Available online at
https://www.asjp.cerist.dz/en/PresentationRevue/134
RIST
Information Processing at the Digital Age Journal

CERIST Natural Language Processing Challenge

March 29th, 2023.

GigaBERT-based Approach for Hate Speech Detection in Arabic Twitter

Bachir Said [1,2] and Mohammed E. Barmati [1,2]

[1] *Kasdi Merbah University, Ghardaia Road, BP.511, Ouargla, 30,000, Algeria*
[2] *Laboratory of Artificial Intelligence and Information Technologies (LINATI)*

Abstract

Natural Language Processing has recently become one of the most trending research areas in Artificial Intelligence, especially in social media-related tasks. This paper describes our participation in the "Hate Speech Detection on Arabic Twitter" task at the CERIST NLP-Challenge 2022 competition. The proposed solution aims to classify the tweets collected in the Arabic ARACOVID19-MFH multi-label and multi-dialect dataset into "Hateful" and "Not Hateful" categories. Based on a pre-trained transformer model known as GigaBERT-v4, our solution outperformed the most common transformer models supporting the Arabic language. Experiments have proved that the GigaBERT-v4 model is more effective than the other models using the previously described dataset, obtaining a 99.46% accuracy and a 98.68% macro F1-score.

Keywords: Arabic Twitter; hate speech detection; COVID'19; multilingual transformers; GigaBERT-v4; XLM-T; AraBERT; mBERT.

---

1. Introduction

In the past four years, the use of transformers in Natural Language Processing (NLP) problem-solving has emerged as the dominant trend in research advancements. Transformer-based pre-trained language models (e.g., BERT Develin et al.,2019, XLNet Yang et al.,2019, RoBERTa Liu et al., 2019, ... are frequently used to solve a variety of NLP problems, including Sentiment Analysis, Machine Translation, and Hate Speech Detection.

Detecting hate speech on social networks has also become a difficult challenge, given the prevalence of hate speech on social media and the lack of an agreed-upon definition of hate speech due to the diversity of languages, dialects, and cultural traditions among societies. On the other hand, since existing transformer models have been pre-trained on a significant quantity of data from numerous web sources, they are the most promising techniques for detecting hate speech.

Detecting hate speech in Arabic text is more challenging than in English since Arabic has a wide vocabulary, distinct traits, and numerous dialects. Instead, the Arabic vocabulary size in most multilingual pre-trained models is smaller than the English vocabulary size.

There are two principal deep learning-based approaches for resolving Arabic NLP challenges in the literature. The first approach combines a word embedding method, such as Tf-Idf or Word2Vec, with a deep learning model, such as CNN, RNN, or LSTM. The second method employs pre-trained Arabic-supporting transformer models. Some transformer models are only fine-tuned in Arabic (i.e., AraBERT Antoun et al., 2020, a pre-trained BERT model specifically for the Arabic language, with a 58k Arabic vocabulary size). However, most of them are multilingual (i.e., XLM-RoBERTa Conneau et al., 2020, which is pre-trained on 2.5TB of filtered data containing 100 languages, with 14k Arabic vocabulary size, and GigaBERT Lan, et al.,2020 a customized BERT for English-to-Arabic cross-lingual transfer, with 26k Arabic vocabulary size).

This article describes our solution to the hate speech detection task submitted to CERIST[1] NLP-Challenge 2022 competition[2]. The proposed solution is based on a promising bilingual transformer model called GigaBERT-v4 Lan et al., which outperformed AraBERT, XLM-ROBERTa$_{base}$, and multilingual BERT (mBERT) Develin et al., 2019 in four Arabic information extraction tasks. Still, as far as we know, it has not yet been fine-tuned with any hate speech dataset. We also conducted a comparative experimental performance evaluation of hate speech detection on the ARACOVID19-MFH multi-label dataset Hadj Ameur and Aliane, 2021.

The remaining sections are organized as follows. Section 2 includes related work on Arabic hate speech detection using transformers. Section 3 then describes the proposed model. Section 4 describes the experiments conducted on the proposed and state-of-the-art models and discusses the results. Section 5 concludes the paper with some concluding observations and research objectives for the future.

2. Transformers-based Arabic Hate Speech Detection

Several Arabic hate speech detection techniques based on transfer learning have been proposed in the past four years. Numerous researchers have utilized existing datasets and corpora for fine-tuning pre-trained transformer models. Others built new single or multiple-task datasets from diverse data sources and social network content for one or multiple tasks.

