BERT for Stock Market Sentiment Analysis

Matheus Gomes de Sousa
FACOM/UFMS
Campo Grande, Brazil
matheusgs.gomes@gmail.com

Kenzo Sakiyama
FACOM/UFMS
Campo Grande, Brazil
kenzosakiyama@gmail.com

Lucas de Souza Rodrigues
FACOM/UFMS
Campo Grande, Brazil
lucas.rodrigues@ifms.edu.br

Pedro Henrique de Moraes
FACOM/UFMS
Campo Grande, Brazil
pedhmoraes@gmail.com

Eraldo Rezende Fernandes
FACOM/UFMS
Campo Grande, Brazil
eraldo@facom.ufms.br

Edson Takashi Matsubara
FACOM/UFMS
Campo Grande, Brazil
edsontm@facom.ufms.br

***Abstract*—When breaking news occurs, stock quotes can change abruptly in a matter of seconds. The human analysis of breaking news can take several minutes, and investors in the financial markets need to make quick decisions. Such challenging scenarios require faster ways to support investors. In this work, we propose the use of *Bidirectional Encoder Representations from Transformers* (BERT) to perform sentiment analysis of news articles and provide relevant information for decision making in the stock market. This model is pre-trained on a large amount of general-domain documents by means of a self-learning task. To fine-tune this powerful model on sentiment analysis for the stock market, we manually labeled stock news articles as positive, neutral or negative. This dataset is freely available and amounts to 582 documents from several financial news sources. We fine-tune a BERT model on this dataset and achieve 72.5% of F-score. Then, we perform some experiments highlighting how the output of the obtained model can provide valuable information to predict the subsequent movements of the Dow Jones Industrial (DJI) Index.**

***Index Terms*—natural language processing, sentiment analysis, stock market**

I. Introduction

News is one of the main drivers of the financial markets for abrupt changes in stock prices. Because of this, financial market analysts need to constantly monitor and evaluate financial news to support stock buying and selling decisions. However, in the financial market stock prices can vary quickly and often the time of reading the text, which can take a few minutes, can cost millions of dollars due to a late decision. Another factor that hinders the individual analysis of the news is the amount of information generated by the hundreds of sources of information. Thus, two problems arise: the quantity of news and time of analysis of the news. The increase in the amount of news requires more prolonged reading and analysis time, and in contrast, a reduction in the time of response time requires analysing less news. The solution to one problem implies worsening the other problem, which makes it difficult to solve using traditional techniques.

A possible solution for quick analysis of a large volume of news is the use of computational algorithms for automatic analysis of text using Natural Language Processing (NLP). One of the tasks in NLP is the sentiment analysis, that seeks to identify the feeling, whether positive, negative or neutral, of texts. In this work, we believed that the sentiment analysis applied to financial news could improve the quality of the decisions of financial agents.

In Mäntylä et al. [1] shows a literature review and the evolution of sentiment analysis. Tubishat et al. [2] presents a literature review on sentiment analysis focusing on aspect level. Zimbra et al. [3] is another literature review using Twitter. According to Zimbra et al. [3], the overall average sentiment classification accuracy is around 61% for a general-purpose system, and domain-specific approaches improve this performance by an average of 11%. This improvement is generally related to domain-specific indicators of sentiment that helps the domain-specific models.

The literature review presented by Zhang et al. [4] reviewed sentiment analysis using deep learning techniques. According to this study, the standard approach for text representation for more modern methods are based on word embeddings [5]. The most frequently learning models are recurrent neural networks such as LSTM [6] and GRU [7]. Also, studies using Attention Mechanism [8], [9] in sentiment analysis start to grow. Studies that performs a welcome combination of embeddings, bidirecional strategies and attention mechanism starts to appear in the literature beating many state of the art algorithms.

A recent paper based on attention mechanism called *Bidirectional Encoder Representations from Transformers* (BERT) [10] obtained state-of-the-art results on eleven natural language processing tasks. The pre-trained BERT model was fine-tuned with one additional output layer to create these state-of-the-art models. Our proposal in this paper is to evaluate BERT on financial news sentiment analysis problem to improve stock market prediction. As a short paper, this research is under development, and we show preliminary results.

Thus, this work aims to experimentally evaluate BERT in the task of stock market sentiment analysis. Further steps of this research will focus on improving stock market prediction. So far, the main contributions of this work are listed below.

