How good BERT based models are in sentiment analysis of Croatian tweets: comparison of four multilingual BERTs

Martina Ptiček
Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, 42000 Varaždin, Croatia
marpticek@foi.hr

Abstract. *Contextual word embeddings like BERT or GPT give the state-of-the-art results in a vast array of tasks in NLP - especially when applied to English datasets, given the fact that these models themselves were trained on numerous data in English language.*
However, the successfulness of these models has not yet been sufficiently researched for low resource languages, as Croatian.
This paper describes a comparison between the application of BERT based multilingual word embeddings (mBERT, DistilBERT, XLM-RoBERTa, CroSloEngual) in sentiment analysis on tweets in Croatian language.
The article shows that BERT based multilingual models give good results in sentiment analysis in Croatian language, particularly the models trained on larger sets of data in Croatian as XLM-RoBERTa and CroSloEngual.

Keywords. Sentiment analysis, contextual word embeddings, multilingual BERT, Croatian language

1 Introduction

Social networks, blogs, comments and reviews provide a valuable source of the large amount of (unstructured) data. More than ever before, prior to making a purchase or deciding which movie to watch, internet users today turn to Internet pages where they can find other users experiences with the items that interests them. Likewise, social networks or comments on news portals are a source of information on the user's opinions on various political events and also about the politicians.

The availability of information as described, has motivated many scientists to start developing systems for automatic recognition of sentiments and opinions that users express and have towards products, events, or public figures.

Sentiment analysis is a research field within the field of Natural Language Processing (NLP) that has the aim of determining "whether a text, or a part of it, is subjective or not and, if subjective, whether it expresses a positive or negative view." (Taboada, 2016, p. 326). Although determination of polarity of a statement is at the heart of sentiment analysis, its methods enable us to also research attitudes in newspaper articles or to determine political perspective or mood in blogs (Pang, B., Lee, L., 2008, p. 23). However, these tasks are not without challenges, first of them being the determination of whether the text is subjective or objective. As explained by Pang and Lee (2008, p. 12), "patterns like "the fact that" do not necessarily guarantee the objective truth of what follows them - and bigrams like "no sentiment" apparently do not guarantee the absence of opinions, either.".

There are currently three dominant approaches in sentiment analysis. The first one is with the use of dictionary (which can be built manually or automatically) and the second one is with using traditional methods of machine learning as Support Vector Machine (SVM) or logistic regression. These two approaches do not necessarily exclude one another so there are also models where machine learning method is combined with the use of dictionaries.

The third approach in sentiment analysis is the use of deep learning methods and word embeddings (Habimana et.al., 2019). With the application of contextual word embedding, as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et.al., 2019) state-of-the-art results have been achieved during last years in performing numerous tasks in NLP so the models have also found their application in text classification, including the sentiment analysis (Sun et.al. 2019, Polignano et.al. 2019, Pota et.al. 2020). The advantage of contextual word embeddings over the static ones is that they situate words in context, depending on the words that appear after of before a particular word.

The aim of this research is to apply BERT based word embeddings on a dataset in Croatian language in

order to learn which one gives the best result in sentiment analysis of tweets in Croatian.

The structure of the paper is as follows: in the chapter Background and Related works, word embedding and BERT contextual word embeddings (mBERT, DistilBERT, XLM-RoBERTa, CroSloEngual) are described. The latter ones will be used in conducting the sentiment analysis on tweets in Croatian in order to evaluate their efficiency. This chapter also gives a short overview of an already implemented research where sentiment analysis of texts in Croatian was conducted by using the mBERT model.

Further on, in the chapter Dataset, information is given on the dataset which is used in this research and the method used for pre-processing the data. The chapter on Methodology provides an overview of methods that were used in sentiment analysis, including the parameters of models used in classification. Finally, in chapters Results and Conclusion, results are presented together with the conclusions based on these results which are accompanied with suggestions for further research.

