https://doi.org/10.51483/IJDSBDA.2.1.2022.20-25

ISSN: 2710-2599

SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Data Science and Big Data Analytics

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Detecting Offensive Language in Multi-Dialectal Arabic Social Media

Ahmed Fahmy[1*]

[1]Computer Science and Engineering Department, The American University in Cairo, Cairo, Egypt. E-mail: awael@aucegypt.edu

Article Info

Volume 2, Issue 1, May 2022
Received : 13 February 2022
Accepted : 09 April 2022
Published : 05 May 2022
doi: 10.51483/IJDSBDA.2.1.2022.20-25

Abstract

Recently, reliance on social media has been steadily increasing from year to year. And as an anonymous medium of communication, people tend to share offensive comments which could be problematic and potentially cause a lot of harm to society. In order to find ways of addressing this issue, researching an automated method that detects offensive text within social media platforms has become important. Research in this field within the Arabic language is not as widely available as in other languages. Due to recent breakthroughs in Arabic Natural Language Processing, we were able to achieve results which are more accurate in detecting offensive content within social media. The Arabic language is in itself a different challenge compared to English, being a morphologically rich language. With the recent breakthrough in transformer based models such as BERT, which have been able to achieve state-of-the-art results in various tasks and building upon the AraBERT pre-training which has been proven to outperform multilingual BERT, as well as utilizing Arabic specific methods of pre-processing, we were able to achieve better results than established approaches for this task. Specifically, the BERT-base model achieved an F1-score of 84.88% on a multi-platform, multi-dialect dataset.

Keywords: *Arabic, Language, Multi-dialect, offensive, BERT*

1. Introduction

Offensive speech on social media has many adverse effects on users, and automated detection of this speech can help to regulate this issue. Research has been much slower in Arabic language NLP tasks relative to English or Latin languages. With social media being able to reach more people in the world quicker than ever, as highlighted by users in the Arab region making up 8.4% of Facebook users (Salem, 2017), research into the issue for the Arabic language is necessary.

This paper is organized into three main sections. Section II discusses previous work on the topic, Section III is an explanation of the approach we took when addressing the issue. Finally, Section IV shows the results and compares with results obtained by other approaches, as well as an analysis of said results.

2. Established Solutions

2.1. Dataset Construction for the Detection of Anti-Social Behavior in Online Communication in Arabic

This research paper mainly studies all types of antisocial behavior such as offensive language and cyber-bullying. The purpose of this research is to collect data, annotate it, add useful features and give a verdict on

* *Corresponding author: Ahmed Fahmy, Computer Science and Engineering Department The American University in Cairo, Cairo, Egypt. E-mail: awael@aucegypt.edu*

whether it would be suitable for usage in machine learning models. They collected Arabic text from YouTube comments which serve the purpose of offensive language detection. They also take into account the different Arabic dialects in their prediction. Furthermore they also take into consideration misspellings, and even text written in various languages. After analyzing the data and while annotating it also adding useful features as discussed, they ultimately realized that this data would be useful for usage in machine learning models for further research towards the detection of offensive language.

2.2. Detecting Offensive Language on Arabic Social Media Using Deep Learning

In this paper the authors look at solving the same topic, attempting four different approaches to the problem, mainly revolving around deep learning. These were Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (Bi-LSTM), Bi-LSTM with attention mechanism, and a combined CNN-LSTM architecture. On all four approaches, they trained and tested on the same dataset of Arabic YouTube comments collected by Alakrot *et al*. (2018) and followed the same preprocessing techniques recommended in the same paper. They took it one step further by applying word embedding. Specifically, they used the AraVec word embedding (Soliman *et al.,* 2017). Ultimately the CNN-LSTM architecture proved to get the best results which was an accuracy of around 87%.