- Corpus of 582 financial news manually labeled with senti-

2375-0197/19/$31.00 ©2019 IEEE
DOI 10.1109/ICTAI.2019.00231

ment[1] from CNBC, Forbes, New York Times, Washington Post, Business Insider and other news websites.
- BERT code extended for fine-tuning on sentiment analysis[2], and additional code needed to reproduce this work[3] are all freely available.
- Experimental evaluation comparing BERT, Support Vector Machines, Naive Bayes, and Convolutional Neural Network.
- Data analysis highlighting the relation between the Dow Jones Industrial index and the developed BERT sentiment classifier.

The rest of the paper is organised as follows. Section II summarises BERT. Section III describes the main three parts of the proposal. Section IV shows the experiments conducted to validate the effectiveness of the proposal. Finally, Section V concludes the study.

II. BERT: Bidirectional Encoder Representations from Transformers

Pre-trained generic language models [10]–[13] have achieved great results on different NLP tasks. Such models are unsupervisedly trained on large amounts of text and may later be applied to potentially any task. BERT [10] is one of the most successful language model available. This model is based on transformer encoder [14]. The Transformer is a sequence-to-sequence architecture based solely on attention mechanisms for both the encoder and the decoder. BERT architecture ignores the decoder network, using only a transformer encoder, since it is not a sequence-to-sequence model (although it can be used in such tasks).

Most language models are based on unidirectional architectures, i.e., outputs are conditioned only on previous words (left context). When applying such models on downstream tasks, fine tuned models are also limited to be left conditioned. This is a limitation for tasks in which the whole text is available during prediction. BERT introduces a bidirectional language model architecture in order to explore such knowledge. Sentiment analysis is modeled as text classification and, thus, can benefit from this aspect.

In Figure 1, we illustrate the basic BERT architecture. The input for the network is the token representation vectors E_i, which is equal to the sum of three representation vectors for each token: a typical word embedding vector, a position embedding vector and a sentence vector. The position embedding provides the model with information about the position of the token within its sentence, since transformer models do not have this notion. The sentence vector is used only when the task requires a context broader than a sentence, which is not the case for sentiment analysis (we consider a document as a sentence).

The attention-based layers produce, for each input token (E_i for the first layer $T_i^{(1)}$, for instance), a representation ($T_i^{(2)}$)

[1]https://drive.google.com/open?id=1eqNwkqb1tnaJm_l975K6LJBic8pMof1x

[2]https://github.com/stocks-predictor/bert

[3]https://github.com/stocks-predictor/stocks-time-series

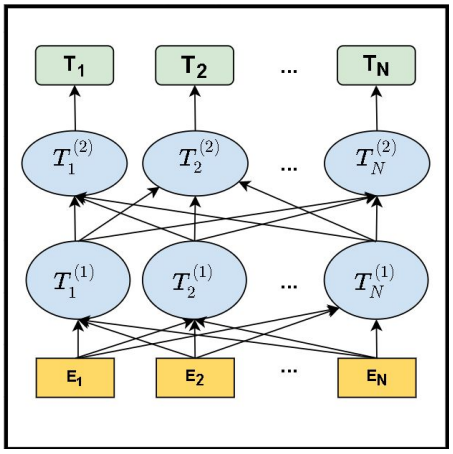

Fig. 1. BERT architecture with a two-layer encoder. E_i, for $i = 1, 2, \ldots, N$, are the input representation vectors (one vector for each input token w_i). $T_i^{(1)}$ are the attention-based representation in the first encoder layer, and $T_i^{(2)}$ are the same in the second encoder layer. And $T_i = T_i^{(2)}$ are the output representation vectors, again, one per token. (Adapted from [10])

computed as a (adaptive) weighted sum of the representations of all tokens within the sentence. This is the main strength of transformer models, i.e., each token representation is based on the representations of all the tokens. Thus, the context is limited only by the input sentence. The output of on attention-based layer is provided as input for the next one. The output of the last attention layer comprises the model output.

In BERT, each input sentence is augmented with an initial artificial token denoted [CLS], as can be observed in Figure 2. When fine-tuning the model on a text classification task, the

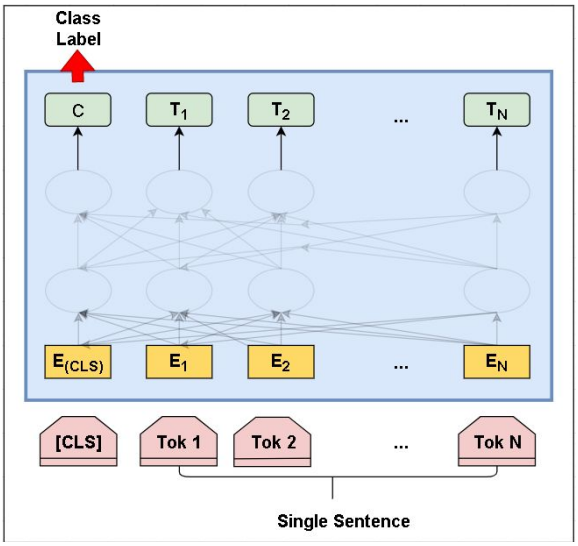

Fig. 2. Sentiment classification using BERT. (Adapted from [10])

output representation of this artificial token is used to feed the classification layer, which is a typical softmax layer.

