Hate speech detection - For Arabic Hate Speech 2022 Shared Task Competition

Prof Ensaf Hussein

Kirollos Hany, Kirollos George, Malak Emad,
Shady Zekry, Seif Hesham, Fady Fayek

Abstract—**Hate speech and offensive language became a crucial problem nowadays due to the large usage of social media by people from different gender, nationality, religion and other types of characteristics allowing anyone to share their thoughts and opinions. In this research paper we experiment the result of using the Arabic pretrained Bert language model MARBERT for feature extraction of the Arabic tweets in the dataset provided by OSACT2022 shared task then feeding the features to classic machine learning models (Logistic Regression, Random Forest). The best results achieved were by the Logistic Regression model with accuracy, precision, recall, and f1-score of 0.8, 0.78, 0.78, and 0.78 for the offensive tweet detection task respectively and for the hate speech tweet detection task results achieved were 0.89, 0.72, 0.8, and 0.76 these results were achieved on the OSACT2022 test dataset.**

I. Introduction

Pre-trained language models based on Transformer [13] such as GPT [12], Bert [7], XLNet [15], and RoBERTa [16] 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 have adopted the methodology of fine tuning a pretrained language model by simply adding a fully connected neural layer specific to the down-stream task the model is being fine tuned for such as sarcasm detection [8] and hate speech detection [5]. Research have 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 pretrained representations of the language [15]. [7] shows that different combinations of different output layers of the Transformers encoder layers result in distinct performance on different 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 top layers are for language modeling [15]. We explore the results of obtaining the text representation from different combination of layers of the Transformers encoder layers then using it as features for classical machine learning models (Logistic Regression, Random Forest) for both of the OSACT2022 shared tasks which is Arabic offensive and hate speech tweets detection. We used the MARBERT pretrained Transformer model as it were trained on a large Arabic tweets corpora and have proved to be efficient in similar tasks such as sentiment analysis where it scored 0.93 F1-score on ArSAS dataset [2]. We experimented on the OSACT2022 shared task dataset, which had a class imbalance problem present in both tasks offensive and hate speech detection we tackled the 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 II gives a brief overview of related work. Section III explains in details our methodology and proposed model. Section IV presents the experiment results and evaluation metrics. Section V concludes our research and our potential future work.

II. 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. There are several research studies conducted to study hate speech and offensive language in online communities and social media over Arabic content. [3] 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 0.73 macro F1-score on the development dataset and 0.69 on the test dataset with word embedding as feature representations. [6] 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 build a hate speech dataset containing 9,316 annotated Arabic tweets and conducted experiments on that dataset and an out-domain dataset showing that the CNN model achieves an F1-score of 0.79 and AUROC of 0.89. [9] proposed a smart deep learning approach for the detection of cyber hate speech. The detection of hate speech on Twitter on 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 0.71 F1-score on a dataset that is collected from the Twitter API. [4] 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 0.82 F1-score.

III. Methods and Materials

A. The dataset

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

The data set we have used we split it into 2 parts. First part is testing data which is 70%, and testing data which represents 30%. Worth mentioning that the data set is from the competition.