HadjAmeur and Aliane, 2021 created a ten-label manually annotated Arabic COVID-19 fake news and hate speech detection dataset. To demonstrate the usefulness of their dataset, they trained and evaluated pre-trained transformer models, such as AraBERT, mBERT, and Distilbert Multilingual, using it. In Barbieri et al.,2022, the authors proposed a multi-task learning classification method for offensive language and hate speech. The BERT-based multi-task learning model was trained using cross-corpora and evaluated on three datasets in modern standard Arabic language and Tunisian and Levantine dialects. Emojis were considered in Althobaiti, 2022 due to their increasing significance in social media content. The authors investigated the use of sentiment analysis and textual emoji descriptions as additional tweet features to improve the performance of the BERT-based model for detecting hate speech. In a departure from previous work, Magnossão et al.,2022 applied two ensemble approaches in addition to six transformer models: majority vote and Highest sum, which outperformed the official dataset baselines.

3. Materials and Methods

The proposed Arabic hate speech detection model shown in Figure 1 is based on GigaBERT-v4-Arabic-and-English Lan et al., 2020 which, as far as we are aware, has not yet been fine-tuned using any hate speech dataset. GigaBERT-v4 has the same configuration as BERT$_{base}$, according to Lan et al., 2020: 12 attention layers, 12 self-attention operations, 768 hidden sizes per layer, and 110 million parameters.

The outperformance of GigaBERT-v4 over AraBERT, XLM-ROBERTa$_{base}$, and mBERT models in named entity recognition (NER), part-of-speech tagging (POS), relation extraction (RE), and argument role labeling (ARL) tasks justified our choice. In addition, GigaBERT-v4 uses 26k Arabic vocabulary size for 4.3B Arabic training data from a variety of sources (Arabic Gigaword Parker et al., 2009, Wikipedia, Oscar Javier et al., 2019 with code-switching).

Detecting hate speech involves three basic steps: input text preprocessing, model fine-tuning with the ARACOVID19-MFH dataset HadjAmeur and Aliane, 2021, and classification of hate speech.

3.1 Data Preprocessing

As previously stated, we utilized the ARACOVID19-MFH multi-label dataset with the single label "Contains hate." The training dataset comprises 8662 rows, including 986 for Hateful tweets and 7676 for Not Hateful tweets.

Data preprocessing is an important and influential stage in the training process. The initial preprocessing step is text cleaning, during which the following operations are performed: eliminating diacritics, punctuation, repeated characters, hashtags, links, usernames, and emojis. The second step is oversampling the imbalanced data. After visualizing the ARACOVID19-MFH dataset, we observed imbalanced classes in the training dataset (7676/8662 Not Hateful tweets = 88.62% and 986/8662 Hateful tweets = 11.38%). The final step involves tokenizing and encoding the thoroughly cleaned text using the Word Piece method. The outcomes of the preprocessing operations on a sample tweet are summarized in Table 1.

[1] http://www.cerist.dz/
[2] http://www.nlpchallenge.cerist.dz/

Table 1: Preprocessing of an input tweet

Input tweet :	@roo222oofa @l__4i @stop_x2 الله يلعن كورونا والصين هي السبب
Cleaned tweet:	الله يلعن كورونا والصين هي السبب
Tokenized tweet:	['الله', 'يل', '##عن', 'كور', '##ونا', 'والصين', 'هي', 'السبب']

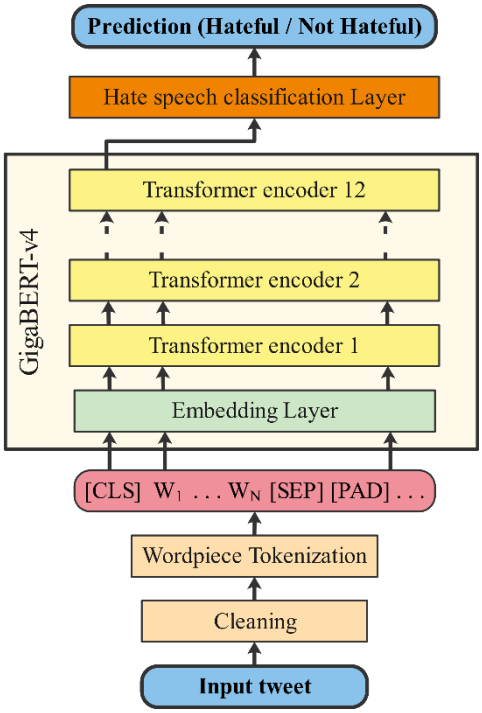

Figure 1: The workflow of the proposed system for hate speech detection of Arabic tweets

3.2 Fine-tuning and hate speech detection

First, we split the training dataset into a training portion of 80% and a validation portion of 20%. Then, we fine-tuned the pre-trained GigaBERT-v4 model using the ARACOVID19-MFH portion (80%) of the training dataset that had been oversampled. We used an AdamW optimizer with a learning rate of lr=5e-5 to optimize the model's parameters. As a prediction function for the binary classification layer, we used *argmax()*.

3.3 Experimental Results and Discussion

The experiments were carried out using Google Collaboratory's[3] python notebook editor. The execution environment consists of a Tesla T4 (15 Gb) GPU, 12,68 Gb of system RAM, and an Intel(R) Xeon(R) 2*CPU running at 2.40GHz.