In the following, we give more details about the pre-trained BERT model employed in this work.

III. Proposal

The objective of the proposal is to indicate the trend of the Dow Jones Index before the opening time. We estimate the sentiment of the market using financial news before its opening, and we used to predict the DJI day trend.

The proposal can be split into three parts: (1) collecting and pre-processing stock news articles; (2) BERT-based model for sentiment analysis; (3) and leveraging the developed model to improve decision making related to stock market prediction. Figure 3 illustrates these three parts so the following sections describe the details of each part.

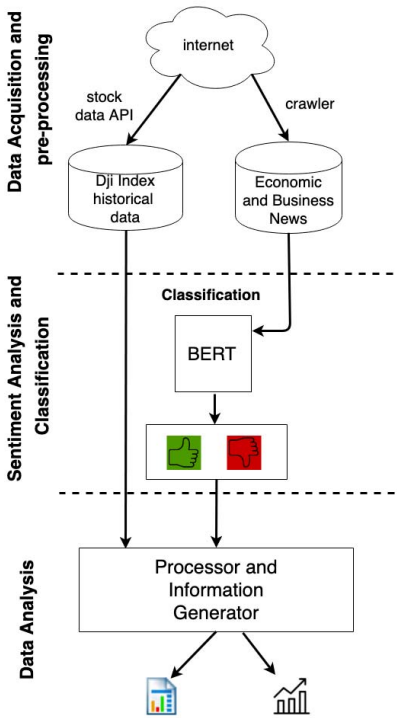

Fig. 3. Proposal

A. Data Acquisition and Pre-Processing

News articles were collected from different website sources showed in Table I. The data was collected from May 26th to February 4th, 2019. We crawled the news articles using Selenium tool [15]. Four volunteered members from our research group to manually labeled dataset as positive, neutral or negative sentiment.

After data acquisition, each document is transformed in a token sequence. The tokenization is made using WordPiece [16] with a 30,000 token vocabulary. On the use of WordPiece, it allows BERT even with a 3000 "words" vocabulary, able to tokenize almost every single word in the English language.

TABLE I
WEBSITES USED TO COLLECT NEWS TO BUILD THE CORPUS.

Source	# Articles	(#pos #neu #neg)	%
Business Insider	51		8,7
CNBC	77		13.2
Forbes	32		5.5
Investopedia	41		7.0
New York Times	45		7.7
Washington Post	31		5.3
Others	305		52.4
Total	582		100.0

In addition to that, the Alpha Vantage API [17] collect the historical data of the Dow Jones Industrial Average (DJI) index in the same time of the news articles.

B. Sentiment Analysis

The creators of BERT proposed two models with different values for the parameters L - layers, H - hidden layer size, A - attention heads: a smaller called **BERT BASE** com L = 12, H = 768 e A = 12 and a bigger called **BERT LARGE** com L = 24, H = 1024 e A = 16. In this research due to our limited computational power we used smaller BERT BASE.

We fine-tuned this pre-trained BERT BASE model using our labeled set. For experimental evaluation purposes, we performed 10 fold cross-validation, and for the running model, we use the model trained using all labeled data.

C. Data Analysis

The data analysis evaluates the financial news mood before the stock market opening time. The idea is to reproduce the scenario where a financial agent is restricted to operate only at the stock market opening time. Therefore, in this part, the system estimates the mood of the available news between OT - HB and OT, where OT is the opening time, and HB (hours before) is a parameter. The proportion of positive news within this time frame is used to indicate the direction of DJI.

IV. Experimental Evaluation

This section evaluates the performance of BERT when compared with naive Bayes, support vector machines (SVM) [18] and TextCNN [19]. The first two algorithms require tabular data format and we converted the texts into bag-of-words (bow) and term frequency inverse document frequency (tfidf) [20] representation. For TextCNN, we used the average vector of word embeddings obtained from fastText [21]. We performed a 10 fold-cross validation procedure to evaluate the learning algorithms. In Table II, we show performance by means of accuracy, precision, recall and F1. The best results are presented in boldface. We adjust the parameters of SVM using Random Search [22] with 20 iterations varying `C` in a exponential scale of 100, `gramma` in a exponential scale of .1, and using a RBF kernel.