2 Background and related works

Traditional approach to sentiment analysis is with the use of dictionaries (which is a demanding task) and with statistical methods of machine learning as SVM and logistic regression and the use of vector semantics as TF-IDF or PPMI. The use of vector semantics is described as "the standard way to represent word meaning in NLP" (Jurafsky and Martin, 2020) which dates to 1950. On the other hand, word embedding is a more recent method but relying on the same concept of presenting the words in a vector space.

Word embeddings can be classified as static ones like word2vec (Mikolov et.al. 2013) or GloVe (Pennington et.al, 2014) and as contextual word embeddings as BERT, ELMo (Peters et.al. 2018) and GPT (Radford et.al., 2018, 2019, Brown et.al. 2021). The main shortage of static word embeddings is that they do not represent meaning in context or put differently, the method itself learns one static embedding for a single word in the vocabulary (Jurafsky and Martin, 2020). On the other hand, contextual embeddings provide for a vector representation of a word which depends on each single context that a particular word has been appearing in. Comparing the results on nine NLP tasks Tenney et.al. (2019) have shown that contextual word representations give better results than non-contextual ones, especially in „syntactic tasks (e.g. constituent labelling) in comparison to semantic tasks (e.g. coreference), suggesting that these embeddings encode syntax more so than higher-level semantics." (Tenney et.al., 2019, p. 10).

BERT and GPT word embeddings both are trained on Transformer architecture that was presented by Vaswani et.al. (2017) as a possible solution to the problem of computational efficiency while using the Recurrent neural networks (RNN) and Long Short Term Memory (LSTM) models. As these authors further note, Transformer model architecture is „relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNN or convolution" (Vaswani et.al., 2017, p. 2). Transformer architecture enables performing parallelization and thus increases the efficiency and significantly reduces training time. In their work the authors have shown that this architecture achieves better results in machine translation than the previous developed models, regarded as state-of the-art. Besides both BERT and GPT being trained on Transformer architecture, the main difference between these two models is that BERT is bidirectional, while GPT is left-to-right architecture.

Although both of the described models give good results in NLP tasks the research presented in this paper was performed by using BERT based models. The reason why BERT models were chosen, is the fact that at the time of writing this article, there are no publicly available GPT models pre-trained on Croatian texts while there were four publicly available BERT based multilingual models pre-trained (also) in Croatian – mBERT, DistilBERT, XLM-RoBERTa and CroSloEngual.

As shown by Devlin et.al. (2019), the BERT model achieves state-of-the-art results in many natural language processing tasks and the best results have been achieved in GLUE (the increase of 7.7%), MultiNLI accuracy (the increase of 4.6%), SQuAD v1.1. question answering (the increase in F1 score of 1.5%) and SQuAD v2.0 (the increase in F1 score of 5.1%).

mBERT (Multilingual Bert)[1] is released by Devlin et.al (2019) as a single language model, pre-trained on 104 different languages with the biggest number of entries on Wikipedia, also including entries in Croatian. Piers et.al. (2019) and Karthikeyen et.al. (2020) have shown that mBERT gives excellent results in cross-lingual models, where a model was trained on one language and then tested on another one.

Relying on BERT and its multilingual version, Sanh et.al (2019) train DistilBERT, which is basically identical to BERT but a smaller, faster, and cheaper model. The model itself was developed due to the costliness and vast negative environmental impact pertinent to the training of BERT model. With the use of knowledge distillation compression technique Sanh et.al. (2019) train this new model with the identical general architecture as the one in BERT model, the difference being in the removal of token-type embeddings and pooler while the number of layers is reduced by a factor of 2. DistilBERT still achieves

[1] https://github.com/google-research/bert

97% of BERT performance, while it is 40% smaller and 60% faster.

XLM-RoBERTa (Conneau et.al., 2020) is another BERT based model. It is pre-trained on texts on 100 different languages and more than two terabytes of data filtered from CommonCrawl data. As stated by the authors "XLM-RoBERTa (XLM-R) outperforms mBERT on cross-lingual classification by up to 23% accuracy on low-resource languages." This model was also trained on Croatian dataset and according to the authors the overall amount of data used for training the model on Croatian was 20.5 GB (3297 million tokens) out of which a smaller part are texts from Wikipedia (approx. 20%) and the rest of data are data from CommonCrawl (approx. 80%).

