling

or undersampling in a multi-modal abusive language detection task.

Furthermore, some researchers have used class weighting (Rani et al., 2020; Banerjee et al., 2021) at the training stage to alleviate the issue of class imbalance. When computing the loss function, class weighting gives greater weights to the minority class instances. Besides, the choice of classification metric is also important in class imbalanced settings. Since macro-average F1 gives equal importance to all the classes unlike micro-average F1 which favors the majority class (Sokolova and Lapalme, 2009; Narasimhan et al., 2016), it has been preferred in such imbalanced settings (Narasimhan et al., 2016). Therefore, in order to reduce the effect of imbalance between classes, Bosco et al. (2018); MacAvaney et al. (2019); Mandl et al. (2020) preferred using macro-average F1 to report the abusive language classification performance. In this thesis, we have thus used *macro-average F1* to report the model performance.

2.4.6 Evolution of Abuse over Time

The abusive phenomenon in social media keeps evolving with time because of language variation (Eisenstein et al., 2014), e.g. the emergence of new words (Grieve et al., 2018), the advent of new events and topics of discussion (Florio et al., 2020; Saha and Sindhwani, 2012), etc. This especially holds true because of the informal and colloquial nature of language on social media. For instance, the outbreak of COVID-19 resulted in a sudden increase in abusive language against East Asians (Vidgen et al., 2020). New slangs, e.g. *covidiot*, emerged in last two years. Such temporal shifts cause degradation in the performance of classifiers trained on a temporally distant corpus (a corpus collected in a comparatively older time frame) when evaluated on newer corpora (Vidgen et al., 2019; Florio et al., 2020). The analysis by Nobata et al. (2016) on the impact of abusive language over time suggested that training with more recent data is preferable to having larger training data. Likewise, Florio et al. (2020) showed that introducing training data that is closer in time with respect to the test data improved the performance of classifiers. Their analysis further revealed that topics shift relatively fast in social media and this is a crucial challenge for building generalizable abusive language detection systems.

2.4.7 Biases in Corpora

The corpus creation process adopted for the available abusive language corpora may induce different types of biases. These biases may be caused due to the sampling strategy, demography of the annotators, time frame of collection, etc. (Wiegand et al., 2019; Yin and Zubiaga, 2021). Such unintended biases may not immediately get reflected in the usual classification performance metrics when models are trained and evaluated on the same corpus (Wiegand et al., 2019). However, the generalizability of a model trained on a biased corpus may get adversely affected. In addition, certain unintended biases may have more serious implications by marginalizing a community that the models intend to protect. We describe the biases usually found in the abusive language corpora and discuss the solutions proposed in the literature to address them.

Sampling/Domain bias

As mentioned previously, true random sampling for the creation of abusive language corpora would lead to a very sparse presence of abusive comments in a corpus. To address this, a common strategy is to sample using keywords known to be prevalent in abusive discussions, which is usually called *focused/biased sampling* in the literature (Wiegand et al., 2019). Wiegand et al. (2019) showed that such focused sampling results in the over-representation of certain terms only in one of the classes. This biases the classifier to make decisions based on those terms rather than learning the underlying concept of abuse. For example, the corpus provided by Waseem and Hovy (2016) was collected using manually identified terms, slurs, and hashtags frequently found in sexist and racist comments. One of the most frequently discussed topics in this corpus is the role of women in sports such as football, which in turn highly correlate with sexist expressions. Wiegand et al. (2019) show that the most frequently correlated terms in the abusive comments were 'commentator', 'comedian', 'football', 'announcer', etc. They used the phrase **topic bias** to describe this particular phenomenon. In the absence of a balanced representation of these terms in the non-abusive comments, a classifier could be prone to detecting an innocuous comment containing these terms as abusive, i.e. such terms would result in *spurious correlations* or artifacts in the data. Davidson et al. (2017) also adopted a key-word based sampling procedure with the *hatebase*[10] lexicon used to collect samples that resulted in an unrealistic proportion of comments with slur terms, with 77.4%, 5.6%, and 16.9% of the comments belonging to 'offensive', 'hate speech', and 'normal' classes.

On the other hand, Founta et al. (2018); Razavi et al. (2010), and the Jigsaw Kaggle challenge[11] applied a *boosted random sampling* strategy that involves random sampling of data, followed by using some heuristics to boost the minority class. Wiegand et al. (2019) showed that the top ten terms correlating with the abusive class in Founta et al. (2018) were mostly slur words such as 'b*tch', 'n*ggas', 'f*cking'. This is in contrast to the observation on the corpus of Waseem and Hovy (2016) and other corpora that used biased sampling (Warner and Hirschberg, 2012a; Kumar et al., 2018), which have non-abusive words highly correlated with abusive comments. This indicated that corpora sampled with procedures closer to random sampling are less affected by data bias.

Wiegand et al. (2019) recommended using cross-corpus classification, i.e. training on one corpus and testing on another, to assess the impact of bias in a corpus. They demonstrated that the corpora involving biased sampling had low cross-corpus performance compared to those sampled with boosted random sampling. This shows that data bias reduces the generalizability of classifiers. They further suggested debiasing the corpora by methods such as augmenting the corpus with additional random comments containing the bias-causing terms (Dixon et al., 2018; Wiegand et al., 2018b). Razo and Kübler (2020) on the other hand reproduced the study of Wiegand et al. (2019) by

[10] https://hatebase.org/
[11] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

comparing the impact of boosted random sampling and biased sampling performed using the same underlying data source. This was done to rule out the possibility of having the data source as a confounding element. Although they found similar trends as Wiegand et al. (2019), the effect was much less pronounced suggesting that the data source can have more effect than the sampling strategy used. van Rosendaal et al. (2020) investigated ways to increase abusive content while still maintaining a wide range of topics when creating the corpus. They focused on controversial topics on Reddit for extracting keywords to sample data from Twitter, rather than manually selecting the keywords.

Author bias

The distribution of authors in the corpora may also lead to a biased corpus. For example, more than 70% of the sexist comments in the corpus provided by Waseem and Hovy (2016) originated from two authors, and 99% of the racist comments came from a single author (Wiegand et al., 2019). Such a skewed distribution of authors is likely to bias the classifiers to predict the class of a comment based on the writing style or topics used by a few authors. In a real scenario, however, comments can originate from many different authors. Nevertheless, Qian et al. (2018) and (Mishra et al., 2018a) incorporated author information as additional features for classifiers. Qian et al. (2018) obtained the tweet history of authors and gathered tweets from a large unlabeled set using textual similarity to model intra-user and inter-user representation to aid the classification. Mishra et al. (2018a) created a community graph with nodes as authors and edges as the connections between them. From this, they obtained representations, referred to as *author profile*, using node2vec (Grover and Leskovec, 2016). However, both these approaches were applied only on the corpus provided by Waseem and Hovy (2016), which has a biased representation of authors, as discussed before. Indeed, classifiers trained on this corpus have shown poor generalizability (Wiegand et al., 2019).

Annotator bias

Annotating abusive corpora, being inherently a subjective task, is not trivial and depends on the level of exposure and expertise of annotators, their beliefs, gender, ethnicity, and other factors (Croom, 2011; Waseem, 2016; Waseem et al., 2021; Sap et al., 2022). Waseem (2016) hired both amateur annotators from a crowd-sourcing platform and expert annotators who were feminist and anti-racism activists to annotate their corpus. They observed a low inter-annotator agreement, with an agreement score of 0.57, for the amateur annotators. They suggested using full agreement rather than majority voting to incorporate the annotations from amateur annotators and seeking expert annotations only when amateur annotators disagree. Moreover, models trained on expert annotations were found to outperform those trained on amateur annotations. Breitfeller et al. (2019) showed the gender-based annotator discrepancy in estimating the degree of offensiveness of 'microaggressions' which are subtle forms of prejudiced expressions toward members

of marginalized communities, essentially being a kind of implicit abuse. They found that female and non-binary annotators were more prone to identifying microaggressions as more offensive than male annotators.

Al Kuwatly et al. (2020) analyzed annotator bias by studying the impact of data annotated by demographically distinct annotator groups on the behavior of classifiers. They obtained multiple interesting observations, some of which are: (i) the inter-annotator agreement of females was significantly lower than that of males but found no significant difference between the two classifiers; (ii) classifiers trained on English data annotated by native English speakers performed substantially better than those trained on data annotated by non-native speakers or a mix of all annotators; (iii) classifiers trained on data labeled by older annotators (over 30) had a better performance than those trained on under-30-labeled data; (iii) the classifiers performed better on data labeled by less educated (below high school) annotators compared to those labeled by better educated (above high school) ones, which is contrary to the usual expectation. This revealed that the demography of the annotators has a crucial influence on the data and the models trained on them reflect those biases. This further suggested how annotating the same data by different groups of annotators can help in estimating the bias induced.

Wich et al. (2020) analyzed annotation bias by building a graph to represent the similarity in the annotation behavior. An annotator was represented by a node in the graph and if at least one comment is labeled by two annotators, an edge was created between them. Moreover, every edge was weighted based on the similarity of the annotations. They applied a community detection algorithm (Blondel et al., 2008) to create annotator groups and similar to Al Kuwatly et al. (2020), they trained classifiers on every group. The classification performance of models trained on annotations from some of the groups differed when evaluated on data annotated by a different group, indicating possible bias in annotations. They suggested weighting the annotations from different groups based on such insights and building classifiers that model the annotator bias.

Sap et al. (2022) took a step further by studying how the identities and the attitudes or the belief systems of annotators influence their annotations from a social psychology standpoint. They measured different types of beliefs in annotators such as advocating free speech, endorsing racist beliefs, traditionalism, etc. on one hand and the belief that abusive language could be harmful, beliefs of empathy, altruism, etc. on the other hand using prior works on social science research. The authors found that some of these beliefs correlated with annotator demographics. They further demonstrated some strong correlations such as annotators with more racist beliefs were less prone to rate anti-Black language as toxic, and more prone to rate African American English as toxic. They recommended reporting annotator attitudes and demographics as a part of the data creation process.

These studies make it evident that the annotation process, quality of the data, and eventually the model performance heavily depend on demography, psychology, and other attributes associated with the annotators. To reduce the annotator bias in the abusive language corpora, it is crucial to provide clear annotation guidelines, select high-quality annotators and give proper training to the annotators (Hsueh et al., 2009; Vidgen and

Derczynski, 2020; Lopez Long et al., 2021).

Unintended social bias

The data creation process may induce unintended social bias against some demographies. For example, Dixon et al. (2018) found that certain group identity terms, like 'gay', 'homosexual', 'islam', etc., were disproportionately present in the toxic class in the Wikipedia toxicity corpus (Wulczyn et al., 2017). This would result in classifiers wrongly classifying a non-toxic comment as toxic because of the presence of these terms. This could lead to unfair applications when such unintended bias is learned by a model as it may further marginalize these groups by flagging any positive discussion involving identity terms as toxic. Since models performing differently for different identity terms can help identify such bias, the authors defined the unintended bias as the following: "*a model contains unintended bias if it performs better for comments containing some particular identity terms than for comments containing others*". To mitigate the bias, they augmented the data with new non-toxic comments containing those identity terms that were observed to have disproportionate distributions across classes. This was done to create a balance in the distributions of these terms. Furthermore, for evaluating this bias, they created a synthetic test set containing both toxic and non-toxic template statements and substituted different identity terms in those templates.

Park et al. (2018) extended this work to specifically reduce gender bias in the data. The adopted three approaches to debias a model: (i) using debiased word embeddings that remove gender stereotypical information (Bolukbasi et al., 2016); (ii) gender swapping, i.e. augmenting the training data by substituting male identities in comments with female identities and vice-versa; and (iii) transfer learning by training a model first on a larger less biased corpus and then fine-tune on the small biased corpus. By employing a mix of debiased embeddings and the data augmentation strategy, they were able to significantly reduce gender bias. Similarly, Nozza et al. (2019) followed the data augmentation strategy of Dixon et al. (2018) to reduce unintended biases against identity terms in misogyny detection.

Davidson et al. (2019) showed the presence of racial bias in some widely used abusive language corpora as the classifiers trained on them are highly prone to predict tweets written in African-American English (AAE) as abusive compared to Standard American English (SAE). This bias would be unfair to African-American users and systematically discriminate against them instead of helping them. Such bias can enter during the data collection process. When keywords are used to extract data, some communities may be disproportionately represented in abusive content and may be far from the true distributions. In some cases, the words used as indicators of abusive language may be associated with the language used by certain communities. For example, the corpus provided by Davidson et al. (2017) has been found to over-represent AAE in abusive comments (Waseem et al., 2018). Moreover, annotator bias can again increase these ethnic disparities. For example, the word 'n*gga' is frequently used by the members of

the African-American community in a harmless manner (Warner and Hirschberg, 2012a). In the corpus of Davidson et al. (2017), this word is frequently associated with abusive language as annotators interpreted the context in their own ways. Sap et al. (2019) gave a similar conclusion in a similar study. They further proposed two ways of mitigating racial bias: '*dialect* and *race* priming'. In these strategies, the annotators were provided with information about the dialect of the comments, and they were asked to take into account the possible race of the authors of the comments based on their dialects while annotating a comment. This made them less likely to label a comment written in AAE as abusive. Zhou et al. (2021) analyzed the performance of different debiasing strategies and showed that they are less effective in mitigating the bias induced by different dialects. They proposed automatic label correction of those comments that were written in AAE and marked as abusive while their corresponding translated SAE comments were predicted as non-abusive. They showed that this method was more effective at reducing such dialectal bias.

Different model training strategies for unintended social bias mitigation have been proposed in the literature. Mozafari et al. (2020) re-weighted the input samples to reduce the effects of n-grams present in the data that have high correlations with the class labels. They showed a significant reduction of racial bias with regard to the disparity in the predicted labels of comments written in AAE versus SAE. Vaidya et al. (2020) proposed a multi-task learning model that jointly learns toxicity classification along with detecting the identity terms present in the comment to alleviate the bias against the identity terms. Liu and Avci (2019) and Kennedy et al. (2020) used different post-hoc model explanation methods to regularize the importance assigned by the classifier to a set of pre-defined identity terms while performing abusive language classification. This was done to enable the abusive language detection models to give less importance to these terms and rather learn from the context associated with these terms in the comments.

Ramponi and Tonelli (2022) studied model fairness (reducing the possibility of classifying non-abusive instances containing identity terms as abusive) and robustness (improving the abusive language detection performance) in both in-distribution and out-of-distribution evaluation settings. The authors differentiated between *authentic artifacts* and *spurious artifacts*. The first one is the set of terms in the data that carry the necessary information for the prediction of abuse, while the second one is the set of terms that are wrongly correlated with the class labels. They further showed that different spurious artifacts may require different treatments, e.g. masking the spurious artifacts comprising identity terms substantially improves the fairness of models while masking non-identity terms does not have the same effect. Moreover, they demonstrated that improving model fairness can reduce the model robustness, i.e. the out-of-distribution classification performance of models and this is an important trade-off to consider. Attanasio et al. (2022) used a method called '*entropy-based attention regularization*' that penalizes terms with low self-attention entropy. This is based on the intuition that while a small entropy indicates that only a few terms are considered important, a large entropy suggests that a larger context contributes to the decision of classifiers. Their method does not require a pre-defined list of identity terms and it can reduce bias induced by any term in the

training data.

Finally, Waseem et al. (2021) mentioned that it is nearly impossible to completely debias abusive language corpora. This is because every step in the machine learning pipeline ranging from data creation, and annotation to modeling involves a subjective choice. They recommended that researchers should be aware of the fact that bias-mitigation approaches can only reduce the effect of a fraction of biases. In this thesis, we tried to address the issue of domain bias to a certain extent that can adversely affect the transferability of classifiers in cross-corpus evaluation settings.

2.5 Transfer Learning in Abusive Language Detection

Traditional machine learning involves training a single model for a specific task using the given dataset. This approach of learning in isolation requires a large number of training instances. However, owing to the time, cost, and effort involved in annotating abusive language corpora (Schmidt and Wiegand, 2017; Malmasi and Zampieri, 2018; Poletto et al., 2019), the sizes of these corpora are usually small. Since deep neural networks require a large amount of data for training (Justus et al., 2018; Huang et al., 2019), it is important to reuse prior available knowledge in related tasks or domains for learning new models for abusive language detection. This can be achieved through transfer learning, as discussed in Chapter 1.

As we have seen in Section 2.4.6, abusive language constantly evolves with time, is targeted against different communities, incorporates corpus-specific data bias, vary in terms of the writing style, vocabulary, topics discussed, and other aspects. Therefore, it is important that abusive language detection models can transfer well across such variations to be practically robust in real-world scenarios. Evaluating a model on a different corpus or domain than the one on which it has been trained forms a more realistic experimental setting and is referred to as *cross-domain* (Wiegand et al., 2019; Karan and Šnajder, 2018) or *cross-dataset (corpus)* (Swamy et al., 2019) evaluation. In this section, we discuss the existing approaches that studied the cross-corpus transferability or generalizability in abusive language detection.

2.5.1 Generalizability Without Exposure to Target

We first discuss the studies where the models learned from the source corpus are not exposed to any data from the target corpus during training but are used directly to detect abusive language on the target. We begin with discussing the works that analyzed the problem of generalizability and then discuss the solutions that have been proposed in the literature to improve it.

Analyzing the problem

Recent studies have highlighted concerns about the generalizability of existing models. Several researchers have observed that despite the impressive performance on the respective test sets from the same corpus, the performance drops significantly when evaluated on a different corpus (Agrawal and Awekar, 2018; Dadvar and Eckert, 2020; Gröndahl et al., 2018; Karan and Šnajder, 2018; Waseem et al., 2018; Swamy et al., 2019; Wiegand et al., 2019; Arango et al., 2019; Fortuna et al., 2021; Yin and Zubiaga, 2021), demonstrating the lack of generalizability. For instance, Arango et al. (2019) reported that the LSTM-based model introduced by Badjatiya et al. (2017) trained on the corpus of Waseem and Hovy (2016) suffered a massive drop from 93% macro-average F1 (macro-F1) score obtained on the corresponding test set of Waseem and Hovy (2016) to 47.5% when evaluated on the test set from Basile et al. (2019). Similarly, Gröndahl et al. (2018) observed that the performance of the LSTM-based model of Badjatiya et al. (2017) trained on the same corpus dropped to 33% macro-average F1 on Davidson et al. (2017), where the offensive and the neutral classes were combined as a single class as opposed to the hate speech class. Agrawal and Awekar (2018) showed that a BiLSTM-based model with an attention mechanism trained on a corpus from Formspring (Reynolds et al., 2011) obtained 3% F1 on Waseem and Hovy (2016) and 35% on a corpus from Wikipedia (Wulczyn et al., 2017). Swamy et al. (2019) performed cross-corpus evaluations with four corpora (Waseem and Hovy, 2016; Davidson et al., 2017; Founta et al., 2018; Zampieri et al., 2019a,b) using the BERT model and the results showed that the drop was less but widely varied ranging from 2 to 35 points in macro-F1.

Gröndahl et al. (2018) argued that the properties of the corpora are more important than the models used for generalizability studies. Wiegand et al. (2019) observed that when fastText models were trained separately on the three corpora provided by Wulczyn et al. (2017), Founta et al. (2018) and Razavi et al. (2010) and evaluated on each other, macro-F1 scores higher than 70% were obtained. However, when they were trained or tested on another set of three corpora provided by Warner and Hirschberg (2012b), Waseem and Hovy (2016) and Kumar et al. (2018), the scores obtained were usually around or below 60% on average in cross-corpus settings. They attributed this observation to the property of the first three corpora, i.e. they comprise a higher proportion of explicit abuse and involve less biased sampling. The other three corpora, on the other hand, involve biased sampling and a larger presence of implicit abuse. Swamy et al. (2019) found that the corpora with similar characteristics and those built using similar search terms performed well on each other. They observed that the corpora having larger proportions of abusive instances showed better generalizability than those with fewer abusive instances, especially when evaluated on a dissimilar corpus. Similar observations were made by Karan and Šnajder (2018) and Waseem et al. (2018). Swamy et al. (2019) further recommended that for improving the generalizability of models, they should be trained on abusive language corpora that have wide coverage of different forms of abusive phenomena rather than having only a few specific types of abuse.

Unlike Gröndahl et al. (2018), Fortuna et al. (2021) found that models do play an im-

portant role in generalization. They reported that BERT along with other Transformer-based models obtained better generalization than models like SVM and fastText in the majority of their experimental configurations. However, they still maintained that the data and especially the labeling scheme across different corpora play an equally important role. Differing from Swamy et al. (2019) and Karan and Šnajder (2018), they found that the model performance depends more on the labeling criteria present in the data, rather than the proportion of the abusive content.

Improving generalizability

Wiegand et al. (2018a) improved cross-corpus abusive language detection through the creation of a new lexicon of abusive words automatically. They first constructed a base lexicon of negative polar expressions, which were annotated for the presence of abusive words. They then expanded the base lexicon by training an SVM using different semantic and linguistic features and applied it to a set of unlabeled negative polar expressions. The authors demonstrated substantially improved generalizability in cross-corpus experiments using an SVM classifier trained on the features derived from their expanded lexicon. Pamungkas and Patti (2019) used features from a different lexicon of abusive words called `HurtLex` (Bassignana et al., 2018) along with linear SVM classifier and LSTM-based models for cross-corpus experiments. They found that LSTM-based models gave better performance, which got further improved when `HurtLex` was used. They also argued that models trained on corpora with general abusive language forms are more robust to detect specific kinds of abuse. Pamungkas et al. (2020) performed further analysis and used BERT along with other models such as SVM and LSTM with and without `Hurtlex`. They concluded that BERT gave the best performance for cross-corpus transfer among all the models. Similarly, Koufakou et al. (2020) combined features obtained from `Hurtlex` with representations obtained from the BERT model and showed improvements in cross-corpus performance.

Nejadgholi and Kiritchenko (2020) used topic modeling and manually analyzed the top ten words and their coherence in every topic to find the domain-specific non-generalizable topics in a corpus from Wikipedia (an extension of the corpus from Wulczyn et al. (2017)). An example of a non-generalizable topic is one that is specific to the Wikipedia platform with top words like 'page', 'article', 'edit', 'talk', etc. that may not occur in a corpus from another platform. They assigned the highest probable topic to every instance and found that removing the instances associated with the non-generalizable topics improved the out-of-domain performance with a BERT model. They recommended that such manual inspection of topics, filtering out the non-generalizable topics, and augmenting data with useful topics should be carried out iteratively during data collection or before annotation for making the corpus robust. Since the Wikipedia corpus comprises general abusive language like 'toxicity', models trained on it were shown to perform well on the out-of-domain instances associated with general abuse. However, the performance dropped when dealing with a sub-category of abuse such as 'hate speech' as the Wikipedia corpus does not represent hateful topics directed toward different communities sufficiently well.

This observation slightly contradicts the argument of Pamungkas and Patti (2019) as mentioned above.

Wang et al. (2020) proposed learning a generic model for different types of abuse by incorporating representations of different aspects of abusive language such as target, content, and linguistic behavior. The authors proposed a 'cross-attention gate flow' mechanism using a cross-transformer encoder to mutually learn and refine these representations. They showed improved performance on abusive language corpora sampled from three different online platforms but they did not perform any cross-corpus experiments. In another attempt to learn a generic model for abusive language, Caselli et al. (2021) continued pre-training the BERT model on a large-scale corpus of Reddit comments from communities banned for being abusive using the Masked Language Model (MLM) objective of BERT and presented a new abuse-inclined model called *HateBERT*. They demonstrated better generalizability using HateBERT in cross-corpus experiments, especially when moving from a corpus with general abusive phenomena to a specific one.

Markov and Daelemans (2021) used a majority-voting ensemble of BERT, RoBERTa, and SVM with different robust sets of features such as part-of-speech (POS) tags, emotion, and abusive lexicons, etc. and demonstrated improvements in cross-corpus performance. Their analysis revealed that the improvement was mostly due to the decrease in false positives, i.e. classifying a non-abusive instance as abusive. Kennedy et al. (2020) also reduced the false positives in out-of-domain data by penalizing the importance assigned to identity terms (like *jews*, *africans*, etc.) using a model explanation method called Sampling and Occlusion (SOC) (Jin et al., 2020).

Chiril et al. (2022) investigated the impact of training on a combination of abusive corpora having specific topics or targets, like the topic of 'racism' in Waseem and Hovy (2016) and testing on an unseen topic like 'misogyny' in Basile et al. (2019). They experimented with many different models and found that the models trained on different specific abusive topics or targets can generalize to a decent extent on a corpus with an unseen topic or target; the best performing model being CNN coupled with the word embedding model of fastText. Toraman et al. (2022) introduced a large-scale corpus from Twitter for English and Turkish having a balanced number of tweets across five domains, where they use the term 'domain' to denote different topics, namely *religion*, *gender*, *racism*, *politics*, and *sports*. They demonstrated that the cross-domain (topic) transfer performance was usually high on average. However, gender-based hate tweets were difficult to classify by other domains and the topic of sports could not generalize well to other domains.

2.5.2 Domain Adaptation With Exposure to Target

In this sub-section, we discuss the scenarios where some amount of labeled or unlabeled data from the target corpus is available. This data is used by the model to transfer or adapt the knowledge learned from a source corpus to improve the cross-corpus performance on the target. We divide this into two parts: '*with target corpus labeled data*' and '*with target corpus unlabeled data*'.

With target corpus labeled data

Agrawal and Awekar (2018) experimented with different flavors of transfer learning, amongst which two strategies used the target corpus labeled data, namely 'feature level transfer learning' and 'model level transfer learning' using a BiLSTM-based model with attention. In the former, they first trained a model on the source corpus and then transferred only the learned word embeddings to the target corpus to train a new model. In the latter, they transferred both the learned word embeddings and the model weights for training another model on the target corpus. They found that transferring only word embeddings resulted in significant improvements over the direct evaluation of the source model on the target. However, transferring the additional model weights gave a similar performance as transferring only the word embeddings, and there were no further gains. This indicated that the learned word embeddings were the most crucial knowledge transferred. For adapting pre-trained word embeddings to the abusive language task, Basile (2020) used the concept of 'weirdness index' (Ahmad et al., 1999) for words and computed the 'polarized weirdness' (Florio et al., 2020) with respect to a specific class label for every word. This is computed as the ratio of the relative frequency of a word (occurrence of a word in class c/total number of words in class c) present in the instances belonging to one class over that in the other class. Basile (2020) adapted the pre-trained word embeddings in such a way that the distance between pairs of words in the embedding space was determined by their respective polarized weirdness scores. Classification using these adapted word embeddings were found to considerably improve recall in abusive language detection task.

Karan and Šnajder (2018) used the *Frustatingly Easy Domain Adaptation* (FEDA) framework (Daumé III, 2007) with linear SVM to augment the given corpus with knowledge from a different corpus which significantly improved the performance in the original corpus. Mozafari et al. (2019) analyzed different fine-tuning methods, as discussed in Section 2.3.2, to transfer knowledge from the pre-trained BERT model to the abusive language corpora. Isaksen and Gambäck (2020) used both 'base' and 'large' versions of the pre-trained BERT model along with further training of BERT on different unlabeled abusive language corpora except the target corpora. Further training was done using the language model pre-training objectives of the Masked Language Model (MLM) and next sentence prediction. All these BERT model variants were then used to transfer knowledge for detecting and differentiating between the hateful and offensive language in the target. Contrary to the expectation, they found that models with general language understanding performed better on the hateful class of the target than models that were further trained on different unlabeled abusive language corpora using the language model objectives.

Waseem et al. (2018) used the multi-task learning (MTL) approach to investigate whether jointly learning a shared representation from two or more corpora, having different annotation schemes, can transfer knowledge in a better way compared to training on a direct combination of all the available corpora. They used BoW and subword embeddings with feed-forward neural networks, where the model comprised both shared

parameters that were learned jointly and task-specific parameters for every corpus. With the MTL approach, the authors obtained improvements compared to training only on a single corpus and testing on another as well as training on a direct combination of the available corpora. Similar approaches using MTL were adopted by Rizoiu et al. (2019) who used ELMo (Embeddings from Language Models) (Peters et al., 2018) as the word embedding model and a bi-LSTM layer, d'Sa et al. (2022) who used the BERT model and Aldjanabi et al. (2021) who used BERT models that were pre-trained on the Arabic language, like AraBERT (Antoun et al., 2020) and MARBERT (Abdul-Mageed et al., 2021).

Corazza et al. (2019) analyzed whether training abusive language classifiers on combined data from multiple social media platforms (using data sampled from Twitter, Facebook, Instagram, Whatsapp) can help in improving their performance on a single platform. They found that training on the combined data using a GRU-based model did improve performance but the gains were not always substantial compared to training only on the data from the corresponding platform. This indicates that social media platforms have their own peculiar use of language that may not be present in a combined generic corpus. Besides, for the corpus sampled from Twitter, the authors showed that training on the single-platform data from Twitter always outperformed the performance from the combined training data. Ozler et al. (2020) compared different multi-domain experimental configurations using the BERT model with a corpus of news comments and two Twitter corpora. These were *single* – a classifier fine-tuned separately for every corpus, *joint* – a classifier fine-tuned on a combination of different corpora, *joint →single* – a classifier first fine-tuned on the combined set and then further fine-tuned on every individual corpus. They found that their *joint* model outperformed *joint →single* model in most cases.

Chiril et al. (2022) investigated the ability of different models in transferring common abusive knowledge from topic-generic abusive corpora (eg. Davidson et al. (2017) with abusive labels of hateful and offensive) to topic-specific corpora having particular types of abuse (eg. Waseem and Hovy (2016) with the abusive labels of racism and sexism). They compared the performance of models trained on a topic-generic corpus and tested on a topic-specific corpus to the models trained on a combination of all the topic-specific corpora and tested on every constituent topic-specific corpus separately. The second approach was found to be more robust than the first one.

Sarwar et al. (2022) proposed a nearest-neighborhood framework using Transformer-based models for cross-lingual abusive language detection in low-resource scenarios, i.e. with limited labeled data from the target. They modeled the interaction between instances from the target corpus and their nearest neighbors from the source corpus based on their label agreement to improve the performance in the target corpus. We have discussed this approach in detail in Chapter 7. Mozafari et al. (2022) applied a meta-learning-based approach called Proto-MAML (MAML: Model-Agnostic Meta-Learning) (Triantafillou et al., 2020), with some modifications, for cross-lingual hate speech and offensive language detection amidst limited labeled data in the target language. A meta-learning algorithm evaluates new tasks at test time after learning to learn on a variety

of training tasks (Thrun and Pratt, 1998).

With target corpus unlabeled data

When there is no availability of labeled instances from the target corpus but there is access to unlabeled target data, Unsupervised Domain Adaptation (UDA) methods can be devised that address the domain shift. There are only a few existing approaches that have adopted UDA for abusive language detection.

Glavaš et al. (2020) studied both cross-domain and cross-lingual abusive language detection where they continued pre-training of Transformer-based models on the abusive corpora in the target language using the Masked Language Model (MLM) objective, as an unsupervised adaptation method. They showed that such domain and language adaptation can provide additional performance gains. Bashar et al. (2021) performed a progressive adaptation of an LSTM-based model using language model pre-training first on a general unlabeled corpus, then on unlabeled corpora that are topically similar to the source and the target domain, and subsequently learned a classifier on the source domain data. Sarwar and Murdock (2022) introduced a data augmentation-based UDA approach that constructed weakly labeled data from a negative emotion corpus (Go et al., 2009) to augment the source corpus. They used different models such as character CNNs, BiLSTMs, and BERT for evaluating their approach.

Nejadgholi et al. (2022) proposed using a model interpretability technique called Testing Concept Activation Vector (TCAV) (Kim et al., 2018) to guide the selection of the most informative out-of-domain instances for augmenting the source training data. Specifically, they used TCAV vectors to quantify the sensitivity of trained classifiers to the explicit and implicit abusive instances and found that augmenting with implicitly abusive instances from the out-of-domain data can improve the cross-domain performance. Subsequently, they formulated a new metric called the *degree of explicitness* to identify a small set of unlabeled and informative implicitly abusive out-of-domain instances to augment the source data, which improved the performance compared to classification confidence-based selection of instances. Ludwig et al. (2022) categorized a domain as the set of instances involving abuse against a specific target community. They investigated different UDA approaches in this setting. We have described UDA approaches in more detail in Chapter 5.

Table 2.2: Summary of the existing approaches for addressing transferability or generalizability in abusive language detection (in chronological order).

Reference	Approach
Wiegand et al. (2018a)	Created a new lexicon of abusive words automatically; trained an SVM using the features derived from their expanded lexicon for improving generalizability.
Agrawal and Awekar (2018)	Investigated three flavors of transfer learning using a BiLSTM-based model: directly evaluating a source model on the target corpus, transferring only the learned word embedding from the source to the target for further training; transferring both learned word embeddings and model weights from source to the target for training.
Karan and Šnajder (2018)	Augmented the given corpus with knowledge from a different corpus using a domain adaptation approach called FEDA with linear SVM.
Waseem et al. (2018)	Investigated an MTL approach that jointly learned a shared representation from two or more corpora, with different annotation schemes, to transfer knowledge across corpora, using BoW and subword embeddings with feed-forward neural networks.
Mozafari et al. (2019)	Analyzed different fine-tuning methods to transfer knowledge from the pre-trained BERT model to abusive language corpora.
Rizoiu et al. (2019)	Proposed an MTL approach based on ELMo embeddings to learn shared information across two corpora.
Corazza et al. (2019)	Analyzed whether training on combined data from multiple social media platforms can help in improving their performance on a single platform; found that social media platforms have their own peculiar language style that may not always be captured through a combined generic corpus.
Pamungkas and Patti (2019)	Infused features from a lexicon of abusive words called `HurtLex` along with linear SVM and LSTM-based models and found improvement in cross-corpus experiments.
Pamungkas et al. (2020)	Compared the performance using BERT and other models with `HurtLex` and found that BERT outperformed other models in cross-corpus transfer for misogyny detection in English.
Ozler et al. (2020)	Compared different multi-domain experimental settings using the BERT model and found that training on a combined corpus from different domains gave the best performance.

Glavaš et al. (2020)	Studied both cross-domain and cross-lingual abusive language detection where they continued pre-training of transformer-based models on the abusive corpora in the target language using the MLM objective.
Koufakou et al. (2020)	Combined features obtained from `HurtLex` with representations from BERT and demonstrated improvements in cross-corpus performance.
Nejadgholi and Kiritchenko (2020)	Used topic modeling and manual analysis to identify domain-specific topics and proposed to remove instances associated with these topics to improve generalizability.
Wang et al. (2020)	Proposed learning a generic model for different types of abuse by incorporating multi-aspect (target, content, and linguistic behavior) representations of abusive language.
Kennedy et al. (2020)	Proposed a regularization method to reduce the importance assigned to identity terms for improving out-of-domain generalization.
Isaksen and Gambäck (2020)	Transferred knowledge using pre-trained BERT model and BERT trained further on unlabeled abusive language corpora different from the target using its language model objectives. Models with general language understanding were found to perform better on the hateful class of the target than models further trained on abusive language corpora.
Caselli et al. (2021)	Retrained BERT on a large-scale corpus of English Reddit comments from communities banned for being abusive, using the MLM objective, and called it HateBERT, which demonstrated better generalizability.
Markov and Daelemans (2021)	Used a majority-voting ensemble of BERT, RoBERTa, and SVM with different robust sets of features and demonstrated improvements in cross-corpus performance due to a reduced number of false positives.
Aldjanabi et al. (2021)	Presented an MTL approach for cross-corpus abusive language detection using BERT models pre-trained on the Arabic language.
Bashar et al. (2021)	Performed unsupervised progressive domain adaptation of an LSTM-based model on different unlabeled corpora and then learned a classifier on the source domain data to transfer knowledge to the target domain.
d'Sa et al. (2022)	Learned shared and corpus-specific information using an MTL approach with BERT for transferring knowledge across different corpora.

Chiril et al. (2022)	Performed an extensive investigation of different models' ability to transfer common abusive knowledge from topic-generic abusive corpora to topic-specific corpora having particular types of abuse, the generalizability of training on a combination of abusive corpora having specific topics or targets, and other analysis.
Toraman et al. (2022)	Introduced a large-scale corpus from Twitter for English and Turkish that have a balanced number of tweets across five different domains or topics, for cross-domain experiments.
Nejadgholi et al. (2022)	Proposed using a model interpretability technique called TCAV to guide the selection of the most informative implicitly abusive out-of-domain instances for augmenting the source training data, demonstrating improved cross-domain performance.
Sarwar and Murdock (2022)	Introduced a data augmentation-based UDA approach that constructed weakly labeled data from a negative emotion corpus to augment the source corpus.
Sarwar et al. (2022)	Proposed a nearest-neighborhood framework using transformed-based models by modeling the interaction between instances from the target corpus and their nearest neighbors from the source corpus based on their label agreement to improve the performance on the target corpus.
Ludwig et al. (2022)	Investigated different UDA approaches in cross-domain settings where they categorized a domain as the set of instances involving abuse against a specific target community.
Mozafari et al. (2022)	Applied a meta-learning-based approach for cross-lingual hate speech detection with limited labeled data in the target language.

Lastly, we present a summary of the existing approaches that addressed transfer learning in the abusive language detection task in Table 2.2. Pamungkas et al. (2021) presents a comprehensive survey of the prior literature on cross-corpus abusive language detection and identified that data bias is one of the major challenges in tackling cross-corpus abuse, which remains an open problem. They concluded that even though the research in the area of transfer learning in abusive language detection is still in its early stage of development, the prior literature stresses the urgency of addressing this problem. In this thesis, we study this important research problem, where we propose solutions to alleviate the issue of data bias caused due to spurious correlations that can disrupt the adaptation to a new corpus, as one of our contributions.

2.6 Summary

In this chapter, we presented the landscape of existing literature on abusive language detection. We began by discussing the nuances in the definition of abusive language, especially that of hate speech. Then a broad overview of the evolution of abusive language classifiers was presented from the pre-deep learning era to the current state-of-the-art transformer-based models used in this task. Subsequently, some of the major challenges across different dimensions of abusive language detection were discussed that make it a complex task. We found that an unclear definition of abusive language reduces the quality of annotation. Moreover, subtle forms of abuse, the subjective nature of the task, non-standard language, the absence of additional contextual information such as the parent comment or images associated with comments, diverse labeling schemes, and other factors contribute to the task complexity.

We also discussed how some of these challenges can adversely affect the transferability of abusive language detection systems. One of the main issues that cause limited cross-corpus performance is the variety of biases induced in the corpus from different sources. These include sampling/domain bias, author bias, social bias resulting from the data creation process, and annotator bias resulting from factors like the annotators' demography and their level of expertise. Moreover, the evolution of the language used in online platforms, the advent of new words, and real-world events that trigger new topics of discussion reduce the generalizability and adaptivity of models. Finally, we presented the existing studies that analyzed and attempted to address this crucial and open problem of transfer learning in abusive language detection.

3 Models and Corpora

3.1 Introduction

This chapter presents the different deep neural network-based models that we have used in the remaining part of the thesis for obtaining pre-trained representations. Furthermore, this chapter provides a detailed description of the different corpora used for our experiments, the text pre-processing performed on them, and the evaluation metric used for reporting the model performance throughout the thesis.

3.2 Deep Neural Network Models

We start with describing the Deep Neural Network (DNN) models used in our proposed approaches either for obtaining the pre-trained representations or as the underlying pre-trained models that are fine-tuned for our tasks. Since all of these models use Transformer layers as their building blocks, we first give an overview of these layers.

3.2.1 Transformer Layers

Transformers were introduced by Vaswani et al. (2017) originally for neural machine translation applications. They demonstrated superior performance on this task, while significantly improving the speed and parallelizability. One of the factors that facilitate this is their ability to process the entire sequence of text at once. This is achieved by the use of the self-attention mechanism, a strategy for capturing the dependencies of a word or token with other words or tokens in the sequence. The Transformer model is built with an encoder and a decoder – each containing a stack of Transformer layers. The representations of the input text sequence are estimated using the encoder. The decoder uses the representations learned by the encoder to yield the output sequence.

A Transformer layer in the encoder consists of two sub-layers: *multi-head self-attention* and *feed forward neural network*. Although a Transformer layer in the decoder has both of those sub-layers, it also features a *multi-head encoder-decoder attention* sub-layer between them that aids in focusing on relevant portions of the input sentence. The multi-head self-attention helps in jointly attending to representations from several sub-spaces, which captures various associations between words or tokens. The outputs of the attention sub-layer are subjected to transformations with a ReLU (Rectified Linear Unit) activation using the feed-forward network. In addition, every sub-layer in both the encoder and decoder has a residual connection (He et al., 2016) surrounding it, which is followed by layer normalization (Ba et al., 2016). The multi-head self-attention sub-layer in the

decoder is actually a *masked multi-head self-attention* sub-layer. This means that the sub-layer is only permitted to focus on earlier positions in the obtained output sequence, which is achieved by masking future positions. Furthermore, to incorporate the order of words in the input sequence, the Transformer adds a vector to each input embedding. This vector, called *positional encoding*, encodes the information about the position of each word. This is also done for the decoder inputs, which are the previous outputs of the decoder. Figure 3.1 provides an illustration of the Transformer layers.

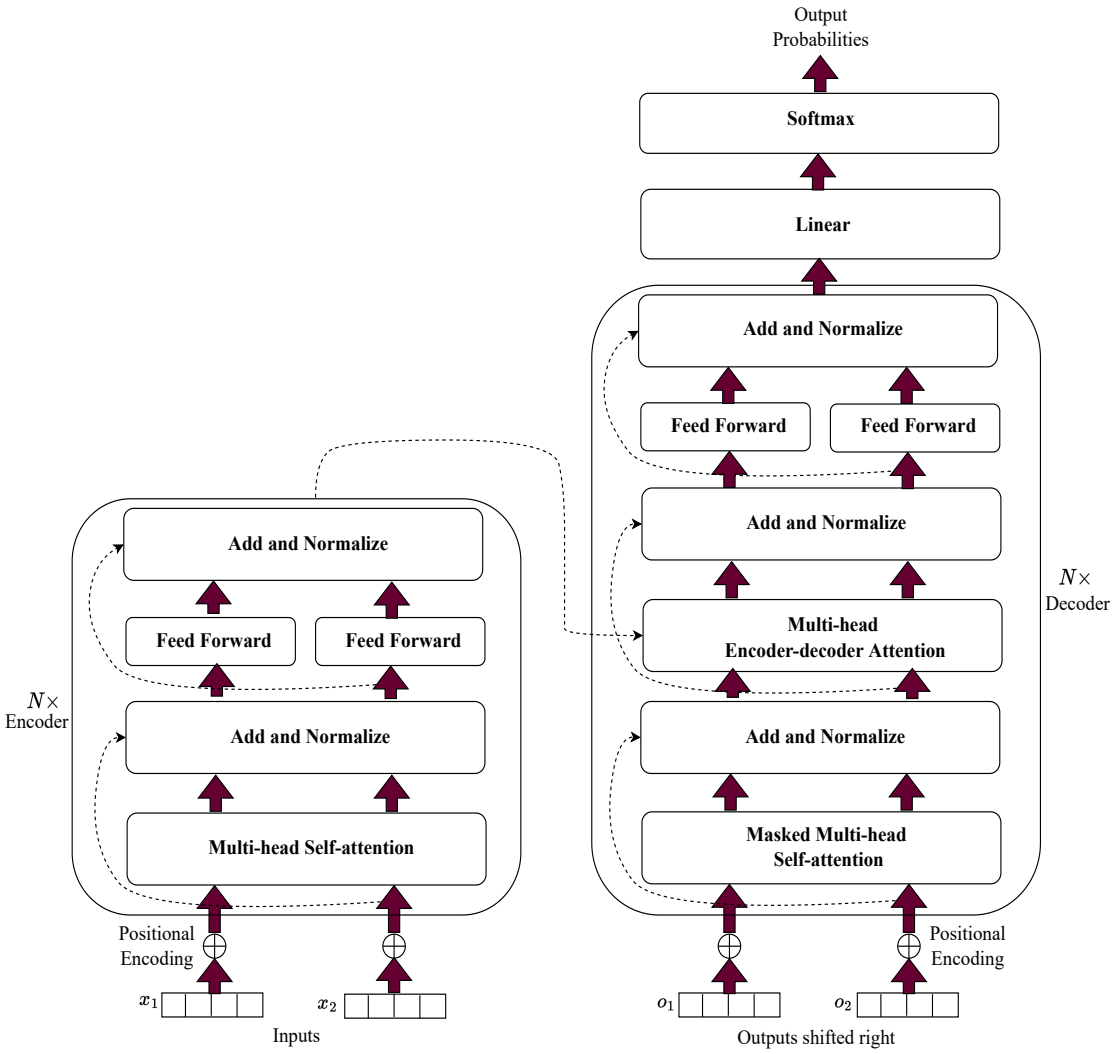


Figure 3.1: Illustration of the Transformer layers with N encoder and decoder layers.

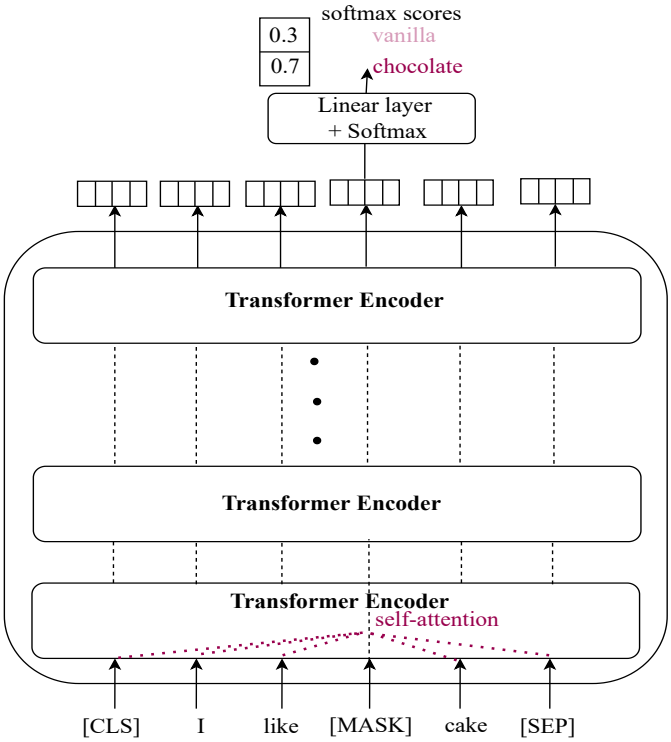


Figure 3.2: Illustration of the Masked Language Model pre-training in BERT.

3.2.2 Bidirectional Encoder Representations from Transformers (BERT)

BERT (Devlin et al., 2019) is a Transformer-based model that uses only the encoder part of the Transformer. It is intended to jointly condition on both left and right contexts in all layers in order to pre-train deep bidirectional representations from the unlabeled text. This is achieved by pre-training through the *Masked Language Model* (MLM) objective, where the self-attention mechanism takes into consideration the context from both the left and right directions of a particular masked token to predict it. It also employs a *next sentence prediction* task that simultaneously pre-trains on text-pair representations in addition to the masked language model.

The MLM pre-training entails corrupting a small percentage of input tokens and predicting those masked tokens at the model's output. To mitigate the mismatch between the pre-training and fine-tuning stages, the "masked" tokens, which form 15% of the tokens in a sequence, are replaced by the `[MASK]` token only 80% of the time, while they are replaced by random tokens 10% of the time and the original token is retained for the remaining 10% of the time. The next sentence prediction task helps in capturing the relationship between two sentences and is particularly helpful for tasks like Question Answering and Natural Language Inference. It involves passing a combination of two sentences `A` and `B`, separated by the `[SEP]` token, as input to the model. The model is then trained to predict whether the sentence `B` follows the sentence `A` or not in the original

document. The pre-training of the BERT model with the MLM objective is illustrated in Figure 3.2.

The BERT model uses WordPieces (Wu et al., 2016) as inputs to the model. This tokenization makes use of a fixed-sized vocabulary to split a word into a set of common subword tokens if the word is not found in the vocabulary. For instance, the word 'playing' may be split into the word-piece tokens 'play' and '##ing' if it is not present in the vocabulary. BERT uses a vocabulary size of 30,000 tokens. A special `[CLS]` token is added at the beginning of a sequence of inputs, whose final state is used as the representation of the entire sequence for classification tasks. This is because the self-attention mechanism enables the output of the `[CLS]` token to capture the aggregated sequence representation. When the pre-trained BERT model is fine-tuned for classification, a new dense layer (classification layer) is added over the model's final representation layer and the parameters of the pre-trained model along with the classification layer are updated using the dataset for the task. The final layer representation of the `[CLS]` token is passed as input to the classification layer. Figure 3.3 illustrates the fine-tuning of BERT for classification.

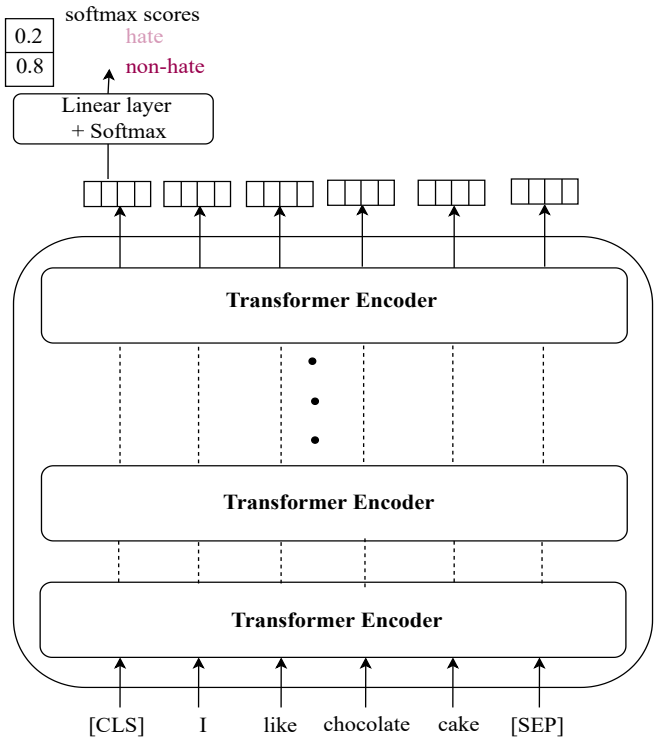


Figure 3.3: Illustration of the fine-tuning of BERT for the classification task.