Clearly, that BERT outperformed the other methods. When performing a paired t-test (p-value = 0.05) we find a significant difference between BERT and TextCNN.

TABLE II
EXPERIMENTAL RESULTS WITH 10 FOLD CROSS VALIDATION. THE NUMBER AFTER ± REPRESENTS STANDARD DEVIATION.

Algorithm	Accuracy	Precision	Recall	F1
NB bow	0.610 ± 0.060	0.593 ± 0.196	0.557 ± 0.069	0.503 ± 0.103
SVM bow	0.628 ± 0.063	0.627 ± 0.074	0.609 ± 0.066	0.601 ± 0.071
NB tfidf	0.610 ± 0.062	0.607 ± 0.102	0.568 ± 0.065	0.542 ± 0.080
SVM tfidf	0.624 ± 0.076	0.631 ± 0.104	0.595 ± 0.083	0.578 ± 0.099
textCNN	0.739 ± 0.05	0.703 ± 0.18	0.500 ± 0.14	0.569 ± 0.12
BERT	**0.825 ± 0.04**	**0.750 ± 0.17**	**0.713 ± 0.16**	**0.725 ± 0.15**

For a more detailed analysis, we constructed the ROC curve of BERT results (Figure 4). The area under the ROC curve (ROC AUC) is 0.87, which indicates how good the model can be to distinguish between positive and negative classes.

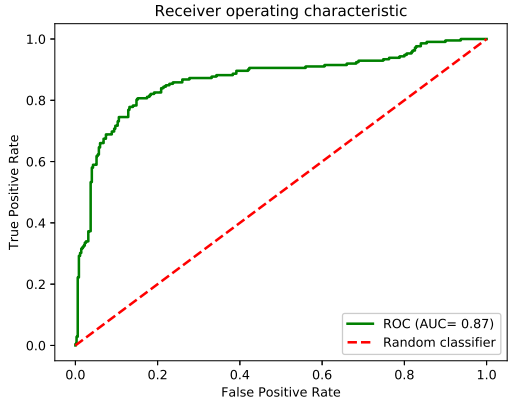

Fig. 4. ROC Curve for Bert Classifier

A. Analysis of the Time Series of the Stock Market

We evaluate the relationship between the market sentiment and the time series of the stock exchange using BERT to classify the news that was collected by the crawler for approximately 1 month and the DJI. Thus, we plotted the time series of the positive news rate during each hour, overlapping the stock exchange variation, and we plotted the moving average considering the previous 10 hours, to smooth the curve, as seen in Figure 5. Note that the DJI index score was normalized using the minmax_scale method of the `sklearn` [23] tool. By the chart, we do not find much correlation with the analysis of feeling with the Dow Jones.

However, to verify if the analysis of feelings can be useful in identifying the trend of falling or rising the index in the day was adopted the following strategy. At the beginning of each day, 5 hours before the stock exchange opened (HB=5), the average sentiment of the news was calculated. We hypothesize that the average sentiment that precedes the stock market opening more strongly indicates the mood of the market in the period that the stock exchange is closed.

This mood was compared to the opening and closing stock market index so that it could assess whether the rating of the news really is indicative of stock market fluctuations. Therefore, positive sentiment was considered, whenever the positive news rate in the period was higher than 50%, otherwise it would be considered negative, see Table III.

TABLE III
DOW JONES INDEX OPENINGS AND CLOSINGS DATA AND THE SENTIMENT OF THE NEWS IN THE PERIOD.

Date	Open DJI	Close DJI	Sentiment	DJI Variation
08-04-2019	0.56	0.40	negative	decrease
09-04-2019	0.40	0.08	negative	decrease
11-04-2019	0.08	0.06	negative	decrease
12-04-2019	0.06	0.52	positive	increase
15-04-2019	0.52	0.49	positive	decrease
16-04-2019	0.49	0.58	positive	increase
17-04-2019	0.58	0.64	positive	increase
18-04-2019	0.64	0.82	negative	increase
22-04-2019	0.82	0.72	positive	decrease
23-04-2019	0.72	0.95	positive	increase
24-04-2019	0.95	0.90	positive	decrease
25-04-2019	0.90	0.68	negative	decrease
26-04-2019	0.68	0.72	positive	increase

Therefore, in the analyzed period, 69 % of the periods between opening and closing the stock market, the sentiment of the news was consistent with the stock exchange variation. However, the collection period was short, and more extended periods must be evaluated to verify if the observed behaviour is significant.