We compared the proposed fine-tuned GigaBERT-v4 to three pre-trained models: AraBERT, mBERT, and the XLM-T language model, Barbieri et al., 2022. Using the validation portion (20%) of the training dataset, we evaluated the accuracy calculated with equation (1) and the f1-score calculated with equation (2) of all models. Based on the use of the oversampling technique, two experiments were conducted. In the first, all models were trained with the training portion of the training dataset (80%), and no oversampling was performed. In the second experiment, they were trained using oversampled training portion. Knowing that the validation portion (20%) was not oversampled in either experiment. Table 2 displays the accuracy and the macro f1-score for the two experiments, computed and averaged over five consecutive runs.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100\%, \quad (1)$$

[3] https://colab.research.google.com/

$$Macro\ F1 = \left(\frac{1}{2} \sum_{i=1..2} \frac{TP_i}{TP_i + \frac{1}{2} * (FP_i + FN_i)} \right) * 100\%, \quad (2)$$

Where *TP* (true positive), *TN* (true negative), *FP* (false positive), *FN* (false negative), and *i* (the class hateful or not hateful).

Table 2.Experimental results of GigaBERT-v4 and three pre-trained transformer models computed and averaged over five consecutive runs

Model	Without oversampling		Using oversampling	
	Accuracy %	F1-score %	Accuracy %	F1-score %
AraBERT	99.22	98.05	99.35	98.41
mBERT	98.80	97.04	99.04	97.68
XLM-Twitter	98.87	97.17	99.07	97.75
GigaBERT-v4	**99.27**	**98.25**	**99.46**	**98.68**

In both experiments, GigaBERT-v4 outperformed other models in terms of accuracy and macro f1-score, as seen in the table above. In the first experiments, we observed that GigaBERT-v4 is better than AraBERT in terms of accuracy by 0.05% and macro f1-score by 0.2%. GigaBERT-v4 outperformed AraBERT more in the second experiment by 0.11% in accuracy and 0.27% in macro f1-score.

For the Arabic hate speech detection final submission, we selected the GigaBERT-v4 with the oversampling technique. Based on the test dataset, our model shows an F1-Score of 0.9700 and an accuracy of 0.9930.

4. Conclusion

As part of the CERIST NLP Challenge competition, this research addresses detecting Arabic hate speech on social networks. We investigated the most prevalent pre-trained transformer models that performed well on multiple Arabic NLP tasks. We fine-tuned the pre-trained GigaBERT-v4 model with the ARACOVID19-MFH training dataset. Since the training dataset was multi-labels, the "Hateful" class size was significantly smaller than the "Not Hateful" class size. Therefore, we utilized the oversampling technique to achieve equilibrium between different classes. The experimental results demonstrated that oversampling the training portion of the training dataset enhanced the models' accuracy and macro f1-score. GigaBERT-v4 outperformed the other models in terms of accuracy and macro-f1-score in every experiment.

Notably, the training dataset contains multiple dialects. Consequently, the fine-tuning of GigaBERT-v4 utilizing a huge multi-dialect Arabic corpus would enhance its performance on many Arabic NLP tasks.

References

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.

Althobaiti, M.J. (2022). BERT-based Approach to Arabic Hate Speech and Offensive Language Detection in Twitter: Exploiting Emojis and Sentiment Analysis. International Journal of Advanced Computer Science and Applications. Vol. 13, N° 5, (2022). DOI:10.14569/IJACSA.2022.01305109

Angel Felipe Magnossão de Paula, Paolo Rosso, Imene Bensalem, and Wajdi Zaghouani. 2022. UPV at the Arabic Hate Speech 2022 Shared Task: Offensive Language and Hate Speech Detection using Transformers and Ensemble Models. In Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection, pages 181–185, Marseille, France. European Language Resources Association.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 258–266, Marseille, France. European Language Resources Association. Arabic Offensive and Hate Speech Detection Using a Cross-Corpora Multi-Task Learning Model.

Hadj ameur M.S., Aliane H., 2021. AraCOVID19-MFH: Arabic COVID-19 multi-label fake news & hate speech detection dataset. Procedia Computer Science 189, 232-241.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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), pages 4171–4186, 2019.

Pedro Javier Ortiz Su´arez, Benoˆıt Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In Proceedings of the 7thWorkshop on the Challenges in the Management of Large Corpora.

Robert Parker, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2009. Arabic Gigaword. URL: https://catalog.ldc.upenn.edu/LDC2011T11.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based Model for Arabic Language Understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pages 9–15, Marseille, France. European Language Resource Association.

Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. 2020. An Empirical Study of Pre-trained Transformers for Arabic Information Extraction. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4727–4734, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov: RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR abs/1907.11692 (2019).

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: generalized autoregressive pretraining for language understanding. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 517, 5753–5763.