Proceedings of the OSACT 2022 Workshop @LREC2022, pages 194–199
Marseille, 20 June 2022

CHILLAX - at Arabic Hate Speech 2022: A Hybrid Machine Learning and Transformers based Model to Detect Arabic Offensive and Hate Speech

Kirollos Hany Makram, Kirollos George Nessim, Malak Emad Abd-Almalak,
Shady Zekry Roshdy, Seif Hesham Salem, Fady Fayek Thabet,
Ensaf Hussein Mohamed
Research Support Center in Computing and Informatics
Department of Computer Science, Faculty of Computers and Artificial Intelligence,
Helwan University, Cairo, Egypt.
kirollos_20180442@fci.helwan.edu.eg,
kirollos_20180437@fci.helwan.edu.eg, Malak_20180615@fci.helwan.edu.eg,
Shady_20180285@fci.helwan.edu.eg, seif_20180284@fci.helwan.edu.eg,
fady_20180413@fci.helwan.edu.eg, ensaf_hussein@fci.helwan.edu.eg

Abstract

Hate speech and offensive language have become a crucial problem nowadays due to the extensive usage of social media by people of different gender, nationality, religion and other types of characteristics, allowing anyone to share their thoughts and opinions. In this research paper, we proposed a hybrid model for the first and second tasks of OSACT2022. This model used the Arabic pre-trained Bert language model MARBERT for feature extraction of the Arabic tweets in the dataset provided by the OSACT2022 shared task, then fed the features to two classic machine learning classifiers (Logistic Regression, Random Forest). The best results achieved for the offensive tweet detection task were achieved by the Logistic Regression model with accuracy, precision, recall, and f1-score of 80%, 78%, 78%, and 78%, respectively. The results for the hate speech tweet detection task were 89%, 72%, 80%, and 76%.
The source code can be found on GitHub here (Hany, 2022)
Keywords: Arabic Tweets, Offensive language, Hate speech, Transformers, Text classification

1 Introduction

Twitter and similar social media platforms users are from every race, religion, nationality, and background communicate and freely share their opinions and beliefs. The down side to this is that it is easy to exploit these social media platforms by sharing offensive and hate speech content that targets and threatens individuals or groups based on common characteristics, or identities. Despite the considerable efforts that social media platforms are making to prevent such content from spreading, it is still threatening the online communities and users are still seeing it on many platforms. It is imperative to detect and prevent such content from appearing on social media platforms, thus motivating our research on Arabic offensive and hate speech detection. Pre-trained language models based on Transformer (Vaswani et al., 2017) such as GPT (Radford et al., 2018), Bert (Devlin et al., 2018), XLNet (Yang and Zhao, 2019), and RoBERTa (Zhuang et al., 2021) have been shown to be effective for learning contextualized language representation achieving state-of-the-art results on a wide variety of natural language processing tasks. Recent research has adopted the methodology of fine-tuning a pre-trained language model by simply adding a fully connected neural layer specific to the downstream task the model is being fine-tuned for, such as sarcasm detection (Farha and Magdy, 2021) and hate speech detection (Aldjanabi et al., 2021). Research has shown that due to the numerous layers present in Transformer models, simply feeding the output of the Transformer's encoder final layer to the fully connected neural layer would restrict the power of the pre-trained representations of the language (Yang and Zhao, 2019). (Devlin et al., 2018) shows that different combinations of different output layers of the Transformers encoder layers result in distinct performance on various tasks like Named Entity Recognition task. It is found that the most contextualized representations of input text tend to occur in the middle layers, while the top layers are for language modeling (Yang and Zhao, 2019). This research used different transformers' encoder layers as feature extractors. Then we used the extracted features to feed into two classical machine learning classifiers; Logistic Regression and Random Forest. We used the MARBERT pre-trained Transformer model as it was trained on a large Arabic tweets corpora and has proved to be efficient in similar tasks such as sentiment analysis, where it scored 0.93 F1-score on the ArSAS dataset ("Abdul-Mageed et al., 2021). Also, we used Logistic Regression as a classifier as it proved superior on binary classification problems as the current subtasks. We conducted our experiments on the OSACT2022 dataset for subtask 1: Arabic offensive, and subtask2: hate speech tweets detection. One of The challenges of the OSACT2022 dataset are that the dataset is imbalanced; the number of tweets on both subtasks aren't equal across the two classes, offensive and hate speech detection. We tackled this problem by using data augmentation techniques to achieve a balanced class distribution in the dataset to prevent the

classifiers from biasing towards the majority class. The research paper is organized as follows. Section 2 gives a brief overview of related work. Section 3 explain in details our methodology and proposed model. Section 4 presents the experiment results and evaluation metrics. Section 5 concludes our research and our potential future work.

2 Related Work

Recently, the interest in detecting hate speech has increased rapidly, attracting the attention of many researchers trying to develop various models and methods to extract hate features and hateful content. Several research studies were conducted to study hate speech and offensive language in online communities and social media over Arabic content. (Abuzayed and Elsayed, 2020) investigate 15 classical and neural learning models with TF-IDF and word embedding as feature representations of the OSACT-2020 dataset; their best classifier (A joint architecture of CNN and RNN) achieved 73% macro F1-score on the development dataset and 69% on the test dataset with word embedding as feature representations. (Alshaalan and Al-Khalifa, 2020) investigate several neural network models that are based on CNN and RNN to detect hate speech in Arabic tweets and also evaluates recent BERT model on the task of Arabic hate speech detection. They built a hate speech dataset containing 9,316 annotated Arabic tweets. They conducted experiments on that dataset and an out-domain dataset showing that the CNN model achieves an F1-score of 79% and AUROC of 89%. (Faris et al., 2020) proposed an innovative deep learning approach for the detection of cyber hate speech. The detection of hate speech on Twitter in the Arabic region, in particular, using a word embedding mechanism for feature representation and fed to a hybrid CNN and LSTM neural network that achieved a 71% F1-score on a dataset that is collected from the Twitter API. (Al-khalifa et al., 2020) collected a 3,000 tweet dataset from Twitter where they experimented BOW and TF-IDF methods for feature representation and classical machine learning models (SVM, NB, RF) and concluded that TF-IDF with SVM achieved the best results of 82% F1-score. So far, related work shows that word embeddings, TF-IDF, and BOW have been used as the feature representation for text experimented with traditional machine learning models as well as deep learning neural networks. In most cases, deep learning neural networks show superior results to classical machine learning models. This further motivates our approach to experiment with the feature representation of the different combinations of hidden layers in pretrained transformer models on classical machine learning models.

3 Methods and Materials

3.1 The dataset

We used the Arabic tweets dataset provided by the OSACT2022 shared task, which contains around 13,000 tweets, where 35% are offensive, and 11% are hate speech. Vulgar and violent tweets represent 1.5% and 0.7% of the dataset, respectively. The dataset was split into three parts train, development, and test, with percentages of 70%, 10%, and 20%, respectively. For the first task, offensive tweet detection, the training dataset contained 5,715 offensive and 3,172 not offensive tweets—figure 1 shows the class imbalance presented in the training dataset for our first task. The training dataset for the second task, hate speech tweet detection, contained 7,928 not hate speech and 959 hate speech tweets, which shows a significant class imbalance. Figure 2 shows the class imbalance presented in the training dataset for the second task.

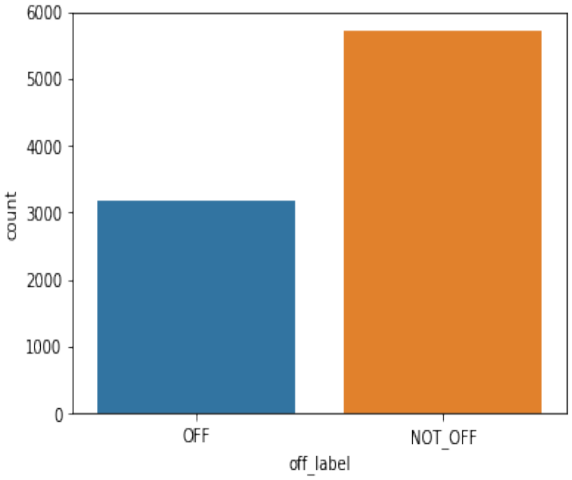

Figure 1: Offensive detection task label count plot

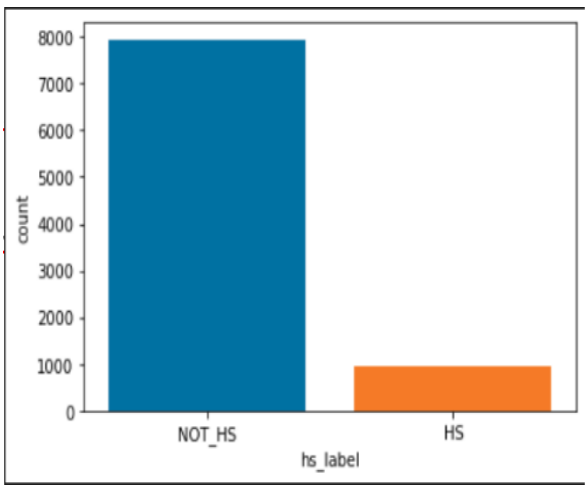

Figure 2: Hate speech detection task label count plot

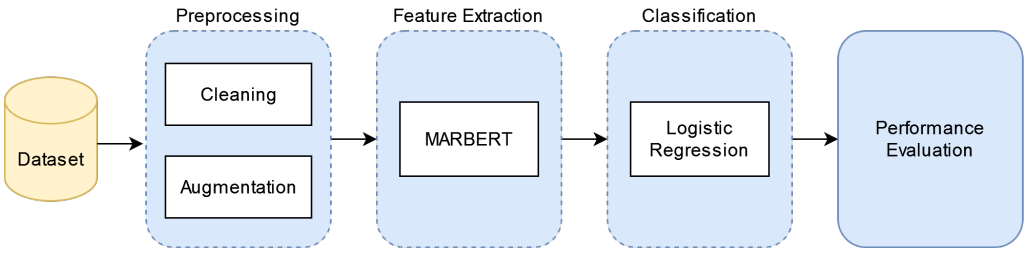

Figure 3: The proposed Model

3.2 Proposed Model

The proposed model consists of three modules; Preprocessing, Feature Extraction, and Classification. We will describe each module in detail in the following subsections. Figure 3 illustrates the proposed model pipeline.

3.2.1 Preprocessing

The preprocessing phase includes two submodules, cleaning and augmentation. In cleaning, all URLs and User mentions were removed. In augmentation, we solved the class imbalance problem in the two tasks. We used the contextual word embedding augmentation technique like the MARBERT Arabic model, which generates new tweets from the minority classes (offensive and hate speech) so that the class distribution in both tasks was balanced to prevent the model from biasing towards the majority classes (not offensive, not hate speech). Some of the augmented tweets had an unknown special token generated; these tokens were removed from the augmented tweets. The NLP aug (Alkhalifa et al., 2020) data augmentation library was used for the data augmentation. Table 1 and 2 show a sample of augmented offensive and hate speech tweets, respectively. It is worth mentioning that traditional text preprocessing methods like stemming, lemmatizing, and punctuation removal were experimented with but resulted in poor results as MARBERT was trained on Twitter text that was not preprocessed with these methods. Simply removing the URL and user mention tags achieved better results.

3.2.2 Feature Extraction

For the feature extraction phase, we used MARBERT pre-trained Arabic language model to extract features which will be later fed to the machine learning models Logistic Regression and Random Forest for training. The MARBERT model is a Bert-based model which consists of 12 hidden layers and a hidden size of 768. The output of the last four hidden layers was obtained, where each layer is of dimensions sentence length x hidden size. The output of each layer is summed to produce a single vector of sentence length x hidden size dimensions. The mean of this vector is computed to create a single vector of hidden size length representing the feature vector for the tweet that will be fed to the machine learning models.

Original Tweets	Augmented Tweets
جمهور الاهلي مضحوك عليهم الحين يصدقو ان عددهم كثير 😂	جمهور وجمهور الاهلي مضحوك عليهم الحين من يصدقو ان الاهلاويين عددهم كثير 😂
#بايع_الكليجا الله يفشلكم فشلتونا يالمراهقات انا استحيت من تصرفات البنات وتعليقاتهم وانا مالي دخل 😂	[UNK] واحد [UNK] سعودي بايع [UNK] يبيع الكليجاالله يفشلكم فشلتونا يالمراهقاتانا استحيت من كثر تصرفات البنات وتعليقاتهم وانا مالي فيها دخل 😂
#بايع_الكليجا من العام اشوفه يبيع ويسترزق الله حاله حال نفسه سالفه التصوير والتشهير فيه قله أدب اللي قاعد يصير البنات م عاد يستحو 😂	[UNK] ▪ بايع [UNK] الكليجا من العام اشوفه يبيع ويسترزق الله حاله حال نفسه ف سالفه التصوير والكلام والتشهير ذي فيه قله وسوء ادب اللي قاعد يصير البنات م عاد يستحو 😂 فيه