CroSloEngual (Ulčar and Robnik-Šikonja, 2020) is also a multilingual BERT model trained on only three languages - Croatian, Slovenian and English where 31% of data (27 million tokens) used in training was in Croatian language (47% in English and 23% in Slovenian language). The data in Croatian are a mix of news articles and general web crawl.

As to the knowledge of this paper's author, the use of the described BERT models on Croatian texts has been very limited. Moreover, only a single research has been done in this field - Pelicon et.al. (2020) perform the zero-shot cross-lingual news sentiment classification, by using the model on texts on Slovenian language and test it on Croatian texts. The results achieved by these authors are F1 66.33% for Slovenian and F1 54.77% for Croatian language. However, their research was performed on significantly different data set then the one that is used in this research and the authors themselves identify as a problem the fact that mBERT model accepts input of a fixed length which led to the need to perform experiments with the sequence length.

As for the sentiment analysis of tweets in other languages, Polignano et.al. (2019) have introduced AlBERTo, BERT based model trained on Italian language and specifically on Italian tweets, which has achieved state of the art results at the time. Pota et.al. surpass this score and manage to achieve even better result by introducing NLP pipeline with pre-processing procedure in first step and using BERT in second step. Nguyen et.al. (2020) represent publicly available pre-trained language model for English Tweets, BERTweet, pre-trained on 850M Tweets. BERTweet outperforms RoBERTa and XLM-RoBERTa models in POS tagging, NER and text classification (i.e. sentiment analysis and irony detection) tasks.

3 Dataset

The dataset used in this research consist of tweets in Croatian language that were collected by a group of researchers in 2016 (Možetić et.al., 2016) and is publicly available[2]. All of the tweets from this dataset were annotated by one single annotator and classified as positive, negative or neutral.

Since the dataset contains URLs of tweets the preliminary step in this research was to fetch them from Twitter. After this was done, it was determined that some tweets were no longer available (they were deleted) so the initial dataset prepared by Možetić et. al. (2016) containing 97.921 tweets was reduced to 47.276 tweets. Furthermore, in the fetched dataset few tweets were doubled due to the fact that in some doubtful cases the annotator annotated the same tweet more than once, each time with a different sentiment. These tweets were removed from the dataset used in this research in order to avoid any confusion for the model. After all above described preparatory steps were done, the final dataset prepared for this research contained 35.880 tweets.

Furthermore, before the tweets were processed, pre-processing was done which involved the removal of URLs, usernames and all special characters with the use of regex, as well as turning all letters in the texts into small letters.

Table 1 Number of tweets in each class and the total number for train set, test dataset and complete dataset

Positive	Negative	Neutral	Total
Train set			
14.517	6.793	7.394	28.704
Test set			
3.629	1.699	1.848	7.176
Complete dataset			
18.146	8.492	9.242	35.880

Finally, in regard to the preparation of the test and the train set, the sentiment of the fetched tweets was taken into account. Namely, as shown in Table 1 the dataset as retrieved from Twitter was not balanced in terms of the tweets' sentiments and the tweets with positive sentiment prevailed (18.146) over the negative ones (8.492) or the ones annotated as neutral (9.242). In order for the test set to be as much as possible similar to the train set, the test set was prepared by using stratified sampling and it was created by taking 20% of tweets from each of the categories (positive, negative and neutral).

4 Methodology

In this research four different multilingual models based on Transformer architecture and BERT (mBERT, DistilBERT, XLM-RoBERTa and CroSloEngual) were tested. Furthermore, Transformers library was used (Wolf et.al., 2020)

[2] http://hdl.handle.net/11356/1054

which is published under Apache 2.0 licence and available on the Github[3]. The models used were the ones that are publicly available retrieved through the Transformers library, as follows: for mBERT *bert-base-multilingual-cased* (12-layer, 768-hidden, 12-heads, 179M parameters), for DistilBERT *distilbert-base-multilingual-cased* (6-layer, 768-hidden, 12-heads, 134M parameters) and for XLM-RoBERTa *xlm-roberta-base* (12-layer, 768-hidden, 8-heads, ~270M parameters). CroSloEngual (12-layer, 768-hidden, 110M parameters) model is also publicly available for download at Clarin.si repository[4]. These models were also used for tokenization of text.

