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.

Transformers and Ensemble methods: A solution for Hate Speech Detection in Arabic language

Angel Felipe Magnossão de Paula[a]*, Imene Bensalem[b,c], Paolo Rosso[a] and Wajdi Zaghouani[d]

[a]Universitat Politècnica de València
[b]MISC Lab – Constantine 2 University
[c]ESCF de Constantine
[d]Hamad Bin Khalifa University

Abstract

This paper describes our participation in the shared task of hate speech detection, which is one of the subtasks of the CERIST NLP Challenge 2022. Our experiments evaluate the performance of six transformer models and their combination using 2 ensemble approaches. The best results on the training set, in a five-fold cross validation scenario, were obtained by using the ensemble approach based on the majority vote. The evaluation of this approach on the test set resulted in an F1-score of 0.60 and Accuracy of 0.86.

Keywords: Hate speech detection ; Transformers ; Ensemble Methods ; Arabic.

**Corresponding author.*
E-mail addresses: adepau@doctor.upv.es (A. F. M. d. Paula); ibensalem@escf-constantine.dz (I. Bensalem); prosso@dsic.upv.es (P. Rosso); wzaghouani@hbku.edu.qa (W. Zaghouani)

1. Introduction

The improvement of automatic hate speech detection is an important factor in diminishing the spread of toxicity online, de Paula et al., 2021a. Despite the recent advances in employing attention mechanisms and other deep learning approaches, Alkomah and Ma, 2022, the detection of hate speech is still considered a major challenge, de Paula et al., 2022a, especially, when dealing with social media text written in low-resource languages, such as Arabic and its various dialects. In fact, most of the naturally occurring Arabic text in social media is written in Dialectal Arabic (DA).

The purpose of this paper is to present our approach to address hate speech detection in Arabic text. To this end, in addition to exploring six transformer-based architectures, Vaswani et al., 2017, Devlin et al., 2019, two ensemble methods are studied, de Paula et al., 2022b, de Paula et al., 2021b. It should be noted that some of these architectures were specifically pre-trained in Arabic. Our code is open source and available on GitHub[†].

To carry out our experiments, we used data shared by the organizers of the CERIST NLP Challenge for task 1.d, Arabic hate speech and offensive language detection on social networks (COVID-19). The task is a binary classification problem where a model has to classify an Arabic tweet as Hateful or Not Hateful. The official evaluation metric for task 1.d is F1-score on the positive class (Hateful). In our experiments on the training set, the two highest F1-scores have obtained by employing the Majority Vote ensemble and AraBERT, respectively. Therefore, the Majority Vote ensemble is the method we have applied on the test set.

The remainder of the paper is structured as follows. Section 2 provides an overview of the problem of hate speech in Arabic. Sections 3 and 4 present the dataset details and the models applied. Finally, we close our paper by discussing the results and drawing some conclusions.

2. Related Works

In the past few years, the number of publications on hate speech detection in the Arabic language has taken a leap, Husain and Uzuner, 2021, especially with the organization of shared tasks addressing this research problem.

The first shared task has been organized within the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT 4), Mubarak et al., 2020. It addressed two binary classification tasks: offensive language detection and hate speech detection. The organizers provided a dataset of 10k tweets, wherein 20% are offensive language and 5% are hate speech. The best approach was obtained with the Support Vector Machine (SVM) model using an extensive pre-processing, Hassan et al., 2020. Another version of OSACT 4 dataset was also used in OffensEval shared task on multilingual offensive language identification, Zampieri et al., 2020. Besides the Arabic dataset, the organizers made available, to the participants, datasets in Danish, English, Greek, and Turkish.

In addition to the binary classification tasks, the shared task organized within OSACT 5 Workshop, Mubarak et al., 2022a, addressed the fine-grained hate speech categorization, where each hateful tweet has to be classified into one of the six following categories: race, religion, ideology, disability, social class and gender. The dataset (which was described in Mubarak et al., 2022b) is composed of more than 12k tweets, with 11% labelled as hate speech. The top-ranked approach, Bennessir et al., 2022, used a multitask model based on MARBERT and QRNN.

The detection of hatred against women (i.e., misogyny) has been addressed in ArMI shared task, Mulki and Ghanem, 2021, which proposed both binary and fine-grained classification subtasks. The dataset is composed

[†] https://github.com/AngelFelipeMP/Arabic-Hate-Speech-Covid-19

of more than 9k tweets, where 61% are misogynistic and labelled with one of 7 categories of misogyny. The best system in this shared task, Mahdaouy et al., 2021, combined the outputs of three different versions of MARBERT model using an ensemble approach.

The following sections are devoted to the description of the dataset and the experiments we conducted as part of our participation in the shared task of hate speech detection organized within the CERIST NLP challenge.

3. Dataset

The dataset of the hate speech detection task, shared by the organizers of the CERIST NLP challenge, was collected from Twitter and split into training (80%) and test (20%) subsets. It consists of 10828 tweets, 11% of which are annotated as hate speech. The domain of the dataset is COVID-19 disinformation. It is a multi-label dataset, which has been annotated not only for hate speech detection but also to tackle other tasks such as fake news detection. Further information on the dataset could be found in, Hadj Ameur and Aliane, 2021.

4. Method

This section presents the transformer models we applied to detect the Arabic hate speech and offensive language in social media (COVID-19) for the challenge of task 1 proposed by the CERIST NLP Challenge 2022 organizers. The main features of our proposed transformed-based models are displayed in Table 1.

Table 1. Applied transformers to task 1.d

Version	Size	Block	Language
AraBERT AraELECTRA	Base	Encoder	Arabic
Albert-Arabic	Large		
AraGPT2	Base	Decoder	
mBERT XLM-RoBERTa	Base	Encoder	Multilingual

A transformer is a massive deep learning model based on the self-attention mechanism, Vaswani et al., 2017, Lin et al., 2022. These models were originally built to handle natural language processing tasks, Ravichandiran, 2021. The self-attention mechanism enables the transformer to focus on the crucial information from the input data, helping the model to achieve impressive results. Different unsupervised tasks are applied during the training process, such as mask language modelling, next sequence prediction, etc., Devlin et al., 2019, Mohammed and Ali, 2021. However, these models require large amounts of data to be trained.

Fortunately, some pre-trained transformers are freely available, and generally, the users can select among three possible model sizes which are related to their number of trainable parameters: (i) Base, (ii) Medium, and (iii) Large. In the Table 1's second column, we can observe that apart from Albert-Arabic, Safaya, 2020, which we could use the Large pre-trained size, we adopted the smaller option (Base) for the models given our computational constraints regarding GPU memory. Albert-Arabic uses parameter reduction techniques to reduce the amount of memory required to allocate the pre-trained model to the GPU, which enables us to use its Large version. The transformer's early architecture, Vaswani et al., 2017, was established based on an encoder and a decoder block. Nevertheless, the modern versions embody just one of those. As shown in Table

1’s third column, to solve task 1.d, we adopt five transformers based on the encoder block, Antoun et al., 2020, Antoun et al., a2021, Conneau et al., 2020, Safaya, 2020, and one transformer based on the decoder block, Antoun et al., b2021.

Regarding the language of the text used for training step, the transformers can be divided into monolingual and multilingual models. The first is trained with monolingual data, which means text data in only one language (e.g., Arabic). The latter is trained with data in more than one language. We used four monolingual models trained in Arabic: AraBERT, Antoun et al., 2020, AraELECTRA, Antoun et al., a2021, Albert-Arabic, and AraGPT2, Antoun et al., b2021. Furthermore, we employed two multilingual models trained with documents in around 100 languages: mBERT, Devlin et al., 2019, and XLM-RoBERTa, Conneau et al., 2020.

5. Results and Discussion

This section describes the transformers’ hyper-parameter selection and the five-fold cross- validation carried out during the training phase. Furthermore, in order to boost our predictions, we proposed the use of two ensemble methods: the Majority Vote and the Highest Sum.

Based on, de Paula et al., 2022, de Paula et al., b2021, de Paula and da Silva, 2022, we used a 0.00001 learning rate and a 0.3 dropout percentage for the transformer’s fine-tuning. We adopted a max length of 128 tokens and a batch size of 18 samples during all experiments. In order to find the suitable number of fine-tuning epochs, we carried out a five-fold cross-validation on the training data based on the task 1.d official metric, F1-score on the Hateful class. Table 2 displays the optimal number of fine-tuning epochs for each transformer model.

Table 2. Models’ best epochs

Model	Size
AraBERT	4
AraELECTRA	3
Albert-Arabic	4
AraGPT2	4
mBERT	3
XLM-RoBERTa	1

In Table 3, we evaluate the best models during the cross-validation in terms of F1-score, Precision and Recall calculated for the Hateful class, and also the Accuracy, which considers the two classes (Hateful and Not Hateful). Additionally, as we previously mentioned, we implemented two ensemble methods. The Highest Sum method aggregates the transformers’ output values separately for each class and then selects the class with the highest sum. The Majority Vote method chooses the most predicted class among the transformers, and if there is a tie, it randomly selects one of the tied classes, de Paula et al., 2022.

Analysing Table 3, we can see that the transformer with the best performance regarding F1-score is AraBERT followed by AraGPT2 and AraELECTRA. The other transformers presented a similar performance. The two ensembles also presented competitive results, achieving the first (Majority Vote) and the third (Highest Sum) best F1-score. The Majority Vote ensemble presented impressive results as it achieved the highest Accuracy and Precision. On the other hand, AraELECTRA, Arabic-ALBERT, mBERT, and XLM-RoBERTa achieved good results when it comes to the Recall while performing poorly in the Accuracy, F1-score, and Precision. During fine-tuning, these models focused on predicting the positive class (Hateful expressions), while sacrificing most predictions for the negative class (Not Hateful). Due to the nature of the recall metric, only the samples that belong to the positive class, i.e., hateful samples, were considered. This explains why, despite obtaining high recall results, the rest of the metrics, which take into consideration the prediction of the negative class samples, presented lower performance scores.

Table 3. Training data experiment results using five-folder cross-validation

Model		F1-score	Acc.	Precision	Recall
Ensembles	Majority Vote	0.76	0.95	0.88	0.69
	Highest Sum	0.62	0.87	0.45	0.96
Transformers	AraBERT	0.68	0.95	0.68	0.68
	AraELECTRA	0.61	0.93	0.80	0.50
	Albert-Arabic	0.21	0.17	0.12	0.96
	AraGPT2	0.20	0.11	0.11	1.00
	mBERT	0.20	0.11	0.11	1.00
	XLM-RoBERTa	0.20	0.11	0.11	1.00

The CERIST NLP Challenge accepts, for each participant in the task, only one submission of the test set predictions. Hence, we utilized the model with the best performance in Table 3 regarding F1-score, the Majority Vote ensemble. The organizers communicated that we achieved a 0.60 F1-score and a 0.86 Accuracy in the official test data, which aligns with our results in the training data.

6. Conclusion

This work addressed the problem of hate speech detection for Arabic language by applying six transformer models: AraBERT, AraELECTRA, Albert-Arabic, AraGPT2, mBERT, and XLM- RoBERTa. We also took advantage of the Majority Vote and Highest Sum ensembles to aggregate the transformer's output and improve our final results. Based on the task 1.d official evaluation metric, AraBERT performed the best among the transformers, and Majority Vote ensemble achieved the highest score among all models using the five-fold cross-validation approach on the training data. Hence, we applied Majority Vote to carry out the official prediction based on the test data. In general, the Majority Vote ensembles presented a more robust performance in this task compared with the single transformers approach.

Acknowledgements

This publication was made possible by the NPRP grant 13S-0206-200281 (Resources and Ap- plications for Detecting and Classifying Polarized and Hate Speech in Arabic Social Media) from the Qatar National Research Fund (a member of Qatar Foundation). The findings achieved herein are solely the responsibility of the authors.

References

Alkomah, F., & Ma, X., 2022. A literature review of textual hate speech detection methods and datasets. Information, 13(6), 273.Antoun, W., Baly, F., & Hajj, H., 2020. AraBERT: Transformer-based Model for Arabic Language Understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection (pp. 9-15).

Antoun, W., Baly, F., & Hajj, H., 2021a. AraELECTRA: Pre-Training Text Discriminators for Arabic Language Understanding. In Proceedings of the Sixth Arabic Natural Language Processing Workshop (pp. 191-195).

Antoun, W., Baly, F., & Hajj, H. 2021b. AraGPT2: Pre-Trained Transformer for Arabic Language Generation. In Proceedings of the Sixth Arabic Natural Language Processing Workshop (pp. 196-207).

Bennessir, M. A., Rhouma, M., Haddad, H., & Fourati, C. (2022, June). icompass at arabic hate speech 2022: Detect hate speech using qrnn and transformers. 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 (pp. 176-180).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., & Stoyanov, V., 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440-8451).

de Paula, A. F. M., & Schlicht, I. B., 2021a. AI-UPV at iberlef-2021 DETOXIS task: Toxicity detection in immigration-related web news comments using transformers and statistical models. arXiv preprint arXiv:2111.04530.

de Paula, A. F. M., da Silva, R. F., & Schlicht, I. B., 2021b. Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models. arXiv preprint arXiv:2111.04551.

de Paula, A. F. M., & da Silva, R. F., 2022. Detection and Classification of Sexism on Social Media Using Multiple Languages, Transformers, and Ensemble Models. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the XXXVIII International Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), CEUR Workshop proceedings (Vol. 3202, pp. 1-11).

de Paula, A. F. M., Rosso, P., Bensalem, I., & Zaghouani, W., 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 (pp. 181-185).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

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.

Hassan, S., Samih, Y., Mubarak, H., Abdelali, A., Rashed, A., & Chowdhury, S. A., 2020. ALT submission for OSACT shared task on offensive language detection. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection (pp. 61-65).

Husain, F., & Uzuner, O., 2021. A survey of offensive language detection for the Arabic language. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 20(1), 1-44.

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. AI Open.

Mahdaouy, A. E., Mekki, A. E., Oumar, A., Mousannif, H., & Berrada, I., 2022. Deep Multi-Task Models for Misogyny Identification and Categorization on Arabic Social Media. arXiv preprint arXiv:2206.08407.

Mohammed, A. H., & Ali, A. H., 2021. Survey of BERT (bidirectional encoder representation transformer) types. In Journal of Physics: Conference Series (Vol. 1963, No. 1, p. 012173). IOP Publishing.

Mubarak, H., Darwish, K., Magdy, W., Elsayed, T., & Al-Khalifa, H., 2020. Overview of OSACT4 Arabic offensive language detection shared task. In Proceedings of the 4th Workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection (pp. 48-52).

Mubarak, H., Al-Khalifa, H., & Al-Thubaity, A., 2022. Overview of OSACT5 Shared Task on Arabic Offensive Language and Hate Speech Detection. In LREC 2022 Workshop Language Resources and Evaluation Conference 20-25 June 2022 (p. 162). a

Mubarak, H., Hassan, S., & Chowdhury, S. A., 2022. Emojis as anchors to detect arabic offensive language and hate speech. arXiv preprint arXiv:2201.06723.

Mulki, H., & Ghanem, B., 2021. ArMI at FIRE2021: Overview of the First Shared Task on Arabic Misogyny Identification. Working Notes of FIRE, 820-830.

Ravichandiran, S., 2021. Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT. Packt Publishing Ltd.

Safaya, A., 2020. Arabic-alBERT.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, Ç., 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In Proceedings of the Fourteenth Workshop on Semantic Evaluation (pp. 1425-1447).