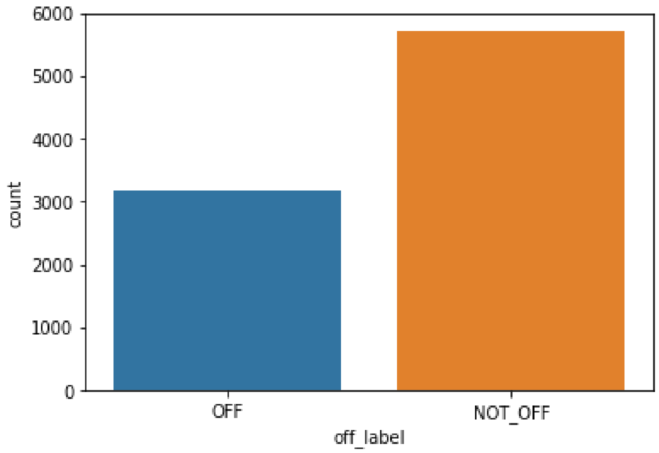


Fig. 1. Offensive detection task label count plot

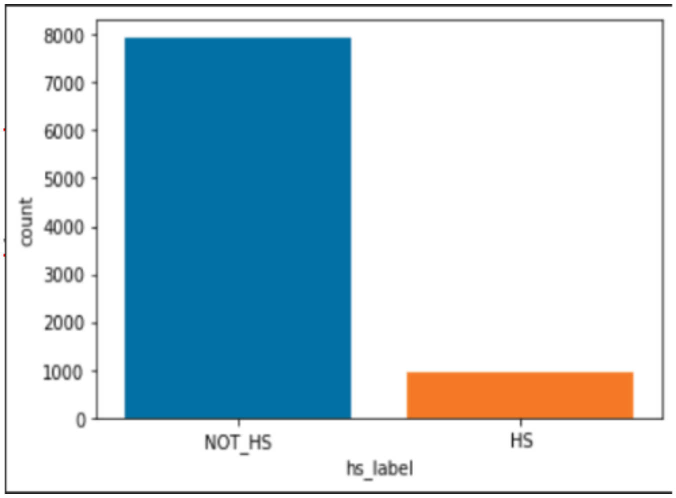
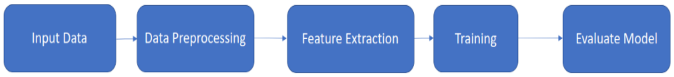


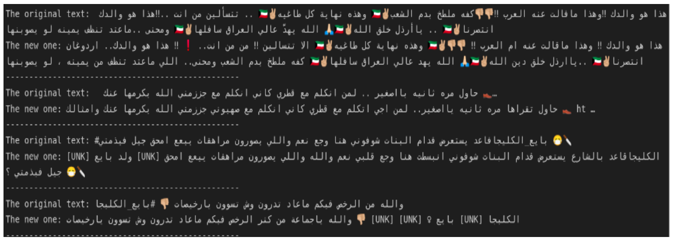
Fig. 2. Hate speech detection task label count plot

IV. Proposed Model

A. Preprocessing

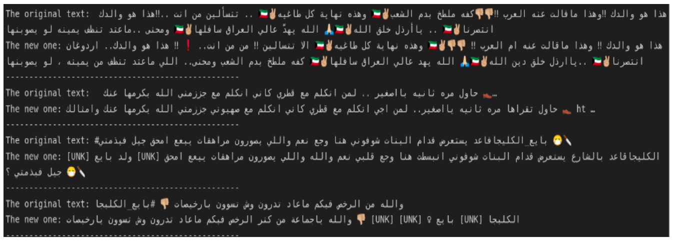
For the preprocessing phase all URLs and user mentions were removed from the tweet text. To tackle the class imbalance problem present in the two tasks contextual word embedding with insert action data augmentation technique using MARBERT Arabic model to generate new tweets of the minority classes (offensive, hate speech) so that the class distribution in both tasks were 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 [4] data augmentation library were used for the data augmentation. Figure 3 and 4 shows a sample of augmented offensive and hate speech tweets respectively.





```
The original text:  هذا هو والدك !!وهذا ماقالت عنه العرب !!👎👎كفه ملطخ بدم الشعب✌🇰🇼 وهذه نهاية كل طاغيه✌🇰🇼 .. تتسألين من انت ..!!هذا هو والدك
انتصرنا✌🇰🇼 .. ياأرذل خلق الله✌🇰🇼🙏 الله يهدّ عالي العراق سافلها✌🇰🇼 ومحنى ..ماعتد تنطق يمينه لو يصوبنها
The new one: هذا هو والدك !! وهذا ماقالت عنه ام العرب !! 👎👎✌🇰🇼 وهذه نهاية كل طاغيه✌🇰🇼 الا تتسالين !! من من انت.. ❗ !! هذا هو والدك.. اردوغان
انتصرنا✌🇰🇼 ..ياارذل خلق دين الله✌🇰🇼🙏 الله يهد عالي العراق سافلها✌🇰🇼 كفه ملطخ بدم الشعب ومحنى.. اللي ماعتد تنطق من يمينه ، لو يصوبنها
------------------------------------------------
The original text:  …👞 حاول مره ثانيه بااصغير .. لمن انكلم مع قطري كاني انكلم مع جززمني الله يكرمها عنك
The new one: … ht 👞 حاول تقراها مره ثانيه يااصغير.. لمن اجي اتكلم مع قطري كاني انكلم مع مهبوني جززمني الله يكرمها عنك وامثالك
------------------------------------------------
The original text: 🔪😷 بايع_الكليجاقاعد يستعرض قدام البنات شوفوني هنا وجع نعم واللي يصورون مراهقات يبيعع امحق جبل فيذمتي#
The new one: الكليجاقاعد بالشارع يستعرض قدام البنات شوفوني انبسطت هنا وجع قلبي نعم والله واللي يصورون مراهقات يبيعع امحق [UNK] ولد بايع [UNK]
🔪😷 جيل فيذمتي ؟
------------------------------------------------
The original text: والله من الرخص فيكم ماعاد تدرون وش تسوون بارخيصات 👎 #بايع_الكليجا
The new one: الكليجا [UNK] بايع ♀ [UNK] [UNK] 👎 والله باجماعة من كثر الرخص فيكم ماعاد تدرون وش تسوون بارخيصات
------------------------------------------------
```