According to the recommendations by Devlin et.al. (2019), each BERT based model was fine-tuned by changing the number of epochs (2, 3 and 4 epochs) and the learning rate (2e-5, 3-e5, 5-e5) while the batch size is always 32 (Devlin et.al recommend batch size to be 16 or 32). AdamW, that is available at Transformers library, was used as the optimizer. Altogether nine experiments were conducted with each of the pre-trained models following the combination of parameters as presented in the Table 2. For each of the models used the best F1 weighted score was monitored and noted, as well as precision, recall and F1 score per class.

For the purpose of making the comparison between the results obtained with these experiments with the results obtained by using traditional machine learning methods and sentiment analysis, an additional step was taken. Classification was done with the use of the machine learning method which is known to give good results in text classification and in sentiment analysis, namely the Support Vector Machine (SVM) (Joachims, 1998) with the linear core. TF-IDF method was used to build the vocabulary, with 10.000 most often used words. Results obtained by SVM were used as baseline data, while keeping in mind that this method is significantly different and that detailed comparison of SVM and neural network approach would require an additional research which is outside of the scope of the research presented in this paper.

All the experiments were done in Google Colaboratory, with the use of GPU while the exception was the SVM where CPU was used.

5 Results

As presented in Table 3 the best results were obtained with the CroSloEngual model – F1 score of 71.42%. Furthermore, even the worst result obtained by CroSloEngual (F1 70.66%) was better than the best result obtained when using the model that has proved to be the second best – XLM-RoBERTa – F1 score 70.33%. XLM-RoBERTa is then followed by mBERT model with F1 score 66.04%. Finally, DistilBERT model gives the least successful result among the models tested, since its best F1 score is 65.93%. All of the results show that the changes of hyper-parameters have effect on the results from 1 to 3.3 points.

Table 2 Hyper parameters of the models – each model was tested with each of the listed parameters.

No.	Learning rate	Epochs	Batch size
1	2e-5	2	32
2	2e-5	3	32
3	2e-5	4	32
4	3e-5	2	32
5	3e-5	3	32
6	3e-5	4	32
7	5e-5	2	32
8	5e-5	3	32
9	5e-5	4	32

Table 3 Review of weighted F1 scores for each of the models used and each of the conducted experiments, according to the hyper-parameters presented in Table 2. Best results are marked in bold.

	mBERT	DistilBERT	XLM-RoBERTa	CroSlo-Engual
1	64.10%	63.25%	69.71%	70.98%
2	65.46%	63.98%	69.84%	**71.42%**
3	65.75%	65.48%	69.93%	71.34%
4	64.68%	63.95%	69.41%	71.10%
5	**66.04%**	65.20%	**70.33%**	71.35%
6	65.70%	65.10%	69.41%	70.86%
7	65.56%	64.90%	68.79%	71.15%
8	64.70%	65.07%	68.80%	70.66%
9	65.38%	**65.93%**	69.41%	70.74%
SVM		62.01%		

In addition to the mentioned results, the SVM model, which was used for making the comparison with the traditional machine learning models and without the use of word embedding gave the F1 score 62.01%. This result is a bit less successful than the least successful one obtained by using the BERT based model, namely DistilBERT with weighted F1 score 63.25% but significantly less than the best result achieved with the use of CroSloEngual model (71.42%)

If we take a closer look into precision, recall and F1 score per class of the best results for each of the tested models presented in Table 4, it can be noticed that best results with all models were achieved for tweets annotated as positive, while poor results were obtained for the neutral tweets. It is worth noting that classification of negative tweets also was less successful than classification of positive ones, however the set for training contained half the amount of neutral and negative tweets in comparison to positive ones. Therefore it is necessary that future

[3] https://github.com/huggingface/transformers
[4] http://hdl.handle.net/11356/1317

research involve tests with balanced datasets that would contain the equal amount of positive and negative examples.

