What do BERT word embeddings learn about the French language?

Ekaterina Goliakova, David Langlois

What do BERT word embeddings learn about the French language?

Ekaterina Goliakova
Université de Lorraine
ekaterina.goliakova6@etu.univ-lorraine.fr
edgolyakova@gmail.com

David Langlois
Université de Lorraine,
LORIA (Laboratoire Lorrain en Recherche en Informatique et ses Applications)
david.langlois@loria.fr

Abstract

Pre-trained word embeddings (for example, BERT-like) have been successfully used in a variety of downstream tasks. However, do all embeddings, obtained from the models of the same architecture, encode information in the same way? Does the size of the model correlate to the quality of the information encoding? In this paper, we will attempt to dissect the dimensions of several BERT-like models that were trained on the French language to find where grammatical information (gender, plurality, part of speech) and semantic features might be encoded. In addition to this, we propose a framework for comparing the quality of encoding in different models.

Keywords: interpretability, word embeddings, intrinsic evaluation, BERT.

1 Introduction

With over 95,000 citations (and counting) since its publication in 2019, BERT (Devlin et al., 2019) can be considered one of the most prominent Large Language Models (LLMs) architectures in the current state of the art, finding applications in fields ranging from text to image generation (Rombach et al., 2022) to protein structure prediction (Jumper et al., 2021), still showing competitive results (Samuel et al., 2023). We can attribute one of the reasons for such a successful and wide-range usage of BERT-like models to the *word embeddings* (multidimensional word representations) they produce.

However, with ever-growing model sizes, the interpretability of dimensions of the learned representations is still a complex task. While the number of parameters is constantly growing, the performance of the models is not improving as rapidly, and we are facing diminishing returns with the increased model and training data scale (Kaplan et al., 2020, van Schijndel et al., 2019). Additionally, the size of available data does not grow at the same rate as the model hyperparameters and all possible training data for Language Models might be exhausted between 2030 and 2040 (Villalobos et al., 2022). Therefore, shifting the research focus from constantly increasing the training datasets towards understanding more about existing models, their parameters (and how they can be improved) might become necessary in the nearest feature.

In this paper, we propose an intrinsic metric evaluating the quality of information encoding (InfEnc) and suggest a framework that allows the identification of the best dimension candidates of word embeddings that potentially encode target information (in our experiments: grammatical number and gender for French nouns and adjectives, part of speech (POS) for French nouns, adjectives, and verbs, and semantic information for French nouns).

2 Related works

The explainability and interpretability of LLMs have become a growing interest for researchers. One of the approaches to the problem is to learn an explainable distributional word embedding model linking each feature to a word (for example, Snidaro et al., 2019). In this work, the representations were learned from a co-occurrence matrix, which allowed for high interpretability of the embeddings. However, the representations were underperforming on similarity tasks such as WS-353 in comparison to other distributional representation models such as GloVe (Lee et al., 2020) and BERT-like embeddings (Chronis and Erk, 2020).

Another approach to interpretability is to probe pre-trained models (Alain and Bengio, 2016; Belinkov, 2022; Tjoa and Guan, 2020). Torroba Hennigen et al., 2020 focus on intrinsic probing that aims not only to identify if a linguistic feature is encoded in representations but additionally how the information is encoded. According to this defi-

nition, our work remains in the scope of intrinsic evaluation. Torroba Hennigen et al., 2020 propose a method to efficiently detect the most relevant subset overall features whereas we pre-select an initial subset and compute a score for each pre-selected feature.

Miaschi et al., 2020 explore the usefulness of word embeddings to predict features at the sentence level (length, depth of the syntactic parsing tree, etc.). The sentence embeddings are obtained by merging word embeddings (by sum, maximum, minimum, and average operators). The authors do not explore which features encode linguistic information. For Ravichander et al., 2021, the objective is also not to dissect word embeddings in order to find features encoding linguistic information, but to study the usefulness of linguistic information for a classification task at sentence level. Contrary to these works, our objective is not to explore linguistic features at the sentence level, but at the word level.

3 Models

In our work, we compared word embeddings of 10 different BERT-like models working with the French language, details about which can be found in Table 1.

mBERT

Being the multilingual version of BERT (Devlin et al., 2019), mBERT can be considered as the baseline for multilingual word embeddings. mBERT is a bi-directional model trained with the Masked Language Modelling (MLM) and the Next Sentence Prediction (NSP) objectives. The data used for the training of mBERT was sourced from the dump multilingual Wikipedias for 104 different languages, including French[1].

DistilBERT

DistilBERT (Sanh et al., 2019) was created in response to the growing sizes of models. It was shown that despite being 40% smaller than the original BERT, DistilBERT has comparable performance on downstream tasks (Sanh et al., 2019, Jia et al., 2021, Abdaoui et al., 2020).

XLM-R

XLM-RoBERTa (or XLM-R) was created to address the underperformance of the traditional BERT architecture on low-resource languages (Conneau et al., 2020). The model was trained on 2.5TB of CommonCrawl corpus (Wenzek et al., 2020), out of 56.8GB (2.3%) were the French data. The model was trained with the MLM objective, using only monolingual data in each of the 100 languages.

CamemBERT

Unlike the models listed above, CamemBERT (Martin et al., 2020) is a monolingual BERT-like model, the architecture of which is similar to that of RoBERTa (Liu et al., 2019). The model was trained on the OSCAR dataset (Ortiz Suárez et al., 2019) (monolingual corpora retrieved from CommonCrawl dataset snapshots), comprising 138GB of raw French data.

FlauBERT

Similarly to CamemBERT, FlauBERT (Le et al., 2020) is a monolingual French BERT-like model. It was trained with the MLM objective on 71GB of diverse French data, out of which 43GB (60%) were obtained from CommonCrawl data, 9GB (13%) from NewsCrawl (Li et al., 2019) corpus, around 7GB (9%) from Wikipedia and Wikisource dumps, with the remaining part of the dataset being constituted from various data sources.

3.1 Extracting word embeddings

Word embeddings of BERT-like models are contextual: if a certain word has multiple senses, its representation might change depending on the sense (Miaschi et al., 2020, Ethayarajh, 2019). In our experiments, we encode every word from the obtained vocabulary using the models and retrieve weights of the last layer to obtain embeddings, assuming that following this procedure we can get the most frequent representation of the word by a given model[2]. Likewise, to avoid ambiguity, we only consider words that are uniquely a noun, an adjective, or a verb.

For similar reasons, if a word is tokenized into multiple pieces, we consider that the tokens might contain ambiguous meanings (compare: *simplement* ('simply') → [simple, ment] and *mentons* ('[we] lie') → [ment, ons][3]). Therefore, to avoid this sort of uncertainty, we only consider embeddings of words that are directly in the learned vocabulary of models (tokenized as one token). This

[1] https://github.com/google-research/bert/blob/master/multilingual.md

[2] CLS and end of the string tokens were discarded and not included in the obtained representations.

[3] For *simplement*, *ment* is the suffix for adverbs whereas for *mentons*, *ment* is the root of the verb *mentir*.

Model	Embedding size	# of parameters	# of layers	Vocabulary size	Tokenization	Training objective
FlauBERT$_{small}$	512	54M	6	68K	Byte-Pair Encoding	MLM
DistilBERT	768	66M	6	119K	WordPiece	MLM
CamemBERT$_{base}$	768	110M	12	32K	SentencePiece	multilingual MLM
mBERT$_{base}$ (cased and uncased)	768	110M	12	105K	WordPiece	MLM, NSP
FlauBERT$_{base}$ (uncased)	768	137M	12	67K	Byte-Pair Encoding	MLM
FlauBERT$_{base}$ (cased)	768	138M	12	68K	Byte-Pair Encoding	MLM
XLM-R$_{base}$	768	270M	12	250K	SentencePiece	multilingual MLM
FlauBERT$_{large}$	1024	373M	24	68K	Byte-Pair Encoding	MLM
XLM-R$_{large}$	1024	550M	24	250K	SentencePiece	multilingual MLM

Table 1: Model sizes for different French BERT-like models (sorted by the number of parameters).

approach also allows us to focus only on the most frequent words that the models have learned during their training.

3.2 Vocabulary and grammatical information

Morphalou (Romary et al., 2004) was used to obtain the initial vocabulary of French words, alongside the grammatical information (gender, number, POS). Morphalou is a lexical resource, that follows the Lexical Markup Framework (LMF) format and contains over 99,000 nouns and their inflected forms, over 14,000 verbs, and over 36,000 adjectives. The corpus contains other POS as well, however, their investigation lies outside of the scope of our project.

For each noun, adjective, or verb entry of Morphalou, we attempt to get a word embedding[4] by a given model. If the word is in the model vocabulary (tokenized as one token), we store this word and the corresponding grammatical information. For each word in the vocabulary of the models we stored its POS, grammatical number and gender (if applicable), and word embedding, afterward applying min-max normalization to the obtained word embeddings for each model.

The sizes of obtained datasets can be found in Table 2, and it can be noticed that the multilingual models have significantly smaller French vocabulary sizes, which can be expected due to French data being only a part of the final vocabulary of the model.

3.3 Semantic information

Additionally, FrSemCor's Sequoia corpus (Barque et al., 2020) was used in order to retrieve semantic information about the obtained words. The corpus contains 12,917 French nouns annotated with 24 "supersenses" (e.g. Act, Person, State, Institution, etc). In our work we focus only on the PERSON[5] and ACT[6] as being one of the most frequent senses in the corpora, as well as in the vocabulary of models. The number of word embeddings of each model associated with the supersenses can be found in Table 3.

Model	Cased	Nouns	Adjectives	Verbs
FlauBERT$_{small}$	✓	12,807	6,504	5,425
DistilBERT	✓	3,858	925	1,079
CamemBERT$_{base}$	✓	8,945	4,584	3,852
mBERT$_{base}$		6,065	2,353	2,163
mBERT$_{base}$	✓	3,858	925	1,079
FlauBERT$_{base}$		15,579	7,590	6,377
FlauBERT$_{base}$	✓	12,807	6,504	5,425
XLM-R$_{base}$	✓	3,401	895	1,233
FlauBERT$_{large}$	✓	12,807	6,504	5,425
XLM-R$_{large}$	✓	3,401	895	1,233

Table 2: Number of word embedding obtained by different models by POS.

Model	Cased	Act	Person
FlauBERT$_{small}$	✓	405	354
DistilBERT	✓	157	139
CamemBERT$_{base}$	✓	311	285
mBERT$_{base}$		182	156
mBERT$_{base}$	✓	157	139
FlauBERT$_{base}$		417	364
FlauBERT$_{base}$	✓	405	354
XLM-R$_{base}$	✓	110	70
FlauBERT$_{large}$	✓	405	354
XLM-R$_{large}$	✓	110	70

Table 3: The number of word embeddings obtained by different models by supersense.