2.3. Multilingual and Multi-Aspect Hate Speech Analysis

In this paper, they widened the task by including hate speech prediction in three languages (Arabic, English, and French) rather than only Arabic for instance. They gathered a dataset of around 13,000 tweets and after labeling the tweets by using the Amazon Mechanical Turk, a crowdsourcing platform. They first tried implementing baseline models such as Logistic Regression which is fed by data from Bag of Words. Then they utilized a Bidirectional LSTM, and attempted training on single languages, and on multiple languages simultaneously in the same dataset. They concluded that in some tasks a multi language model can outperform single language ones. Their best F-score was an 86% average between that of the three languages when detecting the directness of a tweet (F1-score of 84%). When it comes to the Arabic tweets in the dataset, they were able to achieve an F1-score of 56% when detecting the targeted label, yet it must be noted that the labels were either 'Normal' or five other categories of abuse (6 classes).

2.4. Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere

In this research paper they seek developing an automated method towards detection of offensive content within twitter specifically, mainly in Arabic. They focus specifically on religious hate speech. After building their own dataset, they moved towards classification models using mainly deep learning models. They first moved towards pretraining their word embeddings and they chose then to apply that on a Recurrent Neural Network with Gated Recurrent Units.

2.5. L-HSAB: A levantine Twitter Dataset for Hatespeech and Abusive Language

In this paper, they focus on detection of Arabic hate speech. The research is done on the levantine dialect specifically. They create a dataset which consists of a combination of datasets released by former research on the topic, in order to get a bigger more representative dataset. For their implementation they then looked at the most common words within their dataset, assigning hate scores to each of those words. They then used a Support Vector Machine as their supervised model to solve this classification problem and received a final F1-score of 89.6%.

2.6. Arabic Offensive Language on Twitter: Analysis and Experiment

In this piece of research they use another approach, they decided to use Support Vector Machine Techniques. They trained on a dataset of 10,000 tweets which they collected by themselves, they annotated their dataset with four different classes, those being offensive, vulgar, hate speech, or clean. In our implementation we used this dataset, however we operated on binary classes, offensive or clean only. As their pre-processing they used the Farasa Arabic NLP toolkit to apply a series of steps which include tokenization and removing stop words. They experimented with classification models such as the SVM and Logistic Regression coupled with AdaBoost, however ultimately they decided to use a fine-tuned version of the multilingual variant of BERT, which can take almost any language as input, where they added a dense layer and a softmax classifier.

2.7. A multi-Platform Arabic News Comment Dataset for Offensive Language Detection

Here they look at data from various different social media platforms and filter them out those being: YouTube, Reddit, Wikipedia, and Twitter. They combined their own dataset and used crowd sourcing in order to annotate the comments collected. To prepare the data they first tokenized the text removing all punctuations, URLs and stop words. They used a basic classification approach which is SVM to produce their best results.

2.8. Quick and Simple Approach for Detecting Hate Speech in Arabic Tweets

In this piece of research they were seeking to create a quick and easy approach to detecting hate speech. They tried out various models including several classical and neural learning models. They removed all punctuation and normalized the text by using Aravec word embedding (Soliman *et al.*, 2017). They tried out different models such as SVM, Gradient Boosting and Logistic Regression. However they ultimately decided to follow a more complex model, namely, they used a combination of a CNN and an LSTM for their final approach and achieved a fair result of around 73%.

3. Proposed Solution

The solution proposed by this paper is motivated by solving the shortcomings of previous works attempting the same task. These mainly revolve around three main points, namely, the lack of sufficient data, the lack of diverse data, and not deploying state-of-the-art models and methods in many cases discussed in the established solutions sections. To solve these issues we focused on collecting and aggregating a larger, more diverse dataset, applying newly proposed pre-processing methods, and training on the BERT-base model and fine-tuning the AraBERT pre-trained weights.

3.1. Aggregated Dataset

Using a single dataset of the previously constructed ones in Ousidhoum *et al.* (2019); Albadi *et al.* (2018); Alakrot *et al.* (2018); Mubarak *et al.* (2017); and Chowdhury *et al.* (2020) would not feasibly train a complex model such as BERT-base on the task at hand. For this, we propose aggregating datasets in Ousidhoum *et al.* (2019); Albadi *et al.* (2018); Alakrot *et al.* (2018) and Mubarak *et al.* (2017) into one larger, more diverse dataset that is more representative of the Arabic social media landscape. These specific datasets were selected due to the similarity in the methodology of data collection and labeling, and some changes were made to multi-label datasets as the task in this paper is a binary one.