Additionally, we can notice that recall is always higher than precision for positive and negative tweets, which is not the case for neutral ones, where recall is always smaller than the precision.

Out of altogether 7.176 tweets from the test set, all models have accurately classified 3.695 (51.49%), and among those 2.515 were positive, 870 negative and only 310 neutral. Altogether 1002 (13.96%) tweets out of which 619 were neutral, 214 positive and 169 negative were not accurately classified by any of the models.

A more detailed insight into 1049 neutral tweets that were misclassified by at least one model, shows that 505 of them were equally misclassified by all models. In other words, all models consistently classify some neutral tweets as positive and other ones as negative, so there is agreement between models on tweet sentiment. The overlap in misclassification between the models also appears in case of negative and positive tweets, but this is much less often than with neutral tweets. Namely, 99 negative and 140 positive tweets were misclassified, but equally, by all models.

The further analysis of neutral tweets shows that their annotation can in fact be put to question so the accuracy of their classification by models should further be checked by reviewing the initial annotation. Few of the examples of tweets annotated as neutral but equally misclassified by all models are shown in table 5. Although relevant for this type of research, due to the extent of the task of reviewing annotations, this remains to be done in future research. Within the context of this paper, it is interesting to point that all of the models have classified equally a significant number of these tweets annotated as neutral, but whose neutrality can be questioned.

Table 4 Precision, recall and F1 score per class for best results with each of the tested models

Model	Class	Precision	Recall	F1
mBERT	Neg.	62.93%	71.34%	66.87%
	Pos.	73.83%	82.17%	77.78%
	Neut.	53.34%	34.96%	42.24%
DistilBERT	Neg.	61.75%	68.51%	64.96%
	Pos.	74.39%	79.42%	76.82%
	Neut.	52.36%	40.15%	45.45%
XLM-RoBERTa	Neg.	69.03%	74.93%	71.86%
	Pos.	76.37%	86.19%	80.98%
	Neut.	59.87%	40.04%	47.99%
CroSloEngual	Neg.	70.99%	76.93%	73.84%
	Pos.	77.79%	85.12%	81.29%
	Neut.	58.65%	43.29%	49.81%

Also, there is a large quantity of tweets where the results were different depending on the model that was used. Namely, in case of 3.481 tweets at least one of the models gave different classification than the other ones. mBERT and DistilBERT gave different results for 1.464 tweets, which is interesting since these two models are basically the same, except that DistilBERT is smaller, faster, and cheaper than BERT. Furthermore, mBERT and DistilBERT gave different classification than XLM-RoBERTa for 1.543 and for 1.852 tweets respectively, while the results they gave differed from CroSloEngual in 1.639 and 1.874 tweets respectively. XLM-RoBERTa and CroSloEngual differ in classification for 1.268 tweets so the discrepancy between these two models is the minimal. This is expected since these two models give best results and classify accurately the largest number of tweets.

Table 5 Examples of neutral tweets that were equally misclassified by all models, as positive or negative, with English translation.

Tweet	Pred. class
je li to samo meni ili danas i ostalima ne funkcionira ne ucitava nis novoga fb na ios (engl. is it only in my case or others have problems with functioning of loading new things on fb on ios)	Neg.
prehlade i gripe već polako kucaju na vrata pogađate tko će im otvoriti raindrop više (engl. cold and flues are already slowly knocking at the door, you are guessing who will open a raindrop more to them)	Neg.
kako i zbog čega lešinari grade karijere preko nevino ubijene i masakrirane djece (engl. how and why are vultures building their careers over innocently killed and massacred children)	Neg.
suradnja našeg ponajboljeg i najzaposlenijeg producenta Petra Dundova i frankfurtskog techno maga Gregora (engl. cooperation of one of our best and busiest producers Petar Dundov and Gregor techno mage from Frankfurt)	Pos.
Prelijepa Anita Dujić Veljača modna blogerica chinasoulmate komentira kakav nakit voli i blista u kampanji (engl. beautiful Anita Dujić Veljača fashion blogger chinasoulmate comments on the jewelery that she likes and shines in the campaign)	Pos.
dolje kod glavnog ulaza ima aparat ness s cokoladom 4 kn vrijedi svake lipe (engl. downstairs at the main entrance there is a machine with ness chocolate 4 HRK worths every penny)	Pos.