Fig. 3. Augmented offensive tweets sample



```
The original text:  هذا هو والدك !!وهذا ماقالت عنه العرب !!👎👎كفه ملطخ بدم الشعب✌🇰🇼 وهذه نهاية كل طاغيه✌🇰🇼 .. تتسألين من انت ..!!هذا هو والدك
انتصرنا✌🇰🇼 .. ياأرذل خلق الله✌🇰🇼🙏 الله يهدّ عالي العراق سافلها✌🇰🇼 ومحنى ..ماعتد تنطق يمينه لو يصوبنها
The new one: هذا هو والدك !! وهذا ماقالت عنه ام العرب !! 👎👎✌🇰🇼 وهذه نهاية كل طاغيه✌🇰🇼 الا تتسالين !! من من انت.. ❗ !! هذا هو والدك.. اردوغان
انتصرنا✌🇰🇼 ..ياارذل خلق دين الله✌🇰🇼🙏 الله يهد عالي العراق سافلها✌🇰🇼 كفه ملطخ بدم الشعب ومحنى.. اللي ماعتد تنطق من يمينه ، لو يصوبنها
------------------------------------------------
The original text:  …👞 حاول مره ثانيه بااصغير .. لمن انكلم مع قطري كاني انكلم مع جززمني الله يكرمها عنك
The new one: … ht 👞 حاول تقراها مره ثانيه يااصغير.. لمن اجي اتكلم مع قطري كاني انكلم مع مهبوني جززمني الله يكرمها عنك وامثالك
------------------------------------------------
The original text: 🔪😷 بايع_الكليجاقاعد يستعرض قدام البنات شوفوني هنا وجع نعم واللي يصورون مراهقات يبيعع امحق جبل فيذمتي#
The new one: الكليجاقاعد بالشارع يستعرض قدام البنات شوفوني انبسطت هنا وجع قلبي نعم والله واللي يصورون مراهقات يبيعع امحق [UNK] ولد بايع [UNK]
🔪😷 جيل فيذمتي ؟
------------------------------------------------
The original text: والله من الرخص فيكم ماعاد تدرون وش تسوون بارخيصات 👎 #بايع_الكليجا
The new one: الكليجا [UNK] بايع ♀ [UNK] [UNK] 👎 والله باجماعة من كثر الرخص فيكم ماعاد تدرون وش تسوون بارخيصات
------------------------------------------------
```

Fig. 4. Augmented hate speech tweets sample

B. 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-base model which consists of 12 hidden layers and hidden size of 768 the output of the last four hidden layers where each layer is of dimensions sentence length x hidden size were obtained then the the output of each layer is summed to produce a single vector of sentence length x hidden size dimensions then the mean of this vector was computed to produce a single vector of hidden size length which represents the feature vector for the tweet that will be fed to the machine learning models.

C. Training

Scikit-Learn library implementation of Logistic Regression and Random Forest were used in training phase. For the Logistic Regression model a C parameter of 1e-3 and saga solver were used. For the Random Forest Model a max sample parameter of 0.4 was used.

V. Results and Performance Evaluation

Before training the model the train dataset were split into 70% for training and 30% for testing to use for evaluating the model along with the development and test datasets that were provided by the OSACT2022 shared task.

A. Performance Evaluation Metrics

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. Recall is the fraction of relevant tweets that are classified which is formulated in equation 2. F1-score is the mean of precision and recall which is formulated in equation 3. Accuracy the fraction of correct tweets that have been classified from actual classes as shown in Equation 4.

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

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

$$\text{f1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

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

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 actually 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 actually labeled as task class (offensive, hate speech).