Throughout the thesis, while using BERT (Chapter 6 and 7), we fine-tune the pre-trained 'BERT-base-uncased' model for our experiments. This model has a stack of 12 Transformer encoder layers with 110M trainable parameters and uses an embedding di-

mension of 768. BERT is pre-trained over the BooksCorpus (Zhu et al., 2015) and English Wikipedia, comprising 800M and 2,500M words, respectively. For our implementation, we make use of Huggingface's Transformers API (Wolf et al., 2020).

3.2.3 HateBERT

The HateBERT model was introduced by Caselli et al. (2021). This is basically a BERT model 'retrained', i.e. further pre-trained, on a large-scale dataset of Reddit comments using the MLM objective. This dataset is called *RAL-E*: the Reddit Abusive Language English dataset. Amongst a long list of banned communities in English, collected from the official posts by Reddit admins and Wikipedia articles[1], Caselli et al. (2021) chose the communities that were found to host or promote abusive content. They then gathered the posts from these communities by crawling a publicly accessible collection of Reddit comments[2]. The BERT-base-uncased model was subsequently re-trained with the MLM objective over 1,478,348 messages from the RAL-E dataset. As a consequence, the resulting model of HateBERT shifted BERT to incorporate two orientations: (i) language variety inclined towards social media; and (ii) polarity (i.e., offensive, abusive, and hateful). HateBERT was found to be more likely to predict abusive tokens corresponding to the masked tokens, rather than the relatively neutral tokens predicted by BERT. Furthermore, HateBERT demonstrated consistent improvement over BERT for in-dataset evaluations and better generalizability when trained on datasets with generic abusive language phenomena and evaluated on datasets with a special category of abusive language, e.g. hate speech. We fine-tune the pre-trained HateBERT model in Chapter 4 and 5.

3.2.4 Universal Sentence Encoder (USE)

USE (Cer et al., 2018) was introduced to yield pre-trained sentence embedding vectors that give fixed-length numerical representations for a sentence. Two model architectures were explored in USE for encoding sentences into embedding vectors: Transformer encoder and Deep Averaging Network (DAN) (Iyyer et al., 2015). The latter computes an average of the embeddings for words and bi-grams in the sentence. It then passes the average embedding through a feed-forward deep neural network to obtain sentence embeddings. While the Transformer-based sentence encoding model results in higher accuracy in transfer learning tasks, the DAN-based model helps reduce resource consumption at the cost of a little drop in accuracy. Both these variants of USE are trained using multi-task learning, whereby sentence embeddings are passed as input to task-specific layers corresponding to multiple downstream tasks. This is done to obtain more generic sentence embeddings. These downstream tasks include conversational input-response (Henderson et al., 2017), classification, skip-thought (Kiros et al., 2015), etc. Different data sources, such as Wikipedia, web-crawled data, discussion forums, and the Stanford

[1] https://en.wikipedia.org/wiki/Controversial_Reddit_communities

[2] https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/

Natural Language Inference (SNLI) corpus (Bowman et al., 2015) were used for training the models. We employ the Transformer-based pre-trained sentence embeddings from USE for obtaining topic representations in Chapter 4.

3.2.5 Sentence-BERT (SBERT)

SBERT (Reimers and Gurevych, 2019) is an alteration of the pre-trained BERT model that also produces sentence embedding vectors. Reimers and Gurevych (2019) observed that using BERT for the semantic textual similarity task, which requires comparing two sentences, is computationally heavy. This is because BERT accepts a pair of sentences as input to the model to compute their semantic similarity. Therefore, in a collection of $n = 10,000$ sentences, searching for a pair of sentences with the highest similarity requires $n(n-1)/2 \approx 50$M inference computations. This makes BERT sub-optimal or unsuitable for tasks such as large-scale semantic similarity, clustering, and information retrieval using semantic similarity.

SBERT reduces this computation time drastically by computing sentence embeddings that can be simply compared using cosine similarity. For this, it fine-tunes BERT using siamese and triplet network architectures (Schroff et al., 2015). This results in potentially reducing the computation time for finding the most similar pairs of sentences in a collection of 10K sentences from 65 hours with BERT to 5 seconds with SBERT. Reimers and Gurevych (2019) experimented with three different objective functions for learning the network parameters: (i) *classification objective* function that accepts a concatenation of sentence embeddings u, v and their element-wise difference $|u - v|$, multiplies it with a trainable weight, and uses a cross-entropy loss function to update the parameters; (ii) *regression objective* function that computes cosine-similarity between u and v and uses mean-squared loss as the loss function; (iii) *triplet objective* function that takes an anchor sentence a, a positive sentence p, and a negative sentence n, and trains the network such that the distance between a and p is less than a and n. SBERT was trained on a combined dataset from SNLI (Bowman et al., 2015) and the Multi-Genre NLI (Williams et al., 2018). The first one contains 570K pairs of sentences labeled as 'contradiction', 'entailment', and 'neutral'. The second dataset comprises 430K pairs of sentences covering a variety of genres of spoken and written text. We use SBERT for obtaining neighborhood information in Chapter 7.

3.3 Corpus Details

In this section, we describe the different corpora used for our experiments in different chapters of the thesis. We consider five publicly available and widely used abusive language corpora in English and call them *Davidson* (Davidson et al., 2017), *HatEval* (Basile et al., 2019), *Waseem* (Waseem and Hovy, 2016), *Dynamic* (Vidgen et al., 2021c), and *Ethos* (Mollas et al., 2022). The first three were sampled from Twitter. The fourth corpus

was dynamically generated by trained annotators using a platform called `Dynabench`[3]. The fifth corpus was collected from YouTube and Reddit comments. These corpora were sampled during different time periods by the respective authors, with different sampling strategies. The statistics of the individual corpora are presented in Table 3.1. A list of the top ten most frequent words in each of the corpora and those in the hateful/abusive class, except stop words, are present in Table 3.2.

Corpus	**Original classes mapped to Hate/Abuse**	**Original classes mapped to Non hate/Non abuse**	**Average comment length**	**Hate/ Abuse %**
Davidson	hate speech, offensive	neither	14.1	83.2
HatEval	hateful	non-hateful	21.3	42.1
Waseem	racism, sexism	none	14.7	26.8
Dynamic^{v1}	hate	not hate	18.7	54.4
Dynamic^{v2}	hate	not hate	25.2	53.9
Ethos	contains hate speech	no hate speech	21.0	43.4

Table 3.1: Class-mapping and statistics of the corpora used (average comment length is calculated in terms of word numbers).

Corpus	**Frequent words in the corpus**	**Frequent words in abusive/ hateful instances**
Davidson	b*tch, b*tches, like, h*es, p*ssy, h*o, got, ass, f*ck, shit	b*tch, h*e, f*ck, p*ssy, n*gga, ass, f*ck, shit, f*ggot, white
HatEval	b*tch, women, refugees, #buildthewall, immigrant, immigration, illegal, men, migrants, h*e	woman, refugee, immigrant, trump, #buildthatwall, illegal, b*tch, f*ck, immigration, stop
Waseem	#mkr, #notsexist, kat, women, like, andre, get, people, one, think	#notsexist, #mkr, female, girl, kat, men, woman, feminist, call, like
Dynamic^{v1}	people, black, women, f*cking, like, love, think, white, get, want	people, women, black, like, f*cking, white, love, think, want, get
Dynamic^{v2}	people, like, women, black, f*cking, get, think, want, white, one	people, women, like, black, get, white, want, think, men, f*cking
Ethos	people, like, white, get, women, f*cking, black, know, one, f*ck	people, like, white, get, f*cking, f*ck, women, islam, kill, black

Table 3.2: Top ten most frequent words in the corpora after removing the stop-words.

[3] https://dynabench.org/

3.3.1 Davidson

This corpus was collected by selecting tweets that contain terms from the hate speech lexicon *hatebase*[4]. Davidson et al. (2017) used the Twitter API to find Twitter users (around 33K users) who posted these tweets and extracted their timelines resulting in a set of 85.4M tweets. Then they collected a random sample of around 25K tweets from this set that contained terms from the lexicon. They were then manually annotated by crowd-sourcing originally into three labels: 'hate speech' (1.4K), 'offensive' (19.2K) but not hate speech, and 'neither' (4.2K). Hate speech was defined by the authors as '*language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group*' (Davidson et al., 2017). A tweet was annotated by at least three people, and the inter-annotator agreement score, provided by the crowd-sourcing platform CrowdFlower, was reported to be 92%. Waseem et al. (2018) observed that tweets in *Davidson* are mostly written in the United States of America. They also observed that the main targets of racism in these tweets are African Americans.

For our cross-corpus experiments, we merge the classes 'hate speech' and 'offensive' into the class *abuse* and map the remaining tweets into the class *non-abuse* following Wiegand et al. (2019); Swamy et al. (2019), and Karan and Šnajder (2018) to maintain uniformity in the number of classes across corpora. The 'abuse' class covers around 83.2% of the total number of instances. The authors do not provide standard train-validation-test splits for the corpus.

3.3.2 HatEval

This Twitter corpus introduced by Basile et al. (2019) was used to organize a shared task called 'SemEval 2019 task 5', which dealt with the detection of hate against women and immigrants. The task involved two corpora, one in English and the other in Spanish. Most of the tweets were sampled from July to September 2018, with the exception of tweets targeting women. Tweets directed against women were mainly collected from previous challenges on misogyny detection (Fersini et al., 2018b). The English part of the corpus consists of 13K samples and was gathered from Twitter using a combination of sampling strategies. These included extracting tweets from both victims and promoters of hate speech, keyword-based sampling with the use of both neutral and derogatory words against targets, and polarized hashtags. The most frequently occurring words in the gathered tweets are *migrant*, *refugee*, *#buildthatwall*, *b*tch*, *h*e*, *women*. The authors used the corpus for two tasks: a binary task of hate speech detection with classes 'hateful' (5.5K) and 'non-hateful' (7.5K), and a task for identifying the target range (whether the target is an individual or a group) and aggressiveness (if hate speech occurs, whether the tweeter is aggressive or not).

The tweets were annotated using a crowd-sourcing platform and every tweet was annotated by at least thee annotators. The guideline for annotating hateful tweets against

[4] https://www.hatebase.org

immigrants mentioned that two aspects should be jointly taken into consideration for deciding whether a tweet is hateful: (i) the target of hate should be immigrants or refugees, and (ii) it should propagate, instigate, justify, or support violence or hatred toward the target, or should seek to dehumanize, harm, or intimidate the target. Likewise, for the hateful tweets against women, the definition provided to the annotators states that it should be 'a text that expresses hating towards women in particular (in the form of insulting, sexual harassment, threats of violence, stereotype, objectification, and negation of male responsibility)' (Basile et al., 2019). The crowd-sourcing platform reported an 'average confidence score', which combines the inter-annotator agreement and reliability of the annotator, for the task of hate speech detection in English as 83%. After obtaining the crowd-sourcing-based annotated data, Basile et al. (2019) further added two more expert annotators. These expert annotators for the English part of the corpus were native or near-native speakers of British English and had a substantial experience in similar annotation procedures. The final label was provided by taking a majority of the labels assigned by both the crowd-sourced annotators and the expert annotators.

We use the English part of the corpus and the tweets annotated for the binary task of hate speech detection for our experiments. The 'hateful' class covers around 42.1% of the corpus. The corpus provides standard splits, with 9K instances for training, 1K for validation, and 3K for testing. However, we remove the instances containing only URLs (see Section 3.4 on pre-processing), which reduces the train instances from 9000 to 8993.

3.3.3 Waseem

Waseem and Hovy (2016) sampled this Twitter corpus by first doing an initial search using common terms and slurs associated with hate against sexual, gender, religious and ethnic minorities. To create the final corpus, the Twitter API was then queried using keywords that occurred frequently in the tweets obtained from the initial search and their active users. Waseem and Hovy (2016) annotated around 16K tweets first and then got them reviewed by another annotator (a non-activist feminist female student in gender studies). The inter-annotator agreement was reported to be 84%. The following criteria were used to identify sexist and racist hate speech (the points are exactly quoted from Waseem and Hovy (2016)):

1. uses a sexist or racial slur.
2. attacks a minority.
3. seeks to silence a minority.
4. criticizes a minority (without a well founded argument).
5. promotes, but does not directly use, hate speech or violent crime.
6. criticizes a minority and uses a straw man argument.
7. blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims.
8. shows support of problematic hash tags. E.g. "#BanIslam", "#whoriental", "#whitegenocide"

9. negatively stereotypes a minority.
10. defends xenophobia or sexism.
11. contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria.

A major portion of hateful comments in this corpus is subtle and does not involve profane words (Wiegand et al., 2019). Many of these comments involve the context of a cooking show called 'My Kitchen Rules' which frequently resulted in sexist comments thrown towards the participating women. The most frequent topics discussed in the Waseem corpus apart from the cooking show involve sexism related to the role of women in sports (especially football), females as stand-up comedians, their driving skills, etc. The tweets were originally categorized into three classes: 'sexism', 'racism', and 'none' with around 3.4K, 2.0K, and 11.6K instances, respectively. However, since it is made available as tweet IDs, a substantial portion of the comments is observed to be removed as a result of Twitter's filtering strategy to delete hateful content. In fact, in total, we obtain 10.9K tweets, where 2.9K is originally labeled as 'sexism', and 8.0K as 'none'. However, only 20 instances for the 'racism' class could be retrieved. Bodapati et al. (2019) reported a similar number of extracted instances from the *Waseem* corpus. For our experiments, we follow other prior work for cross-corpus evaluations (Wiegand et al., 2019; Swamy et al., 2019), and merge the original 'racism' and 'sexism' classes into the class *abusive/hate*. Around 26.8% of the instances are labeled with this class.

3.3.4 Dynamic

The *Dynamic* corpus was introduced by (Vidgen et al., 2021c). It was dynamically generated using a human-and-model-in-the-loop process aimed at making the corpus robust to different forms of hate. The `Dynabench` platform was used for this dynamic data collection process. This is an open-source online research platform for dynamic dataset collection and benchmarking. 20 trained annotators were employed for creating the data, who were supervised by two expert annotators. The instances generated were annotated into the classes: 'hate' and 'not hate'. The 'hate' class was defined as '*abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation*' (Warner and Hirschberg, 2012b), and this class was further annotated for the type of hate and the targeted social group.

The entire process involves four rounds of data generation, involving $\sim$ 10K data instances in each round, combined with training a hate speech classification model in the loop. The data was generated in every round by exploring the weaknesses of the model trained on the combined data from the previous rounds. First, a RoBERTa (Liu et al., 2019b) model was trained on a combination of 11 corpora for hate speech and toxicity in English taken from `hatespeechdata.com` (Vidgen and Derczynski, 2020), which includes *Davidson*, *Waseem* and *HatEval*. The annotators were asked to create a set of synthetic instances that could mislead the model into misclassifying them through real-time feedback. These included tricky instances like 'I hate the concept of hate',

and 'Tea is f*cking disgusting'. This formed the first round. In the second round, the model was further trained on the synthetically created instances. The annotators were provided with adversarial 'pivots', using which they perturbed some of the instances from the previous round, along with adding original content. Furthermore, they were given some guidelines about the perturbations, which included ensuring that the perturbed instance is realistic, flips the label of the original instances, covers diverse targets of hate, hateful forms, etc., and involves a minimal change in the instance perturbed. For example, the sentiment of a hateful instance 'Hearing a woman speak is the most *tedious* sound' could be changed to create a non-hateful instance 'Hearing a woman speak is the most *wonderful* sound'. Likewise, in the next two rounds, the annotators were guided to enter more realistic perturbations and original content. All the synthetically generated instances were reviewed in every round through a qualitative validation process.

The final set of instances in this corpus comprises different types of hate like derogation, animosity, threatening language, support for hateful entities, and dehumanization. They also cover hate directed against diverse social groups like blacks, women, jews, muslims, immigrants, arabs, disabled people, trans people, and others. We have used two versions of this corpus in our experiments, which we call $Dynamic^{v1}$ and $Dynamic^{v2}$. The first one is an older version of the corpus that contains 40,623 instances (hate: 22K, not hate: 18.6K; 54.4% hate), and the second one corresponding to v0.2.3 of the corpus contains 41144 instances (hate: 22K, not hate: 18.9K; 53.9% hate), where duplicate entries are removed.

3.3.5 Ethos

This is a corpus presented by Mollas et al. (2022). *Ethos* or `ETHOS` stands for multi-labEl haTe speecH detectiOn dataSet. It comprises instances with both binary labels and multi-labels. The corpus was sampled from YouTube and Reddit using an active learning strategy – a learning scheme that poses queries to the user in the form of new informative unlabeled instances to be labeled. It strives to attain high performance with the least amount of labeled instances. This strategy was used to ensure both diversity and balance with respect to different hateful aspects defined. Therefore, even though the corpus involves very few instances (998 in total) compared to the previous corpora, it is diverse and informative. It includes hate directed towards diverse identities, such as gender, race, national origin, disability, religion, and sexual orientation.

In the data creation process, first a set of annotated YouTube comments were used as an initial seed. These instances were provided by the `Hatebusters` platform[5] (Anagnostou et al., 2018) and were annotated by the platform creators. This is a web application that looks for YouTube videos with potentially hateful comments, provides online users with comments that have the highest scores for being hateful, and allows users to voluntarily report hateful comments. Mollas et al. (2022) queried the `Hatebusters'` database and Reddit (using Public Reddit Data Repository[6]) to obtain unlabeled instances for

[5] https://hatebusters.org
[6] https://files.pushshift.io/reddit/comments/

potentially different categories of hate speech. They then performed a grid search to find the best-performing classification model by training on the initially annotated seed data, among a collection of models like SVM, Random Forest (RF), Logistic Regression (LR), and different Neural Network (NN)-based models. The best model was then used to automatically assign probability scores to the unlabeled instances. The instances having probability scores within the range of $[.4, .6]$ were manually annotated and some instances within the range of $[.0, .1] \cup [.9, 1.0]$ were examined to identify misclassification, making use of the active learning strategies of uncertainty sampling (Lewis and Gale, 1994), amongst others (Reyes et al., 2018). Moreover, only one instance of multiple comments with similar meanings was retained. For example, the comments '*I hate white people*' and '*I hate whites*' (Mollas et al., 2022) are similar, and only one of them was added. Almost all the comments were then annotated by five different annotators on a crowd-sourcing platform. An inter-annotator agreement of 81.4% was reported for the binary labels of the presence or absence of hate speech. Initially, a score in the range of $[0, 1]$ is provided for every comment to annotate the presence of hate speech, which was later binarized such that scores ≥ 0.5 was labeled as hate speech. We used the binary version of this corpus with 433 (43.4%) instances annotated as hate speech and 565 (56.6%) do not contain hate speech.

3.4 Text Pre-processing

We pre-process the corpora separately for the training, validation and test set of the data by removing the URLs and newlines, splitting the hashtags into constituent words after removing the '#' symbol (e.g. the hashtags '#buildthewall' is mapped to 'build the wall'), and expanding contractions (e.g. i'll to i will). Furthermore, the maximum number of repeated adjacent characters is set to two as having the same two consecutive characters is common in many words. To illustrate, the tokens 'betttter' and 'bettttttter' are mapped to 'better', and likewise 'looove' and 'loooove' are also mapped to 'loove'. `CrazyTokenizer`[7] is used for these pre-processing steps. We also remove the rarely occurring Twitter handles. This is done by filtering out the Twitter handles whose frequency of occurrence in the corpus is below a threshold frequency. The threshold is computed by taking the range of frequency of occurrence of Twitter handles, i.e. range $R_{freq} = H_{freq} - L_{freq}$ (H_{freq} : highest frequency, L_{freq} : lowest frequency) and setting threshold as $L_{freq} + 0.1 \times R_{freq}$. However, the frequently occurring Twitter handles are retained as they might contain important information. For example, the topic associated with @realDonaldTrump might be informative for the classification model to understand the context and detect the presence of abusive language associated with this topic. The words in the corpora are finally converted into lowercase.

[7] https://redditscore.readthedocs.io

3.5 Evaluation Metric

We use the percentage F1 score, especially the macro-average F1, as the performance metric in this thesis to report the classification performance of the models. The F1 score for a class is measured as the harmonic mean of the precision and recall values for the class. Precision is the ratio of instances correctly predicted as belonging to class c, i.e. True Positives (TP) divided by the total number of instances predicted as c, i.e. True Positives (TP) + False Positives (FP). The recall is the ratio of the correctly predicted instances as class c, i.e. True Positives, divided by the total number of instances that should have been predicted as c, i.e. True Positives + False Negatives (FN). The idea behind the F1 score is that both measurements are equally important and that a good F1 score can only be obtained by combining good precision and good recall. It is given by:

$$\text{F1} = \frac{2 * (precision * recall)}{(precision + recall)}$$

Macro-average F1 or macro-F1 provides the arithmetic mean over all the F1 scores computed for each of the classes, as:

$$\text{macro-F1} = \frac{1}{C} \sum_{i=1}^{C} \text{F1}_i;$$

where C is the total number of classes. Macro-F1 treats all the classes equally unlike micro-F1 (counting the sums of TP, FN, FP across all classes and then calculating the aggregated F1) that favors classes with more number of instances (Sokolova and Lapalme, 2009; Narasimhan et al., 2016). The macro-F1 score ranges between 0 and 1. The higher the score, the better the performance. In this thesis, we multiply the macro-F1 score by 100 to obtain the percentage macro-F1.

3.6 Summary

In this chapter, we provided a description of Transformer layers and the different deep neural networks-based models – BERT, HateBERT, USE, and SBERT – used in the following chapters. These models capture external knowledge from huge corpora. Therefore, they are employed either to provide off-the-shelf pre-trained representations or their pre-trained parameters are further fine-tuned for the abusive language detection task. We also presented different abusive language corpora, namely *Davidson*, *HatEval*, *Waseem*, *Dynamic*, and *Ethos* that are used for our experiments. The first three are sampled from Twitter. *Dynamic* is created using a human-and-model-in-the-loop procedure using the `Dynabench` platform, while *Ethos* is collected from YouTube and Reddit comments. The text pre-processing steps followed for these corpora are presented subsequently. Finally, the evaluation metric used in the thesis for reporting the classification performance is

described. This chapter can, therefore, be referred to while reading the subsequent chapters to obtain further details about the models used in the proposed approaches, corpora, pre-processing steps, and the performance metric used for evaluating the approaches.

4 Analyzing Topic Models for Generalizability in Cross-corpus Abusive Language Detection

4.1 Introduction

In this chapter, we analyze how well abusive language detection models can generalize outside the training corpus and investigate the impact of topic models in improving their generalizability. Particularly, the cross-corpus transfer in abusive language detection task is studied, assuming that *we do not have access to any data from the target corpus during the training phase*. This chapter corresponds to our article Bose et al. (2021a).

In Chapter 2, a detailed overview of the existing studies on cross-corpus analysis in abusive language detection was presented. In fact, the state-of-the-art abusive language detection models have been shown to yield impressive performance when evaluated on a held-out test set from the same corpus as the one on which they have been trained (D'Sa et al., 2020; Mozafari et al., 2019; Badjatiya et al., 2017). However, the performance of these models have been found to degrade considerably, when they encounter abusive comments that differ from the training corpus (Wiegand et al., 2019; Arango et al., 2019; Swamy et al., 2019; Karan and Šnajder, 2018). This is due to the varied sampling strategies used to build different corpora, changing linguistic traits, topical and temporal shifts (Florio et al., 2020), and varied targets of abuse across corpora. Since the content in social media changes rapidly, abusive language detection models with better generalization capabilities are required for effective content moderation in practical scenarios for online platforms (Yin and Zubiaga, 2021). To this end, a cross-corpus analysis and evaluation are essential to estimate the models' generalizability (Wiegand et al., 2019; Swamy et al., 2019).

The latent semantic structures occurring in a text corpus can be discovered using unsupervised topic models, without requiring prior annotations of the samples (documents or comments in our task) (Blei et al., 2003). In topic models, each sample is generally viewed as a mixture of different topics and each topic is viewed as a distribution over all the words. Furthermore, topic models can generalize over unseen samples to infer the latent topic mixtures present in them. We believe that this strength of these models can be exploited to infer topic mixtures in unseen abusive text. Besides, topic models have been explored for generic cross-domain text classification (Jing et al., 2018; Zhuang et al., 2013; Li et al., 2012), demonstrating better generalizability. This inspires us to leverage topic model representations for cross-corpus abuse detection.

Recently, Caselli et al. (2021) showed that the *HateBERT* model, as discussed in Chapter 3, displays better generalizability in cross-corpus experiments for abusive language detection. Furthermore, Peinelt et al. (2020) demonstrated that the combination of topic models and BERT representations leads to better performance at the semantic similarity task. Taking these studies into account, we make the following contributions in this chapter:

- We investigate if combining topic representations with contextualized HateBERT representations can result in better generalizability in cross-corpus abuse detection. Cross-corpus evaluation over three common abusive language corpora supports and demonstrates the effectiveness of this approach.
- We bring some insights into how the association of unseen comments to abusive language topics obtained from original training data can help in cross-corpus abusive language detection.

The rest of the chapter is organized as follows: Section 4.2 gives an overview of topic models. 4.3 describes the architecture of the combination of a topic model and HateBERT. Section 4.4 presents our experimental settings. An analysis of the results obtained is present in Section 4.5, and Section 4.6 concludes the chapter.

4.2 Overview of Topic Models

Topic modeling is an unsupervised statistical approach that expresses each document as a mixture of a set of latent topics. Each topic is characterized by a distribution of words in the corpus. In the Latent Dirichlet Allocation (LDA) topic model (Blei et al., 2001, 2003), a collection of text documents is modeled using a generative probabilistic approach. Every text document d in a corpus is expressed as a multinomial distribution θ_d over topics $\mathcal{T}$, and every topic z is represented by a multinomial distribution ψ_z over words drawn from a fixed vocabulary V, where $d = 1, 2, ..., D$ and $z = 1, 2, ..., \mathcal{T}$. For example, the sentence "*The way the food is grown, distributed and regulated is governed by a complex and interwoven system of local, state and federal food policies*" could be a mixture of about 40% food, 30% economics and 30% politics. The generative process assumed by LDA can be expressed in the graphical model in Figure 4.1.

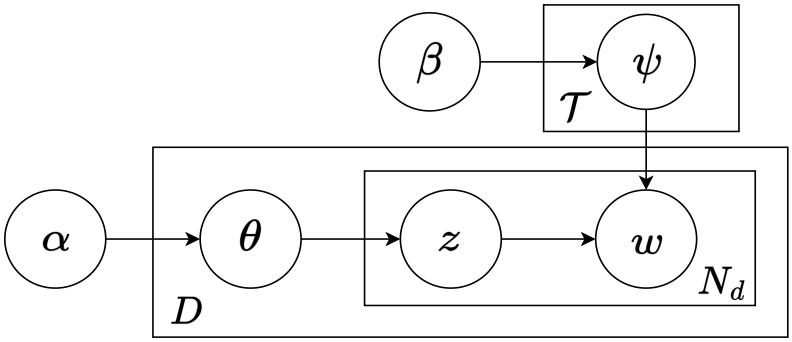


Figure 4.1: Graphical model of LDA

Following a Bayesian framework, the distributions θ_d and ψ_z have corresponding priors α and β with a Dirichlet distribution. z_{d_i} is the indicator of topic for every i^{th} word w_{d_i} in document d. N_d is the number of words in document d. The generative process involves drawing a topic z_{d_i} from the multinomial θ_d, and then drawing a word w_{d_i} from the multinomial $\psi_{z_{d_i}}$ for every w_{d_i} in d. The unsupervised learning of topic distributions entails the process of *posterior inference*: estimating the posterior distributions of the LDA model's latent parameters (θ, ψ, z). Since every word in the corpus vocabulary contributes to every topic with some probability, a new mixture of words in an unseen document can be mapped to the set of latent topics learned from the known corpus. In the context of social media micro-posts or comments, a comment can also be considered a document, and topic modeling can be applied to comments in the same way.

Non-neural Topic Models The parameters of the conventional topic models, like LDA, are learned from the co-occurrence of words within documents (or comments). However, most social media platforms typically involve comments that use few words. As a result, LDA typically experiences a degradation in performance when it is employed for short texts due to sparsity in the co-occurrence of words in these texts (Cheng et al., 2014; Nguyen et al., 2015; Qiang et al., 2020), as the parameters are learned based on little evidence.

In order to handle this data sparsity issue in short texts, different strategies have been proposed in the past years. One such strategy is using the Dirichlet Multinomial Mixture (DMM) model or the mixture of unigrams (Nigam et al., 2000; Yin and Wang, 2014; Zhao et al., 2011) that limits the distribution of document topics so that each short text is sampled from a single topic. Moreover, it has been shown by Phan et al. (2011); Zeng et al. (2018), and Yi et al. (2020) that exploiting external knowledge helps improve topic representations. In Latent Feature Dirichlet Multinomial Mixture (LF-DMM) (Nguyen et al., 2015), pre-trained word vectors that incorporate external knowledge have been used as the latent feature component along with the DMM model. Such external information has also been incorporated into GPU-DMM (Li et al., 2016a) that employs the generalized Pólya urn (GPU) (Mahmoud, 2008) model. It uses a non-parametric probabilistic sampling mechanism that reinforces semantically similar words under the same topic. The single-topic assumption of DMM has been relaxed in Poisson-based DMM (GPU-PDMM) (Li et al., 2017a), and each document is allowed to be generated from a few topics, modeled using Poisson distribution. Biterm Topic Model (BTM) (Cheng et al., 2014), a global word-co-occurrence-based method, treats any two words in the corpus as a biterm and infers topics over the same. Word Network Topic Model (WNTM) (Zuo et al., 2016b) uses LDA to learn word-topic distribution from an undirected word co-occurrence network. Other probabilistic methods like Self-Aggregation-based Topic Modeling (SATM) Quan et al. (2015) and Pseudo-document-based Topic Modeling (PTM) (Zuo et al., 2016a) combine short texts into long pseudo-documents to address the problem of sparsity. Steinskog et al. (2017) explored pooling techniques on Twitter data and combined the tweets sharing the same hashtags or authors. This was aimed at minimizing the difficulties introduced by short texts by tweet aggregation with

the intuition that tweets having the same hashtags or written by the same author are likely to share the same topics.

Neural Topic Models: Similarly, several neural topic models have been proposed in recent years. Since neural networks are effective and flexible for unsupervised representation learning, neural topic models can potentially learn more significant topics from texts (Wang et al., 2021). Cao et al. (2015) explained the standard topic model from the perspective of neural networks and proposed a supervised extension to the topic model to handle supervised tasks. They represented the topic model view of words and documents with an n-gram topic layer and a topic-document layer that makes use of the pre-trained word embeddings. Moody (2016) proposed a modification of the Skip-gram Negative-Sampling (SGNS) objective in Mikolov et al. (2013b) to jointly learn topic vectors, document topic weights, and word vectors.

The Topically Driven Neural Language Model (TDLM) (Lau et al., 2017) makes use of a simple yet elegant attention mechanism to learn the document-topic weights. It computes a document vector for every document from the word embeddings of the constituent words using a convolutional neural network. Then through an attention mechanism, TDLM learns to associate every document with different topic embeddings. The weighted mean of the topic embeddings is then used to predict the words in the document. The authors showed that TDLM generates potentially more coherent topics compared to LDA.

Dieng et al. (2020) proposed the Embedding Topic Model, a generative model that also represents topics as topic embeddings but infers the posteriors of the topic proportions through variational inference (Kingma and Welling, 2014). Neural Variational Document Model (NVDM) (Miao et al., 2016), Product of expert LDA (ProdLDA) (Srivastava and Sutton, 2017), Dual Word Graph Topic Model (DWGTM) (Wang et al., 2021) are some of the other neural topic models that use neural variational inference. Zhao et al. (2021) provided a survey of topic models that are built using deep neural networks.

4.2.1 Topic Models for Cross-domain Text Analysis

There are multiple works on cross-domain topic modeling (Xue et al., 2008; Li et al., 2012; Bao et al., 2013; Zhuang et al., 2013; Zhou et al., 2015; Jing et al., 2018) that have performed cross-domain text classification. Most of these methods divide the topics across domains into shared and domain-specific sub-sets to capture domain-independent and domain-specific information, respectively. Subsequently, the domain-specific topics are aligned to improve the performance obtained by a classifier trained on the source domain data and evaluated on the target domain data. Zhai et al. (2004) and Paul and Girju (2009) proposed different variations of topic models for comparing collections of texts. For instance, Paul and Girju (2009) introduced ccLDA, a cross-collection topic model extending LDA, to analyze the similarities and differences across corpora and apply it to detect the cultural differences in different countries from people's experiences discussed in blogs and forums. Along similar lines, Zuo et al. (2018) proposed a variant

of the cross-collection topic model to perform opinion mining across text collections from online platforms towards subjects such as products or events. All these works leveraged data from both the source and target corpora while learning topics. However, since in this chapter, we do not assume access to unlabeled data from the target corpus during the training phase, we use a topic model trained on the source corpus to improve generalizability across different unseen target corpora.

4.2.2 Topic Models for Examining Abusive Language

Topic models have been shown to potentially discover the biases present in the annotated abusive language corpora. For instance, Davidson and Bhattacharya (2020) analyzed the content of a publicly available abusive language corpus using the structural topic model (Roberts et al., 2014). They demonstrated that some specific topics were disproportionately associated with comments written in African-American English and also annotated as abusive, which could bias the classifiers trained on them. Nejadgholi and Kiritchenko (2020) showed that the removal of platform-specific topics from an abusive language corpus resulted in better generalizability. They recommended that within the data collection process, unsupervised topic models, like LDA, can be used to detect and reduce the corpus-specific content to improve their cross-dataset generalization, prior to the expensive process of annotation. Calderón et al. (2020) used LDA to analyze the representative topics in abusive comments against immigrants. Besides, Evkoski et al. (2021) studied some research questions focusing on the topics present in abusive tweets, the evolution of topics through time on Twitter, and the comparisons of topic distributions across retweet communities: network communities formed by Twitter users who are densely linked through retweets.

4.3 Combining Topic Model and HateBERT

We first give a description and the architecture of the topic model that we have incorporated in our work and then the procedure followed to combine the representations obtained from the topic model and HateBERT.

4.3.1 Universal Sentence Encoder-based TDLM

In this work, we leverage the Topically Driven Neural Language Model (TDLM) (Lau et al., 2017) to obtain topic representations because of its simplicity and as it can employ external knowledge in the form of *pre-trained* embeddings. Such external knowledge is found to be more suitable for short Twitter comments (Yi et al., 2020). Moreover, TDLM is shown to potentially yield more coherent topics compared to LDA. The original model of TDLM, as discussed in Section 4.2, applies a CNN over word embeddings to generate a comment embedding. This comment embedding is used to learn and extract topic distributions. Cer et al. (2018) showed that transfer learning via sentence embeddings performs better than word embeddings on a variety of tasks. Moreover, since comments

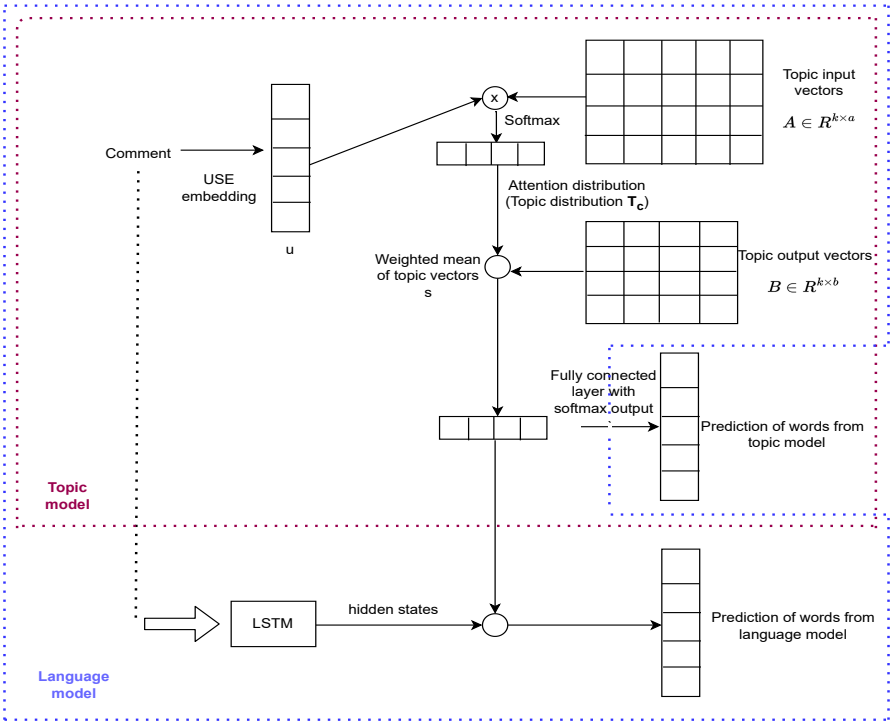


Figure 4.2: Architecture of U-TDLM. As compared to TDLM (Lau et al., 2017), CNN on comment is replaced by USE (Universal Sentence Embedding). k = number of topics.

present in social media are usually much shorter in length as compared to traditional documents and typically involve 1-2 sentences, transformer-based Universal Sentence Encoder (USE) (Cer et al., 2018) (described in Chapter 3) can be directly applied to individual comments to obtain sentence embeddings. Hence, we modify TDLM to accept the USE embeddings extracted from input comments, instead of the comment embeddings from CNN. The modified model is denoted as U-TDLM hereon. The architecture of U-TDLM is present in Figure 4.2.

Following the architecture of TDLM, U-TDLM consists of two components: a *topic model* that uses the comment embedding $u \in R^a$ – obtained from the USE – to learn the topic mixtures present in the comment and a *language model* that captures the sentence level word relation. Here a is the dimension of the embedding vector obtained from the USE model. The topic model component is trained to learn the topic information present in the comments, while the language model component learns the word relations in the comment. In the topic model component, the topic weights for a comment are computed using an attention mechanism that aligns u and individual topic vectors. Given k as the number of topics, two lookup tables $A \in R^{k \times a}$ and $B \in R^{k \times b}$ are maintained for storing the input and output topic vectors, respectively. a and b are the corresponding input and output topic vector dimensions. The dot product of the topic vectors from A and the comment embedding u are then passed through a softmax function to generate the

topic distribution for a comment $T_c \in R^k$. The attention weights from T_c, in turn, are used to obtain the weighted mean s of the topic vectors from B as:

$$T_c = softmax(Au) \quad \text{and} \quad s = B^T T_c \tag{4.1}$$

The topic distribution for a comment embedding u can also be written as $T_c = [p(t_i|u)]_{i=1:k}$, where t_i corresponds to the i^{th} topic. Lau et al. (2017) points out that this topic model component is inspired by the generative process of LDA that defines documents/comments to have a multinomial distribution over topics. Finally, s is connected to a fully connected layer with softmax output for predicting words in the comment. The language model component predicts the next words in the comment using the weighted mean, s, of the topic vectors generated by the topic model. It incorporates the topical information from s into the hidden states of LSTM at each time step. This enables generating related sentences for a topic, providing another way to interpret topics. Even though we jointly train both the components of U-TDLM following Lau et al. (2017), for our task, we are more interested in the topic distribution T_c obtained by the topic model component.

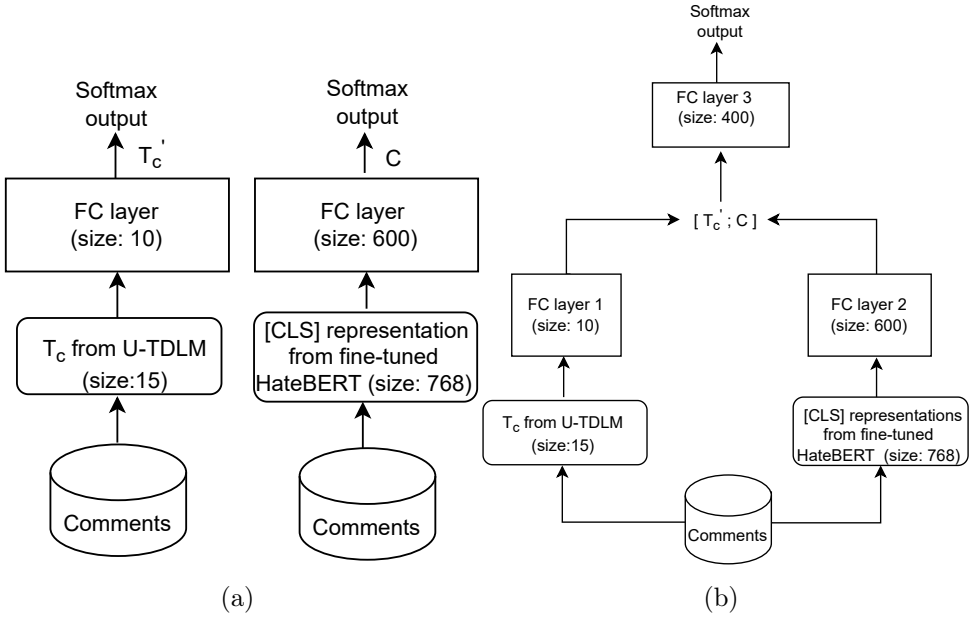


Figure 4.3: Architecture of classifier for individual models (a) U-TDLM and HateBERT, and the combined model (b) HateBERT + U-TDLM; FC: Fully Connected.

4.3.2 Combining the Embeddings

Figure 4.3 presents the architecture of the individual and the combined models. We train U-TDLM to obtain T_c on the train set from the source corpus and use it to infer the T_c on the test set from a different target corpus. T_c is passed through a Fully Connected (FC)

layer to obtain transformed representation $T_c^{'}$. Besides, we first perform supervised fine-tuning of HateBERT[1] on the train set of the source corpus. The vector corresponding to the [CLS] token in the final layer of this fine-tuned HateBERT model is chosen as the HateBERT representation for a given comment. It is transformed through an FC layer to obtain the C vector. Finally, in the *combined model* (HateBERT+U-TDLM), the concatenated vector $[T_c^{'}; C]$ is passed through a final FC layer and softmax activation function.

4.4 Evaluation Set-up

4.4.1 Experimental Settings

The experiments are performed on three different publicly available abusive language corpora from Twitter as they cover different forms of abuse, namely, *HatEval* (Basile et al., 2019), *Waseem* (Waseem and Hovy, 2016), and *Davidson* (Davidson et al., 2017). We target a binary classification task with classes: *abusive* and *non abusive*. Since abusive language is a super-set of hate speech, the different forms of abuse, such as offensive and hate speech are mapped into the class *abusive* following the precedent of previous work on cross corpora analysis (Wiegand et al., 2019; Swamy et al., 2019; Karan and Šnajder, 2018). For *HatEval*, the standard partition of the shared task is used, whereas the other two corpora are randomly split into train (80%), development (10%), and test (10%). The statistics of the train-test splits of these corpora are listed in Table 4.1. A detailed description of each corpus is provided in Chapter 3.

Datasets	**Number of comments**			**Abuse %**
	Train	**Dev**	**Test**	
HatEval	8993	1000	3000	42.1
Waseem	8720	1090	1090	26.8
Davidson	19817	2477	2477	83.2

.

Table 4.1: Statistics of the corpora used (Dev: Development).

We choose a topic number k of 15 for all our experiments based on the initial results for in-corpus classification performance using topic distributions and to maintain a fair comparison. Since the topic input vector size a in U-TDLM is the same as the dimension of embeddings obtained from the USE model, a is set to 512. We use 50 as the topic output vector size, b, following Lau et al. (2017). Besides, the best model checkpoints are selected by performing early stopping of the training using the respective development sets. The FC layers are followed by Rectified Linear Units (ReLU) in the individual as well as the combined models. In the individual models, the FC layers for transforming T_c and the HateBERT representation have 10 and 600 hidden units, respectively. The

[1]Pre-trained model from https://osf.io/tbd58/

final FC layer in the combined model has 400 hidden units. We report the classification performance in terms of mean macro-F1 score and standard deviation over five runs, with random initializations.

4.4.2 Data Pre-processing

The pre-processing details are provided in Chapter 3. In addition, U-TDLM involves the prediction of words from the comments based on the weighted mean s of topic vectors. In this output end of the topic model, our implementation involves the prediction of stemmed words and skipping of stop words. This is done to ensure that the softmax output from s can treat the words like 'raining' and 'rain' equally, which can help them in getting assigned to similar topics.

Train set	**In-corpus performance**			**Cross-corpus test set**	**Cross-corpus performance**			
	HBERT	**U-TDLM**			**HBERT**	**U-TDLM**	**HBERT + Rand**	**HBERT + U-TDLM**
HatEval	53.9±1.7	41.5±0.6		Waseem	66.5±2.2	55.5±2.6	64.6±2.6	**67.8**±2.4
				Davidson	59.2±2.5	**64.4**±2.3	58.2±0.8	60.4±1.4
Waseem	86.1±0.4	73.7±1.4		HatEval	**55.8**±1.4	36.7±0.0	**55.8**±0.9	55.4±0.7
				Davidson	59.8±3.6	28.2±2.4	56.8±1.3	**64.8**±1.8
Davidson	93.7±0.2	75.6±0.8		HatEval	**51.8**±0.2	50.5±1.3	51.4±0.2	**51.8**±0.3
				Waseem	66.6±3.0	48.7±3.3	64.7±2.4	**68.5**±2.1
Average	77.9	63.6			60.0	47.3	58.6	**61.5**

Table 4.2: Macro-average F1 scores (mean±std-dev) for in-corpus and cross-corpus abuse detection. HBERT: HateBERT, Rand: Random vector. The best in each row for the cross-corpus performance is marked in **bold**.

4.5 Results and Analysis

Table 4.2 presents the in-corpus and cross-corpus evaluation of the HateBERT and U-TDLM models, as illustrated in Figure 4.3, in terms of macro-average F1 scores. All models are trained on the train set of the source corpus. The in-corpus performance of the models is obtained on the source corpora test sets, while the cross-corpus performance is obtained on target corpora test sets. It is shown in Table 4.2 that the cross-corpus performance degrades substantially as compared to the in-corpus performance, except for *HatEval* which indeed has a low in-corpus performance. *HatEval* test set is part of a shared task, and similar in-corpus performances have been reported in prior work (Caselli et al., 2021). Overall, comparing the cross-corpus performances of all models, we can observe that the combined model (HateBERT + U-TDLM) either outperforms HateBERT in 4 cases or retains its performance in 2 cases. This hints that incorporating topic representations can be useful in cross-corpus abusive language detection. As an

ablation study, we replaced U-TDLM features with random vectors to evaluate the combined model. Such a concatenation decreased the performance in the cross-corpus setting, yielding an average macro-F1 score of 58.6. This indicates that the topic representations improve generalization along with HateBERT.

4.5.1 Case-studies to Analyse Improvements from U-TDLM

We investigate the cases in Table 4.2 which report relatively large improvements, as compared to HateBERT, either with HateBERT+U-TDLM (train on *Waseem*, test on *Davidson*) or only with U-TDLM (train on *HateEval*, test on *Davidson*). Some of the prominent topics from *Waseem* and *HateEval* associated with abusive comments, and the top words corresponding to these topics are provided in Table 4.3 and Table 4.5, respectively. For better interpretation, topic names are manually assigned based on the top words and the knowledge of the individual corpora. We consider the abusive class as positive, and the non-abusive class as negative in the subsequent discussion.

Topic id	Names	Top words
4	Sexism in sports	football, sex, sport, feminist, drive, woman, call, sexist
9	Feminism	feminist, article, ebook, equality, patriarchy, abuse, freebsd, harass
12	Cooking show	katie, score, mkr, cook, c*nt, blond, less, strategic

Table 4.3: U-TDLM trained on Waseem's train set (topic names are assigned manually for interpretation).

Train on *Waseem* →Test on *Davidson*: In this case, the corresponding scores obtained from three models are – HateBERT: 59.8, U-TDLM: 28.2, HateBERT+U-TDLM: 64.8. U-TDLM shows poor performance due to the large number of *False Negatives* (#FN for U-TDLM: 1824), and less *True Positives* (#TP for U-TDLM: 266). The combined model, on the other hand, has higher *True Positives* compared to those obtained from HateBERT (#TP for HateBERT+U-TDLM: 1556, #TP for HateBERT: 1267). The count of *True Negatives* with the combined model remains similar to that in HateBERT (#TN for HateBERT + U-TDLM: 314, #TN for HateBERT: 340). This suggests that U-TDLM introduces some complementary information in the combined model. However, since U-TDLM may not capture the entire context of the comments, it does not yield decent performance when used alone. This indicates that U-TDLM representations should be combined with the contextual representations obtained from HateBERT for complementing each other.

Source →Target	**Abusive Comments in Target**	**Source topics**
Waseem →Davidson	When women are so proud that they don't like to cook; clean b*tch stop being lazy..It's not cute.	4, 12
	ya girl is a slimy ass h*e. get her under control and tell her to stop spraying bullshit out her mouth all day.	4, 9, 12
HatEval →Davidson	No. Its wrong to try to change f*ggots; There is no "therapy"....sympathize like they are retards.	3, 7
	Naturally, when a shitty leftist rag talks trash about another shitty leftist rag, you better fall in line...	10

Table 4.4: Abusive comments in the target corpus, correctly classified by HateBERT+U-TDLM (Waseem →Davidson) and U-TDLM (HatEval →Davidson). "Source topics": topics that are assigned high weights for the corresponding comments by U-TDLM trained on Source.

Topic id	**Names**	**Top words**
3	Explicit abuse 1	men, c*ck, d*ck, woman, picture, sl*t, s*ck, guy
7	Explicit abuse 2	b*tch, ho*, n*gger, girlfriend, f*ck, shit, s*ck, dumb
10	Politics related	therickwilson, anncoulter, c*nt, commies, tr*nny, judgejeanine, keitholbermann, donaldjtrumpjr

Table 4.5: U-TDLM trained on HatEval's train set (topic names are assigned manually for interpretation).