Overall, the research has shown that better results are obtained with the use of word embeddings, more concretely with using the BERT based models, than with traditional machine learning methods and without word embeddings. In relation to the aim of the research which was to determine which BERT based model gives the best results in sentiment analysis of tweets in Croatian, the research has shown that best results are achieved with the CroSloEngual model. It is noteworthy mentioning that results obtained with XLM-RoBERTa model are very good as well and the use of this model can also be considered when deciding on which model to use in sentiment analysis or in another NLP task in Croatian language.

The results of this research show that the models that have been trained on a bigger dataset on a particular language and a smaller total number of languages will give better results for that particular language. This, of course, is expected and additionally, it has already been set forth by few authors (e.g. Martin et.al., 2020) that monolingual models give better results than multilingual.

Furthermore, in the interpretation of the results it is necessary to take into consideration that the language used on Twitter is colloquial language and that it contains foreign expressions (as English words in Croatian) and the jargon typical of Twitter. On the other hand, mBERT and DistilBERT models that were used in this research were both trained on standard language, which is used in writing articles on Wikipedia. The XLM-RoBERTa model is mostly trained on data from CommonCrawl, which contains texts from web pages that mostly contain texts written in standard language, with the exception of texts from internet forums, comments and reviews which surely use colloquial language. CroSloEngual model is trained on the mix of news articles and general web crawl where probably standard Croatian language prevails, while colloquial language is less present.

All of the specificities mentioned point to the need for further research, also the ones where BERT based models trained on tweets would be used and research where the models used in this research would be applied to sentiment analysis of datasets in standard Croatian language. Likewise, what significantly influences the results is also the domain that the model was trained in. Although BERT word embeddings are contextual, even better results can be expected from the BERT model that would be pre-trained on the texts from the domain that is the same one as the domain of the texts used in sentiment analysis.

6 Conclusion

In this paper an overview of word embedding was given, especially of the contextual word embeddings. Also, description was given of BERT based models trained on many languages.

The central part of the paper presented the methodology and the results of the application of BERT based models to sentiment analysis of tweets in Croatian language. The conducted research has shown that the use of BERT based pre-trained language models gives good results in the sentiment analysis in Croatian language, especially when models like CroSloEngual and XLM-RoBERTa are used, which are the models trained on a bigger dataset in Croatian language.

Further research in this context is necessary – it would be relevant to determine what would be the score achieved in sentiment analysis of tweets with BERT based model pre-trained on Croatian tweets but also to apply models used in this research to various different datasets in Croatian language and to conduct researches in other fields of natural language processing (e.g. POS tagging, NER, question answering etc.), in order to further determine their efficiency.

One of the challenges in performing research as described is surely the lack of publicly available annotated datasets in Croatian. This is why creating quality datasets is necessary as a preliminary step in further testing the models that were used in this research.

References

Brown, B.T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss,A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., J., Clark, Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei D. (2020). *Language Models are Few-Shot Learners*, Retrieved from https://arxiv.org/pdf/2005.14165.pdf

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzman, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440-8451, doi: 10.18653/v1/2020.acl-main.747, Retrieved from https://arxiv.org/pdf/1911.02116.pdf

Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, Retrieved from https://arxiv.org/pdf/1810.04805.pdf

Habmina, O., Li, Y., Li, R., Gu, X., Yu, G. (2020). Sentiment analysis using deep learning approaches: an overview, *Science China*, 63, doi: 10.1007/S11432-018-9941-6

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features, *Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science* (Lecture Notes in Artificial Intelligence), vol 1398. Springer, Berlin, Heidelberg

Jurafsky, D., Martin, J. H. (2020) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Stanford, Retrieved from https://web.stanford.edu/~jurafsky/slp3/