B. Results

The baselines for evaluation provided by OSACT2022 are as following:

Task	Accuracy	Precision	Recall	f1-score
Offensive detection	0.65	0.65	0.65	0.65
Hate speech detection	0.89	0.89	0.89	0.89

The results obtained for each model and dataset for the offensive detection task are as following:

Model	Dataset	Accuracy	Precision	Recall	f1-score
Logistic Regression	Train	0.85	0.85	0.85	0.85
Logistic Regression	Test(30% of train)	0.81	0.81	0.81	0.81
Logistic Regression	Development	0.8	0.77	0.78	0.78
Logistic Regression	OSACT2022-Test	0.8	0.78	0.78	0.78
Random Forest	Train	0.97	0.97	0.97	0.97
Random Forest	Test(30% of train)	0.77	0.77	0.77	0.77
Random Forest	Development	0.75	0.72	0.73	0.72
Random Forest	OSACT2022-Test	0.74	0.72	0.72	0.72

The results obtained for each model and dataset for the hate speech detection task are as following:

Model	Dataset	Accuracy	Precision	Recall	f1-score
Logistic Regression	Train	0.91	0.91	0.91	0.91
Logistic Regression	Test(30% of train)	0.91	0.91	0.91	0.91
Logistic Regression	Development	0.89	0.7	0.81	0.74
Logistic Regression	OSACT2022-Test	0.89	0.73	0.81	0.76
Random Forest	Train	0.98	0.98	0.98	0.98
Random Forest	Test(30% of train)	0.9	0.9	0.9	0.9
Random Forest	Development	0.87	0.67	0.81	0.7
Random Forest	OSACT2022-Test	0.87	0.69	0.79	0.73

VI. Conclusion

We present an approach to detect Hate speech texting detection based on natural language processing.our approach can be applied to detecting hate speech dialogue which is composed of pure text. The challenge was to find a way to detect hate speech in the Arabic language, as the data set imbalanced, and we overcame that by using the data augmentation method and NLP aug [4] tool helped us. Using logistic regression and marbert as feature extraction we got 91%, 90% and 91% for precision, recall and f1-score for hate speech tweets. Furthermore, we aim to develop a more balanced dataset and execute the algorithm in a real environment. The results will indicate if any new algorithms might be necessary to improve detection.

References

[1] Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. Arbert & marbert: deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*, 2020.

[2] Muhammad "Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah" Nagoudi. "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 "2021". "Association for Computational Linguistics".

[3] Abeer Abuzayed and Tamer Elsayed. 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, 2020.

[4] Shaima Al-khalifa, Ibrahim Aljarah, and Mohammad A M Abushariah. Hate speech classification in arabic tweets. *Journal of Theoretical and Applied Information Technology*, 98:1816–1831, 2020.

[5] Wassen Aldjanabi, Abdelghani Dahou, Mohammed AA Al-qaness, Mohamed Abd Elaziz, Ahmed Mohamed Helmi, and Robertas Damaševičius. Arabic offensive and hate speech detection using a cross-corpora multi-task learning model. In *Informatics*, volume 8, page 69. Multidisciplinary Digital Publishing Institute, 2021.

[6] Raghad Alshaalan and Hend Al-Khalifa. Hate speech detection in saudi twittersphere: A deep learning approach. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 12–23, 2020.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Ibrahim Abu Farha and Walid Magdy. Benchmarking transformer-based language models for arabic sentiment and sarcasm detection. In *Proceedings of the sixth Arabic natural language processing workshop*, pages 21–31, 2021.

[9] Hossam Faris, Ibrahim Aljarah, Maria Habib, and Pedro A Castillo. Hate speech detection using word embedding and deep learning in the arabic language context. In *ICPRAM*, pages 453–460, 2020.

[10] Edward Ma. Nlp augmentation. https://github.com/makcedward/nlpaug, 2019.

[11] Hamdy Mubarak, Sabit Hassan, and Shammur Absar Chowdhury. Emojis as anchors to detect arabic offensive language and hate speech. *arXiv preprint arXiv:2201.06723*, 2022.

[12] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[14] Burnap P. Javed A. Liu H. Williams, M. L. and Ozalp. 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.*, 2020.

[15] Junjie Yang and Hai Zhao. Deepening hidden representations from pre-trained language models. *arXiv preprint arXiv:1911.01940*, 2019.

[16] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, 2021.