Table 1: Augmented Offensive tweets sample

Original Tweets	Augmented Tweets
الهندي قاعد يثبت للمعزّب إن الخروف صار صديقه علشان ما يذبحه 🐄 😂	الهندي قاعد يثبت للمعزب ان الخروف صار له صديقه علشان بس ما يذبحه 😂 🐄
قلبنا يابنت خلينا حلوين يعني انتم وش يجملكم غير الميك اب 🐄	قلبنا يابنت خلينا نكون حلوين يعني انتم وش اللي يجملكم من غير الميك اب 🐄
عاهات بزمنكم كانوا يصولون ويجولون شرق اسيا، ختاماً بعالمية ضد سوباولو البرازيلي، وانت بعز مستواك حققت الدوري بعد ٣٢ سنه عاجفه بالضحك والطقطقه عليكم؟	عاهات بزمنكم كانوا يصولون ويجولون شرق قارة اسيا الان ، ختاما فاز بعالمية ضد سوباولو البرازيلي ، تخيل وانت بعز مستواك حققت الدوري بعد ٣٢ سنه سنه عاجفه ، بالضحك عليه والطقطقه شلون عليكم ؟

Table 2: Augmented hate speech tweets sample

3.2.3 Classification

We used two classifiers; The Scikit-Learn library implementation of Logistic Regression and Random Forest were used in the training phase.

For the Logistic Regression model, a regularisation parameter of 1^{-3} was used, and for the optimization algorithm, the Saga solver was used. A max sample parameter of 0.4 was used for the Random Forest model,

where max sample is the number of samples to be drawn from the inputs to train each base estimator.

4 Results and Discussion

Before training the model, the training dataset was split into 70% for training and 30% for testing to evaluate the model, along with the development and test datasets provided by the OSACT2022 shared task.

4.1 Performance Evaluation

The evaluation metrics used are macro averaged Precision, Recall, F1-score, and Accuracy, where Precision is the fraction of classified tweets that are relevant, which is formulated in equation 1. The recall is the fraction of relevant tweets that are classified, which is presented in equation 2. F1-score is the harmonic mean of Precision and Recall, which is presented in equation 3. Accuracy is the fraction of correct tweets that have been classified from actual classes as shown in Equation 4.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{Accuarcy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

where:
True Positive (TP): refers to a set of tweets that have been classified correctly to the task class (offensive, hate speech).
False Positive (FP): refers to a set of tweets that have been classified incorrectly and have been said to be related to the task class (offensive, hate speech) incorrectly.
True Negative (TN): refers to a set of tweets that have not been classified into the task class (offensive, hate speech) and are not labeled as task class (offensive, hate speech).
False Negative (FN): refers to a set of tweets that have not been classified correctly and have been said to be non-related to the task class (offensive, hate speech), but they are labeled as task class (offensive, hate speech).

4.2 Experimental Results

The baselines for evaluation provided by OSACT2022 are as following:

Task	Accuracy	Precision	Recall	F1-score
Offensive detection	65%	65%	50%	39%
Hate speech detection	89%	45%	50%	47%

For the offensive detection task, we applied two experiments. The First one, using MARABERT as a Feature extractor, then fed the feature vector into the Logistic regression classifier and in the second experiment, we used the same feature vector that was produced by MARABERT and then fed it into the Random Forest classifier. Each model was trained on 70% of the training dataset after augmentation and evaluated on the remaining 30% of the training dataset along with the development and the OSACT2022 test dataset. Best results were achieved by Logistic Regression on the test dataset with macro average accuracy, F1-score, of 80%, 78% respectively. The results obtained for each model and dataset for the offensive detection task are as shown in the following Table 3 and Table 4:

Model	Dataset	Accuracy	Precision	Recall	F1-score
Logistic Regression	Train	85%	85%	85%	85%
	Test(30% of train)	81%	81%	81%	81%
	Development	80%	77%	78%	78%
	OSACT-2022-Test	80%	78%	78%	78%

Table 3: Results of detecting offensive language using MARABERT and Logistic Regression

Model	Dataset	Accuracy	Precision	Recall	F1-score
Random Forest	Train	97%	97%	97%	97%
	Test(30% of train)	77%	77%	77%	77%
	Development	75%	72%	73%	72%
	OSACT-2022-Test	74%	72%	72%	72%

Table 4: Results of detecting offensive language using MARABERT and Random Forest

Model	Dataset	Accuracy	Precision	Recall	F1-score
Random Forest	Train	98%	98%	98%	98%
	Test(30% of train)	90%	90%	90%	90%
	Development	87%	67%	81%	70%
	OSACT2022-Test	87%	69%	79%	73%

Table 6: Results of detecting Hate Speech using MARABERT and Random Forest

For the hate speech detection task, we applied two experiments. First one, using MARABERT as a Feature extractor then fed the feature vector into the Logistic regression classifer and in the second experiment, we used the same feature vector that was produced by MARABERT and then fed it into the Random Forest classifier. Each model was trained on 70% of the training dataset after augmentation and evaluated on the remaining 30% of the training dataset along with the development and the OSACT2022 test dataset. Best results were achieved by Logistic Regression on the test dataset with macro averaged accuracy, F1-score of 89%, 76% respectively. The results obtained for each model and dataset for the hate speech detection task are as follows in Table 5, Table 6:

Model	Dataset	Accuracy	Precision	Recall	F1-score
Logistic Regression	Train	91%	91%	91%	91%
	Test(30% of train)	91%	91%	91%	91%
	Development	89%	70%	81%	74%
	OSACT2022-Test	89%	73%	81%	76%

Table 5: Results of detecting Hate Speech using MARABERT and Logistic Regression

5 Conclusion

We proposed a hybrid machine learning and transformers based model for the detection of Arabic offensive and hate speech tweets. We leverage the superiority of transformers for text feature extraction and the excellence of Logistic Regression in binary classification tasks and data augmentation techniques for handling data imbalances. The best results achieved for the offensive tweet detection task were achieved by the Logistic Regression model with accuracy, precision, recall, and f1-score of 80%, 78%, 78%, and 78%, respectively. The results for the hate speech tweet detection task were 89%, 72%, 80%, and 76%. For future work we plan to further investigate the different representations that different combinations of transformer-based models layers have for extracting features of text and also investigate different machine learning classification models such as SVM and Naive Bayes for the binary classification tasks in hope to achieve higher scores and obtaining a more efficient model.

References

Abdul-Mageed, M., Elmadany, A., and Nagoudi, E. M. B. (2020). Arbert & marbert: deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

”Abdul-Mageed, M., Elmadany, A., and Nagoudi, E. M. B. (”2021”). ”ARBERT & MARBERT: Deep bidirectional transformers for Arabic”. In *”Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)”*, pages ”7088–7105”, ”Online”, aug. ”Association for Computational Linguistics”.

Abuzayed, A. and Elsayed, T. (2020). Quick and simple approach for detecting hate speech in arabic tweets. In *Proceedings of the 4th workshop on open-source Arabic Corpora and processing tools, with a shared task on offensive language detection*, pages 109–114.

Al-khalifa, S., Aljarah, I., and Abushariah, M. A. M. (2020). Hate speech classification in arabic tweets. *Journal of Theoretical and Applied Information Technology*, 98:1816–1831.

Aldjanabi, W., Dahou, A., Al-qaness, M. A., Abd Elaziz, M., Helmi, A. M., and Damaševičius, R. (2021). Arabic offensive and hate speech detection using a cross-corpora multi-task learning model. In *Informatics*, volume 8, page 69. Multidisciplinary Digital Publishing Institute.

Alshaalan, R. and Al-Khalifa, H. (2020). Hate speech detection in saudi twittersphere: A deep learning approach. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 12–23.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Farha, I. A. and Magdy, W. (2021). Benchmarking transformer-based language models for arabic sentiment and sarcasm detection. In *Proceedings of the sixth Arabic natural language processing workshop*, pages 21–31.

Faris, H., Aljarah, I., Habib, M., and Castillo, P. A. (2020). Hate speech detection using word embedding and deep learning in the arabic language context. In *ICPRAM*, pages 453–460.

Hany, K. (2022). Osact2022-source-code. `https://github.com/kirollos-hany/OSACT2022-source-code`.

Ma, E. (2019). Nlp augmentation. https://github.com/makcedward/nlpaug.

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*.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

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.

Williams, M. L., B. P. J. A. L. H. and Ozalp. (2020). Hate in the machine: anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology 60(1): 93–117.*

Yang, J. and Zhao, H. (2019). Deepening hidden representations from pre-trained language models. *arXiv preprint arXiv:1911.01940*.

Zhuang, L., Wayne, L., Ya, S., and Jun, Z. (2021). A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227.