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, E. V., Seddah, D., Sagot, B. (2020). *CamemBERT: a Tasty French Language Model*, Retrieved from https://arxiv.org/pdf/1911.03894.pdf

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality, *Proceedings of NIPS, Advances in Neural Information Processing Systems 26*, 2:3111-3119, Retrieved from https://arxiv.org/pdf/1310.4546.pdf

Mozetič, I., Grčar, M., Smailović, J. (2016) *Multilingual Twitter Sentiment Classification: The role of Human Annotators*, PLoS ONE 11(5), doi:10.1371/journal.pone.0155036, Retrieved from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0155036

Nguyen, D.Q., Vu, T., Nguyenn, A.T. (2020). BERTweet: A pre-trained language model for English Tweets, *Proceedings of the 2020 EMNLP (Systems Demonstrations),* pp. 9–14, Retreived from https://aclanthology.org/2020.emnlp-demos.2.pdf

Pang, B., Lee, L. (2008). *Opinion mining and sentiment analysis*, 2008., Foundation and trends in Information Retrieval, 2(1-2):1–135

Pelicon, A., Pranjić, M., Miljković, D., Škrlj, B., Pollak, S. (2020). *Zero-Shot Learning for Cross-Lingual News Sentiment Classification*, Applied Science, doi:10.3390/app10175993, Retrieved from https://www.mdpi.com/2076-3417/10/17/5993

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer. L. (2018). Deep contextualized word representations, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1:2227-2237, doi:10.18653/v1/N18-1202, Retrieved from https://www.aclweb.org/anthology/N18-1202.pdf

Pennington, J., Socher, R., Manning, C. (2014). GloVe: Global vectors for word representation, *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1523-1543, doi:10.3115/v1/D14-1162, Retrieved from https://www.aclweb.org/anthology/D14-1162.pdf

Piers, T., Schlinger, E., Garrette, D. (2019). How multilingual is Multilingual BERT?, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL, 4996-500, doi: 10.18653/v1/P19-1493, Retrieved from https://www.aclweb.org/anthology/P19-1493.pdf

Polignano, M., Basile. P., de Gemmis, M., Semeraro, G., Basile, V. (2019). ALBERTO: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, Volume 2481,* Retrieved from http://ceur-ws.org/Vol-2481/paper57.pdf

Pota, M., Ventura, M., Catelli, R., Esposito, M. (2021). An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian, *Sensors*, 21(1), 133, doi:10.3390/s21010133, Retrieved from https://www.mdpi.com/1424-8220/21/1/133

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*, OpenAI, Retrieved from https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu., J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*, OpenAI, Retrieved from https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis, *Association for Computational Linguistics*, 37(2):267-307, Retrieved from https://direct.mit.edu/coli/article-pdf/37/2/267/1798865/coli_a_00049.pdf

Taboada, M. (2016) Sentiment Analysis: An overview from Linguistics, *Annual Review of Linguistics.* 2: 325-347

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R.T., Kim, N., Van Durme, B., Bowman, S.R., Das, D., Pavlick, E. (2019). *What do you learn from context? probing for sentence structure in contextualized word representations*. Retrieved from https://arxiv.org/pdf/1905.06316.pdf

Sanh, V., Debut, L., Chaumond, J., Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller,faster, cheaper and lighter*, Retrieved from https://arxiv.org/pdf/1910.01108.pdf

Sun, C.; Qiu, X.; Xu, Y.; Huang, X. (2019). How to Fine-Tune BERT for Text Classification?, *Chinese Computational Linguistics-18th China National Conference*, Kunming, China, 18–20

Ulčar, M., Robnik-Šikonja, M. (2020). *FinEst and CroSloEngual BERT: less is more in multilingual models*, Retrieved from https://arxiv.org/pdf/2006.07890.pdf

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.Alexander M. Rush *Transformers: State-of-the-Art Natural Language Processing*, Retrieved from https://arxiv.org/pdf/1910.03771.pdf

Vaswani A, Shazeer N, Parmar N, Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need, *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, Long Beach, 6000–6010, Retrieved from https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf