Proceedings of the OSACT 2022 Workshop @LREC2022, pages 181–185
Marseille, 20 June 2022

UPV at the Arabic Hate Speech 2022 Shared Task: Offensive Language and Hate Speech Detection using Transformers and Ensemble Models

Angel Felipe Magnossão de Paula[1], **Paolo Rosso**[1], **Imene Bensalem**[2,3], **Wajdi Zaghouani**[4]
[1]Universitat Politècnica de València,
[2]MISC-Lab – Constantine 2 University, [3]ESCF de Constantine,
[4]Hamad Bin Khalifa University
adepau@doctor.upv.es, prosso@dsic.upv.es, ibensalem@escf-constantine.dz, wzaghouani@hbku.edu.qa

Abstract

This paper describes our participation in the shared task Fine-Grained Hate Speech Detection on Arabic Twitter at the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT). The shared task is divided into three detection subtasks: (i) Detect whether a tweet is offensive or not; (ii) Detect whether a tweet contains hate speech or not; and (iii) Detect the fine-grained type of hate speech (race, religion, ideology, disability, social class, and gender). It is an effort toward the goal of mitigating the spread of offensive language and hate speech in Arabic-written content on social media platforms. To solve the three subtasks, we employed six different transformer versions: AraBert, AraElectra, Albert-Arabic, AraGPT2, mBert, and XLM-Roberta. We experimented with models based on encoder and decoder blocks and models exclusively trained on Arabic and also on several languages. Likewise, we applied two ensemble methods: Majority vote and Highest sum. Our approach outperformed the official baseline in all the subtasks, not only considering F1-macro results but also accuracy, recall, and precision. The results suggest that the Highest sum is an excellent approach to encompassing transformer output to create an ensemble since this method offered at least top-two F1-macro values across all the experiments performed on development and test data.

Keywords: Deep Learning, Transformers, Offensive Language, Hate Speech, Arabic, Twitter

1. Introduction

The detection of offensive language and hate speech is becoming an important element when it comes to reducing the spread of the toxicity online. Despite some efforts done to address this issue, the automatic detection of hate speech is still considered a challenge, especially when it comes to detecting hate speech written in low-resources languages and varieties such as the several Arabic dialects used in social media nowadays. In this paper, we present our approach to offensive language and hate speech detection for the Arabic language using transformers and ensemble models. To train our models, we used the data set shared by the organizers of the Arabic Hate Speech 2022 shared task on Fine-Grained Hate Speech Detection on Arabic Twitter (Mubarak et al., 2022).

We address the problem of detecting hate speech and offensive language by applying six different transformer models and two ensemble methods. Within the transformers, we tried models based on encoder and decoder blocks and models exclusively trained on Arabic and others trained on several languages. Furthermore, we also combine the transformer results employing the Majority vote and Highest sum ensemble methods. Our code is open and available on Github [1].

The rest of the paper is structured as follows. Section 2 provides an overview on the problem of hate speech and offensive language detection in Arabic. Sections 3 and 4 present the task details and the used dataset. The created models are described in Section 5. Finally, we wrap up our paper with the discussion of the results and some conclusions.

[1]https://github.com/AngelFelipeMP/Transformers-for-Arabic-hate-speech-and-offensive-language

2. Hate Speech and Offensive Language Detection in Arabic

Literature on the identification of hate speech and offensive language in Arabic deals with the two following main tasks:

(i) The identification of the hateful or offensive language. Most of the works in this task propose *binary classification* solutions able to distinguish between two classes: (Hate and Not hate) or (Offensive and Not offensive). See, for example, the methods described in (Albadi et al., 2018; Guellil et al., 2020; Mubarak et al., 2020). Some works proposed datasets that allow addressing the problem as a *multi-class classification* where the hateful or offensive discourse has to be distinguished not only from the clean text but also from other similar discourses categorized as obscene (Mubarak et al., 2017), vulgar (Chowdhury et al., 2020), abusive (Haddad et al., 2019; Mulki et al., 2019), or disrespectful (Ousidhoum et al., 2019).

(ii) The fine-grained categorization of hate speech according to its type or target. To the best of our knowledge, only a few works addressed this task in the Arabic language. Below, we summarized the works that employed Arabic datasets involving fine-grained categories of hate speech.

Mulki and Ghanem (2021) addressed the problem of detecting misogyny, i.e., hatred against women. The authors built a dataset composed of 6550 tweets in the Levantine dialect. They collected them from the accounts of female journalists who were active during the Lebanon protest of October 2019. The tweets have been labelled as misogynistic or not. In addition, seven categories of misogyny have been used to label the misogynistic tweets (for instance, sexual harassment, stereotyping, threat of violence, etc). The authors employed the dataset in a set of experiments where the best results in misogyny identification and categorization, respectively, have been obtained by using AraBert and an ensemble technique combining the predictions of Naïve Bayes, SVM, and Logistic Regression classifiers.

Ousidhoum et al. (2019) created a multilingual dataset comprising more than 3000 tweets in Arabic among others in French and English. Tweets are labelled with five tags indicating (1) the hostility type i.e., whether the tweet is abusive, hateful, offensive, disrespectful, fearful or normal (2) whether the tweet is direct or indirect hate speech, (3) the personal attribute targeted by the hostility such as the origin, the disability, and the gender (4) the target group such as Arabs, refugees, Christians, women among others (5) the feeling that the annotator gets when reading the hateful tweet. The authors used the dataset to evaluate five tasks corresponding to the above five label sets. Apart from the task of classifying tweets as direct or indirect, which is a binary classification, the four other tasks are multi-class classification tasks. The conducted experiments compare a traditional approach based on logistic regression and bag-of-words features with deep learning approaches based on bidirectional LSTM trained under multi-/mono- task and language settings. The results show the outperformance of the deep learning approaches in most of the multi-class classification tasks.

Albadi et al. (2018) created a dataset of 6000 tweets involving religious hate speech referring to the different beliefs in the Middle East. Tweets are labelled as hateful or not. The dataset has been annotated as well with the religious groups targeted by the hateful tweets (Muslims, Jews, Christians, Atheists, Sunnis, Shia, other). The inter-annotator agreement concerning the target-group labels was only 55%. These labels are not available in the published dataset, but have been leveraged by the authors to obtain statistics on the religious groups most targeted by hate speech. The authors conducted binary-classification experiments to identify hateful tweets using three approaches. The first approach is lexicon-based. It consists in summing the sentiment scores of the tweet words. The second approach applies traditional machine learning to character n-grams features. The third approach relies on deep learning. It uses the GRU-based RNN with pre-trained embeddings. Results showed the outperformance of the deep learning approach.

To sum up, the existing approaches to hateful/offensive language identification and categorization use a range of traditional machine learning classifiers (such as SVM, Logistic Regression, Naïve Bayes ... etc) and deep learning methods including transformers, such as AraBert and multilingual Bert (mBert). Some works combine different methods in multitask learning or ensemble settings (Husain and Uzuner, 2021). While the performance of the identification task is promising (most notably, when it comes to a binary classification), the categorization task is still challenging.

3. Task Description

The task presented in this paper is another iteration of the previous similar tasks presented in OSACT 2020. This year, the OSACT 2022 has three subtasks to identify and categorize hate speech in the Arabic language given that the dataset was collected from Twitter with a large amount written in dialectal Arabic. The goal of the first subtask A is to detect whether a tweet is offensive or not with two possible labels: OFF (Offensive) or NOT OFF (Not Offensive).

The second subtask B is similar to task A but with a focus on hate speech. The two possible labels are HS (Hate Speech) or NOT HS (Not Hate Speech). According to the organizers, subtask B is more challenging given the low number of tweets falling into the hate speech class. Finally, the last task is subtask C in which the systems are expected to detect hate speech on fine-grained types this time. The organizers provided six possible labels for this subtask: race, religion, ideology, disability, social class, and gender. The first one is HS1 and is used to label hate speech targeting a specific race. HS2 is reserved for any kind of religious hate targeting either religion or religious groups. HS3 on the other hand is reserved for hate expressions targeting ideologies, while HS4 is reserved for expressions used against people with disabilities. Finally, HS5 and HS6 are the labels for hate targeting people based on their social classes or gender, respectively. The three tasks will be evaluated through the submissions made to the dedicated shared task platform (Codalab).

4. Dataset

The annotated dataset shared by the organizers of this shared task was collected from Twitter. The dataset can be considered among the largest publicly available annotated Arabic datasets released so far for offensiveness, along with fine-grained hate speech types, vulgarity, and violence. The annotation process went through a rigorous process, wherein each tweet is annotated by three annotators to ensure the quality of the annotation. In case of disagreement between the annotators, the label with the majority is taken into consideration. The organizers used a crowdsourcing platform to annotate their data for offensiveness and to classify each tweet into one of the hate speech types (religion, race, disability, ideology, social class, and gender). Moreover,

the data was annotated for containing or not vulgar language or violent terms.

The dataset contains around 13K tweets in total, with 35% annotated as offensive and only 11% marked as hate speech. Furthermore, the tweets annotated as vulgar and violent represent only around 1.5% and 0.7% of the dataset, respectively. According to the organizers, the provided dataset can be considered one of the highest in terms of the percentages of offensive language and hate speech. They claimed also that it is not biased toward specific topics, dialects, or genres since its creation did not rely on specific keywords. The dataset is split into 70% for training, 10% for development, and 20% for testing.

5. Transformer Models

This section introduces the transformer models applied to solve the three OSACT 2022 subtasks. Table 1 shows their main features: (i) transformer's version, (ii) model size, (iii) originating block, and (iv) trained input language

Transformer models are massive deep learning architectures constructed for dealing with natural language processing tasks (Vaswani et al., 2017; Ravichandiran, 2021; Lin et al., 2021). These models are trained, in an unsupervised way, on enormous datasets by performing different tasks, such as mask language modelling, next sequence prediction, and many others (Devlin et al., 2019; Mohammed and Ali, 2021).

Usually, the transformers are available in three different sizes: base, medium, and large. The term size is related to the number of trainable parameters in the model. We used the large size for Albert-Arabic, and, for all the other transformers, we used the base version due to computational constraints.

The original transformer model designed by Google researchers (Vaswani et al., 2017) encompasses encoder and decoder blocks. However, the new versions, currently, contain only either one encoder or one decoder block. We used four models based only on the encoder block: AraBert (Antoun et al., 2020), AraElectra (Antoun et al., 2021a), mBert, and XLM-Roberta (Conneau et al., 2020), and one based only on the decoder block, AraGT2 (Antoun et al., 2021b).

There are transformer models trained on different languages. AraBert, AraElectra, Albert-Arabic, and AraGT2 were trained on a collection of Arabic datasets (Ravichandiran, 2021). mBert and XLM-Roberta belong to a subclass of transformers that we call multilingual. mBert was trained on a dataset including texts written in 104 different languages, and XLM-Roberta was trained on one dataset gathering documents from 100 different languages.

On the top of each transformer model, we added a linear layer classifier which computes a probability distribution based on the possible classes in the subtask, which varied among the three subtasks.

Version	Size	Block	Language
AraBert AraElectra	base	Encoder	Arabic
Albert-Arabic	large		
AraGPT2	base	Decoder	
mBert XLM-Roberta	base	Encoder	Multilingual

Table 1: Transformers used for the OSACT 2022 tasks

6. Results and Discussion

This section explains the hyper-parameter selection and the performance of the models through the validation and test. In addition, we present how we combine the transformer results by means of two ensembles methods: (i) Majority vote; and (ii) Highest sum.

We were concerned about the number of training epochs, learning rate, and dropout percentage for the transformer's fine-tuning. Therefore, we applied a 5-fold cross-validation on the training data to find suitable parameters for each model based on the OSACT 2022 official metric, F1-macro. Table 2 shows the best number of training epochs for the transformers in each subtask. Coincidentally, the appropriate dropout and learning rates found are the same for all the models, respectively equal to 0.3 and 0.00005. We adopted a max length of 64 tokens and a batch size of 32 samples during all experiments.

	Epochs		
Model	Subtask A	Subtask B	Subtask C
AraBert	5	4	4
AraElectra	4	2	5
Albert-Arabic	2	4	5
AraGPT2	5	5	5
mBert	3	5	5
XLM-Roberta	5	1	4

Table 2: Transformer's suitable number of epochs

OSACT 2022 allowed only two submissions of the predictions on the test data. Thus, we trained the transformers on the training data and evaluated them on the development data to find the two best models for each subtask. In addition, we applied two ensemble methods: Majority vote and Highest sum. The Majority vote selects the most predicted class among the transformers, and if there is a tie, it randomly selects one of the classes among the tied classes. The Highest sum aggregates the output values by each transformer separately for each class and selects the class with the highest sum. Table 3 shows the F1-macro results obtained by the two best models in each subtask on the development data.

In order to make the final predictions on the test data for all the subtasks, we applied the two models that obtained the best results on the development data. However, the inferences for subtask C were dependent on the inferences for subtask B. Considering this fact, we must detect whether a tweet has hate speech or not

Subtask	Model	Accuracy	F1-macro
A	**Highest sum**	**0.837**	**0.814**
	AraBert	0.829	0.808
B	**AraElectra**	0.932	**0.795**
	Highest sum	**0.938**	0.794
C	**AraBert**	**0.979**	**0.582**
	Highest sum	0.981	0.513

Table 3: Results of our two best models in the development data

(subtask B), and only in case it belongs to the positive class we detect the type of hate speech (subtask C). Therefore, we used subtask B predictions from our best model to pass the tweets detected with hate content to the subtask C models, with the aim of identifying the type of hate speech in the tweets. Table 4 shows our final results on the test data and the OSACT 2022 baseline for each subtask.

Subtask	Model	F1-macro	Accuracy	Precision	Recall
A	**AraBert**	**0.827**	**0.841**	**0.824**	**0.831**
	Highest sum	0.819	0.837	0.821	0.818
	Baseline	0.394	0.651	0.325	0.500
B	**Highest sum**	**0.792**	**0.932**	**0.858**	**0.751**
	AraElectra	0.757	0.925	0.845	0.711
	Baseline	0.472	0.893	0.447	0.500
C	**AraBert**	**0.423**	**0.920**	**0.542**	**0.369**
	Highest sum	0.325	0.917	0.382	0.294
	Baseline	0.135	0.893	0.128	0.143

Table 4: Final results in the test data

The differences between the results of our worst models and the baselines for the F1-macro are 0.425 subtask A, 0.285 subtask B, and 0.190 subtask C. Thus, we can conclude that all our models (even the worst ones) obtained results significantly superior to the baselines for the OSACT 2022 official metric. The results also suggest that the Highest sum is suitable for aggregating transformers' outputs to create an ensemble. It offered at least the top-two F1-macro across development and test data experiments.

Looking again at table 4, we can see a discrepancy between accuracy and F1-macro for tasks B and C. The F1-macro is computed as the unweighted mean of the F1-score calculated for each class. The recall is one of the factors that compose the F1-score calculation, which is sensitive to false negatives. We hypothesize that because, since the number of positive samples for task B is low - only 11% -, the models achieved high accuracy but had an increased number of false negatives which degraded the F1-macro. Thus, because of the imbalanced proportion of the classes in the training data, the model overfits the distribution and ended up tending to select more the negative class. Task C was affected by the same phenomenon as mentioned for task B, and, besides that, it also suffered a decrease in the F1-macro results because of the multiple labelling of the target variable.

7. Conclusions

In this paper, we proposed to solve the problems of offensive language detection, hate speech detection, and fine-grained hate speech classification by employing six different transformer versions: Arabert, AraElectra, Albert-Arabic, AraGPT2, mBert, and XLM-Roberta. In addition, we also employed two ensemble methods: Majority vote and Highest sum. Our approach outperformed the official OSACT 2020 baselines in all the subtasks.

8. Acknowledgements

This publication was made possible by NPRP grant 13S-0206-200281 (Resources and Applications 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.

9. Bibliographical References

Albadi, N., Kurdi, M., and Mishra, S. (2018). Are they our brothers? analysis and detection of religious hate speech in the Arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Barcelona. IEEE.

Antoun, W., Baly, F., and 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*, pages 9–15, Marseille, France, May. European Language Resource Association (ELRA).

Antoun, W., Baly, F., and Hajj, H. (2021a). AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual), April. Association for Computational Linguistics (ACL).

Antoun, W., Baly, F., and Hajj, H. (2021b). AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207, Kyiv, Ukraine (Virtual), April. Association for Computational Linguistics (ACL).

Chowdhury, S. A., Mubarak, H., Abdelali, A., Jung, S.-G., Jansen, B. J., and Salminen, J. (2020). A multi-platform arabic news comment dataset for offensive language detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6203–6212, Marseille. European Language Resources Association (ELRA).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., and Stoyanov, V. (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 (ACL).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics (ACL).

Guellil, I., Adeel, A., Azouaou, F., Chennoufi, S., Maafi, H., and Hamitouche, T. (2020). Detecting hate speech against politicians in arabic community on social media. *International Journal of Web Information Systems*, 16(3):295–313.

Haddad, H., Mulki, H., and Oueslati, A. (2019). T-HSAB: A tunisian hate speech and abusive dataset. In *International Conference on Arabic Language Processing*, pages 251–263. Springer.

Husain, F. and 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., and Qiu, X. (2021). A survey of transformers. *arXiv preprint arXiv:2106.04554*.

Mohammed, A. H. and Ali, A. H. (2021). Survey of BERT (bidirectional encoder representation transformer) types. *Journal of Physics: Conference Series*, 1963(1):012173.

Mubarak, H., Darwish, K., and Magdy, W. (2017). Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver. Association for Computational Linguistics (ACL).

Mubarak, H., Darwish, K., Magdy, W., and 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. Language Resources and Evaluation Conference (LREC 2020)*, pages 48–52, Marseille. European Language Resources Association (ELRA).

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

Mulki, H. and Ghanem, B. (2021). Let-Mi: An arabic levantine twitter dataset for misogynistic language. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 154–163, Kyiv.

Mulki, H., Haddad, H., Bechikh Ali, C., and Alshabani, H. (2019). L-HSAB: A levantine twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence. Association for Computational Linguistics (ACL).

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., and Yeung, D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Stroudsburg, PA, USA. Association for Computational Linguistics (ACL).

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

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