4 Experimental protocol

As mentioned above, our approach consisted of combining word embeddings with grammatical/semantic information, extracted from human-annotated sources where the final list of tested features consisted of grammatical gender for nouns and adjectives, grammatical number for nouns and adjectives, POS for nouns, adjectives and verbs, and semantic categories for nouns. In order to use the annotations, a decision was made to create a binary vector for each feature: 1 was assigned to if a

[4]For all listed models, we extracted hidden states of the last layer to treat them as the word embeddings of the model, without any fine-tuning.

[5]Examples: *étudiants* ('students'), *neveu* ('nephew'), *femme* ('woman').

[6]Examples: *ablation* ('ablation'), *acceuil* ('reception').

word possessed a certain feature and 0 - otherwise[7].

Due to the relatively small sizes of the datasets of annotated word embeddings, we opted to incorporate 5-fold cross-validation (Fushiki, 2011) instead of the traditional train/test split.

4.1 Intrinsic evaluation of information encoding quality

It can be presumed that the information is encoded either throughout all dimensions of a word embedding or in a smaller subset of dimensions. However, identifying the exact subset is essentially computationally impossible (the number of dimension combinations ranging from 2^{512} for smaller models to 2^{1024} for larger ones). Therefore, we propose an approach to identify dimension subsets' candidates that can potentially contain the target information using the following steps:

- Identify sets of dimensions that are not independent of the feature vector and their intersections;
- Identify sets of important dimensions that are likely to be important for encoding the feature and their intersections;
- Compare the accuracy among each found dimension subset and all dimensions of the word embedding.

4.1.1 Dependent dimensions

The intuition is that if a dimension is classified as dependent from the feature vector, it can be a potential candidate that encodes the information. As a way to find dependent dimensions, we perform two types of tests: Mutual Information (MI) (Kraskov et al., 2004) test and one-way ANOVA test. Additionally, we find the intersection between the sets of dimensions found by the MI test and the ANOVA test, to get another subset of dimension candidates.

As for the MI test, for $i = 1...n$, where n is the number of dimensions in an embedding, f is the feature vector for all words of the vocabulary, we check the condition $MI(d_i, f) > 0$ and if it is fulfilled we add d_i to the list of dependent dimensions detected by the MI test.

The ANOVA test operates with the null hypothesis that both samples have the same population mean, which in our case would represent values of dimensions where $f = 0$ and $f = 1$ have the same distribution. Therefore, for $i = 1...n$, d_i is split into two samples ($d_{i_{f=0}}$ (values of dimension d_i for words that have the associated value $f = 0$) and $d_{i_{f=1}}$ (generated similarly to $d_{i_{f=0}}$), the one-way ANOVA test is performed using the samples and if *p-value* < 0.01, d_i is added to the list of dependent dimensions detected by the ANOVA test.

4.1.2 Important dimensions

Another assumption we operate under is that some dimensions are more important than others in encoding the feature information. For this, we performed a series of tests that allowed us to rank the dimensions and picked only the top $\alpha\%$ of them as important dimensions. The first approach to identifying important dimensions involves training a Logistic Regression (LR) model using all dimensions of an embedding with the feature vector f being used as its target. The absolute weights of the trained LR classifier that are associated with each dimension are sorted in descending order, and top $\alpha\%$ are selected as important dimensions found by the LR test. Similarly, we train a Perceptron classifier following exactly the same approach.

Finally, for $i = 1...n$ we computed $corr(d_i, f)$ where n is the number of dimensions, f is the feature vector and $corr$ is the Point-biserial correlation coefficient. The dimensions were then sorted by the absolute values of the associated correlation score and the top $\alpha\%$ were selected as the ones highlighted by the correlation test. The decision to use the Point-biserial correlation was driven by the fact that it is possible to use the metric with continuous (dimension values) and discrete (the feature vector) and in our case is synonymous with easily computable Pearson correlation.

Moreover, for each α we calculate the intersection between all groups of important dimensions in order to find additional subsets of dimension candidates. All the tests were repeated for the following values of α: [1, 5, 10, 25, 50, 75].

4.1.3 Computing predictions

Having identified a set S consisting of 28 dimension subsets for each model's embeddings (all dimensions, 3 subsets of dependent dimensions, 24 subsets of important dimensions), we use each subset of dimensions to predict values of f on the test set.

To do this for each $s \in S$ we compute the median values of each $d \in s$ associated with $f = 0$

[7] For gender, 1 = feminine and 0 = masculine. For number, 1 = plural and 0 = singular.

and for $f = 1$ on the train set separately. After this process, we obtain two vectors: med_0 and med_1. Following that, for each word embedding w_{test} we select only dimensions $d_{d \in s}$ and compute Mean Absolute Error between w_{test_d} and med_0 (mae_0), as well as Mean Absolute Error between w_{test_d} and med_1 (mae_1). If $mae_0 < mae_1$, the predicted value of the feature vector f_{test} is 0, and 1 otherwise. Having obtained predictions for all words in the test set, we compute prediction accuracy. Finally, we selected the subset s with the highest accuracy to be the *best candidate*. Prediction accuracy for best candidates in each fold is averaged among 5 folds, and the final metric is considered to be *InfEnc*.

4.2 Stable dimensions

The process of 5-fold cross-validation additionally allowed us to validate if a certain dimension appears in the best candidate subset for feature f consistently throughout all 5 folds. If it does, we consider such dimension to be a *stable dimension* for the feature f.

5 Results

Following the protocol above, we calculated InfEnc for all listed models for encoding quality of grammatical gender (for adjectives and nouns), grammatical number (for adjectives and nouns), POS, and semantic supersenses. It is worth noting that for observed experiments, subsets of dimensions appear to achieve higher accuracy in the vast majority of cases.

5.1 Grammatical gender

We performed experiments for gender in 3 parts: nouns (N), adjectives (A), and nouns and adjectives (N+A) combined (the results can be found in Table 4).

Model	Cased	N	A	N+A
FlauBERT$_{small}$	✓	0.805 [2]	**0.95 [1]**	0.794 [2]
DistilBERT	✓	0.605 [4]	0.682 [3]	0.626 [4]
CamemBERT$_{base}$	✓	0.534 [9]	0.538 [9-10]	0.546 [8]
mBERT$_{base}$		0.516 [10]	0.538 [9-10]	0.537 [9-10]
mBERT$_{base}$	✓	0.552 [8]	0.612 [5]	0.537 [9-10]
FlauBERT$_{base}$		0.59 [5]	0.587 [8]	0.585 [6]
FlauBERT$_{base}$	✓	0.669 [3]	0.653 [4]	0.655 [3]
XLM-R$_{base}$	✓	0.557 [7]	0.589 [7]	0.565 [7]
FlauBERT$_{large}$	✓	**0.895 [1]**	0.933 [2]	**0.905 [1]**
XLM-R$_{large}$	✓	0.575 [6]	0.6 [6]	0.6 [5]

Table 4: InfEnc results for the grammatical gender feature, the best results are bolded. The rank is added in square brackets.

We can notice that despite being described as partially trained, only recommended for debugging by the authors[8], and the smallest out of all models tested, FlauBERT$_{small}$ achieves the best score in encoding information about adjectives gender. Similarly, the second smallest model DistilBERT trained on multilingual data performs comparably to FlauBERT$_{base}$, trained only on French data.

As could be expected, generally, multilingual models perform significantly worse in the quality of French gender encoding. Surprisingly, CamemBERT scores in InfEnc are as low as those of multilingual models which could be attributed to its tokenization method, further commented on in Section 6.1. Moreover, one could notice that uncased models appear to perform worse in information encoding than their cased variants (both for mBERT and FlauBERT$_{base}$).

5.2 Grammatical number

Similarly to gender, the evaluation of the quality of encoding number information was conducted in 3 parts: nouns, adjectives, and nouns and adjectives combined (see Table 5). It can be seen that on average multilingual models appear to encode number information better than grammatical gender, one of the explanations for such phenomena could be the fact that the plural form of French nouns is formed similarly to plural forms in other languages, therefore, models could have more exposure to this vocabulary during training.

Model	Cased	N	A	N+A
FlauBERT$_{small}$	✓	0.951 [2]	0.957 [2]	0.943 [2]
DistilBERT	✓	0.698 [4]	0.706 [4]	0.692 [4]
CamemBERT$_{base}$	✓	0.518 [10]	0.551 [10]	0.539 [10]
mBERT$_{base}$		0.562 [9]	0.563 [9]	0.569 [9]
mBERT$_{base}$	✓	0.604 [6]	0.57 [7]	0.588 [8]
FlauBERT$_{base}$		0.651 [5]	0.597 [6]	0.645 [5]
FlauBERT$_{base}$	✓	0.709 [3]	0.736 [3]	0.697 [3]
XLM-R$_{base}$	✓	0.589 [8]	0.641 [8]	0.598 [7]
FlauBERT$_{large}$	✓	**0.956 [1]**	**0.959 [1]**	**0.953 [1]**
XLM-R$_{large}$	✓	0.599 [7]	0.647 [5]	0.64 [6]

Table 5: InfEnc results for the grammatical number feature, the best results are bolded. The rank is added in square brackets.

Similarly, as for gender, we can notice smaller models perform either on par (FlauBERT$_{small}$) or better (DistilBERT performing better than non-distilled mBERT) than their bigger counterparts. Likewise, uncased models show lower InfEnc scores than the cased versions.

[8] https://github.com/getalp/Flaubert

5.3 POS

For POS, the experiments were performed in 3 parts: nouns (N) encoded as 1s and non-nouns as 0s; adjectives (A) encoded as 1s and non-adjectives as 0s; verbs (V) encoded as 1s and non-verbs as 0s. The corresponding results can be found in Table 6. What can be noticed is that for most

Model	Cased	N	A	V
FlauBERT$_{small}$	✓	0.893 [2]	**0.896 [1]**	**0.938 [1]**
DistilBERT	✓	0.641 [5]	0.671 [3]	0.659 [5]
CamemBERT$_{base}$	✓	0.548 [9]	0.573 [8]	0.579 [9]
mBERT$_{base}$		0.539 [10]	0.532 [10]	0.543 [10]
mBERT$_{base}$	✓	0.573 [8]	0.563 [9]	0.608 [8]
FlauBERT$_{base}$		0.689 [3]	0.639 [4]	0.718 [3]
FlauBERT$_{base}$	✓	0.643 [4]	0.611 [6]	0.695 [4]
XLM-R$_{base}$	✓	0.594 [6]	0.596 [7]	0.615 [7]
FlauBERT$_{large}$	✓	**0.901 [1]**	0.889 [2]	0.937 [2]
XLM-R$_{large}$	✓	0.586 [7]	0.618 [5]	0.616 [6]

Table 6: InfEnc results for the encoding of POS information, the best results are bolded. The rank is added in square brackets.

models, except DistilBERT, the score for encoding verb information is the highest among all investigated POS. Also, interestingly, the uncased FlauBERT$_{base}$ model performs better than the cased one in POS information encoding, contrary to the gender and number information encoding. Similarly to previous results, FlauBERT$_{small}$ and DistilBERT show either better or comparable results to bigger models.

5.4 Semantic supersenses

Model	Cased	Act	Person
FlauBERT$_{small}$	✓	**0.809 [1]**	0.868 [2]
DistilBERT	✓	0.699 [3]	0.695 [3]
CamemBERT$_{base}$	✓	0.564 [9]	0.581 [8]
mBERT$_{base}$		0.498 [10]	0.5 [10]
mBERT$_{base}$	✓	0.598 [8]	0.651 [6]
FlauBERT$_{base}$		0.63 [5]	0.659 [4]
FlauBERT$_{base}$	✓	0.666 [4]	0.653 [5]
XLM-R$_{base}$	✓	0.629 [6]	0.618 [7]
FlauBERT$_{large}$	✓	0.806 [2]	**0.87 [1]**
XLM-R$_{large}$	✓	0.624 [7]	0.562 [9]

Table 7: InfEnc results for the encoding of supersense information, the best results are bolded. The rank is added in square brackets.

As we can see in Table 7 the results of InfEnc for the semantic features, the mBERT uncased score is the poorest through all experiments run. Even if FlauBERT$_{large}$ is still showing comparatively high results in the metric, we can notice a big drop in accuracy from the previous experiment results for the model; this could potentially be a sign of the complex nature of semantic features in comparison to grammatical ones.

5.5 Correlation with classification task

To validate if the obtained InfEnc scores are representative of the performance of the embeddings in downstream tasks, for each model, we trained 5 different classifiers: LR, Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), and K-Nearest Neighbors (KNN). For the targets of classification, we used the encoded feature vectors f (gender, number, POS, semantic information). We calculated the mean accuracy for each classifier across 5 folds and additionally mean accuracy among all classifiers, and computed Pearson correlation between obtained accuracies and the InfEnc scores of the models. As can be seen in Figure 1, the obtained

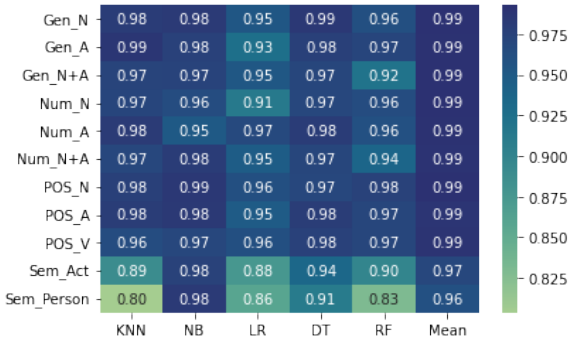
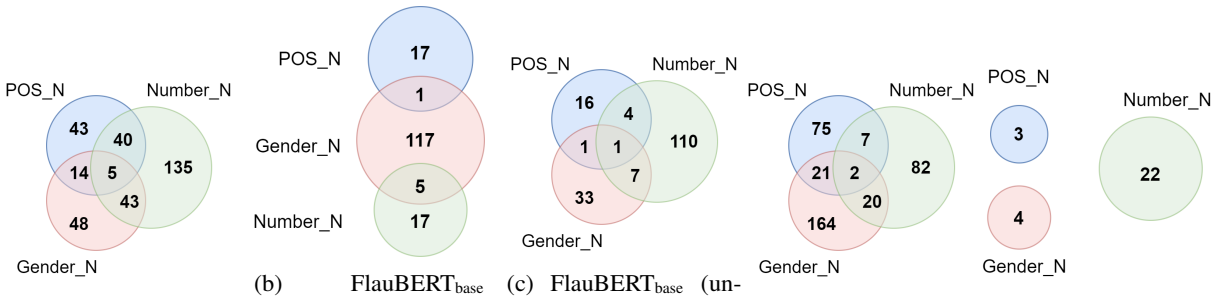


Figure 1: Correlation between accuracies of classifiers and InfEnc scores for each model. "Mean" stands for mean accuracy between all classifiers.

correlation is very high. It can be argued that it is linked to using LR weights as a way to extract meaningful dimensions, however, it can be noticed that the correlation between LR accuracies and InfEnc scores is among the lowest for multiple tasks. Hence, we can assume that InfEnc scores can be a good predictor of classification performance by different classification models. Moreover, it is worth remarking that the correlation with accuracies for classifying semantic features ("Person" and "Act") appears to be lower than for the grammatical features, which can be explained by a more subjective structure of such features.

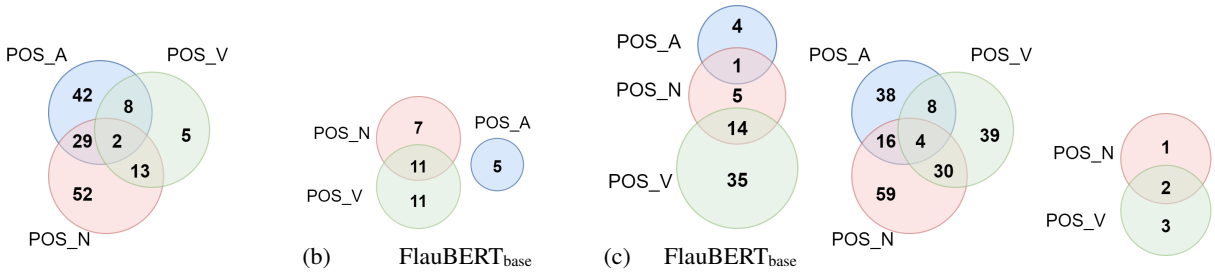
5.6 Stable dimensions

As can be seen in Table 8, except for FlauBERT family models we could not find stable dimensions for all possible features for other models. For DistilBERT, we managed to obtain stable dimensions for all features except for the ones responsible for encoding the POS of adjectives. However, even finding a single stable dimension can be beneficial:



(a) FlauBERT$_{small}$ (b) FlauBERT$_{base}$ (cased) (c) FlauBERT$_{base}$ (uncased) (d) FlauBERT$_{large}$ (e) DistilBERT

Figure 2: Number of overlapping dimensions for noun gender, number, and POS features. Note that when the three sets overlap (a, c, and d subgraphs), the number of overlapping dimensions given in the intersection of the three sets is also included in the other intersections of the two sets.



(a) FlauBERT$_{small}$ (b) FlauBERT$_{base}$ (cased) (c) FlauBERT$_{base}$ (uncased) (d) FlauBERT$_{large}$ (e) DistilBERT

Figure 3: Number of overlapping dimensions for POS features. Note that when the three sets overlap (a and d subgraphs), the number of overlapping dimensions given in the intersection of the three sets is also included in the other intersections of two sets.

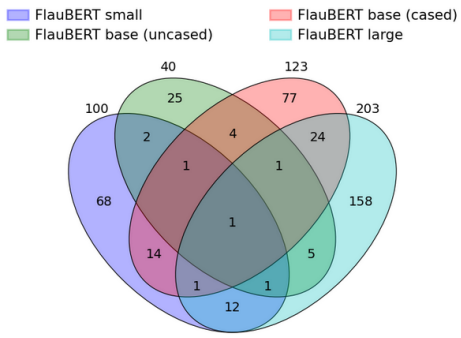


Figure 4: Number of overlapping stable dimensions for noun gender for all FlauBERT models.

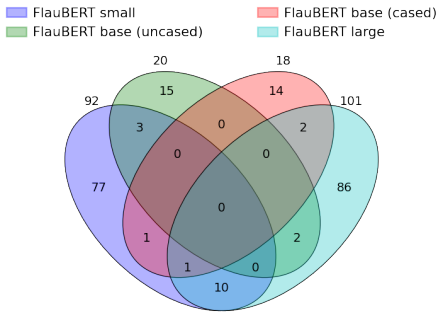


Figure 5: Number of overlapping stable dimensions for noun POS for all FlauBERT models.

for XLM-R$_{base}$ we found only one stable dimension for the adjective number (d_{467}) that has InfEnc score of 0.628 (which is comparable to the score observed in Table 5).

Regarding grammatical information, we noticed that the patterns of encoding are different from model to model (can be seen in Figure 2): some models have dimensions that appear to encode number, gender, and POS simultaneously (like FlauBERT$_{small}$, for example); others, have no overlap between such dimensions (e.g. DistilBERT). On the other hand, for all FlauBERT models and DistilBERT we could find dimensions that appear to encode both information about a word being a noun and a word being a verb (see Figure 3).

Additionally, we could find that some stable dimensions that are shared between FlauBERT$_{small}$ and FlauBERT$_{large}$ (see Figures 4 and 5). This fact could be explained by the highest number of stable dimensions that we observed in these models. We discovered that all models of FlauBERT family share dimension d_{177} in stable dimension corresponding to the gender of a noun and addition-

Model	Cased	Gen_N	Gen_A	Gen_N+A	Num_N	Num_A	Num_N+A	POS_N	POS_A	POS_V	Act	Person
$\text{FlauBERT}_{\text{small}}$	✓	100	175	74	213	107	127	92	77	24	48	95
DistilBERT	✓	4	18	14	22	0	15	3	0	5	3	1
$\text{CamemBERT}_{\text{base}}$	✓	0	1	0	0	0	9	1	0	0	1	0
$\text{mBERT}_{\text{base}}$		0	0	3	0	0	4	1	0	0	0	0
$\text{mBERT}_{\text{base}}$	✓	0	5	0	0	0	0	1	0	0	0	6
$\text{FlauBERT}_{\text{base}}$		40	26	47	120	4	91	20	5	49	2	5
$\text{FlauBERT}_{\text{base}}$	✓	123	22	68	22	4	7	18	5	22	14	1
$\text{XLM-R}_{\text{base}}$	✓	0	0	0	2	1	0	1	1	0	1	1
$\text{FlauBERT}_{\text{large}}$	✓	203	134	227	107	151	195	101	58	73	50	63
$\text{XLM-R}_{\text{large}}$	✓	0	0	1	0	0	1	0	1	1	0	0

Table 8: The number of stable dimensions for each feature.

ally all cased FlauBERT models have d_{55} among dimensions corresponding to POS of nouns[9].

6 Discussion

We believe that identifying dimensions with grammatical or semantic features can be used both during training models, combined with reinforcement training, ensuring a better quality of encoding of information, and as a way to find the best-suited model for the downstream task, as well as evaluating the effect of fine-tuning on the encoding of the target features. Due to time and resource limitations, our experiments were limited in terms of studied features and language which we hope to address in further work. We believe that the research could benefit from incorporating other semantic and grammatical features. However, additionally, we propose the following topics for discussion.

6.1 Differences in tokenization

Investigated models use different tokenization algorithms: BERT and DistilBERT use WordPiece embeddings (Wu et al., 2016), XLM-R and CamemBERT incorporate SentencePiece tokenization (Kudo and Richardson, 2018) and FlauBERT utilizes Byte-Pair Encoding (BPE) algorithm (Sennrich et al., 2016). As can be seen from Tables 1 and 2, the tokenization quite significantly affects the size of the extracted vocabulary. Precisely, CamemBERT is much less likely to tokenize a noun, adjective, or verb as one token than FlauBERT. For example, the tokenization *brunes* ('brown (plural)') → [brune, s] or *gaieté* ('cheerfulness, joy') → [ga, ie, té] signifies that despite the relative commonality of words, CamemBERT gives preference to subword tokenization. This makes the extraction of important dimensions a more complex task for models like this, therefore, additional research needs to be performed where the effect of tokenization on the information encoding is studied.

It is also worth mentioning that in our work we compared the models' performance on vocabularies specific to the model, however, it is worth investigating how the models would perform on a certain basic vocabulary that all models share.

6.2 Changes in representation

Additionally, as was shown in our experiments, smaller models can learn word embeddings that encode target features as effectively as larger models, therefore, saving energy and computation capacity. Investigating dimensions alongside training a model can give a greater insight into how the target features are learned. The researchers have previously found that the F1 score on downstream tasks appears to plateau after a certain number of steps (Müller-Eberstein et al., 2023), which can be expanded in studying how a target feature encoding changes with the number of steps.

7 Conclusion

In this paper, we proposed an intrinsic metric InfEnc and a framework allowing the extraction of stable dimensions that potentially encode grammatical or semantic information from word embeddings of BERT-like models trained on the French language. Our findings include:

1. For all tested features, subsets of dimensions appear to encode the information better than all dimensions of word embedding.
2. Smaller size models can encode the target information on par or better than larger models.
3. Gender information appears to be better encoded in cased models, than in their uncased counterparts.
4. Tokenization affects greatly the encoding of information in word embeddings.

[9] Find the list of all retrieved stable dimensions in Appendix A and the corresponding accuracy of classifiers in Appendix B.

5. There exist the same dimensions that appear to encode target information in different sizes of models (in our case, FlauBERT).
6. There are signs that noun-ness and verb-ness can be encoded in the same dimensions.
7. For multilingual models, one or a small subset of dimensions can encode information better than a large subset of dimensions.

We believe that understanding what information is encoded in each dimension can be beneficial for a multitude of applications: identifying dimensions encoding gender information can potentially help to mitigate gender bias; knowing what dimensions encode the information related to the downstream task can lead to reducing dimensionality and therefore computational cost; contrastive and reinforcement learning in combination with interpretable embedding dimensions can be advantageous for reduction of hallucination of LLMs.

Acknowledgments

The topic of this work was initiated by the work of 3 students: Clémentine Bleuze, Ekaterina Goliakova, and Chun Yang.

References

Amine Abdaoui, Camille Pradel, and Grégoire Sigel. 2020. Load what you need: Smaller versions of mutililingual BERT. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 119–123, Online. Association for Computational Linguistics.

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. In *ArXiv*.

Lucie Barque, Pauline Haas, Richard Huyghe, Delphine Tribout, Marie Candito, Benoit Crabbé, and Vincent Segonne. 2020. FrSemCor: Annotating a French corpus with supersenses. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5912–5918, Marseille, France. European Language Resources Association.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. In *Computational Linguistics*, volume 48, pages 207–219, Cambridge, MA. MIT Press.

Gabriella Chronis and Katrin Erk. 2020. When is a bishop not like a rook? when it's like a rabbi! multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 227–244, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Tadayoshi Fushiki. 2011. Estimation of prediction error by using K-fold cross-validation. In *Statistics and Computing*, volume 21, pages 137–146. Springer.

Qinjin Jia, Jialin Cui, Yunkai Xiao, Chengyuan Liu, Parvez Rashid, and Edward F Gehringer. 2021. All-in-one: Multi-task learning bert models for evaluating peer assessments.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. In *Nature*, volume 596, pages 583–589. Nature Publishing Group.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. In *Physical review E*, volume 69, pages 69–85. APS.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Yang-Yin Lee, Hao Ke, Ting-Yu Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Combining and learning word embedding with wordnet for semantic relatedness and similarity measurement. In *Journal of the association for information science and technology*, volume 71, pages 657–670. Wiley Online Library.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. 2019. Findings of the first shared task on machine translation robustness. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 91–102, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Max Müller-Eberstein, Rob van der Goot, Barbara Plank, and Ivan Titov. 2023. Subspace chronicles: How linguistic information emerges, shifts and interacts during language model training. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13190–13208, Singapore. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Laurent Romary, Susanne Salmon-Alt, and Gil Francopoulo. 2004. Standards going concrete: from LMF to Morphalou. In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, pages 22–28, Geneva, Switzerland. COLING.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. 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 5831–5837, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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.

Lauro Snidaro, Giovanni Ferrin, and Gian Luca Foresti. 2019. Distributional memory explainable word embeddings in continuous space. In *2019 22th International Conference on Information Fusion (FUSION)*, pages 1–7.

Erico Tjoa and Cuntai Guan. 2020. A survey on explainable artificial intelligence (XAI): Toward medical XAI. In *IEEE transactions on neural networks and learning systems*, volume 32, pages 4793–4813. IEEE.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*

(EMNLP), pages 197–216, Online. Association for Computational Linguistics.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Appendix A Stable dimensions

In this section, you will find the retrieved stable dimensions for tested linguistic features for each model.

A.1 FlauBERT$_{\text{small}}$

Feature	Dimension list
Gender: Noun	0, 5, 7, 20, 25, 28, 30, 36, 40, 42, 50, 54, 62, 74, 85, 88, 95, 96, 100, 113, 115, 117, 121, 123, 124, 130, 133, 141, 142, 147, 152, 156, 160, 162, 173, 177, 181, 186, 192, 193, 195, 198, 200, 202, 210, 213, 214, 220, 234, 237, 239, 245, 250, 255, 256, 261, 265, 269, 276, 279, 292, 293, 296, 306, 310, 312, 315, 316, 318, 320, 332, 335, 352, 362, 363, 374, 376, 377, 387, 390, 403, 417, 426, 432, 434, 436, 439, 443, 455, 466, 468, 470, 477, 488, 490, 495, 497, 499, 501, 507
Gender: Adjective	0, 4, 5, 7, 8, 11, 12, 14, 26, 28, 30, 33, 36, 38, 39, 42, 45, 46, 50, 53, 54, 55, 56, 57, 59, 65, 67, 70, 72, 74, 75, 82, 83, 85, 88, 89, 94, 95, 96, 99, 114, 117, 121, 122, 123, 124, 133, 142, 144, 145, 147, 157, 160, 162, 167, 170, 175, 177, 178, 181, 184, 185, 187, 189, 192, 193, 195, 200, 202, 203, 206, 211, 213, 214, 222, 228, 230, 233, 234, 237, 245, 249, 250, 251, 255, 256, 260, 268, 274, 275, 276, 283, 284, 286, 287, 289, 290, 292, 293, 296, 300, 302, 304, 306, 309, 310, 313, 314, 316, 318, 320, 321, 331, 332, 333, 336, 339, 340, 341, 345, 348, 352, 353, 354, 357, 360, 362, 363, 364, 365, 366, 370, 372, 374, 376, 377, 379, 380, 385, 387, 389, 390, 396, 398, 399, 401, 409, 425, 426, 429, 430, 432, 436, 439, 443, 449, 450, 453, 461, 465, 466, 470, 471, 477, 478, 482, 486, 488, 489, 490, 500, 501, 503, 506, 507
Gender: Noun & Adjective	0, 7, 11, 25, 28, 36, 40, 42, 55, 62, 74, 88, 95, 100, 115, 117, 121, 124, 130, 144, 147, 149, 159, 160, 162, 175, 177, 181, 186, 192, 195, 198, 202, 210, 211, 237, 239, 245, 250, 256, 261, 265, 269, 270, 279, 292, 296, 300, 306, 309, 310, 315, 316, 318, 320, 332, 335, 363, 377, 387, 390, 403, 409, 432, 434, 443, 448, 455, 468, 470, 488, 490, 499, 507
Number: Noun	1, 3, 5, 6, 7, 8, 9, 10, 13, 15, 19, 21, 24, 25, 26, 27, 31, 37, 42, 43, 48, 50, 51, 52, 53, 54, 55, 56, 57, 66, 67, 69, 70, 73, 74, 76, 77, 80, 81, 83, 84, 86, 88, 89, 92, 96, 97, 99, 101, 107, 110, 111, 112, 114, 115, 117, 118, 119, 128, 129, 130, 131, 132, 133, 137, 138, 142, 148, 149, 150, 151, 154, 155, 156, 158, 160, 161, 164, 165, 167, 170, 171, 172, 175, 177, 181, 182, 183, 185, 187, 192, 198, 200, 205, 207, 208, 209, 210, 212, 213, 214, 220, 223, 224, 229, 238, 243, 244, 246, 250, 251, 252, 254, 255, 257, 259, 268, 269, 270, 273, 277, 278, 281, 282, 285, 286, 288, 289, 291, 295, 296, 297, 299, 302, 303, 305, 306, 307, 308, 309, 310, 311, 313, 317, 322, 326, 328, 330, 336, 337, 338, 342, 343, 351, 352, 353, 355, 356, 359, 360, 372, 373, 374, 376, 378, 381, 382, 384, 391, 399, 403, 404, 405, 413, 414, 416, 419, 420, 422, 423, 430, 431, 432, 433, 435, 438, 441, 442, 445, 451, 452, 453, 454, 455, 457, 458, 461, 462, 470, 473, 475, 477, 481, 485, 488, 489, 490, 492, 495, 506, 507, 510, 511
Number: Adjective	1, 3, 9, 15, 21, 24, 25, 50, 54, 56, 67, 69, 73, 74, 81, 83, 84, 96, 109, 112, 115, 125, 129, 131, 138, 149, 154, 156, 158, 159, 161, 165, 167, 172, 175, 181, 182, 183, 185, 191, 192, 198, 200, 205, 208, 210, 220, 224, 238, 250, 251, 252, 254, 257, 278, 285, 288, 289, 297, 303, 306, 310, 311, 313, 317, 337, 340, 342, 347, 351, 352, 353, 356, 359, 360, 372, 374, 378, 381, 384, 399, 403, 405, 410, 419, 420, 430, 445, 453, 454, 455, 458, 461, 462, 474, 475, 477, 479, 481, 483, 484, 485, 490, 495, 497, 499, 501
Number: Noun & Adjective	1, 3, 7, 8, 9, 10, 15, 21, 37, 42, 43, 50, 51, 52, 54, 55, 57, 62, 67, 69, 73, 74, 81, 83, 92, 96, 101, 110, 111, 115, 118, 119, 129, 131, 132, 138, 148, 149, 150, 154, 155, 156, 158, 160, 161, 165, 167, 172, 175, 177, 182, 183, 185, 192, 198, 200, 205, 208, 209, 210, 212, 213, 214, 220, 223, 246, 250, 251, 252, 254, 255, 257, 277, 278, 285, 288, 289, 295, 296, 297, 302, 303, 306, 308, 310, 311, 317, 322, 328, 330, 337, 338, 342, 351, 352, 355, 356, 359, 360, 372, 373, 374, 376, 378, 384, 399, 405, 419, 420, 422, 430, 432, 438, 445, 451, 454, 455, 461, 473, 475, 477, 485, 489, 495, 501, 506, 507
POS: Noun	11, 15, 29, 31, 33, 36, 37, 38, 41, 51, 53, 55, 57, 63, 65, 72, 73, 78, 82, 83, 92, 102, 103, 106, 112, 117, 122, 128, 130, 134, 141, 154, 155, 159, 164, 166, 168, 176, 178, 183, 192, 197, 198, 205, 207, 212, 224, 229, 233, 260, 275, 278, 281, 285, 286, 300, 303, 305, 314, 320, 339, 341, 346, 362, 378, 382, 387, 393, 395, 401, 404, 405, 409, 423, 424, 426, 434, 450, 452, 455, 458, 460, 462, 464, 465, 466, 468, 478, 480, 504, 508, 509
POS: Adjective	4, 9, 15, 31, 37, 39, 41, 44, 53, 56, 72, 73, 82, 84, 87, 92, 100, 111, 117, 119, 130, 134, 138, 139, 149, 153, 158, 159, 176, 199, 205, 213, 222, 239, 246, 250, 260, 265, 275, 276, 284, 288, 297, 299, 301, 309, 310, 314, 315, 330, 332, 337, 339, 346, 348, 350, 387, 405, 417, 428, 429, 432, 435, 439, 445, 450, 451, 455, 461, 462, 464, 465, 467, 468, 478, 485, 499
POS: Verb	0, 36, 56, 89, 103, 139, 158, 159, 192, 198, 233, 282, 297, 303, 310, 318, 341, 378, 410, 432, 462, 480, 504, 508
Semantic: Act	25, 26, 33, 39, 41, 102, 113, 114, 119, 128, 146, 166, 170, 172, 173, 177, 182, 187, 201, 202, 218, 227, 245, 255, 273, 274, 281, 284, 294, 313, 339, 357, 366, 380, 381, 399, 400, 406, 417, 423, 455, 461, 466, 468, 485, 489, 499, 500
Semantic: Person	0, 5, 6, 9, 16, 20, 24, 30, 32, 33, 39, 41, 57, 58, 59, 70, 75, 78, 84, 86, 90, 94, 101, 112, 113, 121, 139, 140, 142, 144, 146, 153, 160, 162, 174, 176, 177, 184, 197, 204, 207, 220, 223, 225, 228, 235, 243, 263, 268, 273, 275, 276, 283, 287, 295, 296, 303, 304, 316, 329, 337, 339, 355, 357, 358, 359, 364, 369, 375, 377, 387, 388, 389, 396, 401, 404, 405, 407, 417, 421, 423, 447, 448, 450, 456, 458, 472, 475, 478, 483, 489, 494, 495, 499, 505

A.2 DistilBERT

Feature	Dimension list
Gender: Noun	302, 653, 713, 727
Gender: Adjective	9, 37, 51, 76, 185, 278, 301, 425, 526, 531, 551, 633, 641, 676, 716, 737, 747, 749
Gender: Noun & Adjective	70, 116, 161, 262, 301, 302, 353, 488, 499, 563, 656, 727, 728, 761
Number: Noun	81, 106, 148, 169, 177, 184, 186, 275, 333, 343, 354, 367, 376, 392, 403, 441, 498, 522, 535, 573, 663, 710
Number: Adjective	-
Number: Noun & Adjective	152, 184, 186, 216, 354, 363, 413, 440, 518, 535, 584, 649, 668, 745, 761
POS: Noun	52, 501, 700
POS: Adjective	-
POS: Verb	52, 189, 327, 700, 740
Semantic: Act	398, 446, 621
Semantic: Person	346

A.3 CamemBERT

Feature	Dimension list
Gender: Noun	-
Gender: Adjective	215
Gender: Noun & Adjective	-
Number: Noun	-
Number: Adjective	-
Number: Noun & Adjective	24, 138, 176, 213, 303, 386, 482, 493, 562
POS: Noun	696
POS: Adjective	-
POS: Verb	-
Semantic: Act	343
Semantic: Person	-

A.4 mBERT (uncased)

Feature	Dimension list
Gender: Noun	-
Gender: Adjective	-
Gender: Noun & Adjective	274, 445, 447
Number: Noun	-
Number: Adjective	-
Number: Noun & Adjective	74, 194, 556, 654
POS: Noun	74
POS: Adjective	-
POS: Verb	-
Semantic: Act	-
Semantic: Person	-

A.5 mBERT (cased)

Feature	Dimension list
Gender: Noun	-
Gender: Adjective	9, 223, 519, 540, 575
Gender: Noun & Adjective	-
Number: Noun	-
Number: Adjective	-
Number: Noun & Adjective	-
POS: Noun	529
POS: Adjective	-
POS: Verb	-
Semantic: Act	-
Semantic: Person	143, 319, 447, 504, 585, 715

A.6 FlauBERT$_{\text{base}}$ (uncased)

Feature	Dimension list
Gender: Noun	2, 17, 21, 47, 81, 98, 130, 132, 138, 149, 177, 180, 185, 197, 198, 244, 299, 307, 309, 310, 314, 382, 433, 456, 505, 507, 508, 546, 572, 596, 597, 604, 644, 662, 671, 696, 698, 735, 755, 757
Gender: Adjective	17, 32, 45, 55, 57, 136, 149, 169, 203, 215, 277, 314, 315, 374, 382, 396, 508, 516, 546, 574, 585, 607, 609, 671, 680, 698
Gender: Noun & Adjective	0, 21, 24, 32, 81, 119, 120, 138, 156, 171, 177, 185, 186, 187, 196, 197, 198, 212, 223, 244, 263, 310, 314, 322, 333, 341, 382, 432, 435, 456, 480, 507, 508, 519, 545, 546, 574, 585, 587, 594, 604, 610, 667, 680, 696, 719, 766
Number: Noun	0, 1, 5, 11, 17, 19, 27, 28, 29, 31, 40, 49, 50, 56, 58, 66, 79, 81, 85, 96, 98, 113, 127, 133, 148, 150, 182, 186, 191, 192, 193, 197, 202, 207, 210, 221, 234, 242, 250, 252, 255, 274, 294, 296, 299, 301, 306, 309, 322, 327, 329, 345, 353, 354, 365, 367, 374, 381, 384, 396, 402, 412, 414, 415, 425, 431, 435, 443, 451, 465, 470, 476, 477, 480, 487, 490, 496, 499, 512, 519, 528, 532, 535, 539, 543, 573, 577, 580, 598, 600, 604, 614, 615, 617, 623, 626, 632, 635, 638, 647, 650, 652, 673, 678, 687, 699, 706, 715, 719, 730, 736, 737, 738, 743, 744, 745, 751, 752, 762, 763
Number: Adjective	11, 687, 730, 737
Number: Noun & Adjective	0, 11, 19, 27, 29, 40, 53, 56, 58, 77, 79, 80, 85, 97, 127, 133, 167, 181, 186, 191, 192, 197, 202, 221, 234, 242, 250, 252, 255, 274, 288, 294, 299, 301, 306, 309, 319, 329, 353, 356, 365, 366, 372, 384, 391, 402, 412, 415, 419, 421, 431, 435, 438, 465, 474, 476, 477, 486, 487, 496, 499, 510, 512, 535, 542, 583, 590, 600, 604, 609, 614, 615, 617, 623, 626, 633, 650, 687, 706, 730, 733, 736, 737, 738, 743, 744, 751, 752, 757, 762, 763
POS: Noun	32, 95, 107, 108, 112, 133, 170, 186, 224, 238, 299, 383, 390, 405, 406, 435, 545, 585, 649, 672
POS: Adjective	32, 320, 449, 680, 746
POS: Verb	11, 54, 56, 58, 89, 95, 96, 107, 108, 112, 119, 132, 169, 170, 179, 188, 216, 217, 224, 250, 276, 292, 299, 307, 335, 336, 357, 383, 390, 405, 406, 424, 450, 498, 505, 530, 553, 558, 585, 612, 614, 615, 649, 672, 698, 717, 720, 744, 748
Semantic: Act	48, 752
Semantic: Person	52, 303, 360, 399, 687

A.7 FlauBERT$_{base}$ (cased)

Feature	Dimension list
Gender: Noun	7, 10, 17, 22, 26, 28, 33, 41, 43, 46, 51, 54, 62, 70, 71, 80, 83, 85, 110, 117, 118, 130, 136, 141, 155, 157, 158, 160, 161, 170, 172, 175, 177, 179, 187, 189, 190, 196, 209, 212, 216, 223, 224, 246, 261, 271, 272, 274, 277, 283, 289, 291, 296, 302, 319, 325, 331, 334, 337, 340, 355, 356, 357, 359, 363, 372, 386, 387, 388, 395, 398, 399, 401, 406, 425, 426, 429, 434, 446, 459, 461, 466, 484, 491, 493, 494, 500, 513, 517, 526, 529, 537, 538, 544, 554, 563, 573, 574, 577, 580, 587, 589, 597, 611, 621, 623, 625, 638, 656, 663, 664, 677, 687, 693, 696, 698, 701, 702, 743, 748, 749, 752, 755
Gender: Adjective	71, 85, 118, 130, 155, 170, 189, 225, 309, 331, 337, 340, 359, 398, 425, 484, 491, 611, 621, 661, 663, 698
Gender: Noun & Adjective	30, 33, 43, 46, 62, 68, 70, 83, 85, 106, 117, 130, 155, 157, 170, 172, 177, 187, 189, 209, 216, 223, 224, 272, 274, 277, 283, 296, 319, 331, 337, 349, 357, 359, 363, 372, 395, 398, 406, 425, 429, 435, 484, 493, 494, 500, 506, 512, 517, 526, 537, 544, 570, 571, 573, 580, 587, 589, 597, 600, 611, 655, 663, 698, 725, 746, 749, 752
Number: Noun	34, 53, 125, 127, 176, 180, 196, 205, 238, 268, 279, 290, 343, 398, 449, 466, 500, 571, 594, 623, 672, 760
Number: Adjective	163, 176, 238, 253
Number: Noun & Adjective	34, 176, 180, 238, 325, 466, 594
POS: Noun	55, 81, 87, 106, 119, 162, 243, 248, 345, 413, 417, 454, 564, 569, 602, 688, 698, 720
POS: Adjective	163, 331, 485, 571, 764
POS: Verb	0, 55, 80, 81, 87, 90, 106, 136, 162, 240, 243, 248, 324, 412, 417, 443, 454, 569, 687, 688, 705, 752
Semantic: Act	1, 11, 49, 62, 87, 333, 397, 417, 470, 561, 601, 725, 729, 765
Semantic: Person	611

A.8 XLM-R$_{base}$

Feature	Dimension list
Gender: Noun	-
Gender: Adjective	-
Gender: Noun & Adjective	-
Number: Noun	440, 484
Number: Adjective	467
Number: Noun & Adjective	-
POS: Noun	593
POS: Adjective	100
POS: Verb	-
Semantic: Act	690
Semantic: Person	741

A.9 FlauBERT$_{\text{large}}$

Feature	Dimension list
Gender: Noun	2, 12, 13, 15, 24, 33, 39, 40, 56, 61, 64, 66, 67, 68, 75, 80, 83, 88, 95, 99, 108, 118, 123, 131, 136, 141, 148, 153, 165, 166, 169, 171, 172, 175, 177, 180, 182, 193, 194, 197, 199, 203, 205, 207, 211, 215, 224, 227, 228, 235, 241, 242, 243, 251, 255, 256, 259, 268, 272, 273, 274, 277, 284, 286, 290, 291, 292, 299, 311, 319, 320, 324, 335, 341, 342, 343, 346, 351, 355, 357, 359, 367, 368, 369, 378, 383, 390, 399, 401, 407, 412, 430, 444, 451, 454, 464, 465, 467, 480, 481, 493, 495, 506, 507, 511, 518, 524, 532, 539, 543, 548, 552, 575, 577, 584, 617, 621, 633, 635, 639, 647, 650, 662, 680, 683, 690, 698, 701, 707, 708, 714, 715, 721, 725, 734, 736, 740, 743, 749, 759, 760, 766, 771, 774, 775, 779, 781, 788, 793, 795, 796, 800, 802, 803, 806, 813, 816, 819, 823, 839, 841, 843, 844, 855, 856, 862, 868, 871, 882, 886, 888, 890, 898, 899, 902, 907, 909, 916, 917, 922, 927, 929, 940, 947, 952, 953, 960, 970, 972, 974, 976, 979, 983, 994, 996, 999, 1005, 1007, 1010, 1016, 1018, 1022, 1023
Gender: Adjective	13, 21, 35, 44, 46, 55, 68, 72, 73, 84, 118, 119, 129, 131, 141, 143, 148, 149, 153, 160, 165, 169, 175, 180, 197, 205, 220, 227, 241, 248, 251, 255, 256, 258, 262, 283, 290, 310, 311, 313, 325, 333, 342, 351, 359, 380, 393, 398, 419, 427, 440, 442, 451, 472, 474, 478, 480, 494, 495, 497, 500, 503, 506, 508, 531, 552, 585, 586, 591, 600, 601, 610, 616, 621, 633, 647, 649, 655, 667, 680, 702, 707, 714, 717, 718, 725, 730, 737, 749, 759, 760, 768, 774, 785, 787, 795, 796, 800, 806, 807, 811, 814, 817, 818, 833, 834, 841, 844, 848, 849, 858, 860, 862, 880, 882, 886, 888, 897, 899, 900, 907, 917, 920, 929, 931, 934, 936, 943, 958, 972, 996, 1014, 1017, 1022
Gender: Noun & Adjective	3, 12, 13, 15, 24, 33, 35, 39, 40, 44, 55, 56, 61, 64, 66, 72, 75, 80, 83, 88, 95, 99, 106, 108, 111, 118, 120, 131, 136, 137, 148, 153, 165, 169, 172, 175, 178, 180, 182, 184, 193, 194, 197, 199, 204, 205, 207, 215, 224, 225, 227, 228, 232, 235, 241, 242, 243, 251, 255, 256, 257, 259, 262, 268, 272, 274, 277, 284, 287, 290, 291, 292, 294, 299, 311, 318, 320, 324, 336, 341, 342, 343, 347, 351, 354, 355, 357, 364, 367, 368, 382, 383, 390, 399, 401, 415, 421, 428, 430, 444, 445, 454, 455, 459, 464, 465, 467, 472, 478, 480, 493, 496, 502, 511, 518, 539, 548, 552, 566, 567, 568, 570, 575, 577, 584, 585, 586, 591, 597, 599, 600, 615, 617, 621, 624, 632, 635, 639, 647, 649, 650, 655, 662, 665, 683, 685, 698, 701, 707, 708, 714, 715, 721, 725, 736, 743, 749, 754, 759, 760, 766, 767, 771, 774, 775, 779, 781, 788, 793, 795, 800, 802, 806, 809, 813, 814, 816, 823, 839, 841, 843, 844, 848, 856, 862, 868, 882, 886, 890, 892, 899, 902, 907, 916, 919, 922, 926, 927, 929, 939, 940, 942, 947, 954, 960, 966, 970, 972, 976, 979, 983, 985, 986, 990, 993, 994, 996, 998, 999, 1002, 1005, 1007, 1010, 1014, 1016, 1022, 1023
Number: Noun	7, 28, 29, 34, 55, 59, 93, 103, 121, 123, 136, 138, 139, 147, 150, 171, 184, 185, 191, 194, 209, 223, 234, 246, 250, 259, 262, 278, 281, 305, 315, 330, 334, 352, 357, 358, 370, 373, 387, 389, 398, 404, 421, 435, 436, 469, 476, 480, 485, 486, 488, 491, 497, 505, 508, 517, 532, 545, 546, 554, 556, 561, 565, 576, 583, 587, 602, 606, 634, 638, 641, 653, 660, 678, 680, 690, 691, 707, 721, 724, 729, 730, 774, 775, 783, 787, 814, 851, 874, 877, 898, 911, 921, 927, 928, 930, 937, 939, 947, 967, 978, 991, 994, 996, 1015, 1020, 1022
Number: Adjective	15, 24, 26, 28, 34, 35, 55, 59, 62, 72, 80, 86, 87, 93, 94, 103, 113, 123, 127, 133, 137, 147, 148, 150, 161, 170, 171, 174, 179, 184, 190, 191, 193, 194, 198, 206, 209, 223, 226, 237, 250, 252, 259, 262, 266, 268, 272, 296, 298, 306, 315, 330, 336, 339, 352, 356, 360, 370, 386, 389, 392, 398, 403, 404, 414, 417, 436, 451, 452, 464, 469, 480, 485, 486, 489, 490, 491, 494, 497, 505, 508, 516, 517, 538, 539, 556, 560, 565, 570, 576, 583, 591, 606, 609, 613, 629, 634, 636, 638, 641, 652, 681, 682, 691, 707, 709, 716, 717, 719, 726, 730, 731, 748, 761, 774, 775, 783, 787, 796, 802, 805, 806, 814, 821, 830, 851, 863, 865, 871, 877, 880, 895, 911, 920, 921, 927, 928, 939, 943, 946, 956, 963, 978, 993, 996, 1006, 1010, 1015, 1020, 1021, 1022
Number: Noun & Adjective	7, 9, 10, 20, 28, 29, 34, 39, 47, 55, 56, 59, 74, 85, 86, 87, 93, 96, 98, 103, 110, 111, 119, 121, 123, 133, 136, 137, 139, 140, 147, 150, 153, 161, 179, 184, 185, 191, 193, 194, 198, 203, 206, 209, 219, 223, 234, 238, 243, 246, 250, 259, 262, 266, 278, 293, 294, 296, 301, 305, 306, 315, 320, 330, 345, 352, 357, 365, 370, 373, 386, 387, 389, 397, 398, 404, 416, 417, 419, 421, 422, 428, 435, 438, 439, 451, 459, 469, 476, 477, 480, 485, 486, 489, 497, 505, 508, 514, 516, 521, 539, 545, 546, 554, 556, 565, 567, 572, 575, 576, 583, 584, 585, 587, 591, 602, 606, 608, 609, 634, 636, 638, 641, 644, 648, 649, 651, 655, 672, 680, 683, 690, 691, 703, 707, 709, 712, 716, 721, 724, 726, 729, 730, 731, 732, 748, 772, 774, 775, 783, 787, 793, 802, 808, 814, 818, 830, 842, 851, 858, 861, 873, 874, 877, 881, 883, 895, 901, 911, 920, 921, 925, 927, 928, 930, 937, 939, 947, 955, 956, 966, 967, 978, 988, 990, 991, 994, 996, 1000, 1004, 1015, 1016, 1017, 1020, 1022
POS: Noun	2, 4, 21, 31, 38, 44, 54, 55, 65, 78, 88, 92, 98, 107, 137, 151, 171, 172, 176, 205, 206, 207, 220, 232, 252, 253, 267, 271, 292, 299, 308, 317, 325, 362, 366, 375, 380, 394, 408, 413, 415, 418, 420, 430, 437, 458, 461, 479, 481, 482, 485, 494, 495, 497, 506, 517, 524, 525, 526, 549, 562, 566, 569, 571, 597, 599, 631, 656, 671, 680, 685, 686, 689, 699, 722, 765, 790, 793, 800, 809, 848, 864, 867, 878, 888, 893, 907, 914, 925, 932, 945, 956, 961, 966, 968, 974, 978, 988, 993, 1010, 1016
POS: Adjective	43, 49, 55, 61, 80, 88, 126, 131, 175, 206, 207, 292, 304, 323, 349, 351, 357, 361, 406, 430, 437, 459, 490, 510, 545, 559, 563, 597, 599, 631, 650, 656, 686, 689, 697, 726, 749, 758, 764, 787, 795, 804, 822, 830, 839, 852, 890, 907, 908, 911, 915, 955, 968, 976, 977, 978, 1005, 1013
POS: Verb	2, 22, 31, 44, 88, 98, 99, 106, 172, 182, 207, 208, 232, 233, 249, 252, 253, 266, 267, 271, 274, 278, 299, 333, 362, 374, 375, 404, 406, 416, 424, 427, 436, 437, 439, 441, 473, 479, 481, 490, 506, 529, 539, 580, 664, 667, 683, 689, 724, 748, 765, 775, 793, 800, 803, 809, 848, 853, 864, 886, 895, 904, 909, 912, 915, 925, 927, 932, 974, 976, 988, 991, 1006
Semantic: Act	16, 49, 53, 62, 75, 88, 163, 174, 175, 195, 210, 216, 254, 263, 265, 267, 278, 334, 391, 400, 418, 452, 527, 528, 556, 601, 604, 617, 639, 652, 654, 655, 683, 693, 697, 702, 726, 737, 769, 776, 798, 827, 873, 878, 890, 924, 941, 953, 980, 1013
Semantic: Person	4, 27, 74, 77, 88, 153, 162, 186, 209, 213, 241, 257, 285, 291, 307, 321, 334, 342, 345, 385, 435, 477, 486, 503, 505, 520, 534, 553, 598, 603, 622, 634, 635, 638, 655, 657, 660, 671, 677, 678, 695, 696, 705, 718, 730, 736, 750, 770, 781, 799, 810, 813, 814, 855, 872, 880, 891, 892, 931, 948, 985, 994, 998

A.10 XLM-R$_{\text{large}}$

Feature	Dimension list
Gender: Noun	-
Gender: Adjective	-
Gender: Noun & Adjective	122
Number: Noun	-
Number: Adjective	-
Number: Noun & Adjective	50
POS: Noun	-
POS: Adjective	130
POS: Verb	42
Semantic: Act	-
Semantic: Person	-

Appendix B Classification accuracy

We trained several classifiers to predict the values of feature vectors using all dimensions of word embeddings and the obtained stable dimensions associated with the feature (if any). In this section, you will find the accuracies achieved using stable dimensions only and how they compare to the accuracies of the same classifiers trained using all dimensions of word embeddings.

The used classifiers were K-Nearest Neighbors (KNN), Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT), in the implementation of `scikit-learn`. All accuracies were obtained using 5-fold cross-validation.

In many cases despite sharp decrease in the number of used dimensions, the achieved accuracies are comparable to the whole word embedding vector. However, it is worth noting that for the LR classifier (which was used in the initial setup to retrieve best candidates) the observed accuracies are notably lower in a lot of cases.

B.1 Gender

B.1.1 Gender: Noun

Model	Cased	KNN	NB	LR	RF	DT	Number of stable dimensions
$\text{FlauBERT}_{\text{small}}$	✓	0.763 (↑ 0.066)	0.827 (↓ 0.035)	0.921 (↑0.03)	0.835 (↑0.092)	0.635 (↑0.014)	100
DistilBERT	✓	0.537 (↓ 0.03)	0.59 (↑ 0.035)	0.59 (↓ 0.116)	0.542 (↓ 0.067)	0.521 (↓ 0.006)	4
$\text{CamemBERT}_{\text{base}}$	✓	-	-	-	-	-	0
$\text{mBERT}_{\text{base}}$		-	-	-	-	-	0
$\text{mBERT}_{\text{base}}$	✓	-	-	-	-	-	0
$\text{FlauBERT}_{\text{base}}$		0.561 (↓0.001)	0.6 (↓0.0308)	0.607(↑0.011)	0.603 (↑0.015)	0.543 (↑0.003)	123
$\text{FlauBERT}_{\text{base}}$	✓	0.616 (↑0.031)	0.68 (↑0.041)	0.718 (↓0.041)	0.683 (↑0.019)	0.56(↑0.008)	140
$\text{XLM-R}_{\text{base}}$	✓	-	-	-	-	-	0
$\text{FlauBERT}_{\text{large}}$	✓	0.851 (↑0.072)	0.904 (↑0.018)	0.921(↓0.014)	0.886 (↑0.005)	0.717 (↑0.018)	203
$\text{XLM-R}_{\text{large}}$	✓	-	-	-	-	-	0

Table 9: Accuracies of classifiers trained to predict gender of nouns using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

B.1.2 Gender: Adjective

Model	Cased	KNN	NB	LR	RF	DT	Number of stable dimensions
$\text{FlauBERT}_{\text{small}}$	✓	0.718 (↓0.023)	0.791(↓0.144)	0.911 (↓0.076)	0.804 (↓0.126)	0.622 (↓0.0976)	175
DistilBERT	✓	0.549 (↓0.063)	0.597 (↑0.01)	0.6 (↓0.191)	0.598 (↓0.035)	0.537 (↓0.0168)	18
$\text{CamemBERT}_{\text{base}}$	✓	0.494 (↓0.011)	0.518 (↓0.003)	0.514 (↓0.035)	0.501 (↓0.033)	0.499 (↑0.008)	1
$\text{mBERT}_{\text{base}}$		-	-	-	-	-	0
$\text{mBERT}_{\text{base}}$	✓	0.543(↑0.01)	0.5 (↓0.026)	0.489 (↑0.019)	0.532 (↓0.076)	0.504 (↓0.041)	5
$\text{FlauBERT}_{\text{base}}$		0.529 (↓0.017)	0.571 (↓0.033)	0.582 (↓0.076)	0.575 (↓0.016)	0.528 (↑0.003)	26
$\text{FlauBERT}_{\text{base}}$	✓	0.58 (↓0.025)	0.617 (↑0.05)	0.625 (↓0.197)	0.624 (↓0.066)	0.546 (↓0.036)	22
$\text{XLM-R}_{\text{base}}$	✓	-	-	-	-	-	0
$\text{FlauBERT}_{\text{large}}$	✓	0.815 (↑0.042)	0.863 (↓0.044)	0.896 (↓0.071)	0.859 (↓0.053)	0.685 (↓0.051)	134
$\text{XLM-R}_{\text{large}}$	✓	-	-	-	-	-	0

Table 10: Accuracies of classifiers trained to predict gender of adjectives using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

B.1.3 Gender: Noun & Adjective

Model	Cased	KNN	NB	LR	RF	DT	Number of stable dimensions
FlauBERT$_{small}$	✓	0.775 (↑0.071)	0.844 (↑0.111)	0.908 (↓0.054)	0.848 (↑0.056)	0.715 (↑0.104)	74
DistilBERT	✓	0.59 (0)	0.561 (↓0.01)	0.617 (↓0.088)	0.626 (↓0.003)	0.522 (↓0.043)	14
CamemBERT$_{base}$	✓	-	-	-	-	-	0
mBERT$_{base}$		0.531 (↑0.003)	0.55 (↑0.035)	0.561 (↓0.033)	0.558 (↓0.005)	0.542 (↑0.021)	3
mBERT$_{base}$	✓	-	-	-	-	-	0
FlauBERT$_{base}$		0.548 (↓0.012)	0.588 (↑0.021)	0.592 (↓0.055)	0.593 (↑0.007)	0.540 (↑0.007)	47
FlauBERT$_{base}$	✓	0.646 (↑0.054)	0.606 (↑0.055)	0.72 (↓0.043)	0.693 (↑0.023)	0.595 (↑0.038)	48
XLM-R$_{base}$	✓	-	-	-	-	-	0
FlauBERT$_{large}$	✓	0.84 (↑0.036)	0.92 (↑0.022)	0.964 (↑0.024)	0.909 (↑0.019)	0.744 (↑0.057)	227
XLM-R$_{large}$	✓	0.467 (↓0.117)	0.509(↓0.057)	0.471 (↓0.152)	0.502 (↓0.102)	0.503 (↓0.027)	1

Table 11: Accuracies of classifiers trained to predict the gender of nouns and adjectives using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

B.2 Number

B.2.1 Number: Noun

Model	Cased	KNN	NB	LR	RF	DT	Number of stable dimensions
FlauBERT$_{small}$	✓	0.861 (↑0.07)	0.969 (↑0.033)	0.991 (0)	0.947 (↑0.011)	0.769 (↑0.013)	213
DistilBERT	✓	0.616 (↓0.004)	0.615 (↑0.01)	0.713 (↓0.108)	0.676 (↓0.005)	0.564(↓0.002)	22
CamemBERT$_{base}$	✓	-	-	-	-	-	0
mBERT$_{base}$		-	-	-	-	-	0
mBERT$_{base}$	✓	-	-	-	-	-	0
FlauBERT$_{base}$		0.602 (↑0.032)	0.666 (↑0.041)	0.693 (↓0.017)	0.674 (↑0.032)	0.565 (↑0.018)	120
FlauBERT$_{base}$	✓	0.717 (↑0.033)	0.72 (↑0.115)	0.749 (↓0.065)	0.746 (↑0.004)	0.642 (↑0.003)	22
XLM-R$_{base}$	✓	0.572 (↑0.036)	0.59 (↑0.107)	0.602 (↓0.139)	0.553 (↓0.042)	0.563 (↓0.029)	2
FlauBERT$_{large}$	✓	0.95 (↑0.09)	0.954 (↑0.005)	0.972 (↓0.01)	0.943905 (0)	0.827 (↑0.012)	107
XLM-R$_{large}$	✓	-	-	-	-	-	0

Table 12: Accuracies of classifiers trained to predict the number of nouns using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

B.2.2 Number: Adjective

Model	Cased	KNN	NB	LR	RF	DT	Number of stable dimensions
FlauBERT$_{small}$	✓	0.889 (↑0.148)	0.957 (↑0.021)	0.986 (↓0.002)	0.943 (↑0.017)	0.773 (↑0.06)	107
DistilBERT	✓	-	-	-	-	-	0
CamemBERT$_{base}$	✓	-	-	-	-	-	0
mBERT$_{base}$		-	-	-	-	-	0
mBERT$_{base}$	✓	-	-	-	-	-	0
FlauBERT$_{base}$		0.549 (↑0.003)	0.585 (↓0.019)	0.583 (↓0.075)	0.567 (↓0.027)	0.541 (↑0.018)	4
FlauBERT$_{base}$	✓	0.655 (↑0.05)	0.689 (↑0.122)	0.693 (↓0.129)	0.671 (↓0.022)	0.593 (↑0.015)	4
XLM-R$_{base}$	✓	0.5 (↓0.059)	0.526 (↓0.003)	0.523 (↓0.137)	0.53 (↓0.061)	0.51 (↓0.052)	1
FlauBERT$_{large}$	✓	0.936 (↑0.163)	0.955 (↑0.049)	0.974 (↑0.007)	0.94 (↑0.027)	0.831 (↑0.096)	151
XLM-R$_{large}$	✓	-	-	-	-	-	0

Table 13: Accuracies of classifiers trained to predict the number of adjectives using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

B.2.3 Number: Noun & Adjective

Model	Cased	KNN	NB	LR	RF	DT	Number of stable dimensions
$\text{FlauBERT}_{\text{small}}$	✓	0.888 (↑0.103)	0.966 (↑0.035)	0.989 (↓0.003)	0.945 (↑0.016)	0.774 (↑0.028)	127
DistilBERT	✓	0.606 (↑0.012)	0.653 (↑0.068)	0.674 (↓0.125)	0.662 (↓0.013)	0.577 (↑0.004)	15
$\text{CamemBERT}_{\text{base}}$	✓	0.518 (↓0.009)	0.511 (↓0.002)	0.518 (↑0.009)	0.522 (↓0.012)	0.518 (↑0.011)	9
$\text{mBERT}_{\text{base}}$		0.534 (↓0.008)	0.555 (↑0.007)	0.554 (↓0.097)	0.555 (↓0.035)	0.539 (↓0.002)	4
$\text{mBERT}_{\text{base}}$	✓	-	-	-	-	-	0
$\text{FlauBERT}_{\text{base}}$		0.601(↑0.025)	0.661 (↑0.05)	0.683 (↓0.0195)	0.668 (↑0.032)	0.553 (↑0.011)	91
$\text{FlauBERT}_{\text{base}}$	✓	0.682 (↑0.02)	0.698 (↑0.104)	0.7 (↓0.105)	0.702 (↓0.036)	0.615 (↓0.014)	7
$\text{XLM-R}_{\text{base}}$	✓	-	-	-	-	-	0
$\text{FlauBERT}_{\text{large}}$	✓	0.943 (↑0.074)	0.957 (↑0.014)	0.974 (↓0.007)	0.946 (↑0.004)	0.833 (↑0.027)	195
$\text{XLM-R}_{\text{large}}$	✓	0.517 (↓0.084)	0.566 (↓0.027)	0.58 (↓0.087)	0.545 (↑0.08)	0.545 (↓0.005)	1

Table 14: Accuracies of classifiers trained to predict the number of nouns and adjectives using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

B.3 POS

B.3.1 POS: Noun

Model	Cased	KNN	NB	LR	RF	DT	Number of stable dimensions
$\text{FlauBERT}_{\text{small}}$	✓	0.934 (↑0.01)	0.897 (↑0.037)	0.937 (↓0.027)	0.914 (↓0.0003)	0.794041 (↑0.008)	92
DistilBERT	✓	0.561 (↓0.046)	0.608 (↓0.001)	0.603 (↓0.106)	0.565 (↓0.076)	0.535 (↓0.029)	3
$\text{CamemBERT}_{\text{base}}$	✓	0.507 (↓0.068)	0.531 (↑0.008)	0.532 (↓0.048)	0.492 (↓0.079)	0.493 (↓0.037)	1
$\text{mBERT}_{\text{base}}$		0.506 (↓0.033)	0.54 (↑0.005)	0.547 (↓0.043)	0.503 (↓0.075)	0.501 (↓0.019)	1
$\text{mBERT}_{\text{base}}$	✓	0.526 (↓0.052)	0.502 (↓0.002)	0.468 (↓0.1)	0.505 (↓0.14)	0.503 (↓0.079)	1
$\text{FlauBERT}_{\text{base}}$		0.726 (↑0.008)	0.727 (↑0.09)	0.734 (↓0.051)	0.76 (↓0.006)	0.657 (↑0.011)	20
$\text{FlauBERT}_{\text{base}}$	✓	0.637 (↓0.025)	0.637 (↑0.06)	0.678 (↓0.107)	0.677 (↓0.028)	0.577 (↓0.01)	18
$\text{XLM-R}_{\text{base}}$	✓	0.507 (↓0.053)	0.549 (↑0.011)	0.551 (↓0.119)	0.512 (↓0.106)	0.507 (↓0.046)	1
$\text{FlauBERT}_{\text{large}}$	✓	0.916 (↑0.039)	0.903 (↑0.063)	0.932 (↓0.026)	0.902 (↑0.008)	0.768 (↑0.02)	101
$\text{XLM-R}_{\text{large}}$	✓	-	-	-	-	-	0

Table 15: Accuracies of classifiers trained to predict the POS of noun vs non-nouns using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

B.3.2 POS: Adjective

Model	Cased	KNN	NB	LR	RF	DT	Number of stable dimensions
$\text{FlauBERT}_{\text{small}}$	✓	0.93 (↑0.042)	0.862 (↓0.025)	0.927 (↓0.019)	0.909 (↑0.007)	0.783 (↑0.031)	77
DistilBERT	✓	-	-	-	-	-	0
$\text{CamemBERT}_{\text{base}}$	✓	-	-	-	-	-	0
$\text{mBERT}_{\text{base}}$		-	-	-	-	-	0
$\text{mBERT}_{\text{base}}$	✓	-	-	-	-	-	0
$\text{FlauBERT}_{\text{base}}$		0.598 (↓0.063)	0.593 (↓0.009)	0.607 (↓0.111)	0.62 (↓0.055)	0.562 (↓0.002)	5
$\text{FlauBERT}_{\text{base}}$	✓	0.531 (↓0.075)	0.555 (↓0.003)	0.556 (↓0.186)	0.541 (↓0.120)	0.516 (↓0.045)	5
$\text{XLM-R}_{\text{base}}$	✓	0.524 (↓0.028)	0.562 (↓0.007)	0.562 (↓0.007)	0.522 (↓0.024)	0.519 (↑0.046)	1
$\text{FlauBERT}_{\text{large}}$	✓	0.839 (↓0.018)	0.826 (↓0.039)	0.863 (↓0.08)	0.843 (↓0.051)	0.718 (↓0.026)	58
$\text{XLM-R}_{\text{large}}$	✓	0.52 (↑0.004)	0.503 (↓0.059)	0.498 (↓0.096)	0.504 (↓0.056)	0.504 (↑0.011)	1

Table 16: Accuracies of classifiers trained to predict the POS of adjectives vs non-adjectives using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

B.3.3 POS: Verb

Model	Cased	KNN	NB	LR	RF	DT	Number of stable dimensions
$\text{FlauBERT}_{\text{small}}$	✓	0.875 (↓0.05)	0.792 (↓0.111)	0.835 (↓0.138)	0.848 (↓0.096)	0.754 (↓0.089)	24
DistilBERT	✓	0.596 (↓0.001)	0.618 (↑0.012)	0.621 (↓0.075)	0.616 (↓0.038)	0.545162 (↓0.04)	5
$\text{CamemBERT}_{\text{base}}$	✓	-	-	-	-	-	0
$\text{mBERT}_{\text{base}}$		-	-	-	-	-	0
$\text{mBERT}_{\text{base}}$	✓	-	-	-	-	-	0
$\text{FlauBERT}_{\text{base}}$		0.736 (↓0.033)	0.723 (↑0.041)	0.74 (↓0.092)	0.762 (↓0.044)	0.659 (↓0.02)	49
$\text{FlauBERT}_{\text{base}}$	✓	0.623 (↓0.088)	0.645 (↑0.05)	0.673 (↓0.163)	0.675 (↓0.086)	0.578 (↓0.045)	22
$\text{XLM-R}_{\text{base}}$	✓	-	-	-	-	-	0
$\text{FlauBERT}_{\text{large}}$	✓	0.89 (↑0.022)	0.854 (↓0.023)	0.894 (↓0.075)	0.875 (↓0.056)	0.752 (↓0.029)	73
$\text{XLM-R}_{\text{large}}$	✓	0.506 (↓0.081)	0.576 (↓0.02)	0.579 (↓0.077)	0.496 (↓0.136)	0.496 (↓0.07)	1

Table 17: Accuracies of classifiers trained to predict the POS of verbs vs non-verbs using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

7.4 Semantic supersenses

7.4.1 Act

Model	Cased	KNN	NB	LR	RF	DT	Number of stable dimensions
$\text{FlauBERT}_{\text{small}}$	✓	0.801 (↑0.083)	0.795 (↑0.022)	0.831 (↓0.009)	0.809 (↑0.02)	0.701 (↑0.055)	48
DistilBERT	✓	0.5 (↓0.03)	0.542 (↓0.007)	0.493 (↓0.157)	0.526 (↓0.062)	0.523 (↓0.032)	3
$\text{CamemBERT}_{\text{base}}$	✓	0.512 (↓0.023)	0.515 (↓0.013)	0.498 (↓0.027)	0.487 (↓0.058)	0.487 (↓0.035)	1
$\text{mBERT}_{\text{base}}$		-	-	-	-	-	0
$\text{mBERT}_{\text{base}}$	✓	-	-	-	-	-	0
$\text{FlauBERT}_{\text{base}}$		0.55 (↓0.014)	0.58 (↓0.05)	0.587 (↓0.091)	0.518 (↓0.114)	0.533 (↓0.025)	2
$\text{FlauBERT}_{\text{base}}$	✓	0.615 (↓0.001)	0.616 (↑0.033)	0.638 (↓0.082)	0.626 (↓0.022)	0.557 (↑0.020070)	14
$\text{XLM-R}_{\text{base}}$	✓	0.433 (↓0.086)	0.49 (↓0.009)	0.481 (↓0.19)	0.49 (↓0.076)	0.49 (↓0.052)	1
$\text{FlauBERT}_{\text{large}}$	✓	0.766 (↑0.019)	0.799 (↑0.016)	0.819 (↓0.019)	0.798 (↑0.01)	0.667 (↑0.029)	50
$\text{XLM-R}_{\text{large}}$	✓	-	-	-	-	-	0

Table 18: Accuracies of classifiers trained to classify nouns into 2 categories: `Act` vs non-`Act`. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

7.4.2 Person

Model	Cased	KNN	NB	LR	RF	DT	Number of stable dimensions
$\text{FlauBERT}_{\text{small}}$	✓	0.81 (↑0.052)	0.905 (↑0.04)	0.915 (↑0.010)	0.877 (↓0.002)	0.74 (↑0.015)	95
DistilBERT	✓	0.524 (↓0.057)	0.545 (↓0.052)	0.443 (↓0.292)	0.5 (↓0.134)	0.5 (↓0.093)	1
$\text{CamemBERT}_{\text{base}}$	✓	-	-	-	-	-	0
$\text{mBERT}_{\text{base}}$		-	-	-	-	-	0
$\text{mBERT}_{\text{base}}$	✓	0.634 (↑0.017)	0.508 (↓0.004)	0.468 (↑0.008)	0.675 (↓0.033)	0.618 (↓0.008)	6
$\text{FlauBERT}_{\text{base}}$		0.581 (↑0.067)	0.638 (↑0.047)	0.646 (↓0.008)	0.568 (↑0.003)	0.529 (↓0.022)	5
$\text{FlauBERT}_{\text{base}}$	✓	0.555 (↓0.048)	0.567 (↓0.023)	0.55 (↓0.188)	0.547 (↓0.133)	0.545 (↓0.03)	1
$\text{XLM-R}_{\text{base}}$	✓	0.532 (↑0.022)	0.547 (↑0.023)	0.445 (↓0.095)	0.563 (↑0.001)	0.563 (↑0.118)	1
$\text{FlauBERT}_{\text{large}}$	✓	0.798 (↑0.06)	0.843 (↓0.005)	0.861 (↑0.002)	0.815 (↓0.007)	0.675 (↑0.05)	63
$\text{XLM-R}_{\text{large}}$	✓	-	-	-	-	-	0

Table 19: Accuracies of classifiers trained to classify nouns into 2 categories: `Person` vs non-`Person`. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.