We analyze a few abusive comments in the test set of *Davidson* (target) in Table 4.4, which are wrongly classified by HateBERT, but correctly detected as abusive by the combined model. The topical membership of these abusive comments from *Davidson* indicates that U-TDLM associates high weights to the relevant abuse-related topics from *Waseem*. As indicated in the first example, an abusive comment against women that discusses cooking, in *Davidson*, is mapped to topics 4 (sexism) and 12 (cooking show) from *Waseem*. Similarly, the second comment gets high weight in the three topics 4, 9, and 12 due to its sexist content and use of a profane word. Other pairs of corpora that yield improved performance with the combined model also follow similar trends as above.

Train on *HatEval* →Test on *Davidson*: In this case, while U-TDLM performs considerably well, the combined model only provides a slight improvement over Hate-

BERT, as per Table 4.2 (HateBERT: 59.2, U-TDLM: 64.4, HateBERT+U-TDLM: 60.4). U-TDLM has a higher TP when compared to both HateBERT and the combined model (#TP for U-TDLM: 1924, #TP for HateBERT+U-TDLM: 1106, #TP for HateBERT: 1076), with lower TN (#TN for U-TDLM: 130, #TN for HateBERT+U-TDLM: 373, #TN for HateBERT: 374).

A few abusive comments from *Davidson* that are correctly classified by U-TDLM alone are presented in Table 4.4. The first comment for this case has high weights for the abuse-related topics 3 and 7 from *HatEval* (shown in Table 4.5) due to the presence of the profane word "f*ggot". The second comment only gets a high weight for topic 10, which deals with politics. This is due to the word "leftist", which is associated with a political ideology. As per our analysis, we found that all of these source topics are highly correlated with the abusive labels in the source corpus of *HatEval*. Therefore, these comments from the target corpus of *Davidson* are correctly classified as abusive by U-TDLM. Moreover, since a part of the abusive comments in *HatEval* comprises explicit abuse, U-TDLM yields explicitly abusive topics when trained on *HatEval*, as presented in Table 4.5. We find that more than 82% of the abusive comments of the *Davidson* corpus also involve explicit abuse, with the presence of slur words, as reflected by the most common words in its abusive comments in Chapter 3. We believe that this helps U-TDLM in yielding high TP on the test set from *Davidson* as its explicitly abusive comments trigger the relevant topics from *HatEval*. However, the *HatEval* corpus also comprises comments that use slur words, but are not abusive, which is slightly different from the definition of abuse used for annotating the *Davidson* corpus. Therefore, since HateBERT represents the entire context present in the comments, some TP on *Davidson* yielded by U-TDLM might be converted into FN by the combined model trained on *HatEval*, resulting in decreased performance compared to U-TDLM. Still the combined model performs better than HateBERT in this case.

4.6 Conclusion

In this chapter, we analyzed the generalizability of topic models when they are applied to the task of cross-corpus abusive language detection. An in-corpus and cross-corpus evaluation of HateBERT and U-TDLM has helped us confirm our perspective on generalization in the abusive language detection task. A contextualized representation model like HateBERT can achieve great levels of performance on the abusive language detection task, typically when the evaluation dataset does not differ from the dataset on which it has been trained. The performance of this model degrades drastically on abusive language comments from unseen contexts. Topic models like U-TDLM, which express comments as a mixture of topics learned from a corpus, allow unseen comments to trigger abusive language topics. While topic space representations tend to lose the exact context of a comment, combining them with HateBERT representations can give modest improvements over HateBERT or at the least, retain the performance of HateBERT. These results should fuel interest and motivate further developments in the generalization of abusive language detection models.

5 Unsupervised Domain Adaptation for Abusive Language Detection

5.1 Introduction

In this chapter, we consider the problem of improving the performance of abusive language detection models on a target corpus with a different distribution compared to the source corpus by exploiting the *unlabeled content of the target*. In particular, we aim to analyze the impact of Unsupervised Domain Adaptation, which is a specific type of transfer learning called *transductive transfer learning* (Pan and Yang, 2009), in the task of cross-corpus abusive language detection. The work presented in this chapter is published as Bose et al. (2021b).

Supervised classification approaches for abuse detection require a large amount of expensive annotated data (Lee et al., 2018). Annotating new content for detecting abuse is non-trivial and may require substantial time and effort (Poletto et al., 2019; Ombui et al., 2019). These issues call for approaches that can adapt abusive language detection models to newly seen data out of the original training distribution. Thus, Unsupervised Domain Adaptation (UDA) methods that can perform adaptation without the target domain labels (Ramponi and Plank, 2020), turn out to be attractive in this task. As discussed by Ramponi and Plank (2020); Plank (2011), a coherent type of corpus can typically be considered a domain for tasks such as automatic text classification. Under this condition, UDA approaches can be applied in cross-corpus evaluation setups. Henceforth, we use the terms 'domain' and 'corpus' interchangeably in the remaining part of the thesis.

A task related to abuse detection is sentiment classification (Bauwelinck and Lefever, 2019; Rajamanickam et al., 2020), and it involves an extensive body of work on domain adaptation. We, therefore, analyze if the problem of cross-corpus abusive language detection can be addressed by the existing advancements in domain adaptation. Alongside different UDA approaches, we also evaluate the effectiveness of the HateBERT model, as discussed in Chapter 3. We perform the Masked Language Model (MLM) fine-tuning of HateBERT on the target corpus, which can be considered a form of unsupervised adaptation. Our analysis reveals that domain adaptation in the task of abusive language detection cannot be trivially addressed by the existing advancements in other related tasks, like cross-domain sentiment classification, and therefore it requires solutions specifically suited to this task.

To summarize, the contributions presented in this chapter are:

- We investigate some of the best performing UDA approaches, originally proposed

for cross-domain sentiment classification, and analyze their performance on the task of cross-corpus abusive language detection. We provide some insights on the sub-optimal performance of these approaches. To the best of our knowledge, this is the first work that analyzes UDA approaches for cross-corpus abuse detection.

- We analyze the performance of HateBERT in our cross-corpus evaluation set-up. In particular, the Masked Language Model (MLM) objective is used to further fine-tune HateBERT over the unlabeled target corpus, and subsequently, supervised fine-tuning is performed over the source corpus.

The remaining of this chapter is structured as follows: In Section 5.2, we discuss the shifts across different abusive corpora. In Section 5.3, we give a brief overview of the existing UDA approaches related to our task and describe the approaches that we analyze through our experiments. We then present the experimental settings used in our evaluation in Section 5.4. In Section 5.5, the results of our evaluation and a discussion on the performances of different approaches are presented. Finally, we conclude the chapter in Section 5.6.

5.2 Shifts in Abusive Language Corpora

A discussion about the shifts across abusive language corpora is presented in Chapter 1. We discuss the same here with some more details relevant to the corpora used in this chapter. Florio et al. (2020); Saha and Sindhwani (2012) have detailed the problem of changing topics in social media with time. Hence, temporal or contextual shifts are commonly witnessed across different abusive corpora. For example, the corpora provided by Waseem and Hovy (2016) and Basile et al. (2019) were collected in or before 2016, and during 2018, respectively, and also involve different contexts of discussion, depending on the events that occurred during the time frame of data collection. The most frequent topics from Waseem and Hovy (2016) involve sexism related to the role of women in sports, the role of female participants in a cooking show, females as stand-up comedians, their driving skills, etc. Some examples from this corpus are:

- *Not sexist but even women prefer to watch Men's sports over women playing because it's played at a higher level.*
- *im not sexist but women just cant be comedians not can they be rappers ...*
- *I'm not sexist at all, but I do hold the firm belief that girls/women shouldn't be allowed to drive. No need to explain why.*

In the corpus provided by Basile et al. (2019), a widely discussed topic is the US-Mexico border issues such as:

- *The wall should have names of Americans who died in the hands of illegal immigration. As a reminder for why we are building it in the first place.*
- *#buildthewall U.S. Marshals: Illegal alien wanted for murder in Mexico arrested in SC #endchainmigration via @cbs46*

Moreover, sampling strategies across corpora also introduce bias in the data (Wiegand

et al., 2019), and could be a cause for differences across corpora. For instance, Davidson et al. (2017) sampled tweets containing keywords from a hate speech lexicon, which has resulted in the corpus having a major proportion (83%) of abusive content. As mentioned by Waseem et al. (2018), tweets in Davidson et al. (2017) originate mostly from the United States, whereas Waseem and Hovy (2016) do not have such a geographical constraint in the sampled tweets.

Apart from sampling differences, the targets and types of abuse may vary across corpora. For instance, even though women are targeted both in Waseem and Hovy (2016) and Davidson et al. (2017), the former involves more subtle and implicit forms of abuse, while the latter involves explicit abuse involving profane words. Besides, religious minorities are the other targeted groups in Waseem and Hovy (2016), while African Americans are targeted in Davidson et al. (2017). Owing to these differences across corpora, abusive language detection in a cross-corpus setting remains a challenge. This has been empirically validated by Wiegand et al. (2019); Arango et al. (2019); Swamy et al. (2019), and Karan and Šnajder (2018) with substantial performance degradation across the cross-corpus evaluation settings. Thus, it can be concluded that the different collection time frames, sampling strategies, and targets of abuse would induce a shift in the data.

5.3 Unsupervised Domain Adaptation

5.3.1 Survey of UDA Approaches

Cross-domain sentiment classification: While there are very few UDA methods for our task of abusive language detection, there is a vast body of research on UDA for the related task of cross-domain sentiment classification. Therefore, we first give a brief overview of the prior research on the latter task. Sheoran et al. (2019) provided a survey of different methods proposed for this task. Blitzer et al. (2006) proposed the Structural Correspondence Learning (SCL), which was later extended for cross-domain sentiment classification (Blitzer et al., 2007). SCL typically constructs an aligned feature space using pivot features to find correlations between features across domains. Pan et al. (2010) proposed a Spectral Feature Alignment (SFA) algorithm that uses the domain-independent words as a bridge to align the domain-specific words from different domains into a single cluster. Specifically, it builds a bipartite graph to model the relationship between domain-specific and domain-independent words in terms of co-occurrence. SFA is based on the idea that domain-specific words with connections to more prevalent domain-independent words in the graph are more likely to be aligned, and vice-versa. Topic modeling has also been used to build UDA methods for sentiment classification. Zhou et al. (2015) proposed the Topical Coherence Transfer (TCT) algorithm that aims to reduce the domain gap by using shared topics across domains. It models the relationship between shared topics and domain-specific topics using a joint non-negative matrix factorization framework. Section 4.2.1 in Chapter 4 provides a discussion on the domain adaptation approaches that use topic modeling.

Deep learning-based UDA approaches have received increased attention in the past years. Glorot et al. (2011) introduced the Stacked Denoising Autoencoder (SDA) (Vincent et al., 2008) for domain adaptation. SDA transforms the unlabeled data from different domains to higher-level feature representations by passing the inputs through multiple stacked layers and learning to minimize a denoising reconstruction error loss corresponding to the given inputs. Subsequently, Glorot et al. (2011) trained a linear classifier on the transformed labeled representations from the source domain. To reduce the computational cost and improve the scalability of SDA with high-dimensional data, Chen et al. (2012a) proposed the marginalized SDA (mSDA) that marginalizes the noise by using linear denoisers as the basic building blocks. Besides, Ruder and Plank (2018) proposed the Multi-task Tri-training (MT-Tri) method and demonstrated improvements in the task of cross-domain sentiment classification, while reducing the space and time complexity associated with the classic method of tri-training (Zhou and Li, 2005; Søgaard, 2010). The latter uses the agreement of three independently trained models to decrease the bias in predicting the pseudo-labels for the unlabeled data. MT-Tri instead performs joint training of the three models with shared parameters and model-specific output softmax layers.

One of the most widespread methods for UDA is based on domain adversarial training (Ganin et al., 2016; Ganin and Lempitsky, 2015) that reduces the domain discrepancy between the source and target distributions by maximizing the confusion in domain identification. This, in turn, results in learning domain-invariant representations. More recently, many domain adversarial-based approaches have been explored for cross-domain sentiment classification (Shen et al., 2018a; Rocha and Lopes Cardoso, 2019; Dai et al., 2020; Xue et al., 2020; Ghosal et al., 2020; Du et al., 2020; Ryu and Lee, 2020). However, the representations learned by domain adversarial methods typically do not make use of any linguistic information. Another prominent line of work for domain adaption in the sentiment classification task is the pivot-based approach (see Section 5.3.3), as introduced for SCL by Blitzer et al. (2006). Later many variations of pivot-based representation learning were proposed that use deep neural networks (Ziser and Reichart, 2017, 2018; Ben-David et al., 2020). These approaches are more linguistically informed compared to the adversarial methods. Li et al. (2017b, 2018) proposed an adaptation method based on domain adversarial learning that also performs pivot-based representation learning while automating the extraction of pivots.

Cross-domain abusive language detection: Recently few UDA approaches have been proposed specifically for abusive language detection. However, we do not use these approaches for our experiments in this chapter as they were introduced after publishing this work. Sarwar and Murdock (2022) proposed a data augmentation-based UDA approach, where they constructed a weakly labeled corpus from a negative emotion corpus (Go et al., 2009). First, they trained a sequence tagger on the source domain for predicting potentially abusive tokens in the abusive context of the target domain samples and the samples in the negative emotion corpus. After replacing the tagged tokens with a placeholder token, they performed a TF-IDF-based template matching between the negative

emotion data and the unlabeled target domain data. This was done to find the most similar samples to the target domain from the negative emotion corpus. The placeholder tokens in these samples were then substituted with the tagged tokens from the target domain randomly. Subsequently, weak labels were assigned to these samples. Finally, the adapted weakly labeled data was merged with the source domain data for the final classification. Bashar et al. (2021) adopted an LSTM-based progressive domain adaptation approach. The method involves performing a language model pre-training on general unlabeled corpora from Wikipedia and Twitter, where the Twitter corpora are topically related to the source and the target domains. It then consists of learning an abusive language classifier on the source domain labeled data while keeping the parameters of the lower layers of the pre-trained model frozen to obtain domain invariant representations. Ludwig et al. (2022) performed target group (i.e. the target of abuse) specific adaptation, where they categorize a domain as the set of samples that only comprise abuse against one target group, unlike our setting where a domain can have multiple target groups.

5.3.2 Problem Formulation

Since we assume access to some amount of unlabeled target corpus with no access to the associated annotations, our task of cross-corpus abusive language detection fits well into the formulation of unsupervised domain adaptation. Let a domain $D = \{X, P(X)\}$, with the feature space X and the marginal probability distribution $P(X)$, and a decision function $f = P(Y|X)$, where Y is the label space. UDA methods aim to adapt f learned on the source domain D_S to the target domain D_T, where only the unlabeled target domain samples X_T and the labeled source domain samples X_S are assumed to be available. We denote the source labels by Y_S. In this work, we use the unlabeled samples X_T for adaptation (without using Y_T) and evaluate the performance over the remaining unseen target samples from D_T.

5.3.3 Analyzed UDA Approaches

Owing to their success in cross-domain sentiment classification, we decide to apply the following pivot-based and domain-adversarial UDA approaches to our task of cross-corpus abusive language detection.

5.3.3.1 Pivot-based approaches

Following Blitzer et al. (2006), most pivot-based approaches extract a set of shared features, called pivots, across domains that are (i) frequent in both source X_S and target X_T; and (ii) are highly correlated with the source domain labels Y_S, i.e their mutual information with the source domain labeled data is higher than a pre-defined threshold. Along with pivots, they extract a complementary set of non-pivot features. These approaches use a pivot feature as a bridge for learning the connection between non-pivot features in the two domains, facilitating adaptation. We analyze the following pivot-based approaches in our task:

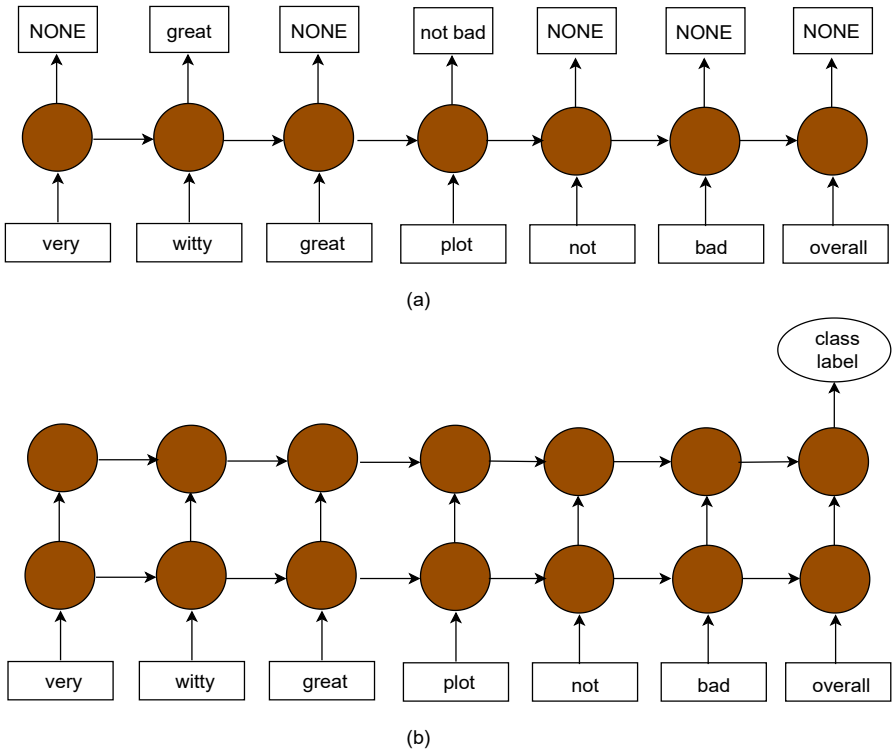


Figure 5.1: (a) Domain adaptation with PBLM representation learning that predicts the next bigram or unigram if one of them is a pivot, predicts NONE otherwise; (b) PBLM-LSTM for the downstream classification task. The figure is adapted from Ziser and Reichart (2018).

Pivot Based Language Modeling (PBLM): This is a sequential pivot-based neural model, proposed by Ziser and Reichart (2018) that has outperformed the prior work on pivot-based representation learning using Autoencoders (AE) (Ziser and Reichart, 2017). Unlike the AE-based methods, which learn a structure-indifferent, single representation for every occurrence of a word in the input corpus, PBLM leverages the sequential nature of the input text and learns context-sensitive representations of words. In particular, it follows a decoupled two-step procedure. Figure 5.1 illustrates the two steps of the PBLM model. In the first step, it employs a Long Short-Term Memory (LSTM) based language model to predict the pivots using other non-pivots features in the unlabeled input samples from both source and target. If the following n-gram is a pivot, it predicts the n-gram; if not, it predicts a NONE tag.

For instance, in the task of sentiment classification, for the domains of movies and kitchen appliances with associated review comments, *great* can be a pivot feature as it is used to describe positive sentiment in both domains. PBLM learns the connection between the non-pivot word *witty* – an adjective commonly used in the domain of books but unlikely to be used for describing kitchen appliances – and the pivot word *great*.

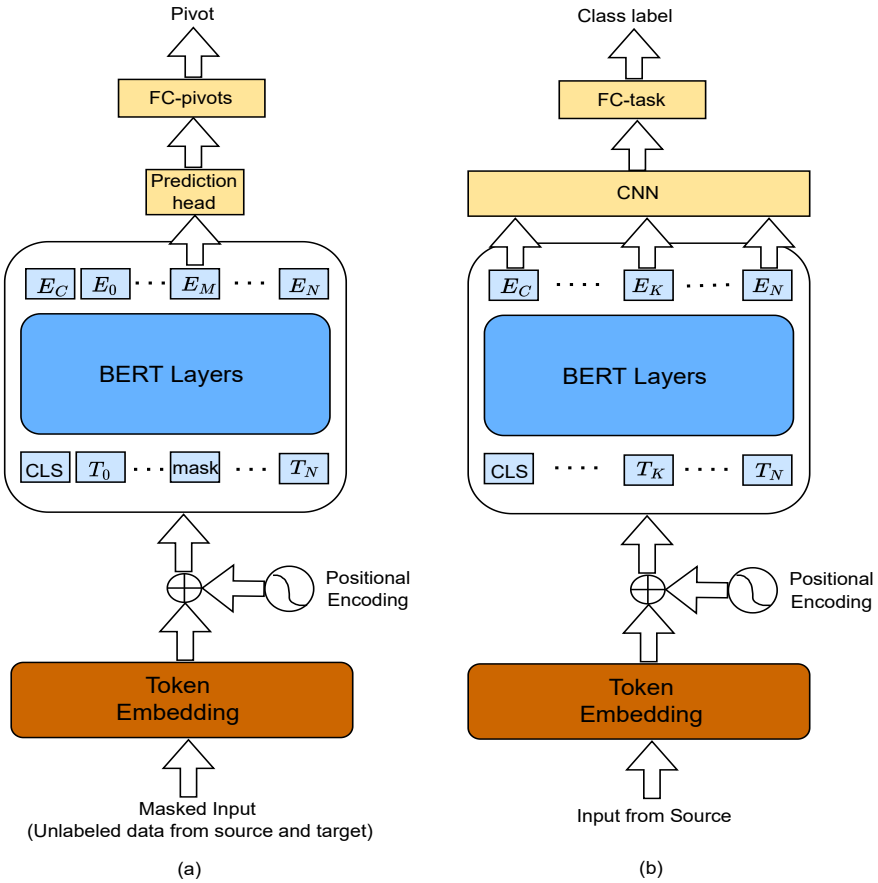


Figure 5.2: Illustration of the (a) MLM training and (b) supervised fine-tuning steps of the PERL model (CNN: Convolutional Neural Network, FC: Fully Connected layer). The figure is adapted from Ben-David et al. (2020).

Similarly, it learns the relation between the non-pivot word *easy* – an adjective usually correlated with positive sentiment in the domain of kitchen appliances but not frequently used for books – and *great*. Thus, PBLM learns the connection between the non-pivot words *witty* and *easy* across domains through the pivot word *great*. Learning to make such connections enable PBLM to perform adaptation between the two domains.

In the next step, it passes the word-level output representations of the source domain samples, obtained from the previous step to CNN (PBLM-CNN) or LSTM-based (PBLM-LSTM) classifiers for the final supervised training with the input texts and their corresponding labels from the source.

Pivot-based Encoder Representation of Language (PERL) (Ben-David et al., 2020): This incorporates pivot-based fine-tuning with the pre-trained BERT (Devlin et al., 2019) model using its Masked Language Model (MLM) objective. An overview of the PERL model is presented in Figure 5.2. To learn the connection between the non-

pivot and pivot words, PERL uses a pivot-based MLM fine-tuning objective. It learns to predict whether the masked unigram/bigram token in the unlabeled input samples from both source and target is a pivot or not; if it is a pivot, the model learns to identify the same. The advantage of such a fine-tuning task is that the BERT parameters are updated to preserve only the minimal information from the non-pivot words that is required to predict the pivots. Since pivots are shared across domains and important for the final classification task, such mapping of non-pivots to pivots can help in improving cross-domain adaptation. The MLM fine-tuning step is followed by supervised task training with a convolution, average pooling, and a linear layer over the encoded representations of the labeled input samples from the source. During the supervised task training, the encoder weights are kept frozen.

Both PBLM and PERL use unigrams and bi-grams as pivots, although higher-order n-grams can also be used.

5.3.3.2 Domain adversarial approaches

The intuition behind these approaches is founded on the theory of domain adaptation by Ben-David et al. (2010), which states that an effective cross-domain transfer can be achieved through feature representations that do not reveal any information about the domain of origin of the input sample.

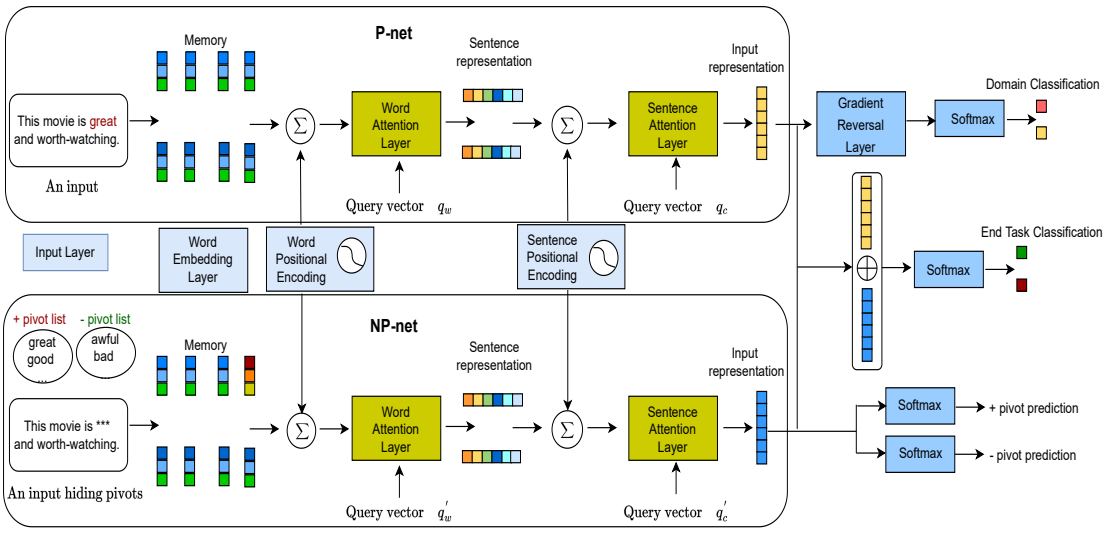


Figure 5.3: Illustration of the HATN framework. The figure is adapted from Li et al. (2017b, 2018).

Hierarchical Attention Transfer Network (HATN) (Li et al., 2017b, 2018): This combines the domain classification-based adversarial training using source and target domain samples with pivot-based representation learning. However, unlike the previously discussed methods, HATN automatically performs pivot construction through an

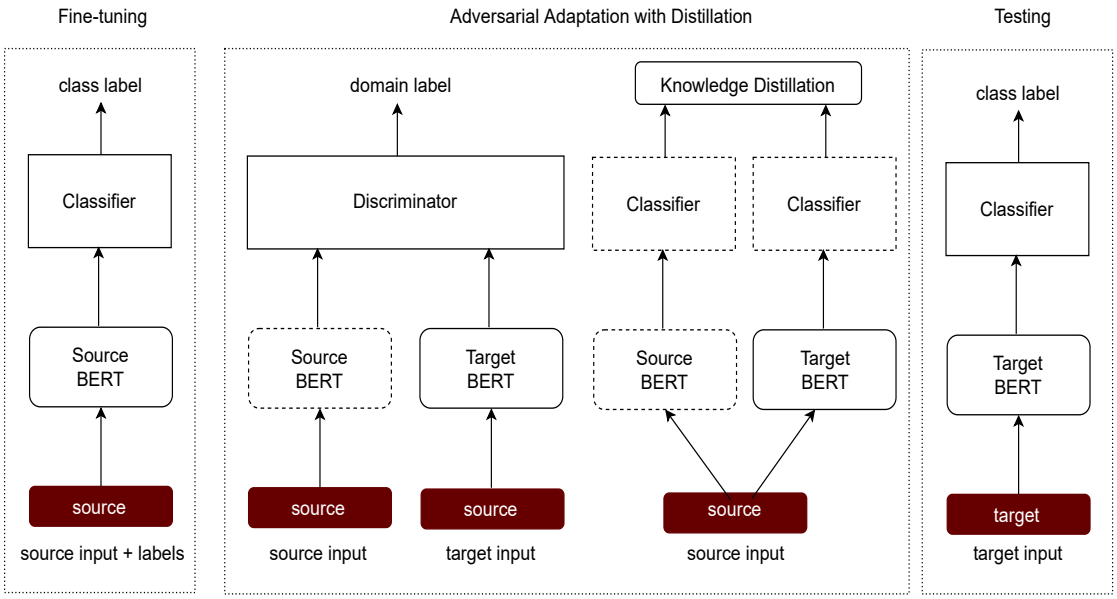


Figure 5.4: Overview of the AAD model. The dashed lines indicate frozen model parameters. The figure is adapted from Ryu and Lee (2020).

adversarial objective. In particular, it comprises two hierarchical attention networks that are trained jointly: *P-net* and *NP-net*, as illustrated in Figure 5.3. P-net extracts the pivots that are both domain-shared and important for the end task. It jointly minimizes a domain adversarial loss and a cross-entropy loss for the end task. The domain adversarial objective aims at obtaining domain-shared representations by training a domain classifier with a Gradient Reversal Layer (GRL) (Ganin and Lempitsky, 2015).

While the domain classifier attempts to identify the domain of origin (source or target) of the input samples from both domains, GRL ensures that the classifier fails to do so by making the representations domain-invariant. The pivots are extracted by identifying the words with the highest word attention present in a sentence with the highest sentence attention in the input samples. NP-net, on the other hand, captures the domain-specific representations and maps them to the domain-shared feature space. It first hides all the pivots in the input samples and passes such samples to the network to learn the existence of positive and negative pivots (based on the positive and negative end-task labels), and the end-task classification. The task-specific cross-entropy loss uses an attention mechanism to train the network for the end task, using the source domain labeled samples.

Adversarial Adaptation with Distillation (AAD) (Ryu and Lee, 2020): This applies domain adversarial training over the pre-trained BERT model. Unlike HATN which uses GRL, AAD incorporates the domain adversarial framework of the Adversarial Discriminative Domain Adaptation (ADDA) (Tzeng et al., 2017). It maintains two

separate BERT encoder models for the source and the target. As the first step, it fine-tunes the source encoder, initialized with the pre-trained BERT model parameters, and the source classifier on the source domain labeled data for the end task. It then initializes the target encoder with the fine-tuned source encoder parameters. The source-encoder parameters and the classifier are kept frozen.

The ADDA framework, as presented in Figure 5.4, basically aims to reduce the distance between the source and target representations by alternately minimizing a domain discriminator loss – that identifies whether the input sample belongs to the source or the target – and an adversarial loss function, inspired by Generative Adversarial Network (GAN) (Goodfellow et al., 2014), which inverts the domain label while updating the target encoder parameters. Ryu and Lee (2020) observed that when BERT is used as the underlying model with the ADDA framework, the discriminative information from the source is catastrophically forgotten (Kirkpatrick et al., 2017), leading to random classification performance. Therefore, they proposed to add a Knowledge Distillation (KD) (Hinton et al., 2015) loss to the adversarial loss function, which aims to transfer knowledge from a teacher model (source) to a student model (target) by minimizing the following objective function L_{KD} (Ryu and Lee, 2020):

$$L_{KD} = t^2 \times \sum_k -\text{softmax}(\boldsymbol{z}_k^T/t) \times \log\big(\text{softmax}(\boldsymbol{z}_k^S/t)\big) \tag{5.1}$$

Here, the temperature value t regulates the amount of knowledge transfer, and z^S and z^T are the logits predicted by the student and the teacher models, respectively.

5.3.4 Adaptation through Masked Language Model Fine-tuning with HateBERT

Rietzler et al. (2020); Glavaš et al. (2020), and Xu et al. (2019) showed that the language model fine-tuning of BERT (using the MLM and the Next Sentence Prediction task) results in incorporating domain-specific knowledge into the model and is useful for cross-domain adaptation. This step does not require task-specific labels. The HateBERT model extends the pre-trained BERT model using the MLM objective over a large corpus of unlabeled abusive comments from Reddit. This is expected to shift the pre-trained BERT model towards abusive language. It is shown by Caselli et al. (2021) that HateBERT is more portable across abusive language corpora, as compared to BERT (see Chapter 3 for details on HateBERT). We, thus, decide to use HateBERT for our task.

In particular, we begin with the HateBERT model and perform MLM fine-tuning incorporating the unlabeled train set from the target corpus. We hypothesize that performing this step should incorporate the variations in the abusive language present in the target corpus into the model. For the classification task, supervised fine-tuning is performed over the MLM fine-tuned model obtained from the previous step, using the source domain labeled data.

5.4 Experimental Setup

5.4.1 Data Description and Pre-processing

We present experiments over the same three abusive language corpora from Twitter, as used in Chapter 4, namely *Davidson* (Davidson et al., 2017), *Waseem* (Waseem and Hovy, 2016) and *HatEval* (Basile et al., 2019). A binary classification task is performed with classes: *abusive* and *non-abusive* and the same train-test splits, as used in Chapter 4, are adopted. Detailed discussions on these corpora and the pre-processing steps followed are provided in Chapter 3.

5.4.2 Evaluation Setup

Given the three corpora listed above, we experiment with all the six pairs of source and target corpora for our cross-corpus analysis. The UDA approaches leverage the respective unlabeled train sets in the target domain D_T for adaptation, along with the train sets in D_S. The abusive language classifier is subsequently trained on the labeled train set in the source domain D_S and evaluated on the test set in D_T. In the "no adaptation" case, the HateBERT model is fine-tuned in a supervised manner on the labeled source corpus train set and evaluated on the target test set. Unsupervised adaptation using HateBERT involves training the HateBERT model on the target corpus train set using the MLM objective. This is followed by a supervised fine-tuning on the source corpus train set.

We use the original implementations of the UDA models[1] and the pre-trained HateBERT[2] model for our experiments. We select the best model checkpoints by performing early stopping of the training while evaluating the performance on the respective validation sets from the source domain D_S. FastText[3] word vectors, pre-trained over Wikipedia, are used for word embedding initialization for both HATN and PBLM. PERL and AAD are initialized with the BERT base-uncased model[4]. In PBLM, we employ the LSTM-based classifier[5]. For both PERL and PBLM, words with the highest mutual information with respect to the source labels and occurring at least 10 times in both the source and target corpora are considered as pivots (Ziser and Reichart, 2018). Following Ben-David et al. (2020), we keep the encoder weights for PERL frozen during the supervised fine-tuning step.

Corpus	Macro-F1	Frequent words in abusive comments
Davidson	93.8±0.1	b*tch, h*e, f*ck, p*ssy, n*gga, ass, f*ck, shit
Waseem	85.5±0.4	#notsexist, #mkr, female, girl, kat, men, woman, feminist
HatEval	51.9±1.7	woman, refugee, immigrant, trump, #buildthatwall, illegal, b*tch, f*ck

Table 5.1: Macro-average F1 (mean ± std-dev) for *in-corpus* classification using supervised fine-tuning of HateBERT.

Source →Target	No-adaptation	Unsupervised Domain Adaptation				
	HateBERT supervised fine-tune only	HateBERT MLM fine-tune on Target	PBLM	PERL-BERT	HATN	AAD-BERT
Hat →Was	66.4±1.1	**68.0**±1.0	57.5±3.4	57.1±1.8	57.3±1.7	60.4±7.8
Was →Hat	**57.8**±0.6	56.5±1.1	51.0±5.2	55.3±0.7	53.5±0.4	55.7±1.3
Dav →Was	**67.5**±0.5	66.7±0.8	57.2±4.8	67.4±1.0	57.5±6.7	41.5±2.8
Was →Dav	60.1±4.4	**67.1**±2.9	46.5±1.3	48.3±1.5	28.0±2.3	35.6±3.7
Hat →Dav	63.8±2.3	**67.8**±1.6	61.8±5.7	62.6±3.8	61.5±5.8	55.2±0.7
Dav →Hat	51.3±0.2	**51.4**±0.4	49.9±0.2	50.3±0.9	50.3±0.5	50.4±3.0
Average	61.2	**62.9**	54.0	56.8	51.4	49.8

Table 5.2: Macro-average F1 scores (mean±std-dev) on different source and target pairs for *cross-corpus* abuse detection (Hat: HatEval, Was: Waseem, Dav: Davidson). The best in each row is marked in bold.

5.5 Results and Analysis

Our evaluation reports the mean and standard deviation of macro-averaged F1 scores, obtained by an approach, over five runs with different random initializations. We first present the in-corpus performance of the HateBERT model in Table 5.1, obtained after supervised fine-tuning on the respective corpora, along with the frequent abuse-related words. As shown in Table 5.1, the in-corpus performance is high for *Davidson* and *Waseem*, but not for *HatEval*. *HatEval* shared task presents a challenging test set and

[1]PBLM: https://github.com/yftah89/PBLM-Domain-Adaptation, HATN: https://github.com/hsqmlzno1/HATN, PERL: https://github.com/eyalbd2/PERL, AAD: https://github.com/bzantium/bert-AAD
[2]https://osf.io/tbd58/
[3]https://fasttext.cc/
[4]https://github.com/huggingface/transformers
[5]CNN classifier obtained a similar performance.

similar performances have been reported in prior work (Caselli et al., 2021). Cross-corpus performances of HateBERT and the UDA models discussed in Section 5.3.1, is presented in Table 5.2. Comparing Table 5.1 and Table 5.2, substantial degradation of performance is observed across the corpora in the cross-corpus setting. This highlights the challenge of cross-corpus performance in abusive language detection.

Cross-corpus evaluation in Table 5.2 shows that all the UDA methods, adopted from the sentiment classification literature, experience a substantial drop in average performance. The average scores remain below 57 when compared to the no-adaptation case of supervised fine-tuning of HateBERT, which yields an overall macro-F1 of 61.2. All these UDA approaches give the worst performance when *Waseem* is used as the source corpus and *Davidson* is the target, with scores remaining below 50. However, the additional step of MLM fine-tuning of HateBERT on the unlabeled train set from target corpus results in an improved performance in most of the cases and improves the average score to 62.9. In the case of *Waseem* →Davidson, the score improves by 7 macro F1 points. In the following sub-sections, we perform a detailed analysis to get further insights into the sub-optimal performance of the other UDA approaches for our task.

5.5.1 Pivot Characteristics in Pivot-based Approaches

To understand the performance of the pivot-based models, we probe the characteristics of the pivots used by these models as they control the transfer of information across the source and target corpora and should ideally exhibit similar behavior across the corpora. As mentioned in Section 5.3.3.1, one of the criteria for pivot selection is their affinity to the available labels. Accordingly, if the adaptation results in better performance, a higher proportion of pivots would have more affinity to one of the two classes. Moreover, this affinity toward the classes should remain similar across the source and target corpora for better adaptation, i.e. if a pivot is more correlated with the abusive class in the source corpus, it should also be correlated with the same class in the target. In the following, we aim to study this particular characteristic across the source train set and the target test set. To compute class affinities, we obtain a ratio of the class membership of every pivot p_i:

$$r_i = \frac{\#\text{abusive comments with } p_i}{\#\text{non-abusive comments with } p_i} \tag{5.2}$$

The ratios obtained for the train set of the source and the test set of the target, for the pivot p_i, are denoted as ${r_i}^s$ and ${r_i}^t$, respectively. A pivot p_i with similar class affinities in both the source train and target test should satisfy:

$$({r_i}^s, {r_i}^t) < 1 - th \text{ or } ({r_i}^s, {r_i}^t) > 1 + th \tag{5.3}$$

Here, th denotes the threshold. The threshold is used to avoid considering the pivots having less prominent affinities to one of the classes, resulting in having ratios close to 1. Ratios less than $(1 - th)$ indicate affinity towards the non-abusive class, while those greater than $(1 + th)$ indicate affinity towards the abusive class. We set the threshold th = 0.3 as we find it to work well for our experiments. For every source →target pair, we

select the pivots that satisfy Equation 5.3 and calculate the percentage of the selected pivots as:

$$\text{perc}_{s \to t} = \frac{\#\text{pivots satisfying Equation 5.3}}{\#\text{Total pivots}} \times 100 \tag{5.4}$$

This indicates the percentage of pivots having a similar affinity towards one of the two classes across the source and target corpora. We now analyze this percentage in the best and the worst case scenarios of PBLM[6].

Worst cases: For the worst case of *Waseem* →*Davidson*, Equation 5.4 yields a low $\text{perc}_{s \to t}$ of 18.8%. This indicates that the percentage of pivots having similar class affinities, across the source and the target, remains low in the worst-performing pair. This is most likely because the pivot selection criteria are not sufficient to extract similarly behaving shared features across these corpora.

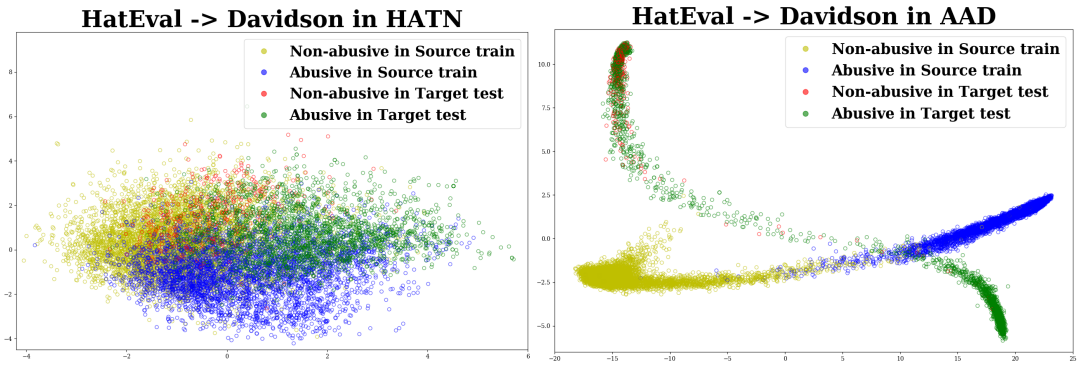
Best case: The best case in PBLM corresponds to *HatEval* →*Davidson*. In this case, Equation 5.4 yields a relatively higher $\text{perc}_{s \to t}$ of 51.4%. This is because the pivots extracted here involve a lot of profane words. Since in *Davidson*, the majority of abusive content involves the use of profane words (as also reflected in Table 5.1), the pivots extracted by PBLM can represent the target corpus well in this case.

5.5.2 Domain Adversarial Approaches

On average, the adversarial approach of HATN performs slightly better than AAD. In order to analyze the difference, we investigate the representation spaces of the two approaches for the best case of HATN i.e. *HatEval* →*Davidson* and the worst case for both the approaches i.e. *Waseem* →*Davidson*. To this end, we apply the Principal Component Analysis (PCA) to obtain the two-dimensional visualization of the feature spaces from the train set of the source corpus and the test set of the target corpus. The PCA plots for *HatEval* →*Davidson* and *Waseem* →*Davidson* are shown in Figure 5.5 and 5.6, respectively. Adversarial training in both the HATN and AAD models tends to bring the representation regions of the source and target corpora close to each other. At the same time, the separation of abusive and non-abusive classes in the source train set seems to be happening in both models.

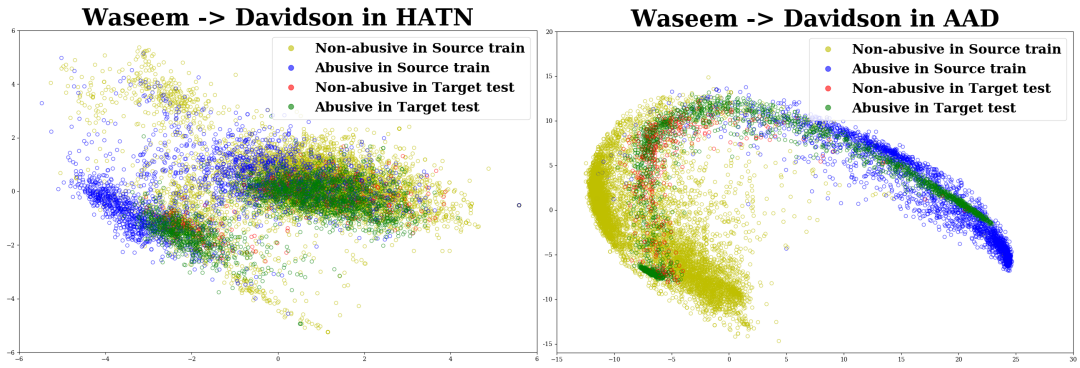
In the case of *HatEval* →*Davidson*, for the representation space of AAD, samples corresponding to abusive and non-abusive classes in the target test set do not follow the class separation seen in the source train set. But in the representation space of HATN, samples in the target test set appear to follow the class separation exhibited by its source train set. Considering the abusive class as positive, this is reflected in the higher number of *True Positives* in HATN as compared to that of AAD for this pair (#TP for HATN: 1393, #TP for AAD: 1105), while the *True Negatives* remain almost the same (#TN for HATN: 370, #TN for AAD: 373). Also, in this case, HATN has a better macro-F1 score (61.5) compared to AAD (55.2).

[6]Pivot extraction criteria are the same for PBLM and PERL and similar percentages are expected with PERL.



(a) (b)

Figure 5.5: (Best viewed in color) PCA based visualization of *HatEval →Davidson* in the adversarial approaches.



(a) (b)

Figure 5.6: (Best viewed in color) PCA based visualization of *Waseem →Davidson* in the adversarial approaches.

For *Waseem →Davidson*, the representation space of HATN shows that the adaptation seems to make the samples in the source train set less discriminative with respect to the classes, and hence the target test samples are also non-discriminative. With AAD, the source train samples remain discriminative with respect to the classes, most likely because it uses two different encoders to encode the source and target domain samples. While the source encoder parameters are kept frozen in AAD once they are fine-tuned on the source, the representations for the source and target instances are learned jointly in HATN during adaptation. The non-abusive sample representations in AAD from the target test set seem to overlap to some extent with the non-abusive representations in the source train samples, but the abusive samples in the target span the source representation spaces of both classes. In terms of macro-F1 scores, HATN performs worse (28.0) in this case compared to AAD (35.6). AAD yields a higher number of *True Positives* compared to HATN (#TP for HATN: 602, for AAD: 746), while the number of *True Negatives*

remain similar (#TN for HATN: 367, for AAD: 370).

One of the limitations of these domain adversarial approaches is the class-agnostic alignment of the common source-target representation space. As discussed in Saito et al. (2018), methods that do not consider the class boundary information while aligning the source and target distributions, often result in having ambiguous and non-discriminative target domain features near class boundaries. Besides, an alignment using the class boundary information can be achieved without having access to the target domain class labels (Saito et al., 2018). Therefore, an effective alignment should also attempt to minimize the intra-class, and maximize the inter-class domain discrepancy (Kang et al., 2019).

Source →Target	**HBERT MLM on all 3 corpora**	**HBERT MLM on Source + Target**	**HBERT MLM on Target**
Hat →Was	**69.7**±0.8	68.9±0.6	68.0±1.0
Was →Hat	**57.2**±1.4	56.8±1.1	56.5±1.1
Dav →Was	60.2±0.7	58.8±0.8	**66.7**±0.8
Was →Dav	63.4±3.9	63.4±3.9	**67.1**±2.9
Hat →Dav	66.6±1.1	66.7±2.1	**67.8**±1.6
Dav →Hat	51.4±0.2	**51.5**±0.1	51.4±0.4
Average	61.4	61.0	**62.9**

Table 5.3: Macro-average F1 scores (mean ± std-dev) for Masked Language Model fine-tuning of HateBERT (HBERT MLM) over different corpora combinations along with supervised fine-tuning on the source; Hat: HatEval, Was: Waseem, Dav: Davidson. The best in each row is marked in bold.

5.5.3 MLM Fine-tuning of HateBERT

It is evident from Table 5.2 that the MLM fine-tuning of HateBERT along with the subsequent supervised fine-tuning over the source corpus results in improved performance in the majority of the cases. We investigate the impact of MLM fine-tuning over different combinations of the source and target corpora, combined with supervised fine-tuning on the source corpus, in order to identify the best configuration. These include a combination of the train sets from all three corpora, combining the source and target train sets, and using only the target train set. Table 5.3 shows that MLM fine-tuning over only the unlabeled target corpus results in the best overall performance. This is in agreement with Rietzler et al. (2020) who observe a better capture of domain-specific knowledge with MLM fine-tuning only on the target domain.

5.5.4 Bridging the Gap between PERL and HateBERT MLM Fine-tuning

Since PERL originally incorporates BERT, Table 5.2 reports the performance of PERL initialized with the pre-trained BERT model. As discussed in Section 5.3.1, PERL applies MLM fine-tuning over the pre-trained BERT model, where only the pivots are predicted rather than all the masked tokens. Following Ben-David et al. (2020), after the BERT encoder weights are learned during the MLM fine-tuning step of PERL, they are kept frozen during supervised training for the classification task. Only the weights of the convolution and the fully connected layers (see Figure 5.1) are updated during this step.

As an additional verification, we try to leverage the HateBERT model for initializing PERL in the same way as BERT is used in the original PERL model, with frozen encoder layers. As shown in Table 5.4, this does not result in substantial performance gains over PERL-BERT on average. Initializing with the HateBERT model increases the average macro F1 performance from 56.8 with BERT to 57.7 with HateBERT. The scores for all the individual pairs remain below the scores obtained from the 'no-adaptation' scenario in Table 5.2. As a further extension, we update all the layers in PERL during the supervised training step and use the same hyper-parameters as those used for HateBERT (Caselli et al., 2021).[7] This results in improved performance from PERL. The average scores increase to 60.8 which is slightly below the average score of 61.2 obtained from the 'no-adaptation' scenario in Table 5.2. However, it still remains behind the best-performing HateBERT model with MLM fine-tuning on target which gives an average score of 62.9.

Source →Target	**PERL-BERT (frozen encoder layers)**	**PERL-HBERT (frozen encoder layers)**	**PERL-HBERT (with layer updates)**
Hat →Was	57.1±1.8	63.2±1.7	**68.3**±0.8
Was →Hat	55.3±0.7	55.0±0.9	**57.8**±0.8
Dav →Was	**67.4**±1.0	65.9±1.3	57.3±3.1
Was →Dav	48.3±1.5	48.1±3.7	**64.4**±2.1
Hat →Dav	62.6±3.8	63.6±0.9	**66.1**±1.8
Dav →Hat	50.3±0.9	50.4±0.6	**51.1**±0.3
Average	56.8	57.7	**60.8**

Table 5.4: Macro-average F1 scores (mean ± std-dev) of PERL initialized with BERT and HateBERT (HBERT) with frozen encoder layers, and PERL initialized with HateBERT with updates across all layers, for all the pairs (Hat: HatEval, Was: Waseem, Dav: Davidson). The best in each row is marked in bold.

[7] Note that the ablation study by Ben-David et al. (2020) discusses the effect of the number of unfrozen encoder layers only in the MLM fine-tuning step, but not in the supervised training step for the end task.

Source →Target	AAD-BERT	AAD-HateBERT
Hat →Was	60.4±7.8	**60.8**±6.4
Was →Hat	**55.7**±1.3	53.7±1.6
Dav →Was	41.5±2.8	**46.1**±2.6
Was →Dav	35.6±3.7	**41.4**±4.1
Hat →Dav	**55.2**±0.7	54.3±1.6
Dav →Hat	**50.4**±3.0	47.7±5.5
Average	49.8	**50.7**

Table 5.5: Macro average F1 scores (mean ± std-dev) of AAD initialized with BERT and HateBERT across all the pairs (Hat: HatEval, Was: Waseem, Dav: Davidson).

5.5.5 AAD with HateBERT

Since AAD also incorporates BERT, we can similarly replace BERT with HateBERT. Table 5.5 reports the performance of AAD initialized with HateBERT. It is observed that the overall performance with AAD-HateBERT (50.7) is close to that obtained with AAD-BERT (49.8).

5.5.6 Source Corpora Specific Behaviour

From our results and analysis, we observe that in general, when models are trained over *HatEval*, they are found to be more robust towards addressing the shifts across corpora. This is reflected in the higher cross-corpus performance scores obtained when *HatEval* is used as the source corpus (see Table 5.2). One of the primary reasons is that *HatEval* captures wider forms of abuse directed towards both immigrants and women. The most frequent words in Table 5.1 also highlight the same. The corpus involves a mix of implicit as well as explicit abusive language.

On the contrary, models trained over *Waseem* are generally unable to adapt well in cross-corpus settings. Since only tweet IDs were made available in *Waseem*, we observe that our crawled comments in this corpus rarely involve abuse directed towards target groups other than women (99.3% of the abusive comments are sexist and 0.6% racist). This is because the majority of these comments have been removed before crawling. Besides, *Waseem* mostly involves subtle and implicit abuse and less use of profane words.

5.6 Conclusion and Future Work

This chapter analyzed the efficacy of some successful Unsupervised Domain Adaptation approaches of cross-domain sentiment classification in cross-corpus abusive language detection. Our experiments highlighted some of the problems with these approaches that render them sub-optimal in the cross-corpus abuse detection task. The extraction of pivots, in the pivot-based models, is not optimal enough to capture the shared space

across domains. This is most likely due to the complexity of abusive language, which makes it difficult to find n-grams pivots whose meaning and behavior remain unchanged in different contexts across corpora. The domain adversarial methods used underperform substantially as the class boundaries learned on the source corpus do not typically generalize well to the target corpus. The analysis of the Masked Language Model fine-tuning of HateBERT on the target corpus displayed improvements in general as compared to only fine-tuning HateBERT over the source corpus, suggesting that it helps in adapting the model towards target-specific language variations. The overall performance of all the approaches, however, indicates that building robust and portable abuse detection models is a challenging problem, far from being solved. This calls for approaches that are designed specifically to suit the challenges of cross-domain abusive language detection.

Future work along the lines of domain adversarial training could explore methods that can learn class boundaries that generalize well to the target corpus while performing alignment of the source and target representation spaces. Such an alignment can be performed without target class labels by minimizing the intra-class domain discrepancy (Kang et al., 2019). Pivot-based approaches could explore pivot extraction methods that account for higher-level semantics of abusive language across the source and target corpora.

6 Penalizing Spurious Correlations with Model Explanations

6.1 Introduction

In the previous two chapters, we considered the detection of abusive language, which covers all forms of toxic, hateful, and offensive language, as discussed in Chapter 2. However, from a practical perspective, moderating all occurrences of toxicity may not often be desirable, especially given the unrestricted nature of the Internet and the right to freedom of expression. On the other hand, hate speech – a subcategory of abusive language – is highly concerning as it promotes discrimination against underrepresented groups or minorities (e.g. *women*, *jews*, *blacks*, etc.) (Jurgens et al., 2019; Waseem and Hovy, 2016). We have discussed in detail what constitutes hate speech in Chapter 2. This creates a pressing need to focus specifically on the task of hate speech detection. Our work in this chapter thus attempts to address cross-corpus hate speech detection. The work presented in this chapter is published as Bose et al. (2022a) and Bose et al. (2022b).

One of the common reasons why a vanilla classifier yields poor out-of-domain performance is that it tends to learn more from domain-specific features than domain-invariant features, which biases the classifier toward the source domain in which it has been trained (Ye et al., 2021; Chen et al., 2019). In this chapter, we thus aim to reduce this source-specific bias and improve the generalizability of the source classifiers to the target domain by applying additional constraints during training using *model explanations*. In particular, *two different domain adaptation approaches* are proposed for improving the cross-corpus performance of hate speech classifiers.

The relative sparsity of hateful content in the real world requires crawling of many of the standard hate speech corpora through keyword-based sampling (Poletto et al., 2021), rather than random sampling. Thus, when hate speech classifiers (D'Sa et al., 2020; Mozafari et al., 2019; Badjatiya et al., 2017) are trained on such corpora, they often learn undesirable corpus-specific features called spurious correlations from the training corpus (Wiegand et al., 2019) leading to substantial performance degradation when evaluated on a corpus with a different distribution (Yin and Zubiaga, 2021; Bose et al., 2021b; Florio et al., 2020; Arango et al., 2019; Swamy et al., 2019; Karan and Šnajder, 2018). For instance, Wiegand et al. (2019) show that in a hate-speech dataset (Waseem and Hovy, 2016), neutral corpus-specific terms, like '*football*', '*commentator*', etc., discussing the role of women in sports, are highly correlated with the hate label, restricting its generalizability.

Target corpus utterances	Actual	Predicted
Genocide is never ok	non-hate	hate
Women are goddesses	non-hate	hate

Table 6.1: Examples of spurious correlations learned by a source classifier between the shaded terms and the hate label.

Recent works have proposed regularization mechanisms to penalize spurious correlations by attempting to explain model predictions using feature attribution methods and align features with prior task-specific knowledge (Ross et al., 2017; Rieger et al., 2020; Adebayo et al., 2020). These are methods to extract post-hoc model explanations, which assign importance scores to input terms that contribute more toward a particular prediction (Lundberg and Lee, 2017). For instance, Liu and Avci (2019) penalized the attributions assigned to terms contained in a manually curated dictionary consisting of group identifiers that are often known to be targets of hate. Kennedy et al. (2020) extracted group identifiers manually from the top terms indicated by a bag-of-words logistic regression model trained on the source corpus. However, regularizing only group identifiers limits the coverage of such approaches, and may not capture other forms of corpus-specific correlations learned by the classifier constraining its performance on a new corpus. Moreover, such manually curated lists may not always remain up-to-date because new terms emerge frequently (Grieve et al., 2018). This calls for methods that can automatically extract such spurious correlations and prevent the classifier from getting biased toward them.

Besides, in the task of detecting objects in images, Zunino et al. (2021) used a domain classifier, trained to differentiate between domains, and visually identified that the irrelevant background information in the images is domain-specific. Thus, they enforce the model explanations to align with the ground-truth annotations highlighting the objects in the image. This inspires us to propose a new domain adaptation approach in hate speech employing a domain classifier, but without having access to such annotations for aligning the attribution scores.

In this chapter, two different domain adaptation approaches are proposed that perform *Dynamic Model Refinement* (*D-Ref*) to reduce the effect of corpus-specific spurious correlations and improve the cross-corpus performance of hate speech classifiers.

- First, we hypothesize that the classification errors in a small labeled subset from the target can reveal spurious correlations between terms and hate speech labels learned from the source (see Table 6.1). To this end, *D-Ref-I*, a new method to identify and penalize spurious terms using feature attribution methods, is proposed.
- Second, we hypothesize that domain-specific terms that are simultaneously predictive of the hate-speech labels are instrumental in restricting the domain invariance of the hate-speech classifier. We propose *D-Ref-II*[1], a method that performs do-

[1]We call this method 'Dom-spec' in our published article. However, we think 'D-Ref-II' is more appropriate here for maintaining uniformity and because both D-Ref-I (referred to as D-Ref in the published article) and D-Ref-II perform dynamic refinement of the source model.

main classification-based source-specific term penalization. It employs a domain classifier to automatically extract the terms that help in identifying the source domain compared to the *unlabeled* target domain, and use feature-attribution scores to identify the subset important for hate-speech classification from the source.

Our methods, through penalization of these terms, automatically enforce the source domain hate speech classifier to focus on relatively more domain-invariant content and the general contextual meaning of instances, rather than on specific terms. Compared to other methods that transform high-dimensional intermediate representations to reduce the domain discrepancy, such as domain adversarial learning (Ryu and Lee, 2020; Tzeng et al., 2017) (discussed in Chapter 5), our proposed approaches make the adaptation more explainable. We demonstrate that both D-Ref-I and D-Ref-II improve the overall cross-corpus performance independently and in combination with pre-defined dictionaries.

The remainder of the chapter is organized as follows: Section 6.2 gives a general overview of feature attribution methods, the prior work, and the attribution methods used in our approaches. Section 6.3 describes the proposed approaches. Section 6.4 presents our experimental setup. Our results, analysis, and discussion corresponding to the results are present in Section 6.5. Finally, Section 6.6 concludes the chapter.

6.2 Feature Attribution Methods

Correctly interpreting the predictions of a model is crucial to increase the end user's trust in the model, study the model's behavior (Adebayo et al., 2020), provide insights for possible improvements, and enhance the understanding of the process that is modeled. In tasks where models assist human decision-making, such as hate speech detection, interpreting the model's output can facilitate informed and hopefully better decisions. Due to the inherently transparent nature of simple models (e.g. linear models), they are more preferred (Lou et al., 2012) in some tasks requiring high-stake decisions (Alemzadeh et al., 2016; Gosiewska et al., 2021) in spite of their lower levels of performance compared to complex models like neural networks, which are harder to interpret. However, given the wide availability of data resources, the benefits of using complex models have expanded. This has driven research on improving the interpretations or explanations of complex models (Bach et al., 2015; Datta et al., 2016; Ribeiro et al., 2016; Sundararajan et al., 2017; Shrikumar et al., 2017).

A model explanation is typically a subset of the input that has the highest influence on the final prediction. Post-hoc methods provide explanations after the model has been trained and are usually model-agnostic (Murdoch et al., 2019). Feature attributions are a class of such methods that assign importance scores to the input features as per their contribution to the model prediction. These scores can then be used to interpret the model decisions. Formally given an input $x = (x_1, x_2, ..., x_n) \in R^n$ a function $F : R^n \rightarrow [0, 1]$ representing the model, and the corresponding prediction y, feature attribution methods assign the vector $a = (a_1, a_2, ..., a_n)$ to the input x, where a_i is the attribution score of x_i for y. Figure 6.1 presents an illustration of the use of feature

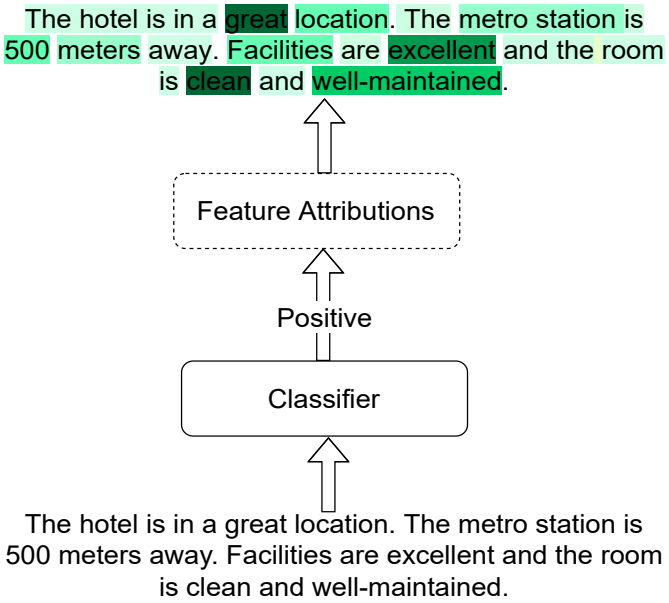


Figure 6.1: An example of feature attributions to find the most important subset of the input (post-hoc explanation) that contributed to the classifier's decision. The darker the shades, the higher the attribution scores.

attributions to find the most important subset of an input contributing to the model decision.

Different feature attribution methods have been proposed in the literature, such as Guided back-propagation (Springenberg et al., 2014), Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016), Layer-wise Relevance Propagation (LRP) (Binder et al., 2016), Deep Learning Important FeaTures (DeepLIFT) (Shrikumar et al., 2017), Integrated Gradients (Sundararajan et al., 2017), Shapley values from game theory (SHAP) (Lundberg and Lee, 2017), Scaled Attention (Serrano and Smith, 2019), InputXGrad (Kindermans et al., 2016; Atanasova et al., 2020), etc. Recent studies (Jain and Wallace, 2019; Serrano and Smith, 2019) showed that attentions on their own cannot always be considered faithful explanations. Explanations are called faithful when they correctly provide the true reasoning behind the model's predictions (Jacovi and Goldberg, 2020). Jain and Wallace (2019) demonstrated that equivalent predictions as that from the original attention distributions can be obtained using alternative attention distributions, indicating that attention weights do not provide robust explanations. On the other hand, Wiegreffe and Pinter (2019) and Vashishth et al. (2019) showed that attention weights, in certain cases, can provide robust explanations. Serrano and Smith (2019) evaluated the faithfulness of attention-based explanations by removing terms in input sequences to observe how quickly decision flips occur. They found that feature attributions in the form of attention weights scaled with their corresponding gradients can potentially give a better indication of term importance for model decisions than using

only attention weights. Chrysostomou and Aletras (2021) proposed a class of methods called Task-Scaling (TaSc) to scale the original attention weights for improving the faithfulness of attention-based explanations. Furthermore, Chrysostomou and Aletras (2022a) performed an empirical study to assess the faithfulness of post-hoc explanations in out-of-domain settings.

Some of these methods like Integrated Gradients, DeepLift, LRP, etc. satisfy a desirable axiom called *completeness*. It posits that the sum of the attributions should be equal to the difference between the output of the model F at the input x and the baseline or reference input $x^{'} \in R^{n}$. The *baseline* is commonly chosen as a background value of the input that is not informative and hence produces a high-entropy prediction representing uncertainty. For image models, the baseline is typically a black image, while for models inputting text, usually the zero embedding vector is considered as the baseline input.

Feature attribution methods are applicable to a wide variety of tasks. Sundararajan et al. (2017) studied the applicability of Integrated Gradients for object recognition, predicting the severity of diabetic retinography using images (a diabetic condition impacting the eyes), question classification, neural machine translation, determining whether a given input molecule is active against a specific target like protein or enzyme, etc. Shrikumar et al. (2017) applied DeepLIFT for digit classification and classification tasks on DNA (Deoxyribonucleic acid) sequence inputs (strings comprising the alphabets {A, C, G, T}) to evaluate the obtained attribution scores.

6.2.1 Model Regularization with Attributions

Feature attribution methods have been used to analyze and correct model decisions by inserting human-provided domain knowledge into models. One of the pioneering works in this direction is by Ross et al. (2017) who proposed the regularization of a model such that its explanations conform with binary annotations provided by domain experts. For this, they used the input gradients-based attribution method. They further showed that in the absence of such domain knowledge, their approach could also help in obtaining models with qualitatively different decision boundaries that yield similar accuracies. These models with different explanations could then be presented to domain experts for inspection and picking the one that is '*right for the right reasons*'.

Other works extended the field by using different strategies to incorporate prior knowledge into the models. Bao et al. (2018) used discrete human provided rationales – features marked by domain experts as justifications for predictions – in a low-resource target task and mapped them to attention scores for improving the performance of the target task. They used resource-rich source tasks to learn the mapping between machine-generated explanations (Lei et al., 2016) and high-quality attentions and transferred this mapping to the target task. Along similar lines, Du et al. (2019) regularized the training process of deep neural networks such that their local explanations are aligned with the rationales provided by domain experts, and proposed to generate sparse explanations when such expert rationales are unavailable. They used an omission-based attribution method (Li et al., 2016b). Rieger et al. (2020) used the contextual decomposition-based attributions

(Murdoch et al., 2018) and aligned them with prior task-specific domain knowledge to prevent spurious correlations. Erion et al. (2020, 2021) introduced the attribution method of expected gradients and encouraged the attributions to incorporate some high-level task-specific priors such as smoothness and sparsity. Weinberger et al. (2020) extended such attribution priors to a deep attribution prior framework using prior information about features, called the meta-features. Adebayo et al. (2020) investigated the efficacy of post-hoc explanation methods in model debugging and found that they could identify spurious correlations but could not conclusively detect the mislabeled training examples. Zunino et al. (2021) used saliency maps to provide periodic feedback to an image classification model to focus on the image regions corresponding to the ground-truth objects rather than the background information. This helped in improved generalization in out-of-domain settings. Mitsuhara et al. (2021) used attention maps to enforce learning from input regions that are known to be important.

6.2.2 Use of Explanations in Hate Speech Detection

There are several prior works that have focused on mitigating unintended bias in hate speech detection models toward certain terms, such as group identifiers. These are terms identifying social groups that are frequent targets of hate. In this direction, Liu and Avci (2019) manually prepared a list of curated group identifiers, and penalized the L_2 distance between their attribution scores (with Integrated Gradients) and a target attribution value of zero. Kennedy et al. (2020) extracted group identifiers manually from the top terms indicated by a bag-of-words logistic regression model trained on the source domain. Yao et al. (2021), on the other hand, collected human-provided compositional explanations regarding different spurious patterns incurred by a source model on the target domain. Then they extended such explanations to the unlabeled target domain data through logic rules. The quantified interactions amongst features and attribution scores from two different methods – Integrated Gradients and Sampling and Occlusion (Jin et al., 2020) – were used to penalize the models. Although Yao et al. (2021) did not use pre-defined lists for refining language models in different target domains, their method still requires refinement advice from human annotators.

A work contemporary to ours, for automatic reduction of lexical overfitting, was proposed by Attanasio et al. (2022). They used entropy-based regularization to penalize terms with low self-attention entropy in order to reduce undesirable bias. However, they did not study the generalization performance in terms of cross-corpus hate speech detection. Another contemporary work by Nejadgholi et al. (2022) proposed the use of the Testing Concept Activation Vector (TCAV), an explanation method from computer vision, to measure the sensitivity of a trained model to the human-defined concepts of implicit and explicit hate. They adjusted TCAV to provide a metric called *degree of explicitness*, which was further used to explain the generalizability of a model on new data and guide data augmentation.

6.2.3 Feature Attribution Methods Used

We now provide a brief overview of the feature attribution methods used in this chapter to identify the contribution of input features to the decision of our Transformer-based model. These attribution methods have been widely used in recent studies (Chrysostomou and Aletras, 2021, 2022b).

Scaled Attention ($\alpha\nabla\alpha$) (Serrano and Smith, 2019): As discussed in the previous section, it has been found that explanations obtained by attention weights may not always be robust. Serrano and Smith (2019) showed that scaling an attention weight with its gradient works better than only using the attention weights to predict the importance of terms for model predictions. Given the i^{th} element of a sequence, here the attention weight α_i corresponding to the final layer of the model is scaled with its corresponding gradient $\nabla\alpha_i = \frac{\delta\hat{y}}{\delta\alpha_i}$, where $\hat{y}$ is the predicted label. So the attribution score is given by $\alpha_i \times \nabla\alpha_i$.

Integrated Gradients (IG) (Sundararajan et al., 2017): This method is based on the notion that the gradient of the prediction function F with respect to input can indicate the sensitivity of the prediction for each input dimension. As such, it aggregates the gradients along a path from an uninformative reference input (e.g. zero embedding vector) toward the actual input such that the predictions change from uncertainty to certainty. Using a path integral, IG accumulates the gradients of F relative to the input along this path. The main reason to use a path integral over an overall gradient at the input is to counteract the possibility of saturation of the gradients around the input. Although there are an endless number of possible paths from a baseline to an input point, IG chooses the straight line path. Formally, the integrated gradient along the i^{th} dimension is defined as:

$$\text{IG}_i(x) ::= (x_i - x_i^{'}) \times \int_{\alpha=0}^{1} \frac{\delta F(x^{'} + \alpha \times (x - x^{'}))}{\delta x_i} \, d\alpha \quad ; \tag{6.1}$$

where x is the actual input, $x^{'}$ is the baseline input, and $\frac{\delta F(x)}{\delta x_i}$ is the gradient of the function $F(x)$ along the i^{th} dimension.

Deep Learning Important FeaTures (DeepLIFT/DL) (Shrikumar et al., 2017): This aims to explain the difference in the output from a reference or baseline output in terms of the difference of the input x and a reference or baseline input $x^{'}$. Given a target output neuron t, a reference activation t^0 of t, $\Delta x_i = x_i - x_i^{'}$, and $\Delta t = t - t^0$, it computes the contribution scores $C_{\Delta x_i \Delta t}$ of each input neuron x_i that are necessary and sufficient to compute t, such that

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t. \tag{6.2}$$

6.3 Proposed Approaches

In this section, a description of the general framework of the proposed approaches, *D-Ref-I* and *D-Ref-II*, is provided. The pre-trained BERT (Devlin et al., 2019) model is used for both approaches. Note that, unlike the last two chapters, HateBERT is not used as our underlying model. This is because, as discussed in Chapter 3, HateBERT is an abusive language-inclined version of BERT, where abusive language is usually considered a superset of hate speech (discussed in more detail in Chapter 2). This might make the model more sensitive toward slur terms and other profanities. As a result, the spurious terms extracted with such a model may not necessarily reflect the undesirable correlations present in the source training corpus, but may also reflect the correlations introduced due to re-training of the model on an abusive language corpus. Since we only deal with hate speech in this chapter and intend to correct spurious correlations in the training corpus, we decide to use a model that is not inclined toward the abusive language phenomenon beforehand.

6.3.1 Dynamic Model Refinement-I (D-Ref-I)

The first approach involves two slightly different problem settings: **(i)** In the first setting, we assume that during training our hate speech classification model has access to the *labeled source domain training data* D_S^{train} *and a small labeled validation set* D_T^{val} *from the target domain*. This follows a low-resource setting, similar to Maharana and Bansal (2020), where the small target domain validation set is used in their policy search approach in the source domain for the task of cross-domain reading comprehension, but full training is unavailable on that data. **(ii)** In the second setting, the model, in addition to D_S^{train} and D_T^{val}, has further access to a *larger set of unlabeled instances from the target domain*, denoted as D_T^{train}.

Our Dynamic Model Refinement-I (D-Ref-I) approach, in both the problem settings, consists of 2 recurring steps across epochs: (a) we first extract a set of spurious terms using D_T^{val} at the end of every epoch; and (b) then the extracted terms are penalized during the next epoch. Note that since D-Ref-I extracts and penalizes spurious terms based on the errors on D_T^{val}, it extracts only the terms that are present both in the source and target corpora but are not predictive of the hate and non-hate labels in the target corpus in the same way as they are in the source-corpus. Figure 6.2 presents the schematic diagram of D-Ref-I and Algorithm 1 provides the algorithm.

(a) Extraction of Spurious Terms

Global term-ranking in source corpus: We first begin with identifying the terms from D_S^{train} that are highly correlated with hate/non-hate labels. These terms are suitable candidates for causing source-specific spurious correlations, restricting generalizability to a new corpus.

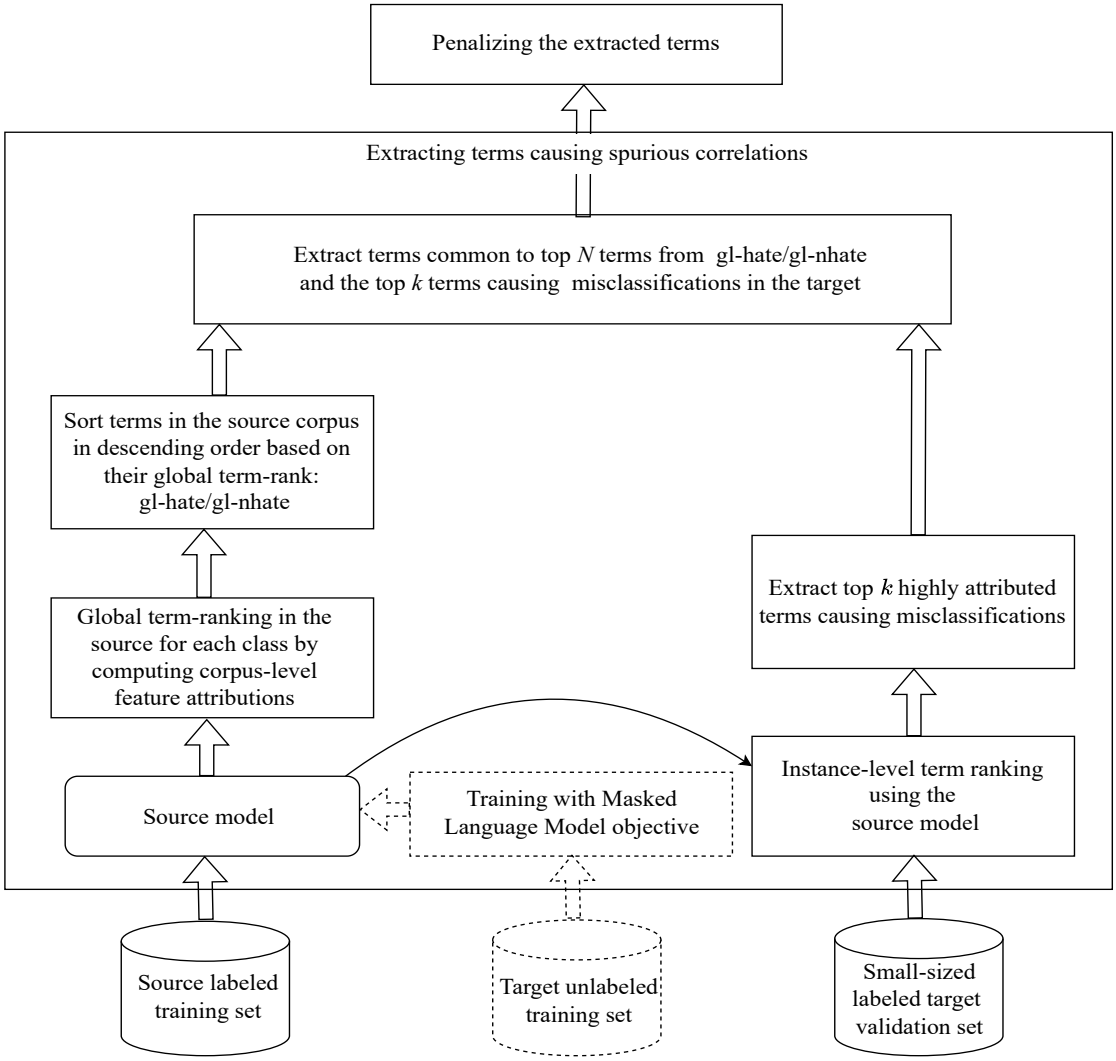


Figure 6.2: Schematic diagram of D-Ref-I. The dotted lines indicate the portion that is used only when the target unlabeled data is available.

For that purpose, at the end of every training epoch ep_i, we first obtain the global class-specific ranked list of terms from D_S^{train}. This is achieved by computing global attributions per term te and class c (gl-atr$^c_{te}$) from its attribution per instance j (loc-atr$^j_{te}$) averaged across all training instances classified as c by the source model trained until epoch ep_i:

$$\text{gl-atr}^c_{te} = \frac{\sum_{j=1}^{|D_S^{train}|} 1_{\hat{y_j}=c} \text{loc-atr}^j_{te} \; \forall \text{occurrence of } te \text{ in j}}{\sum_{j=1}^{|D_S^{train}|} 1_{\hat{y_j}=c} \#(\text{occurrence of } te \text{ in j})} \tag{6.3}$$

Here $c \in \{\text{hate, non-hate}\}$, $\hat{y}$ is the predicted class and 1 is the indicator function. Prior

to this, loc-atr$^{j}_{te}$ are individually normalized using sigmoid to obtain values in a closed range. Rarely occurring terms and stop-words are not considered for the global ranking. The gl-atr$^{c}_{te}$ values are sorted from the highest globally attributed term to the lowest, which yields two ranked term-lists [gl-hate, gl-nhate]$_{ep_i}$.

Instance-level local ranking in target corpus: We hypothesize that terms highly correlated with hate/non-hate classes in the source, but also causing misclassifications in the target, should most likely contribute to spurious source-specific correlations, and may not be important for hate speech labels. Thus, we identify the terms that cause misclassifications in D^{val}_T, and then obtain a list of spurious terms *dynamically* after every epoch ep_i.

The terms in the target instances from D^{val}_T are ranked based on their loc-atr$^{j}_{te}$, starting from the highest attributed term per instance j to the lowest. The top k terms in j is given by $te^{j}_{top_k} = top_k[\text{argsort}_{desc}(\text{loc-atr}^{j}_{te})]$, where k is a hyper-parameter in D^{val}_T. Here argsort$_{desc}$ returns the terms corresponding to the attribution scores sorted in descending order. We treat the two error cases of False Positives (FP) and False Negatives (FN) separately. Here the hate class is considered the positive class.

Since the terms responsible for FP may also be important for the True Positives (TP), only those terms are extracted that have high attributions for FP, but not for TP. Further, another filtering step is applied, where only the terms common to the top N from the ranked *gl-hate* are extracted. This results in discarding the terms that may not be globally correlated with a class with respect to the source model. So $te_{FP} = [te \in te^{FP}_{top_k} \;\; \& \;\; te \notin te^{TP}_{top_k}] \cap top_N(\text{gl-hate})$. Here $te^{FP}_{top_k}$ and $te^{TP}_{top_k}$ comprise top k terms from all the instances j in D^{val}_T that result in FP and TP respectively when evaluated using the source model. Similarly, top k terms corresponding to FN instances are extracted, wherein those common to TN are discarded, and subsequent filtering based on the gl-nhate is performed, i.e. $te_{FN} = [te \in te^{FN}_{top_k} \;\; \& \;\; tok \notin te^{TN}_{top_k}] \cap top_N(\text{gl-nhate})$ $\forall\ j$. This step thus yields a list of possible spurious terms at the end of the epoch ep_i, $S_{ep_i} = [te_{FP}, te_{FN}]_{ep_i}$.

Note that even though it is feasible to split the small-sized target validation set into two parts, with one part used for term extraction and the other used for hyper-parameter tuning, we decide to use the entire target validation set for both purposes. This is because the validation set is already small in size and splitting it further would make it less representative of the target corpus. This could degrade the potential of extracting the error-causing terms from the target. The same validation set can be used for hyper-parameter tuning as the model is not directly trained on this set. In the worst case, the selected hyper-parameters might give better performance on the target validation set compared to the unseen and held-out target test set, a risk that we undertake for better term extraction.

Furthermore, when the larger set of unlabeled target data D^{train}_T is available as assumed in our second problem setting, we first continue pre-training of the BERT model on D^{train}_T using the Masked Language Model (MLM) objective for incorporating the lan-

Algorithm 1 D-Ref-I

Input Source model $\mathcal{M}_{src}$, $D_S^{train}, D_T^{val}, N, k, te_{\text{pre-def}}$ (optional), epochs
Output Refined source model $\mathcal{M}_{src}$

for ep_i in epochs **do**
Compute classification loss $\mathcal{L}^{'}$
/* Extraction of Spurious Terms from $\mathcal{M}_{src}$ trained until ep_{i-1} */
gl-atr$^{c}_{te} \leftarrow \frac{\sum_{j=1}^{|D_S^{train}|} 1_{\hat{y_j}=c} \text{loc-atr}^{j}_{te} \forall \text{occurrence of } te \text{ in j}}{\sum_{j=1}^{|D_S^{train}|} 1_{\hat{y_j}=c} \#(\text{occurrence of } te \text{ in j})}$ ▷ Global term-ranking in source
gl-hate $\leftarrow$ argsort$_{desc}$(gl-atr$^{hate}_{te}$), gl-nhate $\leftarrow$ argsort$_{desc}$(gl-atr$^{nhate}_{te}$)
▷ Sort gl-atr$^{c}_{te}$ in descending order
for instances j in D_T^{val} **do** ▷ Instance-level local ranking in target
if $\mathcal{M}_{src}$ yields FP for j **then**
$te^{FP}_{top_k} \leftarrow te^{FP}_{top_k} \cap top_k[\text{argsort}_{desc}(\text{loc-atr}^{j}_{te})]$
end if
if $\mathcal{M}_{src}$ yields TP for j **then**
$te^{TP}_{top_k} \leftarrow te^{TP}_{top_k} \cap top_k[\text{argsort}_{desc}(\text{loc-atr}^{j}_{te})]$
end if
if $\mathcal{M}_{src}$ yields FN for j **then**
$te^{FN}_{top_k} \leftarrow te^{FN}_{top_k} \cap top_k[\text{argsort}_{desc}(\text{loc-atr}^{j}_{te})]$
end if
if $\mathcal{M}_{src}$ yields TN for j **then**
$te^{TN}_{top_k} \leftarrow te^{TN}_{top_k} \cap top_k[\text{argsort}_{desc}(\text{loc-atr}^{j}_{te})]$
end if
end for
$te_{FP} \leftarrow [te \in te^{FP}_{top_k} \ \& \ te \notin te^{TP}_{top_k}] \cap top_N(gl\text{-}hate)$
$te_{FN} \leftarrow [te \in te^{FN}_{top_k} \ \& \ tok \notin te^{TN}_{top_k}] \cap top_N(gl\text{-}nhate)$
$S_{ep_{i-1}} = [te_{FP}, te_{FN}]_{ep_{i-1}}$

/* Penalizing the Extracted Spurious terms */
Mask terms in $S_{\text{ep}_{\text{i-1}}}$ ▷ Term-mask
OR
$\mathcal{L} \leftarrow \mathcal{L}^{'} + \lambda \mathcal{L}_{\text{atr}}; \mathcal{L}_{\text{atr}} \leftarrow \sum_{\text{t} \in \mathcal{S}_{\text{ep}_{\text{i-1}}}} \phi\,(\text{t})^{2}\,; \text{t} \in \mathcal{S}_{\text{ep}_{\text{i-1}}}$ ▷ Reg

OR

$S_{\text{ep}_{\text{i-1}}} \leftarrow S_{\text{ep}_{\text{i-1}}} \cap te_{\text{pre-def}}$ and follow Reg ▷ Comb
Update the parameters of $\mathcal{M}_{src}$ using $\mathcal{L}$
end for

guage variations of the target domain, following Glavaš et al. (2020). After this step, the same model is used for training on the source labeled corpus D_S^{train} with our dynamic refinement approach.

(b) Penalizing the Extracted Spurious terms

In this step, we attempt to reduce the importance assigned, by the source model, to the extracted spurious terms by penalizing the terms in S_{ep_i} during the next epoch ep_{i+1}. Three different ways for term penalization are proposed:

Term-mask: In this case, the terms from S_{ep_i} present in D_S^{train} are simply masked after every ep_i, by substituting them with the `[MASK]` token, and then the source model is trained during ep_{i+1}.

Reg: Since term masking might eliminate substantial information, we regularize the model using S_{ep_i}. The attributions assigned to these terms are pushed toward zero by the following learning objective on D_S^{train}:

$$\mathcal{L} = \mathcal{L}^{'} + \lambda \mathcal{L}_{\text{atr}}; \mathcal{L}_{\text{atr}} = \sum_{\text{t} \in \mathcal{S}_{\text{ep}_\text{i}}} \phi\left(\text{t}\right)^2; \text{t} \in S_{\text{ep}_\text{i}} \tag{6.4}$$

where $\mathcal{L}^{'}$ is the classification loss and $\mathcal{L}_{\text{atr}}$ is the attribution loss. Here $\phi\left(\text{t}\right)$ is the attribution score for the term t. Intuitively, this should reduce the importance of terms contributing to source-specific patterns and encourage learning more general information. Both losses are computed over D_S^{train}.

Comb: S_{ep_i} is finally combined with the pre-defined group identifiers ($te_{pre\text{-}def}$) from Liu and Avci (2019) and Kennedy et al. (2020) to perform regularization using Equation 6.4.

We surmise that repeating these steps at the end of every epoch should reduce the source-specific correlations while the source model gets trained. Note that while the terms extracted in one epoch are penalized over the next epoch, we do not penalize them throughout the training procedure. After every epoch, our approach results in extracting a new set of terms as the training progresses, which may or may not include the terms extracted previously depending on the training dynamics. This is because once a term is penalized during an epoch, it may not cause spurious correlations for the model in the next epochs. In case a term has not been penalized adequately in a single epoch, the dynamic procedure would automatically extract the same term again in the next epochs; otherwise, it would be dropped from the list of extracted terms. We believe that retaining the extracted terms and over-penalizing them throughout the training procedure can degrade the learning and the model performance.

Three different attribution methods are used with D-Ref-I: Scaled Attention, Integrated Gradients, and DeepLIFT, as discussed in Section 6.2.

6.3.2 Dynamic Model Refinement-II (D-Ref-II)

In our second approach, we assume access to labeled source domain training data D_S^{train} and unlabeled target domain training data D_T^{train}. Here the small set of labeled validation set D_T^{val} from the target domain is used only for hyper-parameter tuning and model selection, following Dai et al. (2020) and Maharana and Bansal (2020), but not for extracting the source-specific terms. This approach for domain adaptation again involves two steps: (a) extraction of source-specific terms and (b) reducing the importance of these terms. Since D-Ref-II does not extract terms by looking into the misclassifications on D_T^{val}, it is expected to penalize terms that may not be present in the target corpus. The schematic diagram of D-Ref-II and the algorithm are provided respectively in Figure 6.3 and Algorithm 2.

(a) Extraction of source-specific terms

Domain classification To identify source-specific terms, we first train a binary domain classifier using D_S^{train} and D_T^{train} that learns to identify whether a candidate instance comes from the source or the target domain. For this, a simple Logistic Regression (LR) with bag-of-words is used, as it is inherently interpretable. Its feature weights are then used to extract the top M most important terms for predicting the source domain class. Each term is tokenized with the BERT WordPiece tokenizer for compatibility with transformer models. The top M terms obtained through domain classification are denoted as S_{LR}.

Attribution-based term ranking Intuitively, the terms from S_{LR} that also contribute highly to the hate-speech labels, are likely to restrict generalization to the target as they could potentially reduce the importance assigned by the classifier to domain-invariant hate-speech terms. Thus, we extract only those source-specific terms that are highly correlated with the labels, given the binary classification task of *hate* versus *non-hate*.

To this end, first the pre-training of BERT is continued on the unlabeled D_T^{train} using the MLM objective. We then perform supervised training on D_S^{train} using this MLM-trained model. After every epoch, two ranked lists of terms are obtained for the two classes, sorted in the order of decreasing importance. This step is the same as the global term-ranking step of D-Ref-I in Section 6.3.1. The ranked lists are constructed using feature attribution methods that yield instance-level local attribution scores loc-atr$_{te}^{j}$ per term te in an instance j: a higher score indicating a higher contribution to the predicted class. As before, the scores of stop-words and the infrequent terms are discarded and loc-atr$_{te}^{j}$ are normalized using the sigmoid function. For obtaining a corpus-level global class-specific attribution score gl-atr$_{te}^{c}$ per term te and per predicted class c, we perform a corpus-level average of all the loc-atr$_{te}^{j}$ for every c using Equation 6.3. The scores gl-atr$_{te}^{c}$ for all te are sorted to obtain the highest attributed (i.e. most important) term per class to the lowest, yielding the ranked lists of terms per class, given by GL = $[gl\text{-}hate, gl\text{-}nhate]_{ep_i}$ after every epoch ep_i.

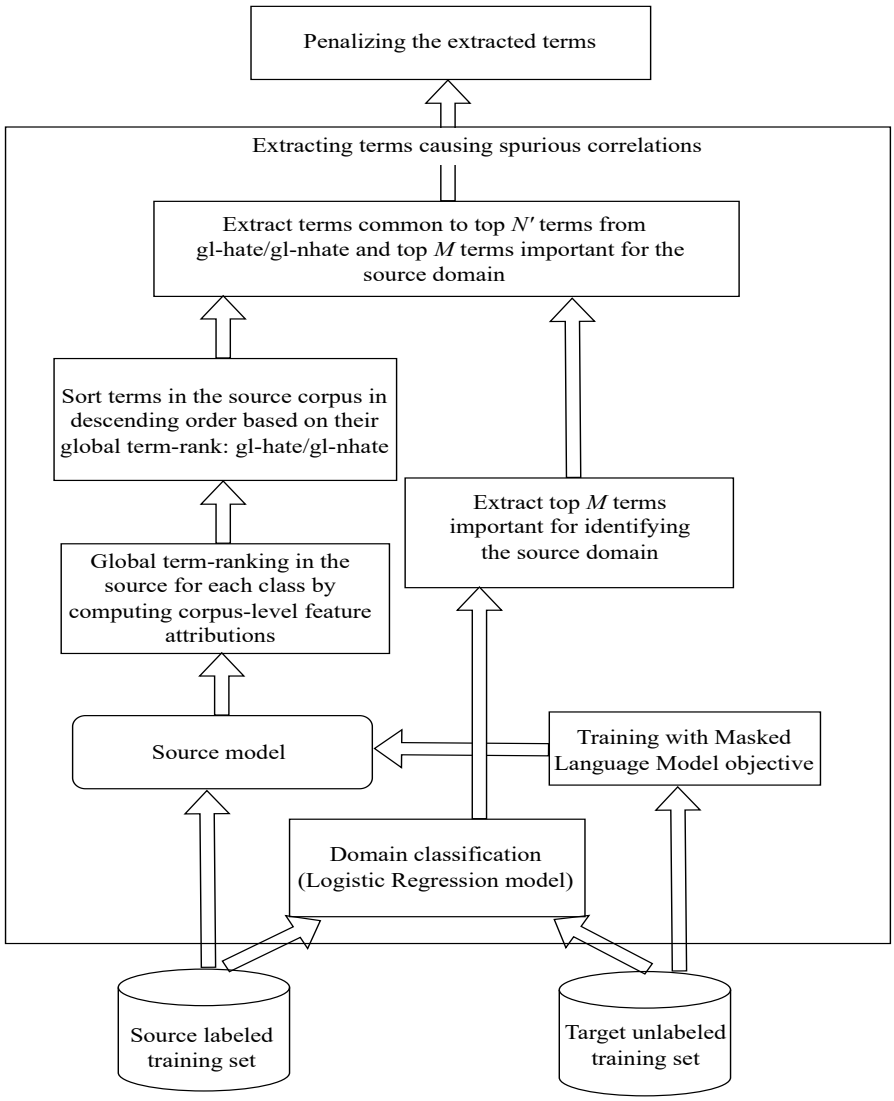


Figure 6.3: Schematic diagram of D-Ref-II.

We extract the source-specific terms te^S that are common to both S_{LR} and the top $N^{'}$ terms from GL, i.e. $\text{te}^S = [te \in S_{LR} \ \& \ te \in top_{N'}(\text{GL})]$. These steps are repeated after every epoch. Note that the list S_{LR} remains constant across the epochs, as it is independent of the hate-speech classification task.

Penalization of Source-specific Terms

We hypothesize that penalizing te^S obtained from the previous epoch during the next epoch should reduce the importance of terms that are both (i) domain-specific and (ii)

Algorithm 2 D-Ref-II

Input Source model $\mathcal{M}_{src}$, D_S^{train}, D_T^{train}, N', M, $te_{\text{pre-def}}$ (optional), epochs
Output Refined source model $\mathcal{M}_{src}$

for ep_i in epochs **do**
Compute classification loss $\mathcal{L}^{'}$
/* Extraction of source-specific terms from $\mathcal{M}_{src}$ trained until ep_{i-1} */
S_{LR} from D_S^{train} through domain classification using D_S^{train} and D_T^{train} and extraction of top-M terms important for identifying source
$\text{gl-atr}_{te}^{c} \leftarrow \frac{\sum_{j=1}^{|D_S^{train}|} 1_{\hat{y_j}=c} \text{loc-atr}_{te}^{j} \forall \text{occurrence of } te \text{ in j}}{\sum_{j=1}^{|D_S^{train}|} 1_{\hat{y_j}=c} \#(\text{occurrence of } te \text{ in j})}$ ▷ Global term-ranking in source
$gl\text{-}hate \leftarrow \text{argsort}_{desc}(\text{gl-atr}_{te}^{hate})$, $gl\text{-}nhate \leftarrow \text{argsort}_{desc}(\text{gl-atr}_{te}^{nhate})$
▷ Sort gl-atr_{te}^{c} in descending order
$\text{GL} \leftarrow [gl\text{-}hate, gl\text{-}nhate]_{ep_{i-1}}$
$\text{te}^{S} \leftarrow [te \in S_{LR} \ \ \& \ \ te \in top_{N'}(\text{GL})]$

/* Penalizing the Source-specific terms */
$\mathcal{L} = \mathcal{L}^{'} + \lambda \mathcal{L}_{\text{atr}}; \mathcal{L}_{\text{atr}} = \sum_{\text{t} \in \text{te}^{\mathcal{S}}} \phi\,(\text{t})^{2}\,; \text{t} \in \text{te}^{\mathcal{S}}$ ▷ Reg
OR
$\text{te}^{S} \leftarrow \text{te}^{S} \cap te_{\text{pre-def}}$ and follow Reg ▷ Comb

Update the parameters of $\mathcal{M}_{src}$ using $\mathcal{L}$
end for

contribute highly to the source labels, and thus, help learn from domain invariant terms. The attribution scores for te^S are minimized with L_2 penalization, in Equation 6.5.

$$\mathcal{L} = \mathcal{L}^{'} + \lambda \mathcal{L}_{\text{atr}}; \mathcal{L}_{\text{atr}} = \sum_{t \in te^{\mathcal{S}}} \phi\left(t\right)^2; t \in te^{\mathcal{S}} \tag{6.5}$$

Here $\mathcal{L}^{'}$ is the classification loss and $\mathcal{L}_{\text{atr}}$ is the attribution loss. λ controls the strength of penalization, and $\phi\left(t\right)$ is the attribution score for t.

We experiment with two variations: (i) **Reg:** penalizing only the terms in te^S; (ii) **Comb**: penalizing the combination of te^S and the terms from Liu and Avci (2019); Kennedy et al. (2020).

For this approach, only two different feature attribution methods are used: Scaled Attention and DeepLIFT/ DL. In this case, we do not use Integrated Gradients (IG) as it is computationally more expensive and DL is most often a good and a faster approximation of IG (Ancona et al., 2018).

6.4 Experimental Setup

6.4.1 Data

We use three standard hate speech corpora: *HatEval* (Basile et al., 2019), *Waseem* (Waseem and Hovy, 2016) and *Dynamic*^{v1} (Vidgen et al., 2021c). Unlike the previous chapters, the *Davidson* (Davidson et al., 2017) corpus is not used for our experiments in this chapter as only 6% of the corpus involves hate speech. This may not be sufficient for our analysis on spurious correlations and could result in additional problems with the extreme class imbalance. Instead, we use *Dynamic*^{v1} that comprises varied forms of hate against many different minorities. *Dynamic*^{v1} is an older version of the corpus provided by Vidgen et al. (2021c). A detailed description of these corpora and the pre-processing steps followed are provided in Chapter 3. Following the precedent of previous works by Wiegand et al. (2019) and Swamy et al. (2019), the detection of *hate* versus *non-hate* is considered, where the hate class covers all forms of hate. *Waseem* (26.8% hate) is split into the train (80%; 8720), validation (10%; 1090) and test (10%; 1090) sets as no standard splits are provided. We use the original splits for *HatEval* (42.1% hate; train: 8993, validation: 1000; test: 3000) and *Dynamic*^{v1} (54.4% hate; train: 32497, validation: 1016, test: 4062). The size of available D_T^{val} in *Dynamic*^{v1} is reduced by randomly sampling 25% of the 4064 instances present in the validation set.

6.4.2 Baselines

We compare our approaches with different baselines in the two settings.

No access to D_T^{train}

The experiments are performed with the vanilla baseline and other comparable baseline methods for term extraction and penalization using attribution scores.

(i) BERT Van-FT (Devlin et al., 2019): vanilla fine-tuning on D_S^{train} without any term-based penalization.

(ii) Liu and Avci (2019): Here Convolutional Neural Network has been used as the base model and regularization of a list of pre-defined group identifier terms have been performed using Integrated Gradients as the feature attribution method. An L_2 distance is penalized between the attribution scores corresponding to the group identifier terms and the target attribution value of zero.

(iv) Kennedy et al. (2020)[2]: Here the BERT model is used and the group identifier terms are manually extracted from the top features of a bag-of-words (BoW) logistic regression trained on each individual corpus. The attribution scores of these terms are then regularized to get close to 0. We experiment with two variations of the approach: **(a)** regularizing all the listed group identifiers in this work, **(b)** regularizing only the group identifiers that are present in top BoW features.

(v) χ^2-test: The χ^2-test with one degree of freedom and Yate's correction (Hofland and Johansson, 1982; Kilgarriff, 2001) is applied to extract terms te from D_S^{train} that reject the null hypothesis, with 95% confidence. The null hypothesis states that in terms of te, both D_S^{train} and D_T^{val} are random samples of the same larger population. We, then, penalize these terms using the attribution scores assigned to them, while BERT is used as the underlying model. For this, we use DeepLIFT (DL) as it yields comparable or higher overall improvements using the proposed approaches taking Table 6.2 and Table 6.3 together.

(vi) Pre-def: In this case, the model is regularized by penalizing the attributions assigned to the combined list of the pre-defined group identifiers provided by Liu and Avci (2019) and Kennedy et al. (2020), where the underlying model is BERT. The same attribution methods as used by our approach are adopted for this baseline.

The complete list of the curated group identifiers ($te_{pre\text{-}def}$) penalized with Pre-def are given below:

{lesbian, gay, bisexual, trans, cis, queer, lgbt, lgbtq, straight, heterosexual, male, female, nonbinary, african, african american, european, hispanic, latino, latina, latinx, canadian, american, asian, indian middle eastern, chinese, japanese, christian, buddhist, catholic, protestant, sikh, taoist, old, older, young, younger, teenage, millenial, middle aged, elderly, blind, deaf, paralyzed, muslim, jew, jews, white, islam, blacks, muslims, women, whites, gay, black, democrat, islamic, allah, jewish, lesbian, transgender, race, brown, woman, mexican, religion, homosexual, homosexuality, africans}

[2]We use Sampling and Occlusion (Jin et al., 2020) following Kennedy et al. (2020).

Access to D_T^{train}

In this case, our approach is compared with the Vanilla baseline (**Van-MLM-FT**) where the BERT model is adapted on D_T^{train} using the MLM objective and the supervised fine-tuning is performed on D_S^{train} without term-based penalization. Furthermore, we again compare our approach with Kennedy et al. (2020) with its two variations as mentioned above. However, since Kennedy et al. (2020) used BERT as their underlying model, the experiments are initialized with MLM adaptation on D_T^{train} in this setting for a fair comparison with our approach. Similarly, Pre-def is also initialized with the MLM adapted BERT on D_T^{train}, and it is called **Pre-def-MLM**. We further compare with the entropy-based attention regularization method proposed by Attanasio et al. (2022) (refer to Section 6.2.2).

Besides, we assess different domain adaptation methods from the sentiment classification task, namely, **BERT PERL** (Pivot-based Encoder Representation of Language) (Ben-David et al., 2020) that adopts the MLM objective of BERT to perform pivot-based fine-tuning; **BERT-AAD** (Adversarial Adaptation with Distillation) (Ryu and Lee, 2020) that performs domain adversarial training; **HATN** (Hierarchical Attention Transfer Network) (Li et al., 2018, 2017b) that extracts pivots using a domain adversarial approach. A detailed description of these domain adaptation approaches is provided in Chapter 5. We also implement and compare with a domain adaptation approach for cross-domain hate-speech detection proposed by Sarwar and Murdock (2022), who adopted a data-augmentation strategy leveraging a negative emotion corpus (Go et al., 2009). The approach learns to tag hate-related terms in the hateful instances of the unlabeled D_T^{train} and the negative emotion corpus using a sequence tagger trained on D_S^{train}. They then constructed a weakly labeled augmented corpus by a TF-IDF-based template matching with the tagged target domain corpus D_T^{train}. Further details about this approach is presented in Chapter 2. For a fair comparison, we use BERT as the underlying model in this case while performing supervised classification.

Finally, as in the previous case, the **χ^2-test** with 1 degree of freedom and Yate's correction is applied. The terms from D_S^{train}, for which the null hypothesis of both D_S^{train} and D_T^{train} being random samples of the same larger population is rejected with 95% confidence, are penalized using their DL scores. All the BERT models are initialized with MLM adaptation on D_T^{train}, except for PERL and AAD, which inherently adapts the model to D_T^{train} as part of their training strategy.

6.4.3 Hyper-parameter Tuning and Implementation Details

As mentioned in Section 6.3.1, we follow two different problem settings - (i) with no access to unlabeled training data D_T^{train} in the target domain; (ii) with access to unlabeled D_T^{train}. For the first setting, our model is only initialized with the pre-trained BERT model. For the second setting, we first continue pre-training of BERT on the unlabeled D_T^{train} using the MLM objective in order to incorporate the language variations of the target domain, following Glavaš et al. (2020). This MLM-trained model is then used

to initialize the supervised classification on D_S^{train} in both D-Ref-I and D-Ref-II. The BERT-base uncased[3] (Wolf et al., 2020) model is used for our experiments. Besides, a decoupled weight decay regularization (Loshchilov and Hutter, 2019) with AdamW optimizer is used, with a weight decay of 10^{-4} for all the BERT models including the Vanilla baselines (Van-FT and Van-MLM-FT). We use a learning rate of 3×10^{-5} for the MLM training and 1×10^{-5} for the supervised fine-tuning, with the epsilon parameter set to 1×10^{-8}. Both the MLM training on the unlabeled target domain training data D_T^{train}, and the supervised fine-tuning on the source domain training data D_S^{train} are run for 6 epochs with a batch size of 8 for all the methods using BERT. The best model for all the baselines and our proposed approaches are selected by tuning over D_T^{val}. We train all the models over D_S^{train} from the source and evaluate over D_T^{test} from the target and report the macro-F1 scores across five random initializations of each experiment.

For D-Ref-I, we set the value of top N terms used from ranked {glist-hate, glist-nhate} as 500. The values of $k \in$ top {10%, 20%, 30%, 40%} of the instance-length in D-Ref-I, and λ in both D-Ref-I, D-Ref-II and Pre-def are selected through hyper-parameter tuning over D_T^{val} using a random seed. For $\alpha\nabla\alpha$ and DeepLIFT, $\lambda \in \{0.1, 0.5, 1, 10, 20, 30, 40, 50, 60\}$ and for IG, $\lambda \in \{1, 10, 20, 30, 40, 50, 60\}$. For D-Ref-II, the value of $N^{'}$ is set to 250 and M to 750 for all our experiments.

For Integrated Gradients, following Liu and Avci (2019), the interpolated embeddings are treated as constants while back-propagating the loss from the regularization term. An all-zero embedding vector is used as the baseline input for both Integrated Gradients and DeepLIFT. For all the prior arts, the original codes are used, as provided by the respective authors. We implement the data-augmentation approach proposed by Sarwar and Murdock (2022) ourselves due to the absence of available implementation. Following the description present in the paper, the training data is prepared for the sequence tagger by labeling all the terms in the hateful instances from the source corpus that are also present in the lexicon from hatebase.org[4]. However, we do not tokenize the lexicon obtained from hatebase.org while searching for the corresponding matching terms in the source corpus. We convert the lexicon into lowercase and look for the exact match in the source corpus.

6.5 Results and Discussion

6.5.1 D-Ref-I

Cross-corpus Performance without D_T^{train}

Table 6.2 presents macro-F1 scores across five random initializations of each experiment using six cross-corpus pairs when there is no access to D_T^{train}. We use macro-F1 because penalization of the terms corrects the misclassifications for both hate and non-hate classes across domains, as the source model can have bias corresponding to both

[3] https://huggingface.co/bert-base-uncased
[4] https://hatebase/org/

Approaches	**H →D**	**D →H**	**H →W**	**W →H**	**D →W**	**W →D**	**Avg**
BERT Van-FT	53.2±1.0	63.3±1.8	67.5±5.1	52.6±2.4	60.3±1.0	46.7±4.0	57.3
Liu and Avci (2019)	45.1±4.5	59.5±0.7	57.2±3.8	52.6±0.8	57.1±2.7	39.6±2.0	51.9
Kennedy et al. (2020) (a)	52.2±1.2	62.0±1.6	62.7±2.9	50.1±6.8	53.5±2.0	45.1±2.3	54.3
Kennedy et al. (2020) (b)	52.0±3.8	61.9±1.7	63.6±3.7	54.8*±1.6	57.0±1.7	46.8±1.9	56.0
BERT χ^2-test	55.4*±1.1	65.0*±1.0	68.1±1.3	53.7±2.1	60.4±2.8	45.2±2.8	58.0
Pre-def ($\alpha\nabla\alpha$)	54.6*±1.3	65.1*±1.1	69.6±3.4	54.4*±1.2	61.9±1.6	47.2±3.1	58.8
D-Ref-I-Term-mask ($\alpha\nabla\alpha$)	53.8±0.6	64.9*±0.7	68.9±3.3	53.6±3.0	59.6±2.2	45.8±3.7	57.8
D-Ref-I-Reg ($\alpha\nabla\alpha$)	54.9*±1.2	65.1*±0.9	68.6±4.0	54.1*±1.0	60.9±1.5	48.7*±4.3	58.7
D-Ref-I-Comb ($\alpha\nabla\alpha$)	55.0*±1.6	64.7*±1.2	69.9±1.6	**55.3***±1.3	61.0±2.8	48.1*±1.0	**59.0**
Pre-def (IG)	55.7*±1.4	63.5±2.8	69.7±2.2	51.7±2.7	60.3±2.2	44.6±3.0	57.6
D-Ref-I-Term-mask (IG)	56.3*±2.3	64.5*±1.8	68.3±2.0	52.3±2.3	59.3±1.3	48.2*±2.1	58.2
D-Ref-I-Reg (IG)	**56.4***±1.4	**65.5***±0.8	69.2±2.5	53.8*±0.7	60.6±1.7	47.7±3.6	58.9
D-Ref-I-Comb (IG)	55.7*±0.8	63.7±2.4	69.1±2.3	52.6±2.3	61.4±2.5	**51.4***±3.6	**59.0**
Pre-def (DL)	54.2±1.6	64.0±1.9	68.1±1.5	52.9±1.2	62.0±1.8	44.5±1.3	57.6
D-Ref-I-Term-mask (DL)	55.1*±1.4	64.9*±1.7	67.2±3.6	52.1±1.9	60.5±2.5	47.2±3.1	57.8
D-Ref-I-Reg (DL)	54.2±1.6	64.8*±0.8	**70.7***±2.7	51.4±0.7	**62.3***±2.5	47.1±5.5	58.4
D-Ref-I-Comb (DL)	55.4*±1.8	64.0±0.9	69.5±3.3	54.0*±0.8	61.5±2.3	48.1*±2.7	58.8

Table 6.2: Macro-F1 (±std-dev) on source →target pairs (H: HatEval, D: Dynamic^{v1}, W: Waseem) with no access to D_T^{train}. **Bold** denotes the best-performing approach in each column and underline denotes the second best. * denotes statistical significance compared to Van-FT with paired bootstrap (Dror et al., 2018; Efron and Tibshirani, 1993), 95% confidence interval. Avg denotes the average performance for each method.

classes. Therefore, the average of the F1 scores with respect to both hate and non-hate classes is important for evaluating the approaches proposed in this chapter. We observe that overall, all feature-attribution methods with D-Ref-I yield improved performance compared to Van-FT and other baselines. While χ^2-test yields improvements over Van-FT, D-Ref-I still displays better performance in most cases. This could be attributed to the fact that although the terms obtained through the χ^2-test from the source indicate differences across domains, they may not necessarily be important for the prediction of hate/ non-hate labels by the source model, and may not contribute to source-specific spurious correlations.

We find that D-Ref-I-Reg with IG and DL achieves better average macro-F1 of 58.9 and 58.4 respectively, compared to the corresponding Pre-def (IG) and Pre-Def (DL) which obtain an average of 57.6. D-Ref-I-Reg ($\alpha\nabla\alpha$) provides an average macro-F1 of 58.7, comparable to Pre-def ($\alpha\nabla\alpha$) with 58.8. However, D-Ref-I-Reg achieves significantly improved scores in more cases, as compared to Pre-def using all the attribution methods, i.e. 4/6 cases ($\alpha\nabla\alpha$), 3/6 cases (IG), and 3/6 cases (DL) with D-Ref-I-Reg, compared to 3/6 ($\alpha\nabla\alpha$), 1/6 (IG) and none (DL) with Pre-def. D-Ref-I-Term-mask exhibits improvements on average ($\alpha\nabla\alpha$: 57.8, IG: 58.2, DL: 57.8) over Van-FT (57.3), demonstrating the effectiveness of the term extraction mechanism of D-Ref-I. Finally, D-Ref-I-Comb displays the best overall performance, with the highest average score of 59. We attribute this improvement from D-Ref-I to its increased coverage with dynamic term extraction, and reduction of spurious source-specific correlations, while the baselines only penalize the group identifiers. A dynamic approach also corrects the model during training before it can get fully biased toward these terms. Finally, it can incorporate

Approaches	H →D	D →H	H →W	W →H	D →W	W →D	Avg
BERT Van-MLM-FT	56.6±1.3	66.2±1.2	70.0±2.5	50.9±2.1	61.4±2.4	43.5±1.9	58.1
MLM + Kennedy et al. (2020) (a)	55.4±2.0	65.5±0.8	64.1±1.4	54.4*±1.3	59.2±1.8	44.5±2.9	57.2
MLM + Kennedy et al. (2020) (b)	54.9±2.9	65.7±0.9	67.3±1.2	54.3*±2.2	62.3±2.7	46.6±3.5	58.5
BERT PERL	54.1±0.7	60.0±0.6	60.1±2.0	**55.2***±0.7	55.5±1.0	37.8±1.2	53.8
BERT-AAD	56.6±1.3	53.9±3.5	68.8±2.5	50.7±1.4	48.3±4.7	**53.0***±1.7	55.2
HATN	48.4±1.6	59.1±0.4	59.7±2.9	51.4±1.8	60.0±2.6	45.4±2.7	54.0
MLM + Sarwar and Murdock (2022)	55.0±1.9	66.2±2.0	68.8±1.1	48.2±3.1	57.9±1.3	36.2±1.1	55.4
MLM + Attanasio et al. (2022)	54.9±1.6	66.5±1.4	64.1±5.0	52.4*±3.7	62.5±0.8	43.5±2.3	57.3
MLM + χ^2-test	57.9±1.6	**67.1**±1.7	69.8±0.8	48.2±3.1	60.4±2.8	44.1±3.4	57.9
MLM + D-Ref-I-Reg ($\alpha\nabla\alpha$)	57.6±1.9	66.2±1.2	**70.7**±1.2	52.5*±4.0	62.8±1.4	48.0*±4.3	59.6
MLM + D-Ref-I-Reg (IG)	58.6*±1.2	66.8±0.5	70.1±1.5	52.1±3.0	62.5±3.0	48.9*±4.4	59.8
MLM + D-Ref-I-Reg (DL)	**58.8***±2.2	66.7±0.6	70.5±1.3	52.4*±3.5	**64.7***±2.1	51.5*±4.9	**60.8**

Table 6.3: Comparison of domain adaptation approaches with D-Ref-I-Reg + MLM. Macro-F1 (±std-dev) on different source →target pairs with access to D_T^{train}. H: HatEval, D: Dynamic^{v1}, W: Waseem. * denotes the significantly improved scores w.r.t. Van-MLM-FT. Avg denotes the average performance for each method.

the pre-defined lists along with the extracted terms in its 'Comb' variant, and further improve the performance.

Cross-corpus Performance with D_T^{train}

D-Ref-I-Reg is further compared with various Domain Adaptation (DA) methods and other baselines mentioned in Section 6.4.2 in Table 6.3. Penalizing pre-defined group-identifier terms (Kennedy et al., 2020) yields a low overall performance even after adapting the BERT model on D_T^{train} with the MLM objective. We note that the average performance of all the other DA approaches in this task is lower than Van-MLM-FT. This follows the analysis and discussion in the previous chapter that these DA approaches from cross-domain sentiment classification perform sub-optimally in tasks similar to cross-corpus hate speech detection. While pivots in BERT PERL are not optimally transferable to the requirements of hate speech, the domain adversarial approach BERT-AAD and HATN are unable to separate the classes well in the target domains. Besides, the DA approach proposed for cross-domain hate-speech detection by Sarwar and Murdock (2022) also yields an overall drop in performance. They perform data augmentation by replacing relevant words from an external negative emotion dataset with tagged hateful terms from the target domain. We find that a major portion of the augmented instances lacks meaning, and this negatively impacts the adaptation. χ^2-test, on average, also fails to surpass the Vanilla baseline.

However, across all feature attribution methods, D-Ref-I-Reg improves the cross-corpus performance compared to Van-MLM-FT and the DA approaches, with an average macro-F1 of 59.6 ($\alpha\nabla\alpha$), 59.8 (IG), and 60.8 (DL), compared to 58.1 from Van-MLM-

Approaches	H →D	D →H	H →W	W →H	D →W	W →D	Avg
BERT Van-MLM-FT	56.6±1.3	66.2±1.2	70.0±2.5	50.9±2.1	61.4±2.4	43.5±1.9	58.1
MLM + Kennedy et al. (2020) (a)	55.4±2.0	65.5±0.8	64.1±1.4	54.4*±1.3	59.2±1.8	44.5±2.9	57.2
MLM + Kennedy et al. (2020) (b)	54.9±2.9	65.7±0.9	67.3±1.2	54.3*±2.2	62.3±2.7	46.6±3.5	58.5
BERT PERL	54.1±0.7	60.0±0.6	60.1±2.0	**55.2***±0.7	55.5±1.0	37.8±1.2	53.8
BERT-AAD	56.6±1.3	53.9±3.5	68.8±2.5	50.7±1.4	48.3±4.7	**53.0***±1.7	55.2
HATN	48.4±1.6	59.1±0.4	59.7±2.9	51.4±1.8	60.0±2.6	45.4±2.7	54.0
MLM + Sarwar and Murdock (2022)	55.0±1.9	66.2±2.0	68.8±1.1	48.2±3.1	57.9±1.3	36.2±1.1	55.4
MLM + Attanasio et al. (2022)	54.9±1.6	66.5±1.4	64.1±5.0	52.4*±3.7	62.5±0.8	43.5±2.3	57.3
MLM + χ^2-test	57.9±1.6	67.1±1.7	69.8±0.8	48.2±3.1	60.4±2.8	44.1±3.4	57.9
Pre-def-MLM ($\alpha\nabla\alpha$)	**58.9***±0.7	67.4±1.5	71.3±1.0	48.9±4.0	60.0±2.0	46.5±4.9	58.8
D-Ref-II-Reg ($\alpha\nabla\alpha$)	58.3±1.8	66.8±0.7	70.1±1.8	52.3*±3.0	60.8±2.2	46.9*±2.5	59.2
D-Ref-II-Comb ($\alpha\nabla\alpha$)	58.7*±2.1	**67.7**±1.0	70.9±1.0	51.5±2.1	59.8±1.5	45.9±3.1	59.1
Pre-def-MLM (DL)	58.5*±1.4	66.5±1.3	70.3±1.7	51.2±1.7	**70.3***±0.5	42.7±2.0	59.9
D-Ref-II-Reg (DL)	58.8*±0.6	66.4±1.2	**72.2**±1.4	52.9*±1.9	63.6*±2.0	48.8*±4.7	60.5
D-Ref-II-Comb (DL)	58.4±1.4	66.7±1.0	71.3±0.9	51.1±2.2	69.5*±2.2	46.6±1.9	**60.6**

Table 6.4: Macro-F1 (±std-dev) on source →target pairs with access to D_T^{train} for D-Ref-II. H: HatEval, D: Dynamic^{v1}, W: Waseem. **Bold** denotes the best score and underline the second best in each column . * denotes statistically significant improvement compared to Van-MLM-FT. D-Ref-II inherently uses MLM, so it is not explicitly denoted in the table. Avg denotes the average performance.

FT. Since D-Ref-I-Reg and Van-MLM-FT use identical MLM adaptation on D_T^{train}, the improvements can be attributed to the dynamic term extraction of our method.

More generally, when the larger set of target domain unannotated instances D_T^{train} are unavailable, D-Ref-I can identify and correct spurious correlations on the source using a small amount of annotated instances from the target D_T^{val}, as demonstrated for the previous setting. When a sufficient number of unannotated instances from the target corpus is available, D-Ref-I can yield further cross-corpus improvements by leveraging the unannotated target instances with the MLM adaptation.

6.5.2 D-Ref-II

Table 6.4 displays the macro-F1 scores obtained, in cross-corpus settings, averaged across five randomly initialized runs. As discussed in Section 6.3.2, we only use $\alpha\nabla\alpha$ and DL as attribution methods with D-Ref-II because IG is much slower than DL and also because DL often approximates IG well (Ancona et al., 2018). D-Ref-II yields improvements on average over all the baselines, using both $\alpha\nabla\alpha$ and DL, not just independently but also in combination (Comb) with Pre-def-MLM, where D-Ref-II-Comb achieves the highest overall performance with DL: 60.6. With DL, D-Ref-II-Reg yields significantly improved performance in 4/6 cases, compared to 2/6 with Pre-def-MLM (DL). This is apparently due to the penalization of relevant source-specific terms that have wider coverage compared to the pre-defined terms in Pre-def-MLM. Besides, since the entropy-based attention regularization by Attanasio et al. (2022) does not use the target domain unlabeled

Non-hate examples from the test set of *Dynamic*^{v1} for *HatEval* →*Dynamic*^{v1}	
FP with Van-FT	**TN with D-Ref-I-Reg**
Wildlife belongs in the f*cking wild .	Wildlife belongs in the f*cking wild.
NHS s*cks	NHS s*cks
There is so much cancer patients in the world but it is mostly the young females who are worstly affected by this disease.	There is so much cancer patients in the world but it is mostly the young females who are worstly affected by this disease.
Hate examples from the test set of *Dynamic*^{v1} for *Waseem* →*Dynamic*^{v1}	
FN with Van-FT	**TP with D-Ref-I-Reg**
Blacks are savages .	Blacks are savages .
Queers can't be truly happy .	Queers can't be truly happy .
Don't get me wrong I don't hate asians but I definitely don't like them	Don't get me wrong I don't hate asians but I definitely don't like them

Table 6.5: Change in attributions with D-Ref-I-Reg.

instances for term-extraction, it may not be optimal for cross-corpus settings.

The large improvement with Pre-def-MLM (DL) for *Dynamic*^{v1} →*Waseem* (70.3) could be attributed to the fact that *Dynamic*^{v1} involves a wide variety of identity terms. Thus, penalizing the pre-defined identity terms might result in a higher emphasis on more generalizable hate-speech content. While only this particular case drives the high average performance with Pre-def-MLM (DL), D-Ref-II-Reg (DL) performs well *consistently* and yields a higher average score (D-Ref-II-Reg: 60.5, D-Ref-II-Comb: 60.6) compared to Pre-def.

6.5.3 Comparing D-Ref-I with D-Ref-II

Table 6.3 and Table 6.4 show that D-Ref-II-Reg yields a slightly lower average performance compared to MLM + D-Ref-I-Reg. The average macro-F1 score of MLM + D-Ref-I-Reg with $\alpha\Delta\alpha$ is 59.6 and that with DL is 60.8. D-Ref-II-Reg yields an average score of 59.2 with $\alpha\Delta\alpha$ and 60.5 with DL. This is likely because D-Ref-I uses the misclassifications on the labeled D_T^{val} from the target domain to extract the source-specific terms, while D-Ref-II does not use D_T^{val} for term extraction. The latter extracts terms using the unlabeled instances from D_T^{train} without knowing anything about the errors incurred by the source model in the target domain.

6.5.4 Qualitative Analysis

D-Ref-I

Table 6.5 shows the change in attributions for some instances in D_T^{test} from $Dynamic^{v1}$ that were misclassified by Van-FT but correctly classified by our D-Ref-I-Reg. The darker the shades, the higher are the attributions assigned by the source classifier. Van-FT wrongly attributes higher importance to 'f*cking', 's*cks' and 'females' for the hate class in the first example, and 'blacks', 'queers', and 'asians' for non-hate in the second due to source-specific correlations. However, D-Ref-I-Reg extracts and penalizes abusive terms like {s*ck, a**hole, d*ck} and group identifier terms like {females, feminist} in the case of *HatEval* $\rightarrow Dynamic^{v1}$ causing False Positives (FP). It penalizes the terms like {africans, dark, queer, asian} for the latter causing False Negatives (FN) in the pair *Waseem* $\rightarrow Dynamic^{v1}$. Our approach not only penalizes the exact terms but also results in reducing the importance of terms with similar meanings (e.g. 'blacks' is contextually close to 'dark', 'africans'), giving more importance to the context around the spurious terms. Since the *Waseem* corpus is made available as tweet IDs, we observed that it mostly contains sexist comments, while most of the racist content must have been removed before we could crawl it. Hence, the terms related to race mostly occur in non-hate contexts causing FN.

The lists of error-causing terms (BERT WordPieces) for FP and FN in D_T^{val} resulting in spurious correlations, extracted for these pairs, are given below. The terms present in the visualization examples of Table 6.5 and ones similar in meaning to them are underlined.

HatEval $\rightarrow$ ***Dynamic***^{v1}

- **Epoch 1: FP:** {idiots, conservative, countries, p*ssy, bloody, americans, move, a**hole, hating, beings, feminist, africans, resources, d*ck, resist, females, attacks, dude, anger } **FN:** {hitler, plague, ##urs, crisis, rescue, funding, gorgeous, treason, journalist, lawyers, agenda, roles, principles, bloody, intern}
- **Epoch 2: FP:** {race, hating, flights, sheep, females, ignorant, feminist, resist, attacks, d*ck, kill, boat, countries, p*ssy, refugee, bloody} **FN:** {president, foreigners, illegal, betrayal, lgbt, riots, gorgeous, treason, joking, chris, intelligent, arguments, humans}
- **Epoch 3: FP:** {countries, race, hating, females, feminist, africans, ridiculous, d*ck, express, comments, organized, s*ck, allow, bloody} **FN:** {illegal, hitler, generally, david, intelligent, secret, chris, equality, dating, yellow, treason, abuses, ##gb, humans, plague, dear, nonsense}
- **Epoch 4: FP:** {isis, genocide, indians, society, supported, females, feminist, attacks, s*ck, destroy, migrants } **FN:** {hitler, opportunities, sister, betrayal, ##ame, gorgeous, ##heads, dating, riots, bank, murders, arguments, humans, fights, plague, influence, targeting, supporters, coordination, lies, ##boys}
- **Epoch 5: FP:** {clean, ignorant, slave, feminist, punish, africans, ##ache, d*ck,

##fs, ars, destroy, status, race, p*ssy, western, send} **FN:** {statement, gross, hitler, sending, yellow, waste, hopefully, trapped, riots, bait, sister, coordination, humans}

Waseem →Dynamic^{v1}

- **Epoch 1: FP:** {female, ##ists, fe, sex, feminist, rap} **FN:** {cast, coward, queer, equality, ##bi , cost, ##sy, born, asian, nazis, kids, cancer, gender, hiring, funded}
- **Epoch 2: FP:** {##ists, her, sex, worse, feminist, ##nt, outraged} **FN:** {welcome, caused, cancer. drag, ##bi, pressure, parent, nazis, troll, cast, trash, ruins, lesbian, attacking, chinese}
- **Epoch 3: FP:** {female, ##ja, might, men, feminist} **FN:** {quoting, govt, referring, nazis, troll, lesbian, rogue, date, chinese, typically}
- **Epoch 4: FP:** {communism, her, openly, intelligent, many, barbie, chicks, females, arguing} **FN:** {date, suggest, ##lat, referring, police, chinese, cancer, voice, native, lesbian}
- **Epoch 5: FP:** {term, f*ck, ##ng, woman, ##ist, feminist, females, prison} **FN:** {removed, educate, freaking, queer, wow, ending, referring, dye, ##wat, issues, africans, vast, chinese, dark}

Few of the extracted terms get repeated in subsequent epochs as a single epoch may not be sufficient to reduce the effect of a term and it may appear in the next epoch as well. Moreover, as the training progresses, the model may learn new patterns, and some extracted terms may reappear and disappear again due to the penalization. However, their overall influence on the source domain classifier should be reduced, even if they reappear, because of penalization that causes improvement in macro-F1.

D-Ref-II

Table 6.6 displays examples of misclassifications for *Waseem →HatEval* and *Dynamic*^{v1} *→Waseem*, yielded by Van-MLM-FT for the respective target domain instances, which are correctly classified by D-Ref-II-Reg. The examples suggest that penalizing source-specific terms results in placing more emphasis on the general contextual meaning of the out-of-domain instances such as 'depression', 'unfortunately...replacing my shoes...', 'wife...shut...make me a sammich', and 'omg... apple... woman... adorable'.

Note that, unlike D-Ref-I, the terms in the examples from the target domain that receive reduced importance with D-Ref-II, compared to Van-MLM-FT, may not be the same or similar terms that are extracted and penalized. This is because D-Ref-II does not consider the misclassifications of D_T^{val} while extracting the source-specific terms. Moreover, the domain classification step results in obtaining terms that are more likely to be infrequent in the target domain. Rather, due to the penalization of source-specific terms, the source domain classifier learns to focus on the wider context of the instances and this indirectly changes the importance of terms from the target domain. For example, we observe that in the case of *Waseem →HatEval*, the automatically extracted te^S includes

Non-hate example from the test set of *HatEval* for *Waseem* → *HatEval*	
FP with Van-MLM-FT	**TN with D-Ref-II-Reg**
Depression is a whole entire b*tch	Depression is a whole entire b*tch
Unfortunately you are in a sticky size my only problem is replacing my shoes has been a b*tch	Unfortunately you are in a sticky size my only problem is replacing my shoes has been a b*tch
Hate example from the test set of *Waseem* for *Dynamic*^{v1} → *Waseem*	
FN with Van-MLM-FT	**TP with D-Ref-II-Reg**
It is good to talk with your wife but it is easier to say shut up n make me a sammich not sexist lol	It is good to talk with your wife but it is easier to say shut up n make me a sammich not sexist lol
Non-hate example from the test set of *Waseem* for *Dynamic*^{v1} → *Waseem*	
FP with Van-MLM-FT	**TN with D-Ref-II-Reg**
Omg I am lisening to an apple genius dude tell this old woman how to use email and it is adorable	Omg I am listening to an apple genius dude tell this old woman how to use email and it is adorable

Table 6.6: Change in attributions with D-Ref-II-Reg.

terms related to the role of women in sports, such as {*sports, sexist, gaming, football, commentary, competition, ...*}. Note that Wiegand et al. (2019) also mention that these terms cause domain or topic bias in *Waseem*, restricting generalizability.

The full list of penalized terms (BERT WordPieces) te^S across epochs for the listed examples in Table 6.6, is given below. Here, there are no separate lists for FP and FN incurred on the target domain validation set because, unlike D-Ref-I, the errors on D_T^{val} are not used for term extraction.

Waseem* → *HatEval

- **Epoch 1:** {college, sports, feminism, la, magnetic, used, unique, ##ava, speech, ##js, tr, ##cking, object, chu, result, ki, bus, ##is, adopt, referring, ##roids, handed, ##em, sh, ##omp, unconscious, anger, gamer, prove, xbox, tri, skill, judgment, tool, block, single, harassment, size, georgia, involved, ##ism, studying, voices, possible, gaming, pl, ##il, helped, ##ke, survey, equality}
- **Epoch 2:** {feminism, used, football, awesome, equal, ##cking, object, ##ification, interest, feminist, ##rra, scientist, ##al, ignorance, bodies, ##work, later, ##nk, troll, ##ss, based, adopt, ##cing, quality, sister, unconscious, criticisms, pro, notch, xbox, tri, unfair, rap, meanwhile, impression, single, harassment, bonus, georgia, constant, sex, ##ist, possible, click, competition, ##per, swedish, ##eral, november, write, eventually, equality}
- **Epoch 3:** {sham, anger, pull, used, focus, speech, ashley, object, interest, bringing, ##na, eye, ##nk, later, quality, ##roids, oppressive, rain, ##omp, statistics, nsw, content, notch, museum, unconscious, typically, tri, ##ol, unfair, writing,

##chan, georgia, constant, annie, ra, weights, click, ##il, furniture, helped, shopping, football, commentary, equality}

- **Epoch 4:** {minded, kat, used, equal, focus, ##hand, tr, ##cking, chu, interest, bringing, thor, fm, ##tag, path, scientist, precious, later, mike, quality, humanist, ##roids, ##el, ##omp, worth, unconscious, nsw, xbox, tri, unfair, nu, kaitlyn, ##ering, pest, fe, camera, giant, constant, weights, gaming, rap, ##il, swedish, opposes, ##thi, november, laughing, survey, equality}
- **Epoch 5:** {feminism, raging, equal, focus, ##hand, ##cking, ##cky, ##tag, ##na, mostly, scientist, ##al, ##rra, adopt, humanist, ft, ##roids, ##el, ##omp, example, unconscious, museum, anger, typically, tri, unfair, impression, yu, single, fe, cu, ##rd, ##ification, constant, grass, gaming, rap, science, ##per, swedish, il, furniture, shopping, november, equality}

Dynamic^{v1} $\rightarrow$ ***Waseem***

- **Epoch 1:** {wheelchair, ##zzi, dali, seekers, ##oons, koreans, ##tos, ##ware, ##ders, handicapped, principles, mac, pregnant, ##tier, ##iers, ##wear, ##bib, barren, ##tite, dyke}
- **Epoch 2:** {customer, pip, principles, ##tos, ##hon, les, ko, vietnamese, teenagers, ##lock, ##sion, ##has, ##gin, ##rmi, poles, buddhist, handicapped}
- **Epoch 3:** {pak, homosexuality, koreans, pleasant, ##tos, mirror, spaniards, ##fs, ro, ##rmi, boom, handicapped}
- **Epoch 4:** {##cky, pak, chin, ##tos, bender, herr, catholics, ro, buddhist}
- **Epoch 5:** {pip, pak, ##tos, yellow, bender, koreans, ##mit, ##sion, ##has, ##rk, ##gin, catholics, ro, arrogance}

6.5.5 In-corpus Performance

Approaches	**HatEval**	**Dynamic**^{v1}	**Waseem**	**Average**
BERT Van-FT	43.3±1.8	85.1±0.5	85.4±0.7	71.3
In-corpus performance on source (left of arrows) while applying domain adaptation for the target (right of arrows)				

	H →D	**H →W**	**D →H**	**D →W**	**W →H**	**W →D**	
D-Ref-I-Reg ($\alpha\nabla\alpha$)	39.7±3.2	38.4±1.7	84.1±1.0	84.2±0.8	84.4±0.7	78.8±8.0	68.3
D-Ref-I-Reg (IG)	40.5±2.0	37.7±2.1	84.0±0.4	84.5±0.4	84.6±1.0	85.3±1.4	69.4
D-Ref-I-Reg (DL)	37.1±1.8	38.1±2.9	84.7±0.6	84.3±1.2	84.4±0.5	80.7±6.4	68.2
D-Ref-II-Reg ($\alpha\nabla\alpha$)	42.4±2.5	42.0±4.1	84.0±0.9	84.5±1.0	85.1±0.7	83.8±0.8	70.3
D-Ref-II-Reg (DL)	41.7±3.7	40.5±4.4	83.9±0.7	82.6±1.5	84.7±1.2	81.1±2.7	69.1

Table 6.7: In-corpus macro F1 (±std-dev), i.e. the source corpus performance, obtained after applying domain adaptation for the target corpus (present at the right-hand side of the arrows) using D-Ref-I-Reg and D-Ref-II-Reg. H: HatEval, D: Dynamic^{v1}, W: Waseem. Model selection and early-stopping are done over the validation set from the target corpus for D-Ref-I and D-Ref-II. For Van-FT, the BERT model evaluated in-corpus *without* adaptation or model selection on the target corpus.

We present the in-corpus performance, i.e. the performance on the source corpus in terms of macro-F1 scores, obtained when the source model is refined for the correspond-

Approaches	HatEval	Dynamic^{v1}	Waseem
BERT Van-FT	1 m 25 s	3 m 52 s	2 m
D-Ref-I-Reg ($\alpha\nabla\alpha$)	1 m 39 s	7 m	3 m 33 s
D-Ref-I-Reg (IG)	9 m 37 s	59 m	19 m 7 s
D-Ref-I-Reg (DL)	4 m 4 s	18 m 36 s	8 m 44 s
D-Ref-II-Reg ($\alpha\nabla\alpha$)	2 m 30s	7 m	3 m 17 s
D-Ref-II-Reg (DL)	4 m	18 m	8 m 16 s

Table 6.8: Per epoch training time on different source corpora; m: minutes, s: seconds.

ing target corpus using D-Ref-I-Reg and D-Ref-II-Reg, in Table 6.7. For D-Ref-I-Reg and D-Ref-II-Reg, the model is tuned over the target corpus validation set. Here BERT Van-FT gives the original performance of the source model, *without* refinement or MLM adaptation on the target, as a reference. In this case, the model is tuned over the source corpus validation set. The *HatEval* corpus is part of a shared task and involves a challenging test set with low in-corpus performance. The drop across in-corpus performance is expected as D-Ref-I and D-Ref-II are aimed at making the model best-suited to the target corpus. The maximum drop in the average performance is 3.1 points of the macro-F1 score by D-Ref-I-Reg (DL).

6.5.6 Computational Efficiency

The per epoch training times for D-Ref-I-Reg and D-Ref-II-Reg while training with different source corpora are presented in Table 6.8. We use one Nvidia GTX 1080 Ti GPU for our experiments. The training times of D-Ref-I-Reg ($\alpha\nabla\alpha$) are less than 2 times of that with Van-FT. With D-Ref-I-Reg (DL), the training time is approximately 4.5 times of that with Van-FT. This demonstrates the computational efficiency of our approach. In the case of D-Ref-I-Reg (IG), the computation time is indeed high. This occurs due to the aggregation of gradients using a path integral and computing gradients over gradients, as also discussed by Kennedy et al. (2020) and Liu and Avci (2019). D-Ref-II-Reg ($\alpha\nabla\alpha$), like D-Ref-I-Reg ($\alpha\nabla\alpha$) takes less than double the time taken by Van-FT to train, and D-Ref-II-Reg (DL) takes roughly 4.5 times of the training time taken by Van-FT. The computation times differ according to the feature attribution method used. However, our approach is not dependent on any particular feature attribution method, as demonstrated by our experiments.

6.6 Conclusion and Future Work

We proposed two domain adaptation approaches for improving cross-corpus hate-speech detection through automatic term extraction with regularization of the source model such that the spurious source-specific correlations are reduced. This alleviates the need for predefined dictionaries and increases coverage. While D-Ref-I extracts terms using a small subset of the labeled target domain validation set, D-Ref-II does not require the labeled

instances from the target domain for term extraction. Therefore, we observed that the kinds of terms extracted by the two approaches are essentially different. D-Ref-I extracts terms that are common to both source and target corpora but are not correlated with the hate and not-hate labels in the target corpus in the same way as they are in the source corpus. D-Ref-II, on the other hand, extracts terms that are predominantly present in the source but may not be present in the target. Our approaches demonstrated consistent cross-corpus performance improvements both independently and in combination with pre-defined terms. We also observed through qualitative analysis that they seem to learn the generalizable context of hate speech to an extent by not focusing on specific terms.

These results should motivate further research on domain adaptation in hate speech and building classifiers that can generalize well to the concept of hate. Indeed, the two approaches can be combined to potentially yield the benefits from both sets of terms if a small labeled subset from the target domain is available. Besides, future work includes applying our methods to other cross-domain text classification tasks as spurious correlations are detrimental in general for any other classification task. This is possible as the methods do not rely on pre-defined knowledge about a given task. Moreover, it is worth exploring how explanation faithfulness – accuracy of the explanations in representing the real reasons behind a model's prediction – can be improved in out-of-domain settings (Chrysostomou and Aletras, 2022a).

7 Neighbourhood-aware Optimal Transport for Low-resource Cross-platform Hate Speech Detection

7.1 Introduction

In this chapter, we consider the problem of cross-platform hate-speech detection in low-resource settings where a small number of labeled instances from the target corpus are available for training. The goal is to effectively transfer knowledge from a different hate speech corpus that has abundant labeled training instances for improving the performance of the low-resource target corpus. In practical scenarios, the labeled source corpus used for transferring knowledge may be sampled from a different online platform compared to the target corpus. Therefore, we study transfer learning in generic cross-platform scenarios. The work presented in this chapter is published as Bose et al. (2022c).

As discussed in Chapter 5, although deep learning-based approaches (Mozafari et al., 2019; Badjatiya et al., 2017) have become the state-of-the-art for hate speech detection, their performance depends on the size of the labeled resources available for training (Lee et al., 2018; Alwosheel et al., 2018). We have also discussed that while annotating a large corpus for hate speech is considerably time-consuming, expensive, and harmful to human annotators (Schmidt and Wiegand, 2017; Malmasi and Zampieri, 2018), models trained on existing labeled hate speech corpora have shown poor generalization on new content (Yin and Zubiaga, 2021) due to multiple differences across corpora. Even though in the previous chapters, mostly unsupervised transfer has been studied, as pointed out by Sarwar et al. (2022), in real-life scenarios most online platforms could invest in obtaining at least some labeled training instances for deploying a hate speech detection system. Thus, here we study a more realistic setting where a limited amount of labeled content for training is available in the target corpus. To address the aforementioned challenges, we aim to devise a strategy that can effectively transfer knowledge from a *resource-rich* source corpus with a higher amount of annotated content to a *low-resource* target corpus with fewer labeled instances amidst the cross-corpus differences.

One popular way to address this problem is sequential transfer learning, where a model is trained or fine-tuned sequentially. It consists of a pre-training phase where generic representations are learned on a source task or domain, and an adaptation phase where the gained information is applied to a target task or domain (Ruder et al., 2019). For instance, Mozafari et al. (2019) fine-tuned the pre-trained BERT model on the limited training examples in hate-speech corpora. Furthermore, following Garg et al. (2020),

a pre-trained model can be first fine-tuned on a resource-rich source corpus and subsequently fine-tuned on the low-resource target corpus. However, since such sequential learning may risk forgetting knowledge from the source, the source and target corpora can be mixed for training (Shnarch et al., 2018). Besides, to learn target-specific patterns without forgetting the source knowledge, Meftah et al. (2021, 2019) augmented pre-trained neurons from the source model with randomly initialized units for transferring knowledge to low-resource domains.

Recently, Sarwar et al. (2022) argued that traditional transfer learning strategies are not systematic. Therefore, they modeled the relationship between a source and a target corpus with a neighborhood framework and showed its effectiveness in transfer learning for content flagging. They modeled the interaction between a query instance from the target and its neighbors retrieved from the source. This interaction was modeled based on their label agreement – whether the query and its neighbors have the same labels – while using a fixed neighborhood size. However, different neighbors may have varying levels of proximity to the queried instance based on their pair-wise cosine similarities in a sentence embedding space. Therefore, intuitively, the neighbors should be weighted according to these similarity scores.

We hypothesize that simultaneously modeling the pair-wise distances between instances from the low-resource target and their respective neighbors from the resource-rich source, along with their label distributions should result in a more flexible and effective transfer. With this aim, we propose a novel training strategy where the model learns to assign varying importance to the neighbors corresponding to different target instances by optimizing the amount of pair-wise transfer. This transfer is learned without changing the model architecture. Such optimization can be efficiently performed using *Optimal Transport* (OT) (Peyré and Cuturi, 2019; Villani, 2009; Kantorovich, 2006) due to its ability to find correspondences between instances while exploiting the underlying geometry of the embedding space. Our contributions in this chapter are summarised as follows:

- We address hate speech detection in low-resource scenarios with a flexible and systematic transfer learning strategy.
- We propose novel incorporation of neighborhood information with joint distribution Optimal Transport. This enables learning of the amount of transfer between pairs of source and target instances considering both (i) the similarity scores of the neighbors and (ii) their associated labels. To the best of our knowledge, this is the first work that introduces Optimal Transport for hate speech detection.
- The effectiveness of our approach is demonstrated through considerable improvements over strong baselines, along with quantitative and qualitative analysis using different hate speech corpora from varied platforms.

The remaining of this chapter is organized as follows: Section 7.2 gives an overview of the relevant prior works on neighborhood frameworks and optimal transport. In Section 7.3, we describe the proposed approach in detail. Section 7.4 provides a description of the corpora and settings used for our experiments. The results along with their analysis

and an ablation study are discussed in Section 7.5. Finally, we conclude the chapter in Section 7.6.

7.2 Related Works

7.2.1 Neighborhood Frameworks

In the past decades, neighborhood information has been used extensively as additional knowledge for text classification. Angelova and Weikum (2006) represented the readily available neighborhood information i.e. the link structure in richly structured data sets (e.g. web pages connected through hyperlinks or scientific papers linked through citations, etc.) with a graph. They modeled the dependencies between a document and its neighbors, with reducing influence as the distance in the graph grows, using a first-order Markov Random Field (MRF) (Pelkowitz, 1990; Stan Z., 2001). For the same kind of data sets, Le et al. (2012) proposed to capture both contents and labels from neighbors for representing their links with a document, and incorporated this link model with the previous MRF-based soft-labeling approach. Besides, k-Nearest Neighbors (kNN)-based approaches have been successfully used in the literature for an array of tasks. Wang and Liu (2010) proposed a graph-based kNN model and a combined feature selection method. These aimed at reducing the computation time for measuring the similarity between a document and its neighbors and improving the classification performance. Salles et al. (2018) performed nearest neighborhood projection to reduce the overfitting issues in a random-forest (Breiman, 2001) classifier by considering only the neighborhood space of test instances for training the classifier.

kNN classifiers (Manning, 2008), which typically predict the class of input through majority voting using its neighbors in the training data, have the advantage of being more transparent compared to deep learning approaches. They have been used for hate speech detection (Prasetyo and Samudra, 2022; Briliani et al., 2019) and other text classification tasks (Trstenjak et al., 2014; Chen et al., 2020b; Kaminska et al., 2021b). Wang et al. (2005, 2006, 2007) introduced the adaptive kNN classifier, which builds an 'influence region' for each training instance x_i by constructing the largest sphere centered on x_i that does not enclose an instance with a different label and based on its statistical confidence level. They used an adaptive and normalized distance measure and weighted each neighbor by the statistical confidence of the corresponding sphere. Balsubramani et al. (2019) proposed another variant of the kNN classifier, where the value of k is determined adaptively for each query based on the properties of every neighborhood. Kaminska et al. (2021b) used the weighted kNN (Dudani, 1976) classification method, which assigns weights to the neighbors based on a distance metric or similarity relation, for emotion detection in tweets. Fuzzy-rough nearest neighbor (FRNN) classifier (Jensen and Cornelis, 2011; Vluymans et al., 2019; Lenz et al., 2019), which is an instance-based classifier based on the fuzzy rough set theory (Dubois and Prade, 1990), has also been used for emotion detection (Kaminska et al., 2021a, 2022).

kNN-based components have been used along with Transformer-based models (Vaswani

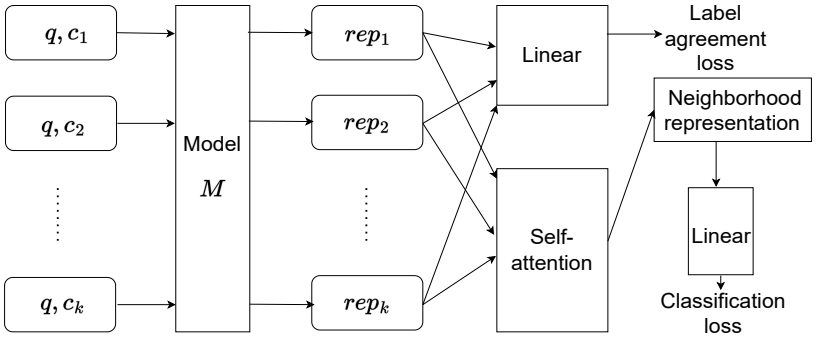


Figure 7.1: Architecture of CE kNN$^+$ model (Sarwar et al., 2022)

et al., 2017) in a variety of tasks, such as language modeling (Khandelwal et al., 2020), question answering (Kassner and Schütze, 2020), dialogue generation (Fan et al., 2021), etc. demonstrating the benefits of incorporating neighborhood information into models. Recently, Sarwar et al. (2022) proposed a neighborhood framework kNN$^+$ for transfer learning in cross-lingual low-resource settings that uses representations from Transformers, unlike the traditional kNN classifier. Since our proposed approach, discussed in this chapter, draws inspiration from this method, we provide an overview of this framework.

Sarwar et al. (2022) showed that a simple kNN classifier is prone to prediction errors as the neighbors may have similar meanings, but opposite labels. They, instead, modeled the interactions between the target corpus instances, treated as queries, and their nearest neighbors retrieved from the source, based on whether a query (q) and its neighbors with the content c_i have the same or different labels. Two model variants were proposed for kNN$^+$: Bi-Encoder kNN$^+$ (BE kNN$^+$) and Cross-Encoder kNN$^+$ (CE kNN$^+$), where CE kNN$^+$ yields the best performance in cross-lingual settings. In the CE kNN$^+$ variant, for every neighbor i, the approach involves obtaining the 'interaction feature' by concatenating q with c_i from every i and obtaining the output representation rep_i from a multilingual language model M as $rep_i = M(q, c_i)$. It performs multi-task learning comprising two losses: *label agreement loss* and *classification loss*. The first loss results in optimizing the neighbor-level agreement indicating whether a query and all of its k neighbors have the same or different labels. The second loss learns to predict the label of q by using a structured self-attention to obtain the neighborhood representation of q from the interaction features rep_i. Fig 7.1 presents the architecture of their CE kNN$^+$ variant.

However, Sarwar et al. (2022) did not consider *the varying levels of the proximity of different neighbors to the query.* Besides, a mini-batch in their framework comprises a query and all its neighbors. For fine-tuning large language models like BERT, the batch size needs to be kept small due to resource constraints. This could limit the neighborhood size in their framework. This is different from our approach, where the neighborhood size is scalable.

7.2.2 Optimal Transport

Optimal Transport is a mathematical theory that allows comparing probability distributions in a geometrically sound manner. It has the ability to transform one probability distribution into another with the minimal possible effort by computing optimal mappings to minimize cost functions. Indeed Optimal Transport (OT) lies at the intersection of various fields like probability theory, geometry, and optimization theory, amongst others. The central idea of OT was first introduced in the work of Monge (1781), which dates back to the 18^{th} century. After Monge's work, a seminal work by Kantorovich (2006, 1942) led to the expansion of OT into several fields, which transformed some aspects of the OT problem into a linear programming problem.

In addition to its various applications in theoretical and applied mathematics (Santambrogio, 2015; Villani, 2009), optimal transport also had great success in machine learning in the last few years (Peyré and Cuturi, 2019). It has been used in a diverse set of applications, such as signal processing (Kolouri et al., 2017), generative modeling (Mémoli, 2011; Bunne et al., 2019; Arjovsky et al., 2017), computing distances for structured data represented as graphs (Titouan et al., 2019), computer vision (Rubner et al., 2000; Solomon et al., 2015; Ge et al., 2021; Birdal et al., 2020), domain adaptation (Courty et al., 2014, 2017b,a; Damodaran et al., 2018) and others. One of the most well-known applications of OT is in generative adversarial networks (GANs), where the so-called Wasserstein GAN (Arjovsky et al., 2017) employed it in its loss function. In the field of NLP, OT has become increasingly popular with applications such as text generation (Li et al., 2020a; Zhao et al., 2019), machine translation (Xu et al., 2021), interpretable semantic similarity (Lee et al., 2022), representation learning (Singh et al., 2020), rationalizing text matching (Swanson et al., 2020), etc.

The application of OT on large-scale datasets is prohibitive due to its computational cost. As a result, there has been much study into ways to speed up the computation of optimal transport. Cuturi (2013) smoothed the classical OT problem of Kantorovich (2006, 1942) by adding an *entropic regularization* term to it and demonstrated that the resulting regularized loss can be effectively computed on GPUs using their novel implementation of Sinkhorn's algorithm (Sinkhorn, 1964; Sinkhorn and Knopp, 1967). This resulted in the computation of the OT distance at a speed that is many orders of magnitude faster than that of the classical transportation solvers. Furthermore, Altschuler et al. (2017) provided the algorithmic complexity of the Sinkhorn distance of Cuturi. In order to decrease the computational cost, further methods based on a mini-batch approximation of OT were proposed (Genevay et al., 2018; Salimans et al., 2018; Damodaran et al., 2018; Fatras et al., 2021).

Optimal Transport for Domain Adaptation: The use of OT for domain adaptation deserves particular mention as it gave state-of-the-art results in this task. In the earlier works of Courty et al. (2014, 2017b), OT was used to discover the coupling between the source and the target domains for unsupervised domain adaptation. The source data and labels were subsequently transported into the target domain using the coupling.

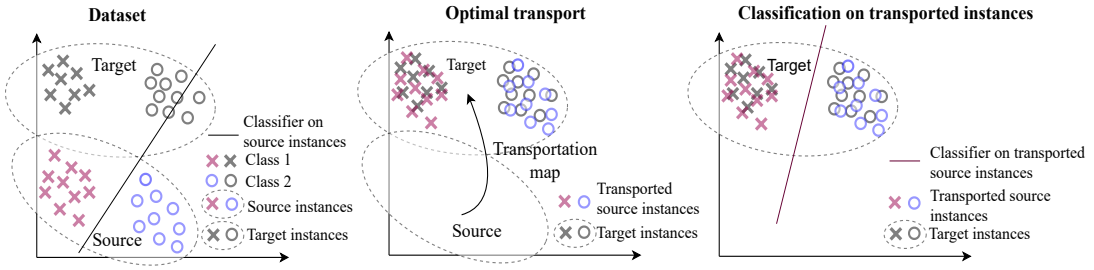


Figure 7.2: Illustration of the method proposed by Courty et al. (2017b). The source and target domain datasets are presented on the left. A classifier trained on the source domain does not fit the target domain data. The source instances are transported onto the target domain by computing a transportation map in the center, usually with a non-linear transformation. The classifier in the target domain is estimated using the transported source instances.

Finally, the classifier was learned using the transferred samples from the target domain. Fig 7.2 presents this training strategy. Different kinds of regularized OT were studied in these works to compute optimal connections between instances. Courty et al. (2014) proposed a regularization scheme for integrating label information in the optimization, while Courty et al. (2017b) encouraged source instances with the same labels to remain close during their transport to the target domain. Figure 7.3 presents an illustration of the optimal connections in OT.

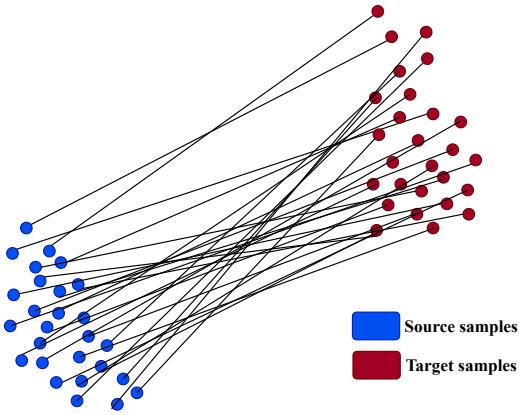


Figure 7.3: Illustration of OT in a discrete setting, where both source and target distributions are empirical. The optimal connections are represented by black lines. The figure is adapted from Fatras (2021).

Given the data and label spaces, denoted by X and Y respectively, in the *covariate shift* assumption, the marginal feature distributions $P(X)$ change across domains, while the conditional distributions $P(Y|X)$ are assumed to remain unchanged. Courty

et al. (2017a) argued that there is no strong reason to support such an assumption about $P(Y|X)$. Therefore, they proposed to address a change in both marginal and conditional distributions by computing the optimal coupling on the joint data and label distributions across domains using their framework called *Joint Distribution Optimal Transport* (JDOT). Damodaran et al. (2018) built over JDOT to propose DeepJDOT that use the representations extracted by the deeper layers of a neural network for computing the optimal connections instead of using the original input space as done in JDOT.

Another work on aligning the source and target domains using OT was proposed by Li et al. (2020b) who used an attention-aware transport distance to compute the domain discrepancy, while Kerdoncuff et al. (2021) optimized the Mahalanobis distance to learn the transport cost. Moreover, OT has been coupled with adversarial learning to learn domain invariant features. These involve optimizing the empirical Wasserstein distance between source and target instances in an adversarial manner (Shen et al., 2018b), encouraging a large margin of separation using an OT-based adversarial approach (Dhouib et al., 2020), among others. Cross-domain alignment has also been formulated as a graph-matching problem with OT (Chen et al., 2020a). Furthermore, Xu et al. (2020) proposed the Reliable Weighted OT (RWOT) that involves using the spatial prototypical information, intra-domain structure, and a weighted OT strategy, among other components, that reduce the negative transfer caused by the instances close to the decision boundaries in the target domain. Recently, Olvera et al. (2021) used the training strategy of Damodaran et al. (2018) for reducing domain mismatch in the task of sound event detection. Besides, Fatras et al. (2021) proposed using the unbalanced version of OT (Benamou, 2003) while computing optimal mapping at the mini-batch level to mitigate the effect of undesired pairings.

We have provided a description of the theory of Optimal Transport (OT) with regards to domain adaptation in Section 7.3.1. In our work, we perform novel incorporation of nearest neighborhood information with OT using the JDOT framework and the work of Fatras et al. (2021). To the best of our knowledge, this is the first work that introduces OT to the hate speech detection task.

7.3 Proposed Approach

Our problem setting involves a low-resource target corpus X^t with a limited amount of labeled training data $(X^t_{train}, Y^t_{train}) = \{x^t_i, y^t_i\}_{i=1}^{n_t}$ and a resource-rich source corpus X^s from a different distribution with a large number of annotated data $(X^s_{train}, Y^s_{train}) = \{x^s_i, y^s_i\}_{i=1}^{n_s}$. Given such a setting, we hypothesize that transferring knowledge from the nearest neighbors in the source should improve the performance on the insufficiently labeled target. Furthermore, to provide additional control to the model, we propose a systematic transfer. With this transfer mechanism, a model can *learn* different weights assigned to the neighbors in X^s_{train} based on their proximity to the instances in X^t_{train} simultaneously in a sentence embedding space and the label space. For this, we incorporate neighborhood information with Optimal Transport (OT), as OT can learn

correspondences between instances from X^s_{train} and X^t_{train} by exploiting the underlying embedding space geometry.

7.3.1 Joint Distribution Optimal Transport

In this work, we use the joint distribution optimal transport (JDOT) framework (Courty et al., 2017a) following the works of Damodaran et al. (2018) and Fatras et al. (2021), proposed for unsupervised domain adaptation in deep embedding spaces. The framework aligns the joint distribution $P(Z, Y)$ of the source and the target domains, where Z is the embedding space through a mapping function $g(.)$, and Y is the label space. For a discrete setting, let $\mu_s = \sum_i^{n_s} a_i \, \delta_{g(x_i^s), y_i^s}$ and $\mu_t = \sum_i^{n_t} b_i \, \delta_{g(x_i^t), y_i^t}$ be two empirical distributions on the product space of $Z \times Y$. Here $\delta_{g(x_i), y_i}$ is the Dirac function at the position $(g(x_i), y_i)$, and a_i, b_i are uniform probability weights, i.e. $\sum_i^{n_s} a_i = \sum_i^{n_t} b_i = 1$.

The 'balanced' OT problem (OT_b), as defined by Kantorovich (2006, 1942), seeks for a transport plan γ in the space of the joint probability distribution $\Pi(\mu_s, \mu_t)$, with marginals μ_s and μ_t, that minimizes the cost of transport from μ_s to μ_t, as:

$$\begin{aligned} OT_b(\mu_s, \mu_t) = \min_{\gamma \in \Pi(\mu_s, \mu_t)} \quad & \sum_{i,j} \gamma_{i,j} c_{i,j} \\ s.t. \quad & \gamma \mathbf{1}_{n_t} = \mu_s, \gamma^T \mathbf{1}_{n_s} = \mu_t \end{aligned} \tag{7.1}$$

Here $c_{i,j}$ is an entry in a cost matrix $C \in R^{n_s \times n_t}$, representing the pair-wise cost (see Section 7.3.2), and $\mathbf{1}_n$ is a vector of ones with dimension n. Each entry $\gamma_{i,j}$ indicates the amount of transfer from location i in the source to j in the target.

The marginal constraint on γ requires that all mass from μ_s is transported to μ_t. However, when there is a presence of outliers in the distributions, this constraint compels the transport plan to transfer the outliers too. This results in an undesirable additional transportation cost. This can be alleviated through relaxation of the constraint on γ, leading to the 'unbalanced' OT (OT_u) (Benamou, 2003). OT_u allows the system to create and destroy mass and is known to be more robust to outliers (Chizat, 2017; Liero et al., 2015). In fact, it would not transport the outliers if such a transfer is prohibitively expensive. The formulation of the unbalanced OT is given below:

$$\begin{aligned} OT_u(\mu_s, \mu_t) = \min_{\gamma \in \Pi(\mu_s, \mu_t)} \quad & \sum_{i,j} \gamma_{i,j} c_{i,j} + \Lambda \, ; \\ \text{where} \quad & \Lambda = \epsilon \, \Omega(\gamma) + \lambda \left(\mathrm{KL}(\gamma \mathbf{1}_{n_t}, \mu_s) + \mathrm{KL}(\gamma^T \mathbf{1}_{n_s}, \mu_t) \right) \\ s.t. \quad & \gamma \geq 0 \end{aligned} \tag{7.2}$$

KL is the Kullback-Leibler divergence that allows the relaxation of the marginal constraint on γ. λ is the marginal relaxation coefficient. $\Omega(\gamma) = \sum_{i,j} \gamma_{i,j}(log(\gamma_{i,j}) - 1)$ corresponds to the entropic regularization term, which allows fast computation of the OT distances (Cuturi, 2013), as discussed in Section 7.2.2. ϵ is the entropy coefficient.

For models with a high-dimensional embedding space like ours, Fatras et al. (2021) and Damodaran et al. (2018) proposed to make the computation of OT losses scalable using

the mini-batch OT. Thus, for every mini-batch, we sample an equal number of instances, given by the batch size m, from X^s_{train} and X^t_{train}, which makes $C \in R^{m \times m}$ and γ square matrices. As discussed by Fatras et al. (2021), since the transport plan at the mini-batch level is much less sparse, it may result in undesired pairings between instances if computed by Equation 7.1. This is because the mini-batches may not necessarily comprise the samples that lie in the support of the full OT plan. Hence, the marginal constraint of the balanced OT may force the pairing of samples that could be considered outliers at the level of the mini-batch. To counteract this effect, we rely on the more robust version of OT as formulated in Equation 7.2. Thus, we adopt the *joint distribution entropy regularized unbalanced mini-batch OT* for our framework, henceforth simply referred to as OT_u. Note that this framework does not modify the underlying model architecture used for classification, but only introduces a new training strategy.

7.3.2 Neighborhood-aware OT_u (OT_u^{NN})

In the above joint distribution framework, the cost matrix C is expressed as the weighted combination of the costs in the embedding and the label spaces:

$$c_{i,j}(g(x^s_i), y^s_i; g(x^t_j), y^t_j) = \alpha\, d(g(x^s_i), g(x^t_j)) + \beta\, L(y^s_i, y^t_j) \tag{7.3}$$

$d(.,.)$ denotes the *embedding distance* (ED) – a squared l_2 distance between the corresponding embeddings, and $L(.,.)$ is *label-consistency loss* (LC) – a cross-entropy loss that enforces a match between the label of the i^{th} source instance and that of the j^{th} target instance. α and β are scalar values. Minimizing the cost in Equation 7.3 results in aligning instances from the source and the target that simultaneously share similar representations and common labels.

We adapt C to account for k nearest neighbors of the target instances in X^t_{train} from the source X^s_{train}. Since BERT is not optimal for semantic similarity search (Reimers and Gurevych, 2019) (discussed in detail in Chapter 3), we extract the neighbors using the Sentence-BERT (SBERT) model (Reimers and Gurevych, 2019). SBERT provides sentence embeddings that can be easily compared using cosine similarity. We hypothesize that allowing transfers to occur only from the corresponding neighbors in the source to the target should result in more effective learning.

For this, we explicitly assign the value $\max(C)$ to $c_{i,j}$ in C whenever the i^{th} source and j^{th} target instances are not neighbors, considering the nearest neighborhood space of k neighbors. Besides, we use the SBERT distances as the embedding distance in Equation 7.3. This distance, in addition to the label consistency term, ensures that γ is learned to allow a higher amount of transfer from neighbors in X^s_{train} that are simultaneously (i) closer in the SBERT space and (ii) share the same label with an instance in X^t_{train}, compared to the neighbors that are further away and/or have opposite labels.

Note that even though we use a neighborhood size of k, the target instances do not attend equally to all of their k neighbors. This is because if the distance between a target instance x^t_j and its top n^{th} neighbor (x^s_i) from the source, within the neighborhood size of k (i.e. $n < k$) is comparatively large, their corresponding (i,j)-th entry in C would have

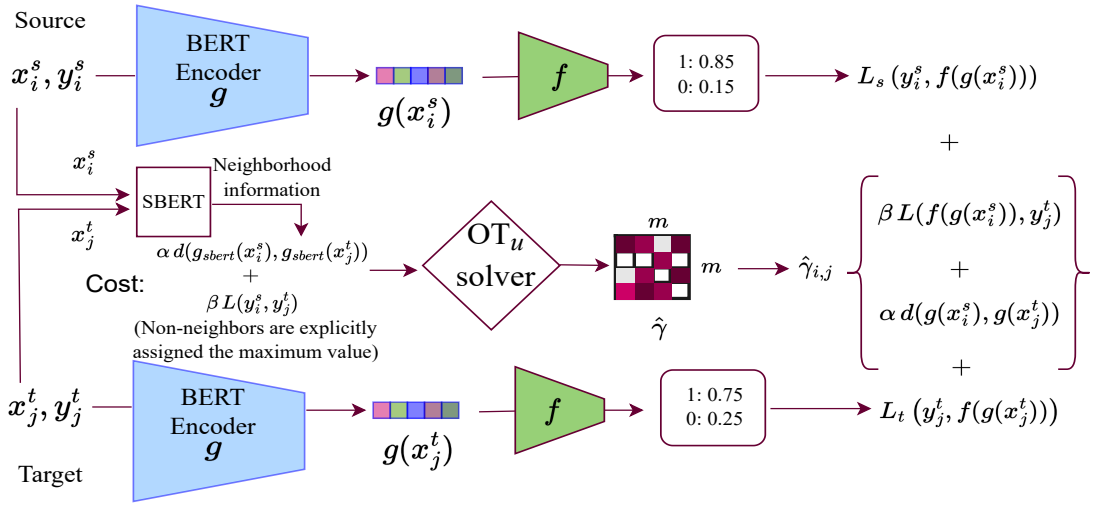


Figure 7.4: Overview of the training strategy in OT_u^{NN}. Even though the BERT encoder g and the classifier f are shared by both corpora, they are illustrated twice for better clarity by representing the two corpora separately. The presented softmax values obtained from f are simply examples provided for illustration. This figure is inspired by Damodaran et al. (2018).

a larger value. This would comparatively reduce the transfer *even if they share common labels*. Thus, for a neighbor with the same label as the target instance, the higher its SBERT distance from the target instance, the lower the amount of transfer. This results in more flexibility where the model can learn from the relevant neighbors corresponding to every target instance.

In addition to the OT_u loss from Equation 7.2, we introduce the cross-entropy losses for the training instances from both X^t_{train} and X^s_{train} in the final loss function, as required by our classification task. Our final loss function[1] for OT_u^{NN} is given by Equation 7.4. Here $g(.)$ encodes a given input using the pre-trained BERT encoder to the BERT embedding space by extracting the `[CLS]` token representation of the last hidden layer. $f(.)$ denotes the classifier, which is one fully connected layer. θ_s and θ_t are the weights assigned to the source and the target cross-entropy losses, respectively.

$$\begin{aligned}\text{OT}_u^{NN} = \min_{\gamma,f,g} \quad & \theta_s \frac{1}{m}\sum_i L_s\,(y_i^s, f(g(x_i^s))) + \sum_{i,j}\gamma_{i,j}c_{i,j} \\ & + \Lambda + \theta_t \frac{1}{m}\sum_j L_t\,(y_j^t, f(g(x_j^t)))\;;\; s.t. \quad \gamma \geq 0\end{aligned} \tag{7.4}$$

[1]The proposed approach is referred to as OT^{NN} in the published article. We added the subscript 'u' here to maintain clarity about the use of the unbalanced version of optimal transport.

Solving the optimization problem: Following Damodaran et al. (2018), we adopt a two-step procedure to solve the above optimization problem at the mini-batch level. The optimal γ is first computed by fixing the model parameters f and g.

$$\min_{\gamma} \sum_{i,j} \gamma_{i,j} \left(\alpha \, d(g_{sbert}(x_i^s), g_{sbert}(x_j^t)) + \beta \, L(y_i^s, y_j^t) \right) + \Lambda \tag{7.5}$$

We use the SBERT embeddings through the mapping function $g_{sbert}(.)$ here instead of the learned BERT embeddings to compute the ED loss. This is done so that the γ is updated based on the semantic proximity in the SBERT space. y_i^s and y_j^t are the ground truth labels for the instances x_i^s and x_j^t from X_{train}^s and X_{train}^t, respectively. In the next step, the model parameters f and g are learned while fixing γ obtained from Equation 7.5, denoted as $\hat{\gamma}$.

$$\begin{aligned} \min_{f,g} \quad & \sum_{i,j} \hat{\gamma}_{i,j} \left(\alpha \, d(g(x_i^s), g(x_j^t)) + \beta \, L(f(g(x_i^s)), y_j^t) \right) \\ & + \theta_s \frac{1}{m} \sum_{i} L_s \left(y_i^s, f(g(x_i^s)) \right) + \theta_t \frac{1}{m} \sum_{j} L_t \left(y_j^t, f(g(x_j^t)) \right) \end{aligned} \tag{7.6}$$

The first part of Equation 7.6 allows the model to learn from the instances in X_{train}^s that are consistent in terms of both the embedding space (ED loss) and the label space (LC loss) with the instances in X_{train}^t. Here we use $g(.)$, instead of $g_{sbert}(.)$, to compute ED so that g learns from the SBERT space through $\hat{\gamma}$. For the LC loss, the predicted labels for x_i^s from the source and the actual labels y_j^t corresponding to x_j^t from the target are used. This is done to update the model parameters f and g based on the target labels and bring source instances that have common labels closer to the target instances. Figure 7.4 provides an illustration of the training strategy of OT_u^{NN}.

We propose different variants of OT_u^{NN}:

OT_u^{NN}: In this variant, we do not use the source cross-entropy loss term in Equation 7.4, thus effectively having $\theta_s = 0$.

$\text{OT}_{u,\text{pre-select}}^{NN}$: Prior to the training, we pre-select the k nearest neighbors from X_{train}^s corresponding to every instance in X_{train}^t, instead of training with all the source instances. Here also $\theta_s = 0$.

OT_u^{NN}+ sloss : This is OT_u^{NN} with source cross-entropy loss (sloss), thus having $\theta_s = 1$.

$\text{OT}_{u,\text{pre-select}}^{NN}$+ sloss : This is similar to the second variant, with $\theta_s = 1$. Here, sloss is computed only on the pre-selected source instances.

7.4 Experimental Settings

7.4.1 Corpus Description

For performing experiments in cross-platform scenarios, we use three hate speech corpora, namely, *Waseem* (Waseem and Hovy, 2016), *Dynamic*^{v2} (Vidgen et al., 2021c), and *Ethos* (Mollas et al., 2022), as they are collected using different sampling strategies from varied platforms (see Chapter 3 for the description of the individual corpora). *Waseem* is a Twitter corpus, whereas *Dynamic*^{v2} is dynamically created using a human-and-model-in-the-loop process where the comments are generated by trained annotators using the `Dynabench`[2] platform. *Ethos*, on the other hand, is sampled from YouTube and Reddit. Similar to the previous chapter, we use the labels of *hate* and *non-hate*, where the former involves all forms of hate. For *Waseem*, we obtain 10.9K tweets in total from the tweet IDs, of which 26.8% instances belong to the *hate* class. *Dynamic*^{v2} has a total of 41144 instances, of which 53.9% is labeled as *hate* and *Ethos* comprises 998 instances of which 43.4% are *hate* instances.

For our experiments, we create two different versions of every corpus depending on its use as the source or the target, as presented in Table 7.1.

Corpus	**Number of comments**		
Source setting			
	Train		
$Waseem_{\text{src}}$	8720		
$Dynamic^{v2}_{\text{src}}$	32924		
$Ethos_{\text{src}}$	998		
Target setting			
	Train	**Validation**	**Test**
$Waseem_{\text{tar}}$	400	100	1090
$Dynamic^{v2}_{\text{tar}}$	400	100	4120
$Ethos_{\text{tar}}$	400	100	200

Table 7.1: Corpus statistics.

Source setting: In the absence of available standard splits, we randomly sample 80% of *Waseem* as the train set, resulting in 8720 instances. For *Dynamic*^{v2}, the original corpus-provided train split of 32924 instances is used. Since *Ethos* has a relatively small size, the entire corpus is used for training, when considered as the source. We call the source versions of these corpora as $Waseem_{\text{src}}$, $Dynamic^{v2}_{\text{src}}$ and $Ethos_{\text{src}}$. Note that the source corpus is only used for training, while its validation set is not employed for our experiments. Instead, the corresponding validation and test sets of the low-resource target corpus are used in the experiments.

Target setting: In order to simulate a low-resource scenario for the target, we down-sample the original training instances of the corpora to 500 instances. This yields three

[2] https://dynabench.org

low-resource target corpora, namely, $Waseem_{\text{tar}}$, $Dynamic^{v2}_{\text{tar}}$ and $Ethos_{\text{tar}}$. Furthermore, each of them is split in the 80-20 ratio to obtain their respective low-resource train (400) and validation (100) sets. For the test set from $Waseem_{\text{tar}}$, we sample 10% of the original data, disjoint from the train and validation sets, given by 1090 instances, and use the original test split of 4120 instances for $Dynamic^{v2}_{\text{tar}}$. For $Ethos_{\text{tar}}$, 20% of the data is randomly sampled, disjoint from the previous set of 500 instances, as the test set.

7.4.2 Baselines

We compare our approach with different baseline methods for transfer learning with or without neighborhood information:

Target-FT: The pre-trained BERT model is fine-tuned on the train set of the low-resource target corpus.

Seq-FT: Here, the BERT model is sequentially fine-tuned first on the resource-rich source corpus and then on the low-resource target corpus.

Mixed-FT: Here, we fine-tune BERT on a mix of the source and target corpora. Since the target instances are limited, they are first over-sampled. Then, for every mini-batch of size m, we randomly sample m training instances each from the source and the target. Their cross-entropy losses are then combined for updating the model parameters, as:

$$\min_{f,g} \quad \theta_s \frac{1}{m} \sum_i L_s(y_i^s, f(g(x_i^s))) + \theta_t \frac{1}{m} \sum_j L_t(y_j^t, f(g(x_j^t))) \tag{7.7}$$

This is similar to Equation 7.4 without the OT_u^{NN} losses.

***k*NN-FT:** For every target instance, top-k neighbors are retrieved from the source, ranked with cosine similarities over SBERT embeddings. This yields a subset of source instances that are neighbors to the target instances. We then fine-tune the BERT model with the strategy used for Mixed-FT.

***k*NN ranking:** Here, the labels of the target instances are predicted using a majority voting strategy. This voting is done over the labels associated with the top-k retrieved neighbors from the source based on their cosine similarities.

Weighted *k*NN: This uses a weighted voting of the top-k neighbors. Here we compute the sum of cosine similarities of neighbors associated with every class. The class with the highest score is returned as the predicted label of the target instance.

CE kNN$^+$ + SRC: This is the Cross-Encoder-based neighborhood framework kNN$^+$, proposed by Sarwar et al. (2022), as discussed in Section 7.2.1. For a fair comparison, we use the pre-trained BERT as the base representation. The CE kNN$^+$ is first trained on the source (SRC) and then with the target instances and their neighbors from the source.

PretRand: This is a transfer learning strategy proposed by Meftah et al. (2021) for low-resource domain adaptation. They jointly learn a pre-trained branch in the target model with a normalized, weighted, and randomly initialized branch. This is done so that the model can learn target-specific patterns while retaining the source knowledge. For a fair comparison, we use the pre-trained BERT as the base model, which is first fine-tuned on the source. For the random branch, following the approach, we add a BiLSTM layer and a Fully Connected layer over the final hidden layer from BERT. The final predictions are obtained using an element-wise sum of the predictions from the two branches.

OT$_u$: Finally, OT$_u$ is used to transfer knowledge from the source to the target using both the ED and LC losses, similar to Equation 7.4. However, this is done *without* incorporating any neighborhood information in both the cost matrix and the computation of γ.

7.4.3 Evaluation Metric

We report most of the performance scores in this chapter using the F1 score for the *hate* class, following Sarwar et al. (2022) and Attanasio et al. (2022). This is because, in a practical scenario, scores with respect to the hate class might be more important, compared to the non-hate class. In addition, we also report the performance using the macro-F1 score as done in the previous chapters. A description of the F1 metric is provided in Chapter 3.

7.4.4 Hyper-parameters and Implementation Details

All the models are trained for 10 epochs initialized with the pre-trained BERT, with a maximum sequence length of 128 tokens. The Adam optimizer is used with a learning rate of 5×10^{-5}. Besides, we perform the hyper-parameter tuning for k with a random seed and model selection using the best F1 scores for the hate class over the respective target corpus validation sets. However, while reporting the macro-F1 scores, the same hyper-parameter tuning and model selection are done using the best macro-F1 scores, in a similar way, over the respective target corpus validation sets. For the baselines of kNN-FT, kNN ranking, weighted kNN, and the OT$_u^{NN}$ variants, the number of neighbors (k) are selected from the range {10, 30, 50, 70, 100, 200, 300, 400, 500}.

After the preliminary experiments, we set $\alpha = 0.05$, $\beta = 10$, $\epsilon = 0.2$, $\lambda = 0.5$, and $\theta_t = 10$ for all our experiments. We set $\theta_s = 1$ for OT$_u^{NN}$ / OT$_{u,\text{pre-select}}^{NN}$ + sloss and $\theta_t = 10$ in Equation 7.4, 7.6 and 7.7 for all the experiments. For OT$_u^{NN}$ without sloss, we

set $\theta_s = 0$. A batch size of 32 is used for OT_u^{NN} and the baselines, except CE-kNN$^+$. The latter inherently requires the batch size to be equal to the neighborhood size, as it provides query-neighborhood pairs as inputs to the model.

For implementing the proposed OT_u^{NN} framework, the pre-trained BERT-base uncased model, implemented by `HuggingFace`[3] (Wolf et al., 2020), having 110 million parameters, is fine-tuned with the JDOT framework[4]. We encode an instance into the embedding space by obtaining the representations of the `[CLS]` token from the last hidden layer of BERT. For incorporating the neighborhood information, the pre-trained SBERT sentence embeddings are used from '`all-mpnet-base-v2`'[5] model, which is a sentence transformer model. For computing γ, the entropic regularized unbalanced OT_u solver is used with the Python Optimal Transport package[6] (Flamary et al., 2021) at the mini-batch level.

For CE kNN$^+$ + SRC, the experiments are performed with the implementation provided to us by the authors and report the results for the neighborhood size of 10 in Table 7.2. We implement PretRand ourselves following the description provided by Meftah et al. (2021). This approach is evaluated by the authors on the tasks of part-of-speech tagging, chunking, named entity recognition and morphosyntactic tagging. Therefore, the approach uses a sequence labeling model with pre-trained word embeddings and a BiLSTM-based feature extractor. However, for a fair comparison with our approach, we use the pre-trained BERT model as the feature extractor instead of the BiLSTM model for the pre-trained units. For the randomly initialized units, we follow the approach and add a BiLSTM layer over the last hidden layer of the BERT model. The pre-trained BERT model is first fine-tuned, without the randomly initialized units, on the source corpus. Then it is fine-tuned with the additional randomly initialized units on the target corpus. We use the Adam optimizer with a learning rate of 5×10^{-5} for the pre-trained BERT parameters. For the randomly initialized units, the Adam optimizer with a learning rate of 1.5×10^{-2} is used following Meftah et al. (2021).

7.5 Results

7.5.1 Discussion

Table 7.2 shows the performance obtained with the baselines and the OT_u^{NN} variants across the test sets of three low-resource target corpora using different resource-rich source corpora. We also present the performance with Target-FT for reference. The reported F1 scores are computed as averages over five runs of the same experiments with different random initializations.

The results show that transferring knowledge from a resource-rich corpus to a low-

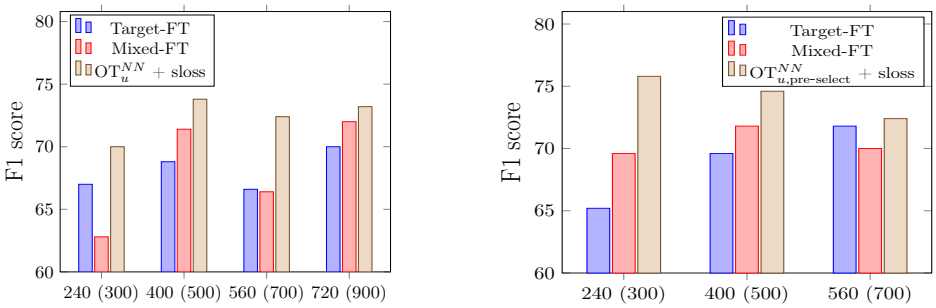
[3] https://github.com/huggingface/transformers
[4] https://github.com/bbdamodaran/deepJDOT
[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[6] https://pythonot.github.io/gen_modules/ot.unbalanced.html#ot.unbalanced.sinkhorn_unbalanced

Target corpus	$\boldsymbol{W}_{\mathbf{tar}}$		$\boldsymbol{D}^{v2}_{\boldsymbol{tar}}$		$\boldsymbol{E}_{\boldsymbol{tar}}$	
Target-FT	64.0±2.1		68.8±3.2		69.6±6.4	
Source corpus	$\boldsymbol{D}^{v2}_{\mathbf{src}}$	$\boldsymbol{E}_{\mathbf{src}}$	$\boldsymbol{W}_{\mathbf{src}}$	$\boldsymbol{E}_{\mathbf{src}}$	$\boldsymbol{D}^{v2}_{\mathbf{src}}$	$\boldsymbol{W}_{\mathbf{src}}$
Seq-FT	63.2±2.1	65.0±1.1	67.0±2.2	70.8±3.9	**79.8**±0.7	70.2±3.1
Mixed-FT	61.2±2.7	66.6±2.2	69.8*±1.6	71.4±3.9	77.6±2.1	71.8±3.5
kNN-FT	62.2±1.2	65.6±0.8	69.4*±2.3	70.8±1.9	77.2±1.5	70.6±3.4
kNN ranking	57.0	60.0	40.0	73.0*	77.0	49.0
Weighted kNN	57.0	60.0	37.0	73.0*	77.0	47.0
CE kNN$^+$ + SRC	59.8±1.8	**68.4***±0.8	65.6±1.6	68.8±3.9	76.8±0.7	67.6±2.8
PretRand	59.6±5.1	63.2±2.9	71.0*±0.6	72.2*±2.0	77.6±2.2	71.4±3.7
OT$_u$	65.4*±1.5	66.6±1.0	70.0*±2.8	71.4±5.2	73.6±3.6	**74.6***±2.9
OT$^{NN}_u$	**65.6***±2.9	67.4*±1.6	**71.6***±1.4	73.2*±0.7	73.8±2.3	72.6*±3.1
OT$^{NN}_{u,\text{pre-select}}$	64.2±1.5	67.0±2.1	**71.6***±2.7	72.6*±1.0	75.4±1.4	73.2*±1.9
OT$^{NN}_u$ + sloss	62.8±2.2	**68.4***±0.8	69.2*±3.2	**73.8***±1.6	76.8±1.9	73.4*±0.8
OT$^{NN}_{u,\text{pre-select}}$ + sloss	65.2*±1.7	66.6±1.6	70.2*±3.7	72.2*±1.3	77.2±1.3	**74.6***±2.5

Table 7.2: F1 score for the hate class (±std-dev) on the target corpus. The last four are the proposed OT$^{NN}_u$ variants. W: *Waseem*, D^{v2}: *Dynamic*^{v2}, E: *Ethos*. **Bold** denotes the best, underline denotes the second-best scores in each column. * denotes the significantly improved scores compared to Seq-FT using the McNemar test (Dror et al., 2018; McNemar, 1947).

resource corpus is generally helpful. The best scores in the 6 respective settings of Table 7.2 are substantially higher than those from Target-FT. Furthermore, while the baseline methods show inconsistent performance across different settings, the proposed OT$^{NN}_u$ variants yield the best performance in 5 out of 6 cases, and the second-best in 3 cases. The baselines of Mixed-FT, kNN variants and CE kNN$^+$ achieve significant improvements compared to the vanilla Seq-FT for only 1 case, and PretRand achieves it for 2 cases. OT$^{NN}_u$ variants, on the other hand, yield significant improvements in most cases; for instance, OT$^{NN}_u$ has significantly improved scores in 5 out of 6 cases. Besides, the best scores from OT$^{NN}_u$ improve over OT in 5 settings, while staying on par with OT in the remaining setting. This demonstrates that incorporating neighborhood information results in more effective transfer.

When *Dynamic*$^{v2}_{\text{src}}$ is used for transferring knowledge to *Ethos*$_{\text{tar}}$, Seq-FT yields the highest score (79.8). This is apparently because *Dynamic*$^{v2}_{\text{src}}$ comprises wide range of hateful forms directed toward different social groups. Since *Ethos*$_{\text{tar}}$ also involves hate against a variety of social groups, pre-training on all the source instances from *Dynamic*$^{v2}_{\text{src}}$ for transfer learning, instead of training with the nearest neighbors, seems to be more helpful in this case. However, this is not the case when the transfer occurs from *Ethos*$_{\text{src}}$ to *Dynamic*$^{v2}_{\text{tar}}$. This is likely because the *Dynamic*^{v2} corpus involves adversarial instances that can easily fool a hate speech detection system trained on a different corpus. Besides, *Ethos*$_{\text{src}}$ has a subset of hateful forms and social groups covered by *Dynamic*^{v2}. Therefore, a nearest neighborhood framework for transferring knowledge from *Ethos*$_{\text{src}}$ to *Dynamic*$^{v2}_{\text{tar}}$ yields an improved performance, the highest score being 73.8 obtained by OT$^{NN}_u$ + sloss, compared to 70.8 from Seq-FT.



(a) $Ethos_{\text{src}}$ to $Dynamic^{v2}_{\text{tar}}$ (b) $Waseem_{\text{src}}$ to $Ethos_{\text{tar}}$

Figure 7.5: Performance (F1 score for the hate class) with different sizes of the target train set. The total number of labeled instances available from the target are mentioned within the brackets, where the remaining instances are used as the target validation set.

Target corpus	$\boldsymbol{W_{tar}}$		$\boldsymbol{D^{v2}_{tar}}$		$\boldsymbol{E_{tar}}$	
Source corpus	$\boldsymbol{D^{v2}_{src}}$	$\boldsymbol{E_{src}}$	$\boldsymbol{W_{src}}$	$\boldsymbol{E_{src}}$	$\boldsymbol{D^{v2}_{src}}$	$\boldsymbol{W_{src}}$
Seq-FT	**63.2**±2.1	65.0±1.1	**67.0**±2.2	**70.8**±3.9	**79.8**±0.7	**70.2**±3.1
CE kNN$^+$ + SRC						
$k = 10$	59.8±1.8	68.4±0.8	65.6±1.6	68.8±3.9	76.8±0.7	67.6±2.8
$k = 20$	61.2±1.5	67.6±1.5	64.8±1.6	69.2±3.2	76.8±1.0	67.4±3.3
$k = 30$	60.3±1.6	68.1±1.0	64.4±1.9	69.9±2.8	76.8±0.5	68.5±1.7
$k = 40$	61.6±1.6	68.6±1.4	64.6±1.0	**70.8**±3.5	76.2±1.2	68.2±2.6
$k = 50$	60.8±2.0	**68.8**±0.7	62.8±2.6	68.4±4.8	75.8±0.4	68.4±0.5

Table 7.3: Performance of CE kNN$^+$ + SRC (F1 for hate) with different neighborhood sizes, compared with Seq-FT. W: *Waseem*, D^{v2}: $Dynamic^{v2}$, E: *Ethos*. F1 score (±std-dev) is reported on the low-resource target corpus with 400 labeled training instances (total 500 labeled instances from the target) available. **Bold** denotes the best score in each column.

Varying the size of X^t: We vary the size of the labeled target corpus available for training. We illustrate the cases of transferring knowledge from $Ethos_{\text{src}}$ to $Dynamic^{v2}_{\text{tar}}$ in Figure 7.5(a), and from $Waseem_{\text{src}}$ to $Ethos_{\text{tar}}$ in Figure 7.5(b), with different OT_u^{NN} variants. For $Dynamic^{v2}_{\text{tar}}$, 300, 500, 700, and 900 instances are sampled. We use 80% for training, given by 240, 400, 560, and 720 instances, respectively, and the remaining 20% for validation. Since the *Ethos* corpus is small, only 300, 500, and 700 instances are sampled as $Ethos_{\text{tar}}$, with the same proportions for training and validation. The target test set remains the same as in Table 7.1 for different training sizes. We observe that the OT_u^{NN} variants consistently improve the performance, with larger improvements obtained when the size of available target instances is lower. Mixed-FT, on the other hand, is inconsistent, and in some cases performs worse than Target-FT.

The improvements with OT_u^{NN} can be attributed to the fact that it can systematically *learn* the amount of transfer based on both the embedding distance and label consistency.

CE kNN$^+$ + SRC: Even though Sarwar et al. (2022) use 10 as the neighborhood size in their task of transfer learning in a cross-lingual set-up, we experiment with different

Target corpus	$\boldsymbol{W}_{\text{tar}}$		$\boldsymbol{D}^{v2}_{\text{tar}}$		$\boldsymbol{E}_{\text{tar}}$	
Source corpus	$\boldsymbol{D}^{v2}_{\textbf{src}}$	$\boldsymbol{E}_{\textbf{src}}$	$\boldsymbol{W}_{\textbf{src}}$	$\boldsymbol{E}_{\textbf{src}}$	$\boldsymbol{D}^{v2}_{\textbf{src}}$	$\boldsymbol{W}_{\textbf{src}}$
OT^{NN}_{u} + sloss	62.8±2.2	**68.4**±0.8	**69.2**±3.2	**73.8**±1.6	**76.8**±1.9	**73.4**±0.8
OT^{NN}_{u} + sloss (without ED)	**63.8**±1.3	65.8±1.7	68.0±0.0	70.0±2.4	76.4±0.8	71.8±2.5
OT^{NN}_{u} + sloss (without LC)	62.0±2.1	66.4±2.2	67.6±2.7	72.4±1.4	75.2±2.6	67.8±3.9
$\text{OT}^{NN}_{u,\text{pre-select}}$ + sloss	**65.2**±1.7	66.6±1.6	**70.2**±3.7	**72.2**±1.3	**77.2**±1.3	**74.6**±2.5
$\text{OT}^{NN}_{u,\text{pre-select}}$ + sloss (without ED)	64.4±1.5	**67.6**±1.4	67.6±4.3	70.8±2.3	75.6±2.7	74.2±5.6
$\text{OT}^{NN}_{u,\text{pre-select}}$ + sloss (without LC)	62.2±2.6	63.8±1.5	67.2±5.0	71.8±1.5	74.6±4.1	67.2±5.2

Table 7.4: Ablation study without the Embedding Distance (ED) /Label Consistency (LC) losses. F1 for hate (±std-dev) on low-resource target corpus. **Bold** denotes the best, underline denotes the second-best score for each OT^{NN}_{u} variant. W: *Waseem*, D^{v2}: *Dynamic*^{v2}, E: *Ethos*.

Target corpus	$\boldsymbol{W}_{\textbf{tar}}$		$\boldsymbol{D}^{v2}_{\textbf{tar}}$		$\boldsymbol{E}_{\textbf{tar}}$	
Source corpus	$\boldsymbol{D}^{v2}_{\textbf{src}}$	$\boldsymbol{E}_{\textbf{src}}$	$\boldsymbol{W}_{\textbf{src}}$	$\boldsymbol{E}_{\textbf{src}}$	$\boldsymbol{D}^{v2}_{\textbf{src}}$	$\boldsymbol{W}_{\textbf{src}}$
Target-FT	77.5±0.7		63.7±1.7		73.4±1.4	
Mixed-FT	75.5±0.8	**78.4**±1.0	63.2±1.3	**66.8**±1.3	79.0±2.0	71.2±2.4
OT^{NN}_{u}	**78.0**±0.8	77.9±0.9	64.1±1.1	**66.8**±1.8	78.6±2.1	74.4±1.9
$\text{OT}^{NN}_{u,\text{pre-select}}$	77.5±1.5	77.8±0.4	64.2±0.6	66.5±1.2	78.9±2.2	**76.6**±1.4
OT^{NN}_{u} + sloss	75.5±1.6	77.9±0.8	64.5±0.9	66.7±1.0	78.8±1.5	72.6±3.4
$\text{OT}^{NN}_{u,\text{pre-select}}$ + sloss	76.7±0.8	77.7±0.6	**64.9**±1.3	66.5±1.4	**79.3**±1.6	74.8±3.3

Table 7.5: Macro-F1 (± std-dev) on the target corpus for different settings. **Bold** denotes the best scores, underline the second best in each column when both the resource-rich source and the low-resource target corpora are used for training.

neighborhood sizes (k values). The results are reported in Table 7.3. However, we could not increase the neighborhood size beyond 50 because of resource constraints. This is because a mini-batch in their framework comprises a query instance from the target and all its k neighbors from the source. Thus, the number of neighbors is limited by the mini-batch size, which usually needs to be kept small when fine-tuning large language models like BERT. We can observe from Table 7.3 that the performance obtained with different neighborhood sizes is similar.

7.5.2 Ablation Study

To analyze the importance of the constituent losses in OT^{NN}_{u}, we present an ablation study in Table 7.4 by removing the ED and the LC losses from the OT^{NN}_{u} variants. The best performances for each variant are obtained in 5 out of 6 cases when both the ED and the LC losses are incorporated. The second-best performances are obtained, in most cases, when the ED loss is removed. This suggests that while both losses are essential for an effective transfer, the LC loss contributes more toward the final performance than the ED loss.

7.5.3 Macro-F1 Scores:

We present the macro-F1 scores averaged over five random initializations of every experiment in Table 7.5. The OT_u^{NN} variants yield the best and second-best scores in 5 out of 6 cases compared to Mixed-FT demonstrating the efficacy of the proposed approach.

7.5.4 Analysis of OT_u^{NN} Representations

Target instance: rt i am not sexist but girls who smoke are a big turn off even if your supermodel status its still a turn off.	
Instances	**GT**
Nearest neighbors from source using SBERT	
i do not actually care much for looks but the inflated sense of self importance that girls seem to have nowadays is rather obnoxious.	hate
totally agree. i could never find a woman or man attractive to go out with. .... i could nt be in a relationship with them as i am aromantic.	non-hate
every guy i know at work smoke at least 10 fags a day and i call them out. do not like it then do not do it.	non-hate
wow the truth is i did not get turned on until you mentioned cigar smoking men pissing on other men	non-hate
women have no redeeming features	hate
i am not sexist but women are dumb	hate
weed is stoner trash. if you are not a stoner guy, smoking this estrogen-filled garbage makes you a honorary one.	non-hate
maybe i am sexist because women are dumb	hate
women are less attracted to looks. they are far more attracted to high status and a dominant personality. the good news is you can acquire both of these.	non-hate
women without big t*ts are not attractive	hate
Nearest neighbors from source using OT_u^{NN}	
i am not sexist but women are dumb	hate
maybe i am sexist because women are dumb	hate
thrilled to have been born as the superior male gender	hate
men are superior than women in every single aspect	hate
women of colour are not pleasant to look at	hate
leadership positions should only be occupied by men. that is the way to get things done.	hate
all women are created as equal b*tches	hate
all women should be suppressed	hate
this is the fact feminism is girlf*rts you cant unsmell	hate
women are devious	hate

Table 7.6: Qualitative analysis of an example with its top 10 nearest neighbors extracted using the SBERT and the learned OT_u^{NN} representations, where the source is $Dynamic^{v2}_{\text{src}}$ and the target is $Waseem_{\text{tar}}$; GT: Ground Truth class.

We analyze the effect of training with OT_u^{NN} on the representation space by extracting the nearest neighbors of target instances. These neighbors are ranked with cosine similarity over the learned OT_u^{NN} representations, and we check their ground truth classes. They are compared with the nearest neighbors obtained using SBERT representations. Table 7.6 contains an example of a hateful instance from the training set of $Waseem_{\text{tar}}$, and its top 10 nearest neighbors from $Dynamic^{v2}_{\text{src}}$. We observe that the neighbors retrieved using the SBERT representations belong to both hate and non-hate classes. This is because SBERT is optimized mainly for semantic similarity, while they are sub-optimal in differentiating hateful instances from non-hateful ones. On the other

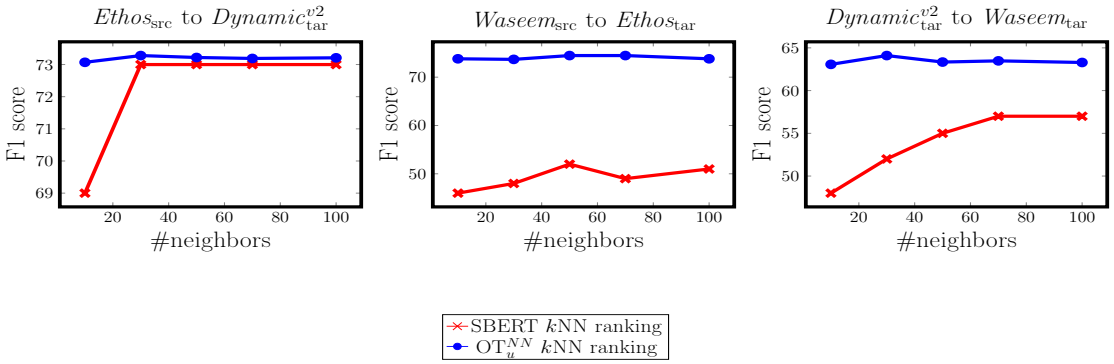


Figure 7.6: F1 (hate) using the majority voting of the k-Nearest Neighbors retrieved from SBERT and OT^{NN}_u representations.

Target corpus	$\boldsymbol{W}_{\textbf{tar}}$		$\boldsymbol{D}^{v2}_{\textbf{tar}}$		$\boldsymbol{E}_{\textbf{tar}}$	
Source corpus	$\boldsymbol{D}^{v2}_{\textbf{src}}$	$\boldsymbol{E}_{\textbf{src}}$	$\boldsymbol{W}_{\textbf{src}}$	$\boldsymbol{E}_{\textbf{src}}$	$\boldsymbol{D}^{v2}_{\textbf{src}}$	$\boldsymbol{W}_{\textbf{src}}$
Mixed-FT	17.8 m	0.4 m	4.7 m	0.5 m	14.0 m	4.7 m
OT^{NN}_u	18.9 m	0.4 m	5.1 m	0.6 m	14.2 m	5.0 m
$\text{OT}^{NN}_{u,\text{pre-select}}$	3.7 m	0.3 m	1.1 m	0.6 m	6.5 m	3.4 m
OT^{NN}_u + sloss	18.9 m	0.4 m	5.0 m	0.6 m	14.5 m	4.9 m
$\text{OT}^{NN}_{u,\text{pre-select}}$ + sloss	11.7 m	0.4 m	3.8 m	0.6 m	5.5 m	3.9 m

Table 7.7: Per epoch training time in minutes (m) for different settings. W: *Waseem*, D^{v2}: $Dynamic^{v2}$, E: *Ethos*.

hand, the neighbors obtained from OT^{NN}_u representations indicate that OT^{NN}_u brings instances across corpora, which are both semantically similar (the topic of women) and belong to the same class closer in the representation space, compared to those belonging to the opposite class.

In addition, we study the effect of the OT^{NN}_u representations by performing a simple majority voting of the top k nearest neighbors retrieved from the source with SBERT versus OT^{NN}_u. Figure 7.6 demonstrates the performance obtained on the target test set. Here the neighbors from the two representation spaces are ranked using cosine similarities. We can see that majority voting using the OT^{NN}_u representations achieves higher performance compared to that using the SBERT representations for different numbers of neighbors.

7.5.5 Computational Efficiency

We present the per epoch training time of Mixed-FT and OT^{NN} variants for different settings of the source and target corpora in Table 7.7. Mixed-FT is a baseline that involves training of the pre-trained BERT model on the combination of the source and target corpora. For every mini-batch of size m, there are m instances sampled from each of the source and target corpora (Equation 7.7). This is the same mini-batch sampling

that is followed in OT_u^{NN}. We use one Nvidia GTX 1080 Ti GPU for our experiments. It can be observed that OT_u^{NN} results in approximately the same computation time as taken by Mixed-FT in most of the settings as it does not change the model architecture, but only introduces a new training strategy. With the 'pre-select' variant, the computation time gets further reduced in a few settings. This is because, in this variant, the model only gets trained on a subset of pre-selected source instances based on the neighborhood size.

7.6 Conclusion and Future Work

In this chapter, we proposed a framework for transferring knowledge to a low-resource hate speech corpus by incorporating neighborhood information with Optimal Transport. It allowed the model to flexibly learn the amount of transfer from the nearest neighbors based both on their proximity in a sentence embedding space and label consistency. For this, we assigned a maximum cost for entries associated with non-neighbors to the cost matrix representing the pair-wise cost between source and target instances. Moreover, for computing the optimal transport plan, distances in the sentence embedding space were used. We jointly minimized the embedding distance and label consistency losses to facilitate the alignment of the joint embedding and label distributions across domains. Our framework yielded substantial improvements across hate speech corpora from varied platforms in low-resource settings. Besides, the qualitative analysis of its learned representations demonstrated that they incorporate both semantic and label similarities. This is different from sentence embedding representations, where semantically similar instances may have opposite labels. In the future, the framework can be applied for knowledge transfer in other tasks, such as sentiment classification, bragging detection (Jin et al., 2022), etc., as the methodology is not restricted to only hate speech detection.

8 Conclusion and Future Research Directions

We begin this chapter with an overview of the main contributions of this thesis and some associated limitations in Section 8.1. Based on the knowledge acquired in our work, we then venture into the potential directions for future research in Section 8.2. This chapter ends with an ethics statement in Section 8.3.

8.1 Summary of Contributions

Human interactions in social media undergo rapid changes over time and across different social and cultural contexts. This is reflected in terms of varying topics of discussion that are often triggered by real-world events, changes in writing style, the vocabulary used, and other variations even within the same language. This results in distributional shifts across different abusive language corpora sampled at different time frames or using varied generation procedures. Furthermore, the sampling strategy used for creating these corpora may result in the disproportionate presence of certain terms across the annotated classes resulting in corpus-specific correlations that impede their generalization capabilities. Considering all these factors, automatic abusive language detection systems typically suffer from degraded performance when evaluated on new out-of-distribution instances. This thesis focused on the problem of domain shift and proposed strategies for transferring knowledge in a more robust manner across corpora such that the annotation effort in the new corpus can be minimized.

In **Chapter** 4, we analyzed the impact of topic model representations on the generalizability of abusive language classifiers, without having access to a target corpus during training. The topic distributions of comments were obtained using the neural network-based TDLM model and these representations were combined with the contextualized representations obtained from the HateBERT model fine-tuned on the source corpus. We further investigated if the association of unseen comments from a new corpus with the topics present in the source corpus can contribute additional knowledge for cross-corpus abusive language detection.

- Our experiments highlighted the problem of generalizability in abusive language detection through substantially poor cross-corpus performances of classifiers compared to the in-corpus performances.
- We observed that the topic representations themselves were not sufficient for abuse

detection as they tend to lose the complete context in comments. Nonetheless, the combination of these representations with those from HateBERT gave moderate improvements in cross-corpus performance.

- Our analysis showed that some of the target corpus abusive comments, misclassified when using only HateBERT, could find associations with a few prominent topics present in the abusive comments of the source corpus. This was a potential reason for them being correctly classified by incorporating information from the topic model.
- Overall, our experiments and analysis indicated that the topic representations provide some complementary information to the contextual representation space, thereby improving the generalizability of classifiers.

A limitation of this approach can be seen in a scenario where certain topics are prominently present in the comments belonging to the abusive class in the source corpus but an out-of-distribution corpus includes comments discussing similar topics in a non-abusive context. This can also happen if some topics are prominently associated with the non-abusive comments in the source but with abusive comments in the target. In such cases, including topic information might not yield improvements or, in the worst case, it might adversely affect the performance on the target corpus.

In **Chapter** 5, we considered the problem of adapting the classifiers trained on a source corpus to a target corpus, assuming the availability of unlabeled instances from the target corpus. The performance of some popular pivot-based and adversarial approaches from a related task of sentiment classification was investigated for unsupervised domain adaptation in cross-corpus abusive language detection. We compared them with the adaptation of the pre-trained HateBERT model on the target corpus using the MLM objective and its subsequent fine-tuning on the source.

- Our experiments demonstrated the sub-optimal performances of the pivot-based and adversarial domain adaptation approaches in our task. They demonstrated poor performance, even when compared with the vanilla-supervised fine-tuning of HateBERT on the source corpus without adaptation.
- Our analysis revealed that the pivot selection criteria used in the pivot-based approaches resulted in extracting pivots that typically do not exhibit similar class affinities in the source and target corpora; around 18.8% of pivots had a similar behavior across corpora for the worst performing case and 51.4% in the best performing case. This suggests the difficulty of obtaining n-gram-based shared features that retain similar meaning and behavior in different contexts in the complex task of abuse detection.
- The PCA plots of the transformed representation spaces obtained by the domain adversarial approaches showed that the class separation learned in the source corpus does not usually reflect the separation in the target corpus sufficiently well.

- The MLM adaptation of the HateBERT model on the target corpus improved the cross-corpus performance compared to the vanilla no-adaptation baseline, indicating that such adaptation helps in incorporating the language variations of the target corpus.
- Our analysis brought important insights into the applicability of existing domain adaptation approaches and highlighted the need to build adaptation methods specifically tailored to address the challenges in abusive language detection.

In this work, we have adopted a practical scenario, where within a corpus, the abusive language is directed toward different targets. However, in order to have a deeper analysis of the impact of existing domain adaptation approaches, it could be worth considering a domain as a collection of comments dealing with one specific target when such annotations are available. In that case, domain adaptation methods would attempt to bridge the domain gap between abuse directed against two different target groups, rather than that between two different corpora. This would be a more controlled setting and reduce the effect of additional challenges brought by the presence of multiple targets. This direction has been explored by Ludwig et al. (2022) who constructed multiple target group-specific domains from a single corpus.

Chapter 6 addressed the problem of spurious corpus-specific correlations that hinder the cross-corpus performance. We focused on hate speech detection, considering hate speech as a sub-category of abusive language. Two domain adaptation approaches D-Ref-I and D-Ref-II were proposed that use feature attribution methods to automatically extract and penalize the source-specific terms that restrict domain invariance. D-Ref-I uses the misclassifications in a small set of labeled instances from the target corpus and the set of highly attributed terms for the predicted labels in the source corpus for source-specific term extraction. On the other hand, D-Ref-II uses the unlabeled instances from the target and performs domain classification to extract the terms that help in identifying the source corpus. It further penalizes the subset of these terms that have high attributions for the predicted labels in the source. We used the feature attribution methods of Integrated Gradients, Scaled Attention, and DeepLIFT for our experiments. The two proposed approaches perform term extraction and penalization of their attribution scores dynamically in every epoch while fine-tuning the BERT model on the source corpus.

- Significant improvements were obtained in cross-corpus performance with both approaches compared to a range of baselines.
- When the pre-trained BERT model was first adapted to the unlabeled target corpus instances using the MLM objective and then fine-tuned on the source corpus with our approaches, the performance improved even further. The best overall performance and more consistent improvements were achieved using DeepLIFT which gave a macro F1 score ≥ 60.5 compared to the macro F1 of 58.1 obtained using the vanilla baseline with MLM adaptation over the target.

- The qualitative analysis of some target corpus instances displayed the effect of the proposed approaches in changing the attributions towards more relevant terms compared to the vanilla baseline. It indicated that the penalization enables the model to potentially learn from wider domain invariant context in the comments.
- The two approaches extract different kinds of terms. D-Ref-I extracts terms that are present in both source and target corpora but are correlated to the classes in the target in a different way compared to that in the source, which leads to misclassifications in the target. D-Ref-II, on the other hand, typically extracts terms that are more predominantly present only in the source corpus and not in the target due to the domain classification step and also have high attributions for the predicted labels in the source.

While we have demonstrated the cross-corpus performance improvements from the two proposed approaches, it might be interesting to see whether the combination of the two approaches helps in improving the performance even further as an immediate future work. Besides, since we have focused on domain adaptation, the proposed approaches adapt the source model to the target corpus by minimizing the effect of those spurious correlations that degrade its performance only on the specific target. However, such adaptation may not guarantee model fairness through the mitigation of unintended social biases (Dixon et al., 2018) present in the models. Ramponi and Tonelli (2022) showed that improving out-of-distribution fairness results in decreased robustness or out-of-distribution model performance. Therefore, ensuring both the robustness and fairness of hate speech detection models requires further investigation.

In **Chapter** 7, we focused on the research question of effective knowledge transfer from a resource-rich source corpus to a low-resource target corpus amidst domain shift. For simulating the low-resource scenario, we assumed the availability of a small number of labeled instances in the target corpus. Since OT can find correspondences between pairs of source and target instances in a geometrically sound manner and compute the optimal amount of transfer required, OT was adopted for transferring knowledge in our task. In particular, we used the framework of *joint distribution entropy regularized unbalanced mini-batch OT* (OT_u) that (i) aligns the joint distribution of the embedding and label spaces across domains using the embedding distance loss and label consistency loss (ii) alleviates undesirable transport cost in the presence of outliers (iii) enables faster computation of OT distances with entropic regularization. Furthermore, deriving inspiration from recent work on transfer learning using neighborhood framework, we incorporated the neighborhood information with this framework of OT_u leading to our proposed approach of OT_u^{NN}. BERT was used as the underlying model. OT_u^{NN} allows flexible learning of the amount of transfer based simultaneously on the proximity of the target instances and their neighbors in a sentence embedding space obtained from the SBERT model, and their corresponding labels. We also experimented with four slightly different variants of OT_u^{NN} depending on whether or not the neighbors of the target instances were pre-selected from the source and whether or not an additional cross-

entropy loss over the source corpus was included. Our cross-corpus experiments involved hate speech corpora from different platforms.

- We observed that transferring knowledge from a different source corpus to a low-resource target was generally helpful and substantially improved the performance on the target corpus.
- OT_u^{NN} achieved significantly improved performance over the vanilla baseline of sequential transfer in 5 out of 6 cases. Incorporating neighborhood information helped in improving the performance compared to using OT_u without neighborhood information.
- OT_u^{NN} variants consistently improved the performance when the number of available labeled target instances was varied. Larger improvements were observed when the number of labeled target instances was lower.
- The ablation study showed that both label consistency loss and the embedding distance loss are important for achieving improved performance but the label consistency loss has a higher contribution.
- The qualitative analysis of the representation space learned by OT_u^{NN} demonstrated that OT_u^{NN} brings the instances, which are both semantically similar and belong to the same class, closer in the embedding space. This is different from SBERT representations which can have opposite labels associated with the nearest neighbors.

Since our framework uses neighborhood information for transferring knowledge, it relies on the degree of proximity of the neighbors. However, if all of the source and target instances are very distant semantically, all the nearest neighbors from the source may have very low cosine similarity to the corresponding target instances. In such scenarios, the framework may yield limited improvements over the vanilla fine-tuning as the available neighborhood information would be much weaker. In such cases, the performance would mainly depend on the label consistency of the neighbors.

8.2 Perspectives

The findings from this thesis open avenues for further interesting directions of research. Some of our envisioned directions that could potentially enhance the proposed approaches as well as long-term research directions are listed below.

8.2.1 Short-term

Domain-shared topics to bridge domain gap: As discussed in Chapter 4, topic models have been used for cross-domain text classification using the unlabeled data from the target (Xue et al., 2008; Li et al., 2012; Bao et al., 2013; Zhuang et al., 2013; Zhou et al., 2015; Jing et al., 2018). This is usually done by aligning domain-specific topics through domain-shared i.e. common topics across domains. However, because of topic

bias (Wiegand et al., 2018a) caused by corpus-specific correlations, a common topic across domains might not be a generalizable predictor of abusive language classes. For example, in the corpus provided by Waseem and Hovy (2016), the topic of women in sports is spuriously associated with the abusive class (Wiegand et al., 2018a). If a similar topic is present in a different corpus, comments involving a discussion on sports or on women may not necessarily be abusive. Therefore, it is likely that even common topics across the abusive language corpora are associated with opposite labels. Nonetheless, if a small amount of labeled data from the target is available, the ground truth label information from the target corpus can be used to find the common topics that are associated with the same class in both corpora. These common topics can then be used to bridge the domain gap. This can be done in a similar manner as other works on cross-domain topic modeling or treating words having higher membership to these topics as pivot words. These pivots can then be masked and predicted using the other words in the corpora to learn domain invariant representations following the pivot-based domain adaptation approaches (Ziser and Reichart, 2017, 2018; Ben-David et al., 2020).

Generalizability in the wild: The domain adaptation approaches D-Ref-I/D-Ref-II (D-Ref-X) proposed in Chapter 6 aim to adapt a model trained on a source corpus to a single target corpus. Therefore, they mitigate spurious correlations that help in improving the domain invariance with respect to a single target. However, cross-corpus adaptation may not necessarily ensure generalization 'in the wild', i.e. the ability of models to generalize to multiple unseen corpora. One potential way to address this is using a combined set of instances from more than one out-of-distribution abusive language corpora, to extract and penalize undesirable corpus-specific correlations and refine the source model using the D-Ref-X approaches. This could help in capturing other forms of spurious patterns that might not be captured by a single target corpus. Subsequently, the model can be evaluated on completely different unseen corpora to assess its generalizability.

Concept Activation Vectors instead of feature attributions: Nejadgholi et al. (2022) used the Testing Concept Activation Vector (TCAV) (Kim et al., 2018), a model interpretability method, with an abusive language detection model to analyze the influence of some pre-defined human-friendly concepts on the model prediction. They used concepts like COVID-19, anti-Asian abuse, etc., and obtained their representations by averaging the representations of instances from held-out subsets that express these concepts. These instances were manually annotated for the relevant concepts. Unlike feature attributions that measure the importance given to input terms for the prediction, TCAV can explain the sensitivity of a model to different high-level concepts useful for the task. Another advantage of TCAV is that it is not restricted to the concepts defined on the basis of the training examples and it can be used for any concept. Besides, it provides global explanations for models. Therefore, it can be used to replace feature attributions for extracting spurious corpus-specific correlations and improve the generalizability of an abusive language detection model. Such concepts can potentially be discovered using clustering methods, topic models, and other similar approaches. The top words corre-

sponding to the topics of interest can be used to manually define the concepts present in the corpus. TCAV scores can then be leveraged for automatically extracting the concepts that cause spurious correlations in a corpus, e.g. finding associations between a neutral concept like sports with the abusive class. Moreover, this can be used to examine the sensitivity of the trained model to a new concept. e.g. abuse against a new target group, which may not have been seen during training.

Examining fine-grained label space using OT_u^{NN} representations: As discussed in Chapter 7, the representation space learned using the proposed OT_u^{NN} framework brings semantically similar instances belonging to the same class closer to one another across corpora. This characteristic of the OT_u^{NN} representation space could be used to gather additional knowledge about the fine-grained categories of the unannotated instances during evaluation, without a need for re-training. In Chapter 7, we merged the original fine-grained labels in the corpora to binary labels of 'hate' and 'non-hate' for maintaining uniformity in the label spaces across corpora. However, the target corpus instances could be mapped into the original finer categories present in the source corpus (if the source corpus has fine-grained annotations) through its retrieved nearest neighbors from the source. This could be particularly useful to identify the kind of hate present in the unannotated instances. For example, the fine-grained labels of 'racism' and 'sexism' in the corpus of Waseem and Hovy (2016) or the different types of hate such as 'derogation', 'threatening language', 'dehumanization', etc. in the corpus of Vidgen et al. (2021c) could be used to obtain useful inferences about the unannotated instances. This could be done through a majority voting of the retrieved top k nearest neighbors from the source in terms of these fine-grained labels. Since the nearest neighbors obtained from the OT_u^{NN} representations are learned to be both semantically similar and belong to the same class, the unannotated instances are more likely to include the same fine-grained characteristics of hate as their annotated neighbors. For example, if the majority of top k neighbors from the source corpus belong to the class 'sexism', it is highly likely that the queried target also involves a sexist comment. This property of the OT_u^{NN} representation space can be explored in the future.

Transferring knowledge to low-resource languages: Sarwar et al. (2022) used their nearest neighborhood framework in cross-lingual settings. Similarly, the OT_u^{NN} framework can be explored for transferring knowledge from resource-rich languages, such as English, to low-resource languages. This can be done by extracting the cross-lingual neighbors using multilingual sentence embedding models like LaBSE (Feng et al., 2022) and their corresponding labels. It would, therefore, be interesting to analyze the impact of OT_u^{NN} training strategy for knowledge transfer to low-resource languages.

8.2.2 Long-term

In order to build truly robust and generalizable abusive language detection models, future studies need to explore and address multiple broader challenges with this task in the long

term. We present some of those emerging research directions below:

Incorporating knowledge from social psychology into computational models: Abusive language, especially hate speech, in most cases is a result of deeper social stereotypes and biases in human minds. Such stereotypes keep evolving with time due to the ever-changing world politics and social structure across different cultures. This makes the detection of subtle expressions of abuse extremely challenging for computational models. For this, it is important that a broader knowledge about the stereotypes existing in society be incorporated into the machine-leaning models so that they can identify them.

The psychology literature involves extensive research on stereotypes, social behavior, and the associated concepts of bias and discrimination (Dovidio et al., 2010). However, these insights have only been partially leveraged by the computer science community (Balayn et al., 2021). Recently, Fraser et al. (2021) discussed that a causal theory from social psychology characterizes stereotypes across two dimensions: warmth and competency (Fiske et al., 2018). Warmth refers to the characteristics of sociability and morality (answering questions like '*are they helpful or harmful to me?*') and could have positive (e.g. friendly, trustworthy) or negative (e.g., cold, dishonest) associations. Similarly, competence refers to agency and ability (answering questions like '*If they are helpful or harmful, how capable are they to enact that?*') and could be attributed positively (e.g., confident, intelligent) and negatively (e.g. fearful, stupid). Stereotypes have been found to be related to different combinations of these dimensions. Fraser et al. (2021) further incorporated warmth and competency in semantic embedding spaces and eventually used them to analyze how anti-stereotypical statements can be automatically generated to reduce biased thinking. Kennedy et al. (2022) derived definitions from previous works in sociology and psychology for formulating their annotation procedure while creating a hate speech corpus.

Therefore, an interesting direction would be incorporating knowledge and insights from social psychology into the data creation process as well as building more informed and robust computational models for addressing the dynamic nature of abusive language. This would be greatly benefited from machine learning and social science researchers coming together for addressing the task of abusive language detection. Insights from social psychology could potentially help in building more generalizable abuse detection models that can capture the abusive intent even when the language, topics, or targets change.

Contextually informed abuse detection: Most abusive language datasets are a collection of abusive comments without providing additional contextual information about the comments, such as the conversational thread, parent comments, and associated media content like images, videos, etc. However, in certain cases, social media comments can only be well understood within their conversational context or in the presence of additional modalities beyond text. We have made a detailed discussion in Chapter 2 about the role of contextual information in abusive language detection. Comments considered

abusive in one context may not be considered abusive in another and vice-versa (Vidgen et al., 2019). However, only a handful of the available datasets provide such contextual information (Pavlopoulos et al., 2020; Vidgen et al., 2021b). This calls for the creation of more datasets comprising additional context. Furthermore, when such information is available, abusive language detection models should be able to make proper use of this contextual information. Since social media comments are sometimes associated with media content, models that are capable of using both textual and multi-modal information, whenever the latter is available, need to be built. Zhu et al. (2022) studied the multi-modal detection of abusive memes in a zero-shot setting for identifying unseen types of memes. In the future, there should be more exploration of the impact of contextual awareness through conversational threads, images, or video content on the generalizability of abusive language detection systems.

Lifelong learning or continual learning: Human beings learn by constructing memories about previously acquired information and applying that learning to new experiences. On the other hand, neural networks usually find it difficult to both retain prior knowledge and adapt to new knowledge (Biesialska et al., 2020; Sodhani et al., 2020). Domain adaptation approaches, as proposed in this thesis, can help in adapting a model trained on an older dataset to a new set of comments amidst a distributional shift. However, model adaptation for a new domain is usually prone to forgetting the learned domains in the past, a phenomenon referred to as the *catastrophic forgetting* (McCloskey and Cohen, 1989; French, 1999) The rapidly changing social media content calls for abusive language detection models that can continuously learn and accumulate knowledge from a continuous data stream across time, while also retaining the knowledge acquired in the past. This ability is called lifelong learning or continual learning (Ring et al., 1994). In lifelong learning settings, a task is typically referred to as learning on each portion of the data. The goal is to learn continuously from a sequence of tasks while minimizing catastrophic forgetting of the knowledge learned from the previous tasks.

Even though there has been a considerable amount of work on lifelong learning for computer vision, there are limited works for NLP (Greco et al., 2019; Sun et al., 2020; Biesialska et al., 2020). To the best of our knowledge, there is only one work by Qian et al. (2021) on lifelong learning in the task of abusive language detection. Qian et al. (2021) showed that the common lifelong learning methods do not work well for abusive language as the similarity between different tasks, i.e. portion of the continuous stream of abusive language data (can also be considered domains), is unstable and comparatively low. The abrupt changes in the data make the problem of catastrophic forgetting in abusive language more severe, which is difficult to address with the existing methods. Qian et al. (2021) proposed using Variational Representation Learning (VRL) for distilling knowledge from each task into a latent variational distribution. They further incorporated a memory module based on Load-Balancing Self-Organizing Incremental Neural Network (LB-SOINN) (Zhang et al., 2013).

Existing lifelong learning methods, however, usually learn from fully labeled data available in each domain, which requires substantial annotations in the newly observed

domains (Rostami, 2021). Thus, it is important to adapt to new domains in the lifelong learning setting to address the input distributional shifts over time without requiring a large number of annotations in the new domains. Rostami (2021) proposed lifelong unsupervised domain adaptation with lifelong learning using only unannotated data in the new domain and experimented with image datasets. Indeed, lifelong or continual learning is an interesting research direction for abusive language detection that can potentially make systems more robust in practical applications.

At the same time, the creation of annotated abusive language datasets with newly emerging topics for providing updated inputs to lifelong learning systems is itself a challenge. To address this, human-in-the-loop strategies including active learning can be employed in parallel with continual learning systems to gather new datasets. Active learning, as used by Mollas et al. (2022), can ensure the creation of diverse abusive language data with less annotation effort required. Other human-in-the-loop dynamic data generation processes, as adopted by Vidgen et al. (2021c), could also be incorporated to increase the amount of training data on diverse new topics, targets, and abusive forms. Such data creation strategies along with lifelong learning can help in building truly robust abusive language detection systems that can handle the rapid and continuous evolution of abusive language in online platforms and improve the model performance on new abusive language forms.

Re-formulation of the abusive language detection task: Finally, since annotating abusive language is inherently subjective, the abusive language detection task needs rethinking and reformulation (Sap et al., 2022). This inevitable subjectivity is a major challenge for building a truly generalizable system that is useful for addressing the social implications of abusive content. This is because models tend to learn the notion of abuse based on the limited information and spurious correlations present in the labeled datasets they are trained on, which restricts their generalizability. Moreover, social media companies typically provide their content moderators with detailed guidelines or policies for determining what constitutes abuse and to avoid inconsistencies in the manual decisions. Since abusive language detection models are not provided with these guidelines and different datasets use their own definitions for annotations, they are typically unable to develop a good understanding of what is meant by abuse and thus have limited scope for performance improvements on new data.

Calabrese et al. (2022) instead reformulated the abusive language detection task with the idea of '*policy-aware abuse detection*'. This task requires detecting whether a comment violates a certain policy where the policy is incorporated into the model through its machine-friendly representation. They decomposed a policy, such as '*Posts containing dehumanizing comparisons targeted to a group based on their protected characteristics violate the policy...*', into a collection of intents (e.g. dehumanization) and associated slots (e.g. target, protected characteristic, and dehumanizing comparison). Their task entailed that all the slots must be filled in order to judge that a comment violates the policy and hence should be moderated. This made the abusive language detection system explainable as one can understand why the model has flagged a comment as abusive based

on its filled slots. To this end, they re-annotated the dataset provided by Vidgen et al. (2021c) for intent and slots. Another work by Sap et al. (2020) reformulated the abusive language detection task by introducing the concept of '*social bias frames*' to understand the nuanced implications of abusive language. They introduced a new dataset capturing multiple information about a comment such as the perceived intention of a comment, the implied meaning, whether it comprises sexual references, whether it targets a group, etc. Along similar lines, future research should invest more into rethinking the abusive language detection paradigm in order to build models that can better understand the nuances of abusive language and develop a generalizable understanding of abuse.

8.3 Ethical Considerations

This thesis serves as a means to build more robust abusive language detection models that can make proper use of the existing curated abusive language resources and adapt well to new resources or social media comments, which have not been well-annotated due to time and cost constraints. The abusive language resources used for the work are publicly available and cited appropriately, wherein the authors have discussed the sampling techniques and annotation guidelines in detail. Some of these details are discussed in Chapter 3. We acknowledge that annotating abusive content can have negative effects on the mental health of the annotators and, therefore, proper precautionary measures should be taken by creators of such resources.

The abusive examples presented in the thesis are only intended for research purposes and for better analysis of the models explored. We have replaced a subset of vowels in the profane words presented as examples with * wherever possible. The terms extracted and penalized in this work (Chapter 6) are not meant to be used off-the-shelf, but the approaches should serve as a starting point for research on model debugging and building more generalizable abusive language classifiers. Indeed, encountering and dealing with abusive content is an inevitable part of the research on abusive language detection. However, we believe that confronting the problem, rather than deliberate avoidance, would help in addressing it and reduce its negative implications on society. Alongside, while deploying abusive language classifiers, their potential side effects, such as censorship (Ullmann and Tomalin, 2020), social biases (Dixon et al., 2018; Sap et al., 2019), etc. should also be taken into account. Besides, the detection of abusive language could be accompanied by the promotion of positive and constructive interactions such as counter-speech (Chung et al., 2019; Qian et al., 2019).

While identifying individual abusive instances in social media can aid manual content moderators, completely preventing the occurrence and spread of online hate speech is more challenging. This is because abusive language results from deeper social stereotypes and coordinated attacks from communities with vested interests. Different abusive language forms evolve with time on online platforms, resulting from the ever-changing world politics and social structure. This requires that a broader knowledge about the stereotypes existing in society be incorporated into the machine-leaning models so that

they can identify them. This is because it is the presence of such biases and stereotypes in human minds that facilitate the origin and spread of abuse.

9 Résumé étendu

La croissance phénoménale d'internet au cours des dernières décennies a radicalement transformé presque tous les aspects de notre vie quotidienne. En particulier, l'avènement des média sociaux, qui comprennent les blogs, les wikis, les sites de réseaux sociaux, les sites de partage de vidéos et d'autres plates-formes (Kaplan and Haenlein, 2010; Kane et al., 2014), a entraîné des changements majeurs dans le mode de communication, car il est porteur de la promesse fondamentale d'une plus grande facilité d'échange entre les individus. Les médias sociaux ont fourni des plateformes permettant de s'exprimer publiquement et d'atteindre potentiellement un large public en publiant des contenus variés. Les utilisateurs peuvent facilement entrer en contact avec d'autres personnes partageant les mêmes intérêts sur la base de leur profil personnel et rejoindre des communautés qui leur ressemblent. En effet, les médias sociaux encouragent un engagement prolongé et productif et offrent une multitude d'avantages.

Cependant, l'expérience vécue sur ces plateformes est parfois gâchée (Papacharissi, 2004; Jurgens et al., 2019). En particulier, les médias sociaux sont devenus un terrain fertile pour des propos abusifs, qui sont une forme de communication antisociale, blessante et agressive de la part de certains groupes de personnes. Cela a empêché ces plateformes d'offrir un environnement sûr pour leurs utilisateurs. Les propos abusifs rendent les fils de commentaires empoisonnés et improductifs, car les insultes échangées dissuadent les participants qui seraient prêts à contribuer positivement à la conversation. Les discours de haine, une catégorie de propos abusifs, ciblent particulièrement les minorités (Herring et al., 2002; DeAngelis, 2009; Waseem and Hovy, 2016; Jurgens et al., 2019) et diffusent des opinions préconçues et des stéréotypes à leur encontre, marginalisant ainsi davantage les groupes sous-représentés. Ils peuvent même avoir des conséquences graves, comme un traumatisme durable (Vidgen et al., 2021a; Hinduja and Patchin, 2019) pour les victimes et conduire à des incidents violents (Burnap and Williams, 2014; Alnazzawi, 2022).

Ces dernières années, l'accent a été mis sur la réglementation des propos injurieux pour les plateformes de médias sociaux, en raison de l'application de nouvelles lois dans différents pays. Cependant, la modération de l'énorme flux de contenu en ligne généré chaque jour est une tâche énorme. L'examen et la suppression manuels des commentaires abusifs prennent du temps, sont coûteux et ont également des effets psychologiques néfastes sur les modérateurs[1], notamment face aux contraintes de temps et à la toxicité de ces contenus. Cela crée une forte motivation pour la détection automatique de propos abusifs. Les systèmes de détection automatique peuvent analyser l'énorme volume de texte et signaler les contenus détectés aux autorités compétentes, ce qui rend le proces-

[1] https://www.theguardian.com/technology/2019/sep/17/revealed-catastrophic-effects-working-facebook-moderator

sus de modération beaucoup plus rapide. Ces systèmes peuvent être développés à l'aide de techniques du traitement automatique des langue (TAL).

9.1 Motivation

La détection des propos abusifs peut être considérée comme une tâche de classification. La plupart des modèles d'apprentissage automatique reposent sur l'hypothèse de base selon laquelle les données utilisées pour l'apprentissage et l'évaluation sont échantillonnées à partir de la même distribution de probabilité sous-jacente. Cependant, cette hypothèse ne se vérifie pas dans de nombreuses applications du monde réel lorsque les données de test sont issues d'un processus de génération différent de celui des données d'apprentissage (Daumé III and Marcu, 2006). En particulier, la nature des conversations dans les médias sociaux est sujette à des changements au fil du temps. Cela conduit souvent à un changement considérable dans la distribution des données échantillonnées de différentes manières ou à différentes périodes de temps. Le caractère spontané de ces discours en ligne aboutit fréquemment à la création de nouveaux termes. En outre, Eisenstein et al. (2010, 2014) ont remarqué que les formes écrites du langage utilisé dans les médias sociaux varient en fonction de facteurs tels que le contexte social et géographique de l'intervenant. Une autre dimension est l'évolution rapide des sujets dans le temps, causée par les événements sociopolitiques en constante évolution à travers le monde (Florio et al., 2020; Saha and Sindhwani, 2012). Les propos abusifs, en particulier, sont très sensibles à cette dynamique, car les utilisateurs réagissent souvent impulsivement aux nouvelles de dernière minute et à d'autres événements qui déclenchent des explosions soudaines de contenu abusif. Pour des raisons similaires, les propos abusifs sont souvent dirigés contre différentes communautés ou individus à différentes périodes. Par conséquent, la plupart des corpus de propos abusifs échantillonnés dans un certain laps de temps présentent un vocabulaire limité en ce qui concerne la diversité de l'utilisation de la langue, les sujets, les cibles, etc. Par exemple, un corpus de propos abusifs qui ne comporte que des termes misogynes peut ne pas être suffisant pour détecter des contenus xénophobes.

De plus, les corpus de propos abusifs eux-mêmes comportent différents biais, principalement dus à la stratégie d'échantillonnage utilisée pour leur création (Wiegand et al., 2019). Certains biais spécifiques au corpus se produisent généralement en raison de l'utilisation de certains mots-clés servant à échantillonner le contenu abusif, ce qui entraîne une représentation disproportionnée de certains termes dans le corpus. Par exemple, Wiegand et al. (2018a) a signalé qu'une telle stratégie d'échantillonnage a entraîné des biais spécifiques concernant les thèmes ou les styles des utilisateurs dans un corpus populaire fourni par Waseem and Hovy (2016).Les classificateurs appris sur de tels corpus sont susceptibles de capturer ces biais plutôt que d'apprendre le concept généralisable d'abus.

En TAL, le terme *domaine* fait généralement référence à un type de corpus cohérent, en ce qui concerne le sujet, le genre, le style, le registre linguistique, etc. (Plank, 2011;

Ramponi and Plank, 2020). Étant donné que les différents corpus de propos abusifs sont échantillonnés à partir de différentes distributions sous-jacentes, les termes 'corpus' et 'domaine' peuvent être utilisés de manière interchangeable. Par conséquent, ces facteurs entraînent un *glissement de domaine* à travers ces corpus. Afin d'évaluer la manière dont les classificateurs de propos abusifs se généralisent à de nouvelles données, il est recommandé d'analyser leurs performances inter-corpus ou inter-domaine, c'est-à-dire entrainés sur un corpus et évalués sur un autre (Wiegand et al., 2018a; Karan and Šnajder, 2018).

Cependant, en raison des variations précédemment mentionnées des propos abusifs, il est difficile de développer un corpus qui soit robuste à de nouveaux cas d'abus. Les systèmes de détection de propos abusifs appris sur d'anciens corpus fonctionnent parfois mal sur des commentaires récents, ce qui réduit leur intérêt pratique. Récemment, de nombreux travaux ont montré que les classificateurs de propos abusifs obtiennent de bons résultats sur leurs ensembles de test respectifs mais leurs performances se dégradent considérablement lorsqu'ils sont évalués sur des données provenant de corpus différents (Karan and Šnajder, 2018; Swamy et al., 2019; Wiegand et al., 2019; Yin and Zubiaga, 2021). En outre, l'annotation de nouveaux corpus contenant des propos abusifs nécessite beaucoup d'efforts, d'argent et du temps, et a également un effet psychologique négatif sur les annotateurs (Schmidt and Wiegand, 2017; Poletto et al., 2019). Par conséquent, il peut être souhaitable de transférer les connaissances des corpus étiquetés existants de la meilleure façon possible lors de la construction d'un nouveau modèle. Le but de cette thèse est de définir des stratégies qui permettent de transférer efficacement les connaissances pour minimiser l'effet négatif du changement de domaine ainsi que de réduire l'effort d'annotation nécessaire pour atteindre des niveaux de performance satisfaisants pour un nouveau domaine.

L'apprentissage par transfert est un concept utilisé dans l'apprentissage automatique où les connaissances acquises précédemment pour un domaine ou une tâche sont appliquées pour résoudre un problème concernant un domaine ou une tâche différente mais connexe(Pan and Yang, 2009; Ruder et al., 2019; Mozafari et al., 2020). DDans notre application, puisque la tâche reste la même mais que le domaine change, nous cherchons soit à construire des modèles généralisables sans utiliser les données du nouveau domaine, soit à adapter nos modèles à un nouveau domaine lorsque certaines données de ce domaine sont disponibles. Ce dernier scénario, appelé adaptation au domaine, est une catégorie particulière de l'apprentissage par transfert, appelée *apprentissage par transfert transductif*. (Pan and Yang, 2009). Contrairement à l'apprentissage par transfert inductif, où les tâches source et cible diffèrent, dans l'apprentissage par transfert transductif, les tâches source et cible restent inchangées, tandis que les domaines source et cible sont différents en termes de leurs distributions de probabilité sous-jacentes. Dans cette thèse, nous nous sommes concentrés sur le problème du transfert de domaine et avons proposé des stratégies pour transférer les connaissances d'une manière plus robuste de sorte que l'effort d'annotation de ce nouveau corpus puisse être minimisé.

9.2 Modèles Thématiques pour L'analyse de la Généralisabilité

Dans le Chapitre 4, nous avons analysé l'impact des représentations du modèle thématique sur la généralisabilité des classificateurs de propos abusifs, sans avoir accès au corpus cible pendant l'apprentissage. Nous avons obtenu les distributions thématiques des commentaires à l'aide du modèle TDLM (Topically Driven Language Model) basé sur un réseau neuronal et nous avons combiné ces représentations avec les représentations contextualisées obtenues à partir du modèle HateBERT ajusté (finetuned) sur le corpus source. Nous avons également cherché à savoir si l'association des commentaires non vus d'un nouveau corpus avec les thèmes présents dans le corpus source peut apporter des connaissances supplémentaires pour la détection des propos abusifs.

- Nos expériences ont mis en évidence le problème de la généralisation de la détection des propos abusifs : les performances inter-corpus sont nettement plus faibles que celles intra-corpus.
- Nous avons observé que les représentations des thèmes à elles-seules n'étaient pas suffisantes pour la détection des propos abusifs car elles ont tendance à ne pas prendre en compte le contexte complet des commentaires. Néanmoins, la combinaison de ces représentations avec celles de HateBERT a permis une légère amélioration des performances inter-corpus.
- Notre analyse a montré que certains commentaires du corpus cible, mal classés en utilisant uniquement HateBERT, avaient des relations avec certains sujets importants présents dans les commentaires abusifs du corpus source. C'est une raison possible pour laquelle ils ont été correctement classés en incorporant les informations du modèle thématique..
- Globalement, nos expériences et nos analyses ont montré que les représentations thématiques fournissent des informations complémentaires à la représentation contextuelle, améliorant ainsi la généralisation des classificateurs.

9.3 Adaptation Non Supervisée du Domaine

Dans le Chapitre 5, nous avons étudié le problème de l'adaptation des classificateurs appris sur un corpus source à un corpus cible, en supposant la disponibilité d'instances non étiquetées du corpus cible. Nous avons étudié les performances de certaines approches populaires basées sur la méthode des mots pivots et la méthode des réseaux antagonistes (adversarial networks) pour l'adaptation non supervisée au domaine dans une tâche connexe de classification des sentiments pour la détection des propos abusifs. Nous les avons comparées à l'adaptation du modèle HateBERT pré-entraîné sur le corpus cible en utilisant l'objectif MLM (masked language model) suivi d'un ajustement sur le corpus source.

- Pour notre tâche, nos expériences ont montré les performances sous-optimales des approches d'adaptation de domaine basées sur les pivots et antagonistes. Elles

ont montré des performances médiocres, même lorsqu'on les compare au classique ajustement de HateBERT sur le corpus source sans adaptation.

- Notre analyse a révélé que les critères de sélection des mots pivots utilisés ont permis d'extraire des pivots qui ne correspondent pas aux mêmes classes dans les corpus source et cible; seulement 18.8 % des pivots avaient un comportement similaire dans les corpus source et cible dans le cas le pire cas et 51.4 % dans le cas le plus favorable. Cela indique la difficulté d'obtenir des caractéristiques partagées basées sur les n-grammes qui conservent une signification et un comportement similaires dans différents contextes dans la tâche complexe de détection des propos abusifs.
- Les graphiques issus de l'analyse en composantes principales des espaces de représentation obtenus par les approches antagonistes ont montré que la séparation des classes apprises dans le corpus source ne coïncide pas suffisamment bien avec la séparation dans le corpus cible dans la plupart des cas.
- L'adaptation MLM du modèle HateBERT sur le corpus cible a amélioré les performances inter-corpus par rapport au modèle sans adaptation, ce qui indique qu'une telle adaptation permet d'incorporer les variations linguistiques du corpus cible.
- Notre analyse a apporté un éclairage important sur l'applicabilité des approches existantes d'adaptation au domaine et a mis en évidence la nécessité de construire des méthodes d'adaptation spécifiquement adaptées pour relever les défis de la détection des propos abusifs.

9.4 Explications du Modèle pour la Pénalisation des Corrélations Fallacieuses

Le Chapitre 6 aborde le problème des corrélations fallacieuses spécifiques à un corpus qui nuisent aux performances inter-corpus. Dans ce chapitre, nous nous sommes concentrés sur la détection des discours de haine, en considérant ces derniers comme une sous-catégorie de propos abusifs. Nous avons proposé deux approches d'adaptation au domaine qui effectuent un raffinement dynamique, à savoir D-Ref-I et D-Ref-II, qui utilisent des méthodes d'attribution de caractéristiques pour extraire automatiquement et pénaliser les termes spécifiques à la source qui limitent l'invariance du domaine. D-Ref-I utilise les erreurs de classification sur un petit ensemble d'instances étiquetées du corpus cible et les termes qui obtiennent la plus grande attribution pour les étiquettes de haine et de non-haine dans le corpus source sont extraits. D'autre part, D-Ref-II utilise les instances non étiquetées de la cible et effectue une classification de domaine pour extraire les termes qui ont aidé à obtenir une bonne identification dans le corpus source. Cette méthode pénalise ensuite le sous-ensemble de ces termes qui sont fortement attribués aux étiquettes obtenues sur le corpus source. Pour nos expériences, nous avons utilisé les méthodes d'attribution de caractéristiques suivantes: Integrated Gradients, Scaled Attention et DeepLIFT. Les deux approches proposées procèdent à l'extraction de termes et à la pénalisation de leurs scores d'attribution de manière dynamique à chaque époque, tout en ajustant le modèle BERT sur le corpus source.

- Nous avons obtenu des améliorations significatives et cohérentes des performances inter-corpus avec les deux approches par rapport à des approches de la littérature (baselines).
- Lorsque le modèle BERT pré-entraîné est d'abord adapté aux instances non étiquetées du corpus cible à l'aide de l'objectif MLM, puis affiné sur le corpus source avec nos approches, les performances se sont encore améliorées. La meilleure performance globale et des améliorations consistantes ont été obtenues en utilisant DeepLIFT qui a donné un score macro F1 de ≥ 60.5 % par rapport au macro F1 de 58.1 % obtenu avec une méthode baseline d'adaptation MLM sur la cible.
- L'analyse qualitative de certaines instances du corpus cible a montré l'effet des approches proposées dans le changement des attributions en faveur de termes plus pertinents. Elle a indiqué que la pénalisation permet au modèle d'apprendre à partir d'un contexte plus large et invariant au domaine.
- Les deux approches ont extrait différents types de termes. D-Ref-I a extrait des termes qui étaient présents à la fois dans les corpus source et cible, mais qui étaient corrélés aux classes dans la cible d'une manière différente de celle de la source, ce qui a entraîné des erreurs de classification dans la cible. En revanche, D-Ref-II extrayait typiquement des termes qui étaient présents de manière plus prédominante uniquement dans le corpus source et non dans la cible en raison de l'étape de classification de domaine, et qui ont également contribué à la prédiction des étiquettes de la source.

9.5 Transport Optimal en Fonction du Voisinage

Dans le Chapitre 7, nous nous sommes concentrés sur la question de recherche du transfert efficace des connaissances d'un grand corpus vers un corpus de taille limité dans un contexte de changement de domaine. Pour simuler un scénario à faibles ressources, nous avons supposé la disponibilité d'un petit nombre d'instances étiquetées dans le corpus cible. Puisque le transport optimal (OT optimal transport) peut trouver des correspondances entre des paires d'instances source et cible d'une manière géométriquement adéquate et calculer la quantité optimale de transfert requise, nous avons adopté l'OT pour le transfert de connaissances. En particulier, nous avons utilisé le cadre de l'OT à mini-lots, non équilibrés et régularisés par l'entropie de la distribution conjointe (OT_u) qui

(i) aligne la distribution conjointe des espaces d'embeddings et d'étiquetage en utilisant la distance entre les embeddings et la perte de cohérence d'étiquetage;

(ii) élimine le coût de transport indésirable en présence de valeurs aberrantes;

(iii) permet un calcul plus rapide des distances OT avec la régularisation de l'entropie.

En outre, en nous inspirant des travaux récents sur l'apprentissage par transfert utilisant le voisinage, nous avons incorporé les informations de voisinage au formalisme OT_u, ce qui a conduit à l'approche proposée de OT_u^{NN}. Nous avons utilisé BERT comme modèle sous-jacent. Notre approche a permis un apprentissage flexible de la quantité de trans-

fert, basé simultanément sur la proximité des instances cibles et de leurs voisins dans l'espace d'embeddings obtenu à partir du modèle SBERT et de leurs étiquettes correspondantes. Nous avons également expérimenté quatre variantes légèrement différentes de OT_u^{NN} selon que nous présélectionnions ou non les voisins des instances cibles dans la source et que nous incluions, ou pas, une perte d'entropie croisée supplémentaire sur le corpus source. Nos expériences inter-corpus ont porté sur des corpus de discours haineux provenant de différentes plateformes.

- Nous avons observé que le transfert de connaissances d'un corpus source vers une cible à faibles ressources était généralement utile et améliorait les performances sur le corpus cible.
- OT_u^{NN} a permis d'améliorer significativement les performances par rapport à la méthode baseline dans 5 cas sur 6. L'incorporation d'informations de voisinage a permis d'améliorer les performances par rapport à l'utilisation d'OT_u sans informations de voisinage.
- Les variantes de OT_u^{NN} ont constamment amélioré les performances lorsque le nombre d'instances cibles étiquetées disponibles variait. Des améliorations plus importantes ont été observées lorsque le nombre d'instances de cibles étiquetées était plus faible.
- Les études d'ablation ont montré que les fonctions objectifs concernant la cohérence des étiquettes et la distance d'embeddings étaient toutes deux importantes pour l'obtention de meilleures performances, mais que la perte de cohérence des étiquettes avait une contribution plus importante.
- L'analyse qualitative de l'espace de représentation appris par OT_u^{NN} a montré que OT_u^{NN} rapproche les instances qui sont à la fois sémantiquement similaires et qui appartiennent à la même classe dans l'espace d'embedding. Ceci est différent des représentations SBERT qui peuvent avoir des étiquettes différentes associées à des plus proches voisins.

Notre étude et les stratégies proposées apportent des contributions importantes pour mieux comprendre et traiter le problème crucial de l'apprentissage par transfert dans la détection des propos abusifs.

Bibliography

Abdul-Mageed, M., Elmadany, A., and Nagoudi, E. M. B. (2021). ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Adebayo, J., Muelly, M., Liccardi, I., and Kim, B. (2020). Debugging tests for model explanations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 700–712. Curran Associates, Inc.

Agrawal, S. and Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. In *European conference on information retrieval*, pages 141–153. Springer.

Ahmad, K., Gillam, L., and Tostevin, L. (1999). University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In *TREC*.

Al Kuwatly, H., Wich, M., and Groh, G. (2020). Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Alatawi, H. S., Alhothali, A. M., and Moria, K. M. (2021). Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT. *IEEE Access*, 9:106363–106374.

Aldjanabi, W., Dahou, A., Al-qaness, M. A. A., Elaziz, M. A., Helmi, A. M., and Damaševičius, R. (2021). Arabic offensive and hate speech detection using a cross-corpora multi-task learning model. *Informatics*, 8(4).

Alemzadeh, H., Raman, J., Leveson, N., Kalbarczyk, Z., and Iyer, R. K. (2016). Adverse events in robotic surgery: A retrospective study of 14 years of fda data. *PLOS ONE*, 11(4):1–20.

Alnazzawi, N. (2022). Using twitter to detect hate crimes and their motivations: The hatemotiv corpus. *Data*, 7(6).

Alorainy, W., Burnap, P., Liu, H., and Williams, M. L. (2019). "the enemy among us": Detecting cyber hate speech with threats-based othering language embeddings. *ACM Trans. Web*, 13(3).

Altschuler, J., Niles-Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in neural information processing systems*, 30.

Alwosheel, A., van Cranenburgh, S., and Chorus, C. G. (2018). Is your dataset big enough? sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling*, 28:167–182.

Anagnostou, A., Mollas, I., and Tsoumakas, G. (2018). Hatebusters: A web application for actively reporting youtube hate speech. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 5796–5798. AAAI Press.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2018). Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Angelova, R. and Weikum, G. (2006). Graph-based text classification: Learn from your neighbors. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, page 485–492, New York, NY, USA. Association for Computing Machinery.

Antoun, W., Baly, F., and Hajj, H. (2020). AraBERT: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Arango, A., Pérez, J., and Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 45–54, New York, NY, USA. Association for Computing Machinery.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

Atanasova, P., Simonsen, J. G., Lioma, C., and Augenstein, I. (2020). A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Attanasio, G., Nozza, D., Hovy, D., and Baralis, E. (2022). Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL2022*. Association for Computational Linguistics.

Ba, J., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *ArXiv*, abs/1607.06450.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46.

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 759–760, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Balayn, A., Yang, J., Szlavik, Z., and Bozzon, A. (2021). Automatic identification of harmful, aggressive, abusive, and offensive language on the web: A survey of technical biases informed by psychology literature. *Trans. Soc. Comput.*, 4(3).

Balsubramani, A., Dasgupta, S., Moran, S., et al. (2019). An adaptive nearest neighbor rule for classification. *Advances in Neural Information Processing Systems*, 32.

Banerjee, S., Raja Chakravarthi, B., and McCrae, J. P. (2020). Comparison of pretrained embeddings to identify hate speech in indian code-mixed text. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 21–25.

Banerjee, S., Sarkar, M., Agrawal, N., Saha, P., and Das, M. (2021). Exploring transformer based models to identify hate speech and offensive content in english and indo-aryan languages. In *FIRE*.

Bao, Y., Chang, S., Yu, M., and Barzilay, R. (2018). Deriving machine attention from human rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1903–1913, Brussels, Belgium. Association for Computational Linguistics.

Bao, Y., Collier, N., and Datta, A. (2013). A partially supervised cross-collection topic model for cross-domain text classification. In *Proceedings of the 22nd ACM International Conference on Information amp; Knowledge Management*, CIKM '13, page 239–248, New York, NY, USA. Association for Computing Machinery.

Bashar, M. A., Nayak, R., Luong, K., and Balasubramaniam, T. (2021). Progressive domain adaptation for detecting hate speech on social media with small training set and its application to covid-19 concerned posts. *Social Network Analysis and Mining*, 11(1):1–18.

Basile, V. (2020). Domain adaptation for text classification with weird embeddings. In *CLiC-it*.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Bassignana, E., Basile, V., and Patti, V. (2018). Hurtlex: A multilingual lexicon of words to hurt. In *CLiC-it*.

Bauwelinck, N. and Lefever, E. (2019). Measuring the impact of sentiment for hate speech detection on twitter. In Folds, D., Lefever, E., Gera, R., and Hoste, V., editors, *Proceedings of HUSO 2019, The fifth international conference on human and social analytics*, pages 17–22. IARIA, International Academy, Research, and Industry Association.

Ben-David, E., Rabinovitz, C., and Reichart, R. (2020). Perl: Pivot-based domain adaptation for pre-trained deep contextualized embedding models. *Transactions of the Association for Computational Linguistics*, 8:504–5221.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Mach. Learn.*, 79(1–2):151–175.

Benamou, J.-D. (2003). Numerical resolution of an "unbalanced" mass transport problem. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 37(5):851–868.

Biesialska, M., Biesialska, K., and Costa-jussà, M. R. (2020). Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6523–6541, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., and Samek, W. (2016). Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, pages 63–71. Springer.

Binns, R., Veale, M., Kleek, M. V., and Shadbolt, N. (2017). Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *International Conference on Social Informatics (SocInfo)*.

Birdal, T., Arbel, M., Simsekli, U., and Guibas, L. J. (2020). Synchronizing probability measures on rotations via optimal transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Blei, D., Ng, A., and Jordan, M. (2001). Latent dirichlet allocation. *Advances in neural information processing systems*, 14.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia. Association for Computational Linguistics.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Bodapati, S., Gella, S., Bhattacharjee, K., and Al-Onaizan, Y. (2019). Neural word decomposition models for abusive language detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 135–145, Florence, Italy. Association for Computational Linguistics.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Bosco, C., Felice, D., Poletto, F., Sanguinetti, M., and Maurizio, T. (2018). Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9. CEUR.

Bose, T., Aletras, N., Illina, I., and Fohr, D. (2022a). Domain classification-based source-specific term penalization for domain adaptation in hate-speech detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6656–6666, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Bose, T., Aletras, N., Illina, I., and Fohr, D. (2022b). Dynamically refined regularization for improving cross-corpora hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 372–382, Dublin, Ireland. Association for Computational Linguistics.

Bose, T., Illina, I., and Fohr, D. (2021a). Generalisability of topic models in cross-corpora abusive language detection. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 51–56, Online. Association for Computational Linguistics.

Bose, T., Illina, I., and Fohr, D. (2021b). Unsupervised domain adaptation in cross-corpora abusive language detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 113–122. Association for Computational Linguistics.

Bose, T., Illina, I., and Fohr, D. (2022c). Transferring knowledge via neighborhood-aware optimal transport for low-resource hate speech detection. In *Proceedings of the Asia-Pacific Chapter of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Botelho, A., Hale, S., and Vidgen, B. (2021). Deciphering implicit hate: Evaluating automated detection algorithms for multimodal hate. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1896–1907, Online. Association for Computational Linguistics.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breitfeller, L., Ahn, E., Jurgens, D., and Tsvetkov, Y. (2019). Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Briliani, A., Irawan, B., and Setianingsih, C. (2019). Hate speech detection in indonesian language on instagram comment section using k-nearest neighbor classification method. *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, pages 98–104.

Bunne, C., Alvarez-Melis, D., Krause, A., and Jegelka, S. (2019). Learning generative models across incomparable spaces. In *International conference on machine learning*, pages 851–861. PMLR.

Burnap, P. and Williams, M. L. (2014). Hate speech, machine classification and statistical modelling of information flows on twitter: interpretation and communication for policy decision making. In *Proceedings of IPP 2014*, pages 1–18.

Calabrese, A., Ross, B., and Lapata, M. (2022). Explainable abuse detection as intent classification and slot filling. *arXiv preprint arXiv:2210.02659*.

Calderón, C. A., de la Vega, G., and Herrero, D. B. (2020). Topic modeling and characterization of hate speech against immigrants on twitter around the emergence of a far-right party in spain. *Social Sciences*, 9(11).

Cao, Z., Li, S., Liu, Y., Li, W., and Ji, H. (2015). A novel neural topic model and its supervised extension. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2210–2216. AAAI Press.

Caselli, T., Basile, V., Mitrović, J., and Granitzer, M. (2021). HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., and Granitzer, M. (2020). I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Cer, D., Yang, Y., yi Kong, S., Hua, N., Limtiaco, N. L. U., John, R. S., Constant, N., Guajardo-Céspedes, M., Yuan, S., Tar, C., hsuan Sung, Y., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder. In *EMNLP demonstration*, Brussels, Belgium.

Cercas Curry, A., Abercrombie, G., and Rieser, V. (2021). ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, 6(1):1–6.

Chen, C., Xie, W., Huang, W., Rong, Y., Ding, X., Huang, Y., Xu, T., and Huang, J. (2019). Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 627–636.

Chen, L., Gan, Z., Cheng, Y., Li, L., Carin, L., and Liu, J. (2020a). Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pages 1542–1553. PMLR.

Chen, M., Xu, Z., Weinberger, K. Q., and Sha, F. (2012a). Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th International Conference on Machine Learning*, ICML'12, page 1627–1634, Madison, WI, USA. Omnipress.

Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012b). Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80.

Chen, Z., Zhou, L. J., Li, X. D., Zhang, J. N., and Huo, W. J. (2020b). The lao text classification method based on knn. *Procedia Computer Science*, 166:523–528. Proceedings of the 3rd International Conference on Mechatronics and Intelligent Robotics (ICMIR-2019).

Cheng, X., Yan, X., Lan, Y., and Guo, J. (2014). Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941.

Chiril, P., Pamungkas, E. W., Benamara, F., Moriceau, V., and Patti, V. (2022). Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation*, 14(1):322–352.

Chizat, L. (2017). *Unbalanced optimal transport: Models, numerical methods, applications*. PhD thesis, Université Paris sciences et lettres.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014a). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014b). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Chong, Z. (2018). Germany kicks year off with strict online hate speech law. https://www.cnet.com/tech/services-and-software/german-hate-speech-law-goes-into-effect-on-1-jan/. Online; accessed July 2022.

Chrysostomou, G. and Aletras, N. (2021). Improving the faithfulness of attention-based explanations with task-specific information for text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 477–488, Online. Association for Computational Linguistics.

Chrysostomou, G. and Aletras, N. (2022a). An empirical study on explanations in out-of-domain settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6920–6938, Dublin, Ireland. Association for Computational Linguistics.

Chrysostomou, G. and Aletras, N. (2022b). Flexible instance-specific rationalization of NLP models. AAAI Conference on Artificial Intelligence.

Chung, Y.-L., Kuzmenko, E., Tekiroglu, S. S., and Guerini, M. (2019). CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to

fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Conneau, A., Schwenk, H., Barrault, L., and Lecun, Y. (2017). Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1107–1116, Valencia, Spain. Association for Computational Linguistics.

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., and Villata, S. (2019). Cross-platform evaluation for italian hate speech detection. In *CLiC-it*.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017a). Joint distribution optimal transportation for domain adaptation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3733–3742, Red Hook, NY, USA. Curran Associates Inc.

Courty, N., Flamary, R., and Tuia, D. (2014). Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2017b). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865.

Croom, A. M. (2011). Slurs. *Language Sciences*, 33(3):343–358.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Dadvar, M. and Eckert, K. (2020). Cyberbullying detection in social networks using deep learning based models; a reproducibility study. *ArXiv*, abs/1812.08046.

Dadvar, M., Trieschnigg, D., Ordelman, R., and Jong, F. d. (2013). Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer.

Dai, Y., Liu, J., Ren, X., and Xu, Z. (2020). Adversarial training based multi-source unsupervised domain adaptation for sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7618–7625.

Damodaran, B. B., Kellenberger, B., Flamary, R., Tuia, D., and Courty, N. (2018). Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463.

Datta, A., Sen, S., and Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617.

Daumé III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Daumé III, H. and Marcu, D. (2006). Domain adaptation for statistical classifiers. *J. Artif. Int. Res.*, 26(1):101–126.

Davidson, T. and Bhattacharya, D. (2020). Examining racial bias in an online abuse corpus with structural topic modeling. *arXiv preprint arXiv:2005.13041*.

Davidson, T., Bhattacharya, D., and Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

de Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

DeAngelis, T. (2009). Unmasking racial micro aggressions. *Monitor on Psychology*, 40(2):42.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dhouib, S., Redko, I., and Lartizien, C. (2020). Margin-aware adversarial domain adaptation with optimal transport. In *International Conference on Machine Learning*, pages 2514–2524. PMLR.

Dieng, A. B., Ruiz, F. J. R., and Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Dovidio, J. F., Glick, P., and Hewstone, M. (2010). The sage handbook of prejudice, stereotyping and discrimination. *The SAGE Handbook of Prejudice, Stereotyping and Discrimination*, pages 1–672.

Dror, R., Baumer, G., Shlomov, S., and Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

d'Sa, A. G., Illina, I., and Fohr, D. (2020). BERT and fasttext embeddings for automatic detection of toxic speech. In *2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA)*, pages 1–5. IEEE.

D'Sa, A. G., Illina, I., and Fohr, D. (2020). Towards non-toxic landscapes: Automatic toxic comment detection using DNN. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 21–25, Marseille, France. European Language Resources Association (ELRA).

d'Sa, A. G., Illina, I., Fohr, D., and Akbar, A. (2022). Exploration of multi-corpus learning for hate speech classification in low resource scenarios. In *TSD 2022-25th International Conference on Text, Speech and Dialogue*.

Du, C., Sun, H., Wang, J., Qi, Q., and Liao, J. (2020). Adversarial and domain-aware BERT for cross-domain sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, Online. Association for Computational Linguistics.

Du, M., Liu, N., Yang, F., and Hu, X. (2019). Learning credible deep neural networks with rationale regularization. *2019 IEEE International Conference on Data Mining (ICDM)*, pages 150–159.

Dubois, D. and Prade, H. (1990). Rough fuzzy sets and fuzzy rough sets. *International Journal of General System*, 17(2-3):191–209.

Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(4):325–327.

Duggan, M. (2017). *Online harassment 2017*. Pew Research Center.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA.

Eger, S., Şahin, G. G., Rücklé, A., Lee, J.-U., Schulz, C., Mesgar, M., Swarnkar, K., Simpson, E., and Gurevych, I. (2019). Text processing like humans do: Visually attacking and shielding NLP systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1634–1647, Minneapolis, Minnesota. Association for Computational Linguistics.

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. Association for Computational Linguistics.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA. Association for Computational Linguistics.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2014). Diffusion of lexical change in social media. *PloS one*, 9(11):e113114.

ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., and Yang, D. (2021). Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., and Lee, S.-I. (2020). Learning explainable models using attribution priors.

Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., and Lee, S.-I. (2021). Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7):620–631.

Espinosa Anke, L., Declerck, T., Gromann, D., Zhang, Z., Luo, L., Gromann, D., Espinosa Anke, L., and Declerck, T. (2019). Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semant. Web*, 10(5):925–945.

Evkoski, B., Ljubešić, N., Pelicon, A., Mozetič, I., and Kralj Novak, P. (2021). Evolution of topics and hate speech in retweet network communities. *Applied Network Science*, 6(1):1–20.

Fan, A., Gardent, C., Braud, C., and Bordes, A. (2021). Augmenting Transformers with KNN-Based Composite Memory for Dialog. *Transactions of the Association for Computational Linguistics*, 9:82–99.

Fatras, K. (2021). *Deep learning and optimal transport: learning from one another*. PhD thesis, Université de Bretagne Sud.

Fatras, K., Séjourné, T., Flamary, R., and Courty, N. (2021). Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pages 3186–3197. PMLR.

Fauzi, M. A. and Yuniarti, A. (2018). Ensemble method for indonesian twitter hate speech detection. *Indonesian Journal of Electrical Engineering and Computer Science*, 11(1):294–299.

Fehn Unsvåg, E. and Gambäck, B. (2018). The effects of user features on Twitter hate speech detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 75–85, Brussels, Belgium. Association for Computational Linguistics.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Fersini, E., Rosso, P., and Anzovino, M. (2018a). Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

Fersini, E., Rosso, P., and Anzovino, M. (2018b). Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@SEPLN*.

Fiske, S. T., Cuddy, A. J., Glick, P., and Xu, J. (2018). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Social cognition*, pages 162–214. Routledge.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. (2021). Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Florio, K., Basile, V., Polignano, M., Basile, P., and Patti, V. (2020). Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12).

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Fortuna, P., Soler, J., and Wanner, L. (2020). Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Fortuna, P., Soler-Company, J., and Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing Management*, 58(3):102524.

Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Fraser, K. C., Nejadgholi, I., and Kiritchenko, S. (2021). Understanding and countering stereotypes: A computational approach to the stereotype content model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 600–616, Online. Association for Computational Linguistics.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Gage, P. (1994). A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Gambäck, B. and Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada. Association for Computational Linguistics.

Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. PMLR.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030.

Gao, L. and Huang, R. (2017). Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Gao, L., Kuppersmith, A., and Huang, R. (2017). Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 774–782, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Garg, S., Vu, T., and Moschitti, A. (2020). Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. In *34th AAAI Conference on Artificial Intelligence*.

Ge, Z., Liu, S., Li, Z., Yoshie, O., and Sun, J. (2021). Ota: Optimal transport assignment for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 303–312.

Genevay, A., Peyré, G., and Cuturi, M. (2018). Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR.

Ghosal, D., Hazarika, D., Roy, A., Majumder, N., Mihalcea, R., and Poria, S. (2020). KinGDOM: Knowledge-Guided DOMain Adaptation for Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3198–3210, Online. Association for Computational Linguistics.

Ghosh, A. and Veale, T. (2016). Fracking sarcasm using neural network. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 161–169.

Gitari, N. D., Zuping, Z., Damien, H., and Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Glavaš, G., Karan, M., and Vulić, I. (2020). XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning*, ICML'11, page 513–520, Madison, WI, USA. Omnipress.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Gomez, R., Gibert, J., Gomez, L., and Karatzas, D. (2020). Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Gosiewska, A., Kozak, A., and Biecek, P. (2021). Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. *Decision Support Systems*, 150:113556. Interpretable Data Science For Decision Making.

Greco, C., Plank, B., Fernández, R., and Bernardi, R. (2019). Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3601–3605, Florence, Italy. Association for Computational Linguistics.

Grieve, J. G. D., Nini, A., and Guo, D. (2018). Mapping lexical innovation on american social media. *Journal of English Linguistics*, 46:293 – 319.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., and Asokan, N. (2018). All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, AISec '18, page 2–12, New York, NY, USA. Association for Computing Machinery.

Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

Hall, M. A. and Frank, E. (2008). Combining naive bayes and decision tables. In *FLAIRS conference*, volume 2118, pages 318–319.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Henderson, M., Al-Rfou, R., Strope, B., Sung, Y.-H., Lukács, L., Guo, R., Kumar, S., Miklos, B., and Kurzweil, R. (2017). Efficient natural language response suggestion for smart reply. *ArXiv*, abs/1705.00652.

Hern, A. (2019). Revealed: catastrophic effects of working as a Facebook moderator. https://www.theguardian.com/technology/2019/sep/17/revealed-catastrophic-effects-working-facebook-moderator. Online; accessed July 2022.

Herring, S., Job-Sluder, K., Scheckler, R., and Barab, S. (2002). Searching for safety online: Managing" trolling" in a feminist forum. *The information society*, 18(5):371–384.

Hinduja, S. and Patchin, J. W. (2019). Connecting adolescent suicide to the severity of bullying and cyberbullying. *Journal of school violence*, 18(3):333–346.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hofland, K. and Johansson, S. (1982). *Word frequencies in british and american english*. Norwegian computing centre for the Humanities.

Hosseini, H., Kannan, S., Zhang, B., and Poovendran, R. (2017). Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*.

Hosseinmardi, H., Mattson, S. A., Ibn Rafiq, R., Han, R., Lv, Q., and Mishra, S. (2015). Analyzing labeled cyberbullying incidents on the instagram social network. In *International conference on social informatics*, pages 49–66. Springer.

Hsueh, P.-Y., Melville, P., and Sindhwani, V. (2009). Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 27–35, Boulder, Colorado. Association for Computational Linguistics.

Hu, R. R., Dorris, W., Vishwamitra, N., Luo, F., and Costello, M. (2020). *On the Impact of Word Representation in Hate Speech and Offensive Language Detection and Explanation*, page 171–173. Association for Computing Machinery, New York, NY, USA.

Huang, K., Hussain, A., Wang, Q.-F., and Zhang, R. (2019). *Deep learning: fundamentals, theory and applications*, volume 2. Springer.

Isaksen, V. and Gambäck, B. (2020). Using transfer-based language models to detect hateful and offensive language online. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 16–27, Online. Association for Computational Linguistics.

Iyyer, M., Manjunatha, V., Boyd-Graber, J., and Daumé III, H. (2015). Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China. Association for Computational Linguistics.

Jacovi, A. and Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Jain, S. and Wallace, B. C. (2019). Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Jensen, R. and Cornelis, C. (2011). Fuzzy-rough nearest neighbour classification and prediction. *Theoretical Computer Science*, 412(42):5871–5884.

Jha, A. and Mamidi, R. (2017). When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16.

Jin, M., Preotiuc-Pietro, D., Doğruöz, A. S., and Aletras, N. (2022). Automatic identification and classification of bragging in social media. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3945–3959, Dublin, Ireland. Association for Computational Linguistics.

Jin, X., Wei, Z., Du, J., Xue, X., and Ren, X. (2020). Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *Proceedings of the 2018 International Conference on Learning Representations, ICLR*.

Jing, B., Lu, C., Wang, D., Zhuang, F., and Niu, C. (2018). Cross-domain labeled LDA for cross-domain text classification. In *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018*, pages 187–196. IEEE Computer Society.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Jurgens, D., Hemphill, L., and Chandrasekharan, E. (2019). A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

Justo, R., Corcoran, T., Lukin, S. M., Walker, M., and Torres, M. I. (2014). Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*, 69:124–133.

Justus, D., Brennan, J., Bonner, S., and McGough, A. S. (2018). Predicting the computational cost of deep learning models. In *2018 IEEE international conference on big data (Big Data)*, pages 3873–3882. IEEE.

Kaminska, O., Cornelis, C., and Hoste, V. (2021a). Fuzzy-rough nearest neighbour approaches for emotion detection in tweets. In *International Joint Conference on Rough Sets*, pages 231–246. Springer.

Kaminska, O., Cornelis, C., and Hoste, V. (2021b). Nearest neighbour approaches for emotion detection in tweets. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 203–212, Online. Association for Computational Linguistics.

Kaminska, O., Cornelis, C., and Hoste, V. (2022). LT3 at SemEval-2022 task 6: Fuzzy-rough nearest neighbor classification for sarcasm detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 987–992, Seattle, United States. Association for Computational Linguistics.

Kane, G. C., Alavi, M., Labianca, G., and Borgatti, S. P. (2014). What's different about social media networks? a framework and research agenda. *MIS Q.*, 38(1):275–304.

Kang, G., Jiang, L., Yang, Y., and Hauptmann, A. G. (2019). Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4893–4902.

Kantorovich, L. V. (1942). On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201.

Kantorovich, L. V. (2006). On the translocation of masses. *Journal of Mathematical Sciences*, 133(4):1381–1382.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68.

Karan, M. and Šnajder, J. (2018). Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Kassner, N. and Schütze, H. (2020). BERT-kNN: Adding a kNN search component to pretrained language models for better QA. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3424–3430, Online. Association for Computational Linguistics.

Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., Coombs, K., Havaldar, S., Portillo-Wightman, G., Gonzalez, E., Hoover, J., Azatian, A., Hussain, A., Lara, A., Cardenas, G., Omary, A., Park, C., Wang, X., Wijaya, C., Zhang, Y., Meyerowitz, B., and Dehghani, M. (2022). Introducing the gab hate corpus: Defining and applying hate-based rhetoric to social media posts at scale. *Lang. Resour. Eval.*, 56(1):79–108.

Kennedy, B., Jin, X., Mostafazadeh Davani, A., Dehghani, M., and Ren, X. (2020). Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Kerdoncuff, T., Emonet, R., and Sebban, M. (2021). Metric learning in optimal transport for domain adaptation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 2162–2168.

Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., and Lewis, M. (2020). Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., and Testuggine, D. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International conference on machine learning*, pages 2668–2677. PMLR.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Kindermans, P.-J., Schütt, K., Müller, K.-R., and Dähne, S. (2016). Investigating the influence of noise and distractors on the interpretation of neural networks. *ArXiv*, abs/1611.07270.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In Bengio, Y. and LeCun, Y., editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. *Advances in neural information processing systems*, 28.

Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59.

Kottasová, I. (2017). Europe says Twitter is failing to remove hate speech. https://money.cnn.com/2017/06/01/technology/twitter-facebook-hate-speech-europe/index.html. Online; accessed July 2022.

Koufakou, A., Pamungkas, E. W., Basile, V., and Patti, V. (2020). HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018). Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kwok, I. and Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, page 1621–1622. AAAI Press.

Lau, J. H., Baldwin, T., and Cohn, T. (2017). Topically driven neural language model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 355–365, Vancouver, Canada. Association for Computational Linguistics.

Le, N. D.-H., Tran, T.-S., and Tran, M.-T. (2012). Exploring neighborhood influence in text classification. In *2012 Fourth International Conference on Knowledge and Systems Engineering*, pages 79–85.

Lee, S., Lee, D., Jang, S., and Yu, H. (2022). Toward interpretable semantic textual similarity via optimal transport-based contrastive sentence learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5969–5979, Dublin, Ireland. Association for Computational Linguistics.

Lee, Y., Yoon, S., and Jung, K. (2018). Comparative studies of detecting abusive language on twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 101–106, Brussels, Belgium. Association for Computational Linguistics.

Lei, T., Barzilay, R., and Jaakkola, T. (2016). Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Lenz, O. U., Peralta, D., and Cornelis, C. (2019). Scalable approximate frnn-owa classification. *IEEE Transactions on Fuzzy Systems*, 28(5):929–938.

Lewis, D. D. and Gale, W. A. (1994). A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer.

Li, C., Duan, Y., Wang, H., Zhang, Z., Sun, A., and ma, Z. (2017a). Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM Transactions on Information Systems*, 36:1–30.

Li, C., Wang, H., Zhang, Z., Sun, A., and Ma, Z. (2016a). Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 165–174, New York, NY, USA. Association for Computing Machinery.

Li, J., Li, C., Wang, G., Fu, H., Lin, Y., Chen, L., Zhang, Y., Tao, C., Zhang, R., Wang, W., Shen, D., Yang, Q., and Carin, L. (2020a). Improving text generation with student-forcing optimal transport. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9144–9156, Online. Association for Computational Linguistics.

Li, J., Monroe, W., and Jurafsky, D. (2016b). Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Li, L., Jin, X., and Long, M. (2012). Topic correlation analysis for cross-domain text classification. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, page 998–1004. AAAI Press.

Li, M., Zhai, Y.-M., Luo, Y.-W., Ge, P.-F., and Ren, C.-X. (2020b). Enhanced transport distance for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13936–13944.

Li, Z., Wei, Y., Zhang, Y., and Yang, Q. (2018). Hierarchical attention transfer network for cross-domain sentiment classification. In *AAAI Conference on Artificial Intelligence*.

Li, Z., Zhang, Y., Wei, Y., Wu, Y., and Yang, Q. (2017b). End-to-end adversarial memory network for cross-domain sentiment classification. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2017)*.

Liero, M., Mielke, A., and Savaré, G. (2015). Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211:969–1117.

Liu, F. and Avci, B. (2019). Incorporating priors with feature attribution on text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6274–6283, Florence, Italy. Association for Computational Linguistics.

Liu, P., Li, W., and Zou, L. (2019a). NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

Lopez Long, H., O'Neil, A., and Kübler, S. (2021). On the interaction between annotation quality and classifier performance in abusive language detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 868–875, Held Online. INCOMA Ltd.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lou, Y., Caruana, R., and Gehrke, J. (2012). Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 150–158, New York, NY, USA. Association for Computing Machinery.

Ludwig, F., Dolos, K., Zesch, T., and Hobley, E. (2022). Improving generalization of hate speech detection systems to novel target groups via domain adaptation. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 29–39, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., and Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Magu, R., Joshi, K., and Luo, J. (2017). Detecting the hate code on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 608–611.

Maharana, A. and Bansal, M. (2020). Adversarial augmentation policy search for domain and cross-lingual generalization in reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3723–3738, Online. Association for Computational Linguistics.

Mahmoud, H. (2008). *Pólya urn models*. Chapman and Hall/CRC.

Malmasi, S. and Zampieri, M. (2018). Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30:187 – 202.

Mandl, T., Modha, S., Kumar M, A., and Chakravarthi, B. R. (2020). Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 29–32, New York, NY, USA. Association for Computing Machinery.

Manning, C. D. (2008). *Introduction to information retrieval*. Syngress Publishing,.

Markov, I. and Daelemans, W. (2021). Improving cross-domain hate speech detection by reducing the false positive rate. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 17–22, Online. Association for Computational Linguistics.

Matsuda, M. J. (2018). *Words that wound: Critical race theory, assaultive speech, and the first amendment*. Routledge.

McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Meftah, S., Semmar, N., Tamaazousti, Y., Essafi, H., and Sadat, F. (2021). Neural supervised domain adaptation by augmenting pre-trained models with random units. *ArXiv*, abs/2106.04935.

Meftah, S., Tamaazousti, Y., Semmar, N., Essafi, H., and Sadat, F. (2019). Joint learning of pre-trained and random units for domain adaptation in part-of-speech tagging. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4107–4112, Minneapolis, Minnesota. Association for Computational Linguistics.

Mémoli, F. (2011). Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487.

Menini, S., Aprosio, A. P., and Tonelli, S. (2021). Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection. *arXiv preprint arXiv:2103.14916*.

Miao, Y., Yu, L., and Blunsom, P. (2016). Neural variational inference for text processing. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1727–1736. JMLR.org.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Mikolov, T., Corrado, G., Chen, K., and Dean, J. (2013b). Efficient estimation of word representations in vector space. In *International Conference on Learning Representations@articlenigam2000text, title=Text classification from labeled and unlabeled documents using EM, author=Nigam, Kamal and McCallum, Andrew Kachites and Thrun, Sebastian and Mitchell, Tom, journal=Machine learning, volume=39, number=2, pages=103–134, year=2000, publisher=Springer* , pages 1–12.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mishra, P., Del Tredici, M., Yannakoudakis, H., and Shutova, E. (2018a). Author profiling for abuse detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1088–1098, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mishra, P., Yannakoudakis, H., and Shutova, E. (2018b). Neural character-based composition models for abuse detection. In *Proceedings of the 2nd Workshop on Abusive*

Language Online (ALW2), pages 1–10, Brussels, Belgium. Association for Computational Linguistics.

Mishra, P., Yannakoudakis, H., and Shutova, E. (2019). Tackling online abuse: A survey of automated abuse detection methods. *CoRR*, abs/1908.06024.

Mitsuhara, M., Fukui, H., Sakashita, Y., Ogata, T., Hirakawa, T., Yamasita, T., and Fujiyoshi, H. (2021). Embedding Human Knowledge into Deep Neural Network via Attention Map. In *International Conference on Computer Vision Theory and Applications*.

Moh, M., Moh, T.-S., and Khieu, B. (2020). No "love" lost: Defending hate speech detection models against adversaries. In *2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pages 1–6.

Mohammad, S., Shutova, E., and Turney, P. (2016). Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33.

Mollas, I., Chrysopoulou, Z., Karlos, S., and Tsoumakas, G. (2022). ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*.

Monge, G. (1781). *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale.

Moody, C. (2016). Mixing dirichlet topic models and word embeddings to make lda2vec, arxiv:1605.02019.

Mou, G., Ye, P., and Lee, K. (2020). Swe2: Subword enriched and significant word emphasized framework for hate speech detection. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management*, CIKM '20, page 1145–1154, New York, NY, USA. Association for Computing Machinery.

Mozafari, M., Farahbakhsh, R., and Crespi, N. (2019). A BERT-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer.

Mozafari, M., Farahbakhsh, R., and Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS one*, 15(8):e0237861.

Mozafari, M., Farahbakhsh, R., and Crespi, N. (2022). Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, 10:14880–14896.

Murdoch, J., Liu, P. J., and Yu, B. (2018). Beyond word importance: Contextual decomposition to extract interactions from lstms. In *Proceedings of the 2018 International Conference on Learning Representations, ICLR*.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*.

Narasimhan, H., Pan, W., Kar, P., Protopapas, P., and Ramaswamy, H. G. (2016). Optimizing the multiclass f-measure via biconcave programming. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 1101–1106. IEEE.

Nejadgholi, I., Fraser, K., and Kiritchenko, S. (2022). Improving generalizability in implicitly abusive language detection with concept activation vectors. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5517–5529, Dublin, Ireland. Association for Computational Linguistics.

Nejadgholi, I. and Kiritchenko, S. (2020). On cross-dataset generalization in automatic detection of online abuse. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 173–183, Online. Association for Computational Linguistics.

Nguyen, D. Q., Billingsley, R., Du, L., and Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2):103–134.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Nockleby, J. T. (2000). *Encyclopedia of the American Constitution*. Macmillan, 2nd edition.

Nozza, D., Volpetti, C., and Fersini, E. (2019). Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, page 149–155, New York, NY, USA. Association for Computing Machinery.

Oboler, A. and Connelly, K. (2014). Hate speech: A quality of service challenge. In *2014 IEEE Conference on e-Learning, e-Management and e-Services (IC3e)*, pages 117–121.

Olvera, M., Vincent, E., and Gasso, G. (2021). Improving sound event detection with auxiliary foreground-background classification and domain adaptation. In *DCASE 2021-6th Workshop on Detection and Classification of Acoustic Scenes and Events*.

Ombui, E., Karani, M., and Muchemi, L. (2019). Annotation framework for hate speech identification in tweets: Case study of tweets during kenyan elections. *2019 IST-Africa Week Conference (IST-Africa)*, pages 1–9.

Ozler, K. B., Kenski, K., Rains, S., Shmargad, Y., Coe, K., and Bethard, S. (2020). Fine-tuning for multi-domain and multi-label uncivil language detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 28–33, Online. Association for Computational Linguistics.

Pal, D., Chaudhari, K., and Sharma, H. (2022). Combating high variance in data-scarce implicit hate speech classification.

Pamungkas, E. W., Basile, V., and Patti, V. (2020). Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing Management*, 57(6):102360.

Pamungkas, E. W., Basile, V., and Patti, V. (2021). Towards multidomain and multilingual abusive language detection: a survey. *Personal and Ubiquitous Computing*, pages 1–27.

Pamungkas, E. W. and Patti, V. (2019). Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.

Pan, S. J., Ni, X., Sun, J.-T., Yang, Q., and Chen, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 751–760, New York, NY, USA. Association for Computing Machinery.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New media & society*, 6(2):259–283.

Park, J. H., Shin, J., and Fung, P. (2018). Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Paul, M. and Girju, R. (2009). Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1408–1417, Singapore. Association for Computational Linguistics.

Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., and Androutsopoulos, I. (2020). Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

Peinelt, N., Nguyen, D., and Liakata, M. (2020). tBERT: Topic models and BERT joining forces for semantic similarity detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055, Online. Association for Computational Linguistics.

Pelkowitz, L. (1990). A continuous relaxation labeling algorithm for markov random fields. *IEEE Transactions on Systems, Man, and Cybernetics*, 20(3):709–715.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Perea, M., Duñabeitia, J. A., and Carreiras, M. (2008). R34d1ng w0rd5 w1th numb3r5. *Journal of Experimental Psychology: Human Perception and Performance*, 34(1):237.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Phan, X.-H., Nguyen, C.-T., Le, D.-T., Nguyen, L.-M., Horiguchi, S., and Ha, Q.-T. (2011). A hidden topic-based framework toward building applications with short web documents. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):961–976.

Plank, B. (2011). *Domain adaptation for parsing*. PhD thesis, University of Groningen.

Poletto, F., Basile, V., Bosco, C., Patti, V., and Stranisci, M. (2019). Annotating hate speech: Three schemes at comparison. In Bernardi, R., Navigli, R., and Semeraro, G., editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55:477–523.

Prabhakaran, V., Waseem, Z., Akiwowo, S., and Vidgen, B. (2020). Online abuse and human rights: WOAH satellite session at RightsCon 2020. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 1–6, Online. Association for Computational Linguistics.

Prasad, N., Saha, S., and Bhattacharyya, P. (2021). A multimodal classification of noisy hate speech using character level embedding and attention. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Prasetyo, V. R. and Samudra, A. H. (2022). Hate speech content detection system on twitter using k-nearest neighbor method. In *AIP Conference Proceedings*, volume 2470, page 050001. AIP Publishing LLC.

Pratiwi, N. I., Budi, I., and Jiwanggi, M. A. (2019). Hate speech identification using the hate codes for indonesian tweets. In *Proceedings of the 2019 2nd International Conference on Data Science and Information Technology*, DSIT 2019, page 128–133, New York, NY, USA. Association for Computing Machinery.

Qian, J., Bethke, A., Liu, Y., Belding, E., and Wang, W. Y. (2019). A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Qian, J., ElSherief, M., Belding, E., and Wang, W. Y. (2018). Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 118–123, New Orleans, Louisiana. Association for Computational Linguistics.

Qian, J., Wang, H., ElSherief, M., and Yan, X. (2021). Lifelong learning of hate speech classification on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2304–2314, Online. Association for Computational Linguistics.

Qiang, J., Qian, Z., Li, Y., Yuan, Y., and Wu, X. (2020). Short text topic modeling techniques, applications, and performance: A survey. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1.

Quan, X., Kit, C., Ge, Y., and Pan, S. J. (2015). Short and sparse text topic modeling via self-aggregation. In *IJCAI*.

Rajamanickam, S., Mishra, P., Yannakoudakis, H., and Shutova, E. (2020). Joint modelling of emotion and abusive language detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4270–4279, Online. Association for Computational Linguistics.

Ramponi, A. and Plank, B. (2020). Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ramponi, A. and Tonelli, S. (2022). Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3027–3040, Seattle, United States. Association for Computational Linguistics.

Rani, P., Suryawanshi, S., Goswami, K., Chakravarthi, B. R., Fransen, T., and McCrae, J. P. (2020). A comparative study of different state-of-the-art hate speech detection methods in Hindi-English code-mixed data. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 42–48, Marseille, France. European Language Resources Association (ELRA).

Rathpisey, H. and Adji, T. B. (2019). Handling imbalance issue in hate speech classification using sampling-based methods. In *2019 5th International Conference on Science in Information Technology (ICSITech)*, pages 193–198.

Razavi, A. H., Inkpen, D., Uritsky, S., and Matwin, S. (2010). Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer.

Razo, D. and Kübler, S. (2020). Investigating sampling bias in abusive language detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 70–78, Online. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Reyes, O., Altalhi, A. H., and Ventura, S. (2018). Statistical comparisons of active learning strategies over multiple datasets. *Knowledge-Based Systems*, 145:274–288.

Reynolds, K., Kontostathis, A., and Edwards, L. (2011). Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, volume 2, pages 241–244. IEEE.

Rezvani, N., beheshti, A., and Tabebordbar, A. (2020). Linking textual and contextual features for intelligent cyberbullying detection in social media. In *Proceedings of the 18th International Conference on Advances in Mobile Computing Multimedia*, MoMM '20, page 3–10, New York, NY, USA. Association for Computing Machinery.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Rieger, L., Singh, C., Murdoch, W., and Yu, B. (2020). Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8116–8126. PMLR.

Rietzler, A., Stabinger, S., Opitz, P., and Engl, S. (2020). Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4933–4941, Marseille, France. European Language Resources Association.

Ring, M. B. et al. (1994). Continual learning in reinforcement environments.

Rizoiu, M.-A., Wang, T., Ferraro, G., and Suominen, H. (2019). Transfer learning for hate speech detection in social media. *arXiv preprint arXiv:1906.03829*.

Rizos, G., Hemker, K., and Schuller, B. (2019). Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 991–1000, New York, NY, USA. Association for Computing Machinery.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American journal of political science*, 58(4):1064–1082.

Rocha, G. and Lopes Cardoso, H. (2019). A comparative analysis of unsupervised language adaptation methods. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 11–21, Hong Kong, China. Association for Computational Linguistics.

Ross, A. S., Hughes, M. C., and Doshi-Velez, F. (2017). Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2662–2670.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2016). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In Dipper, S., editor, *NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, Bochumer Linguistische Arbeitsberichte, pages 6–9, Germany. Ruhr-Universitat Bochum.

Rostami, M. (2021). Lifelong domain adaptation via consolidated internal distribution. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 11172–11183. Curran Associates, Inc.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121.

Ruder, S., Peters, M. E., Swayamdipta, S., and Wolf, T. (2019). Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Ruder, S. and Plank, B. (2018). Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1054, Melbourne, Australia. Association for Computational Linguistics.

Ryu, M. and Lee, K. (2020). Knowledge distillation for BERT unsupervised domain adaptation. *arXiv preprint arXiv:2010.11478*.

Saha, A. and Sindhwani, V. (2012). Learning evolving and emerging topics in social media: A dynamic nmf approach with temporal regularization. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, page 693–702, New York, NY, USA. Association for Computing Machinery.

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732.

Salimans, T., Zhang, H., Radford, A., and Metaxas, D. (2018). Improving GANs using optimal transport. In *International Conference on Learning Representations*.

Salles, T., Gonçalves, M., Rodrigues, V., and Rocha, L. (2018). Improving random forests by neighborhood projection for effective text classification. *Information Systems*, 77:1–21.

Salminen, J., Veronesi, F., Almerekhi, H., Jung, S.-G., and Jansen, B. J. (2018). Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 88–94.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Samory, M., Sen, I., Kohne, J., Flöck, F., and Wagner, C. (2021). " call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In *ICWSM*, pages 573–584.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., and Choi, Y. (2020). Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., and Smith, N. A. (2022). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Sarwar, S. M. and Murdock, V. (2022). Unsupervised domain adaptation for hate speech detection using a data augmentation approach. In *Proceedings of the 16th International Conference on Web and Social Media*.

Sarwar, S. M., Zlatkova, D., Hardalov, M., Dinkov, Y., Augenstein, I., and Nakov, P. (2022). A neighborhood framework for resource-lean content flagging. *Transactions of the Association for Computational Linguistics*, 10:484–502.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

Schwartz, O. (2016). In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation. https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation. Online; accessed July 2022.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Serrà, J., Leontiadis, I., Spathis, D., Stringhini, G., Blackburn, J., and Vakali, A. (2017). Class-based prediction errors to detect hate speech with out-of-vocabulary words. In *Proceedings of the First Workshop on Abusive Language Online*, pages 36–40, Vancouver, BC, Canada. Association for Computational Linguistics.

Serrano, S. and Smith, N. A. (2019). Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Shen, J., Qu, Y., Zhang, W., and Yu, Y. (2018a). Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Shen, J., Qu, Y., Zhang, W., and Yu, Y. (2018b). Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Sheoran, A., Joshi, A., and Bhattacharyya, P. (2019). Domain adaptation for sentiment analysis : A survey.

Shnarch, E., Alzate, C., Dankin, L., Gleize, M., Hou, Y., Choshen, L., Aharonov, R., and Slonim, N. (2018). Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.

Singh, S. P., Hug, A., Dieuleveut, A., and Jaggi, M. (2020). Context mover's distance & barycenters: Optimal transport of contexts for building representations. In *International Conference on Artificial Intelligence and Statistics*, pages 3437–3449. PMLR.

Sinkhorn, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879.

Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348.

Sodhani, S., Chandar, S., and Bengio, Y. (2020). Toward training recurrent neural networks for lifelong learning. *Neural computation*, 32(1):1–35.

Søgaard, A. (2010). Simple semi-supervised training of part-of-speech taggers. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 205–208, Uppsala, Sweden. Association for Computational Linguistics.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Solomon, J., de Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. (2015). Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4).

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Srivastava, A. and Sutton, C. (2017). Autoencoding variational inference for topic models. In *International Conference on Learning Representations*.

Stan Z., L. (2001). Markov random field modeling in image analysis.

Stanković, R., Mitrović, J., Jokić, D., and Krstev, C. (2020). Multi-word expressions for abusive speech detection in Serbian. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 74–84, online. Association for Computational Linguistics.

Steinskog, A., Therkelsen, J., and Gambäck, B. (2017). Twitter topic modeling by tweet aggregation. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 77–86, Gothenburg, Sweden. Association for Computational Linguistics.

Sui, J. (2015). *Understanding and fighting bullying with machine learning*. PhD thesis, The University of Wisconsin-Madison.

Sun, F., Ho, C., and Lee, H. (2020). LAMOL: language modeling for lifelong language learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Swamy, S. D., Jamatia, A., and Gambäck, B. (2019). Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China. Association for Computational Linguistics.

Swanson, K., Yu, L., and Lei, T. (2020). Rationalizing text matching: Learning sparse alignments via optimal transport. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5609–5626, Online. Association for Computational Linguistics.

Thrun, S. and Pratt, L. (1998). Learning to learn: Introduction and overview. In *Learning to learn*, pages 3–17. Springer.

Titouan, V., Courty, N., Tavenard, R., Laetitia, C., and Flamary, R. (2019). Optimal transport for structured data with application on graphs. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6275–6284. PMLR.

Toraman, C., Şahinuç, F., and Yilmaz, E. (2022). Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Trager, G. L. and Bloch, B. (1941). The syllabic phonemes of english. *Language*, pages 223–246.

Triantafillou, E., Zhu, T., Dumoulin, V., Lamblin, P., Evci, U., Xu, K., Goroshin, R., Gelada, C., Swersky, K., Manzagol, P.-A., and Larochelle, H. (2020). Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations*.

Trstenjak, B., Mikac, S., and Donko, D. (2014). Knn with tf-idf based framework for text categorization. *Procedia Engineering*, 69:1356–1364. 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013.

Tufekci, Z. (2019). YouTube's Recommendation Algorithm Has a Dark Side. https://www.scientificamerican.com/article/youtubes-recommendation-algorithm-has-a-dark-side/. Online; accessed July 2022.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971.

Ullmann, S. and Tomalin, M. (2020). Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology*, 22(1):69–80.

Vaidya, A., Mai, F., and Ning, Y. (2020). Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 683–693.

van Aken, B., Risch, J., Krestel, R., and Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. In *2nd Workshop on Abusive Language Online (ALW2)*.

Van Deursen, A. J. and Helsper, E. J. (2018). Collateral benefits of internet use: Explaining the diverse outcomes of engaging with the internet. *new media & society*, 20(7):2333–2351.

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., and Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PloS one*, 13(10):e0203794.

van Rosendaal, J., Caselli, T., and Nissim, M. (2020). Lower bias, higher density abusive language datasets: A recipe. In *Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language*, pages 14–19, Marseille, France. European Language Resources Association (ELRA).

Vashishth, S., Upadhyay, S., Tomar, G. S., and Faruqui, M. (2019). Attention interpretability across nlp tasks. *ArXiv*, abs/1909.11218.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vidgen, B., Burden, E., and Margetts, H. (2021a). *Understanding online hate: VSP regulation and the broader context.* Ofcom.

Vidgen, B. and Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Vidgen, B., Hale, S., Guest, E., Margetts, H., Broniatowski, D., Waseem, Z., Botelho, A., Hall, M., and Tromble, R. (2020). Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., and Margetts, H. (2019). Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Vidgen, B., Nguyen, D., Margetts, H., Rossini, P., and Tromble, R. (2021b). Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Vidgen, B., Thrush, T., Waseem, Z., and Kiela, D. (2021c). Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682. Association for Computational Linguistics.

Villani, C. (2009). Optimal transport: Old and new. volume 338. Springer.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103.

Vluymans, S., Mac Parthaláin, N., Cornelis, C., and Saeys, Y. (2019). Weight selection strategies for ordered weighted average based fuzzy rough sets. *Information Sciences*, 501:155–171.

Wang, J., Neskovic, P., and Cooper, L. N. (2006). Neighborhood size selection in the k-nearest-neighbor rule using statistical confidence. *Pattern Recognition*, 39(3):417–423.

Wang, J., Neskovic, P., and Cooper, L. N. (2007). Improving nearest neighbor rule with a simple adaptive distance measure. *Pattern Recogn. Lett.*, 28(2):207–213.

Wang, J.-G., Neskovic, and Cooper (2005). An adaptive nearest neighbor algorithm for classification. In *2005 International Conference on Machine Learning and Cybernetics*, volume 5, pages 3069–3074 Vol. 5.

Wang, K., Lu, D., Han, C., Long, S., and Poon, J. (2020). Detect all abuse! toward universal abusive language detection models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6366–6376, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wang, Y., Li, X., Zhou, X., and Ouyang, J. (2021). Extracting topics with simultaneous word co-occurrence and semantic correlation graphs: Neural topic modeling for short texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 18–27, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wang, Z. and Liu, Z. (2010). Graph-based knn text classification. In *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, volume 5, pages 2363–2366. IEEE.

Warner, W. and Hirschberg, J. (2012a). Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

Warner, W. and Hirschberg, J. (2012b). Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, page 19–26, USA. Association for Computational Linguistics.

Waseem, Z. (2016). Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Waseem, Z., Lulz, S., Bingel, J., and Augenstein, I. (2021). Disembodied machine learning: On the illusion of objectivity in nlp. *arXiv preprint arXiv:2101.11974*.

Waseem, Z., Thorne, J., and Bingel, J. (2018). Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In Golbeck, J., editor, *Golbeck J. (eds) Online Harassment. Human–Computer Interaction Series*, pages 29–55, Cham. Springer International Publishing.

Weinberger, E., Janizek, J. D., and Lee, S.-I. (2020). Learning deep attribution priors based on prior knowledge. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Wich, M., Al Kuwatly, H., and Groh, G. (2020). Investigating annotator bias with a graph-based approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 191–199, Online. Association for Computational Linguistics.

Wiegand, M., Ruppenhofer, J., and Eder, E. (2021). Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.

Wiegand, M., Ruppenhofer, J., and Kleinbauer, T. (2019). Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Wiegand, M., Ruppenhofer, J., Schmidt, A., and Greenberg, C. (2018a). Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.

Wiegand, M., Siegel, M., and Ruppenhofer, J. (2018b). Overview of the germeval 2018 shared task on the identification of offensive language.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Witten, I. H. and Frank, E. (2002). Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J. R., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G. S., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Würschinger, Q. (2021). Social networks of lexical innovation. investigating the social dynamics of diffusion of neologisms on twitter. *Frontiers in Artificial Intelligence*, 4.

Xu, H., Liu, B., Shu, L., and Yu, P. (2019). BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Xu, J., Zhou, H., Gan, C., Zheng, Z., and Li, L. (2021). Vocabulary learning via optimal transport for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7361–7373, Online. Association for Computational Linguistics.

Xu, R., Liu, P., Wang, L., Chen, C., and Wang, J. (2020). Reliable weighted optimal transport for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, Z. and Zhu, S. (2010). Filtering offensive language in online communities using grammatical relations. In *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, pages 1–10.

Xue, G.-R., Dai, W., Yang, Q., and Yu, Y. (2008). Topic-bridged plsa for cross-domain text classification. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 627–634, New York, NY, USA. Association for Computing Machinery.

Xue, Q., Zhang, W., and Zha, H. (2020). Improving domain-adapted sentiment classification by deep adversarial mutual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9362–9369.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yao, H., Chen, Y., Ye, Q., Jin, X., and Ren, X. (2021). Refining language models with compositional explanations. *Advances in Neural Information Processing Systems*, 34.

Ye, Y., Huang, Z., Pan, T., Li, J., and Shen, H. T. (2021). Reducing bias to source samples for unsupervised domain adaptation. *Neural Networks*, 141:61–71.

Yi, F., Jiang, B., and Wu, J. (2020). Topic modeling for short texts via word embedding and document correlation. *IEEE Access*, 8:30692–30705.

Yin, J. and Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 233–242, New York, NY, USA. Association for Computing Machinery.

Yin, W. and Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7.

Yu, N., Pan, S., Yang, C.-c., Tsai, J.-Y., et al. (2020). Exploring the role of media sources on covid-19–related discrimination experiences and concerns among asian people in the united states: cross-sectional survey study. *Journal of Medical Internet Research*, 22(11):e21684.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019a). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019b). SemEval-2019 task 6: Identifying and categorizing offensive language in social media

(OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Zeng, J., Li, J., Song, Y., Gao, C., Lyu, M. R., and King, I. (2018). Topic memory networks for short text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3120–3131, Brussels, Belgium. Association for Computational Linguistics.

Zhai, C., Velivelli, A., and Yu, B. (2004). A cross-collection mixture model for comparative text mining. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 743–748, New York, NY, USA. Association for Computing Machinery.

Zhang, H., Xiao, X., and Hasegawa, O. (2013). A load-balancing self-organizing incremental neural network. *IEEE Transactions on Neural Networks and Learning Systems*, 25(6):1096–1105.

Zhang, Z. and Luo, L. (2019). Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.

Zhang, Z., Robinson, D., and Tepper, J. A. (2018). Detecting hate speech on twitter using a convolution-gru based deep neural network. In Gangemi, A., Navigli, R., Vidal, M., Hitzler, P., Troncy, R., Hollink, L., Tordai, A., and Alam, M., editors, *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 745–760. Springer.

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., and Buntine, W. L. (2021). Topic modelling meets deep neural networks: A survey. *CoRR*, abs/2103.00498.

Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., and Eger, S. (2019). MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. In *ECIR*.

Zhong, H., Li, H., Squicciarini, A. C., Rajtmajer, S. M., Griffin, C., Miller, D. J., and Caragea, C. (2016). Content-driven detection of cyberbullying on the instagram social network. In *IJCAI*, volume 16, pages 3952–3958.

Zhou, G., Zhou, Y., Guo, X., Tu, X., and He, T. (2015). Cross-domain sentiment classification via topical correspondence transfer. *Neurocomputing*, 159:298–305.

Zhou, X., Sap, M., Swayamdipta, S., Choi, Y., and Smith, N. (2021). Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.

Zhou, Z.-H. and Li, M. (2005). Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11):1529–1541.

Zhu, J., Lee, R. K.-W., and Chong, W. H. (2022). Multimodal zero-shot hateful meme detection. In *14th ACM Web Science Conference 2022*, WebSci '22, page 382–389, New York, NY, USA. Association for Computing Machinery.

Zhu, Y., Kiros, R., Zemel, R. S., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

Zhuang, F., Luo, P., Yin, P., He, Q., and Shi, Z. (2013). Concept learning for cross-domain text classification: A general probabilistic framework. In *IJCAI International Joint Conference on Artificial Intelligence*, pages 1960–1966.

Ziser, Y. and Reichart, R. (2017). Neural structural correspondence learning for domain adaptation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 400–410, Vancouver, Canada. Association for Computational Linguistics.

Ziser, Y. and Reichart, R. (2018). Pivot based language modeling for improved neural domain adaptation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1241–1251, New Orleans, Louisiana. Association for Computational Linguistics.

Zunino, A., Bargal, S. A., Volpi, R., Sameki, M., Zhang, J., Sclaroff, S., Murino, V., and Saenko, K. (2021). Explainable deep classification models for domain generalization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3227–3236.

Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., and Xiong, H. (2016a). Topic modeling of short texts: A pseudo-document view. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 2105–2114, New York, NY, USA. Association for Computing Machinery.

Zuo, Y., Wu, J., Zhang, H., Wang, D., and Xu, K. (2018). Complementary aspect-based opinion mining. *IEEE Transactions on Knowledge and Data Engineering*, 30(2):249–262.

Zuo, Y., Zhao, J., and Xu, K. (2016b). Word network topic model: A simple but general solution for short and imbalanced texts. *Knowl. Inf. Syst.*, 48(2):379–398.