V. CONCLUSION

The results indicate that BERT has superior performance than the convolutional neural networks and word embeddings approach in the order of 8.6% when compared to the hit rate(accuracy). The results comparing the time series of sentiment analysis of the news and the Down Jones index are very noisy and difficult to analyze. We use the sentiment analysis of economic news as an indicator of falling or rising in the day. The proposal achieved 69% hit rate in the prediction of stock exchange variation. Although the data collection period was short, the data presented in the III table gives a good indication of the effectiveness of the implemented predictor.

Down Jones Industrial Average index includes several stocks such as MSFT (Microsoft), INTC (Intel), BA (Boeing), among others. Only observing the DJI can be challenging to evaluate which stocks have better chance to change, as there may be cases where the significant increase in the price of a particular stock may not increases the index. This discussion shows us that, as future work, one can extract specific news from individual companies and do data processing and analysis on the value of the shares of those companies. Also, as an extension of this work, one could observe news about a company, collect its accounting data and build a more precise predictor.

REFERENCES

[1] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—a review of research topics, venues, and top cited papers," *Computer Science Review*, vol. 27, pp. 16–32, 2018.

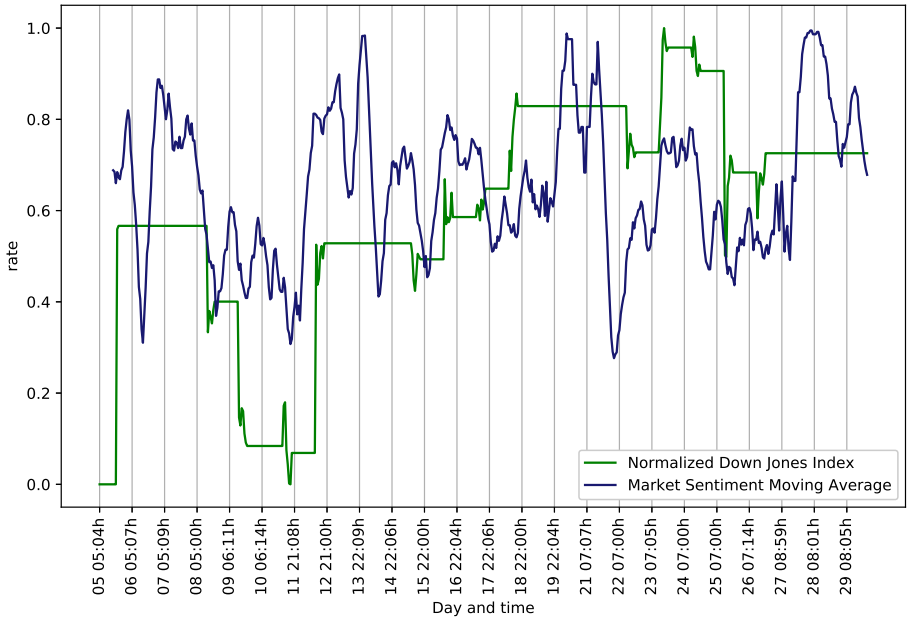

Fig. 5. Variation of Normalized Down Jones Index (Green) and Market Sentiment Moving Average (Blue).

[2] M. Tubishat, N. Idris, and M. A. Abushariah, "Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges," *Information Processing & Management*, vol. 54, no. 4, pp. 545–563, 2018.

[3] D. Zimbra, A. Abbasi, D. Zeng, and H. Chen, "The state-of-the-art in twitter sentiment analysis: a review and benchmark evaluation," *ACM Transactions on Management Information Systems (TMIS)*, vol. 9, no. 2, p. 5, 2018.

[4] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[8] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[11] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

[12] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. of NAACL*, 2018.

[13] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *CoRR*, 2019.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, 2017.

[15] SeleniumHQ, "Selenium automates browsers," https://www.seleniumhq.org/.

[16] Z. C. Q. V. L. M. N. Yonghui Wu, Mike Schuster, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.0814*, 2016.

[17] A. V. I. . 2019, "Free APIs for realtime and historical financial data, technical analysis, charting, and more!" https://www.alphavantage.co/.

[18] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[19] S. Rosenthal, N. Farra, and P. Nakov, "Semeval-2017 task 4: Sentiment analysis in twitter," in *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, 2017, pp. 502–518.

[20] A. Aizawa, "An information-theoretic perspective of tf–idf measures," *Information Processing & Management*, vol. 39, no. 1, pp. 45–65, 2003.

[21] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[22] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281–305, 2012.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.