The aggregated dataset, after ensuring the homogeneity of the data, amounted to a total of 24,242 instances for training. This dataset contains multi-platform and multi-dialectical social media posts. A large part of the performance gained is due to the larger dataset, and it would have been a larger gain if it were not for the increased complexity of the task after having added multiple dialects to the dataset. Making this change while maintaining high accuracy metrics is only possible since AraBERT is pre-trained on multi-dialect data.

3.2. Pre-Processing

The aforementioned combined dataset was pre-processed to remove specific characters and artifacts that could hinder the model performance. Various pre-processing techniques are utilized in English language models, however, a one to one import of the same methods do not typically provide the best results (Antoun *et al.*, 2020). This is because of the Arabic language's system of concatenation, meaning that different words could share the same meaning yet have different forms. For example, the "Al" prefix which is equivalent to "the" in English is attached to the word, yet it is not a part of its meaning. This inherent difference in the Arabic language often leads to worse performance when attempting to apply the same techniques used for Latin languages.

To address this issue we used the approach proposed by Abdelali *et al.* (2016) where words are first segmented into stems, prefixes and suffixes. In their paper, Abdelali *et al.* found that this method was faster and more accurate than other implementations of Arabic segmentation. This example further illustrates how the Farasa segmenter works, where this word is broken down into three:

3.3. Model

To be able to solve the problem at hand, we fine-tuned the AraBERT pre-trained Arabic representation model (Antoun *et al.*, 2020) which is based on BERT-base (Devlin *et al.*, 2018).

The BERT model, which has pushed the state-of-the-art since it's introduction, is at its core a series of Bidirectional Transformer Encoders that are stacked. The base version, has 12 encoder blocks, 12 attention heads and 768 hidden dimensions (Devlin *et al.*, 2018). The maximum sequence capacity for BERT-base is 512 words, which enabled us to use the full dataset without altering via sliding window techniques or dropping any instances, since all posts were well below this threshold. The number of parameters in the model amounted to 110M (Devlin *et al.*, 2018). Model pre- training is done by both Masked Language Modeling, where a randomly selected word is predicted and the objective is to predict the masked word via context, as well as the self describing next sentence prediction. BERT's bidirectionality allows better understanding of context and was found to achieve much better results on similar tasks (Devlin *et al.*, 2018). A model overview is found in Figure 1.

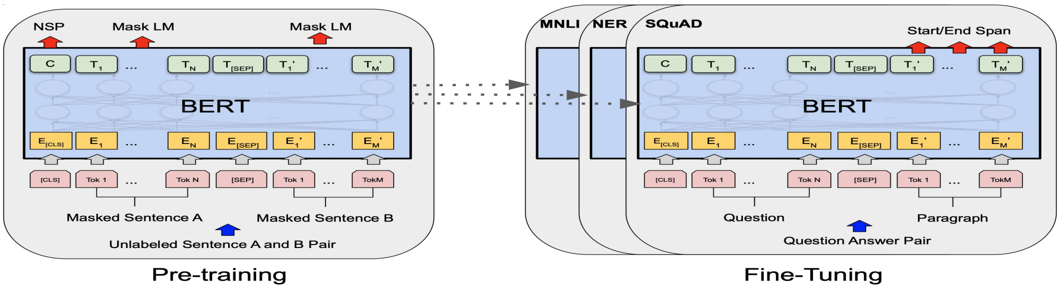

Figure 1: Aspect Ratio

The need for an increased dataset is highlighted since the number of parameters in this model exceeds other models in previous research. The model was implemented using the PyTorch framework with the HuggingFace library for Natural Language Processing (NLP). The Adam optimization algorithm was used with a warm-up linear schedule, and an initial learning rate of $2e^{-5}$. The full sequence length of 512 was utilized, with a batch size of 8.

4. Experimental Results

Figure 2 shows the Receiver Operating Characteristic (ROC) curve obtained by training the model.It must be noted that 20% of the dataset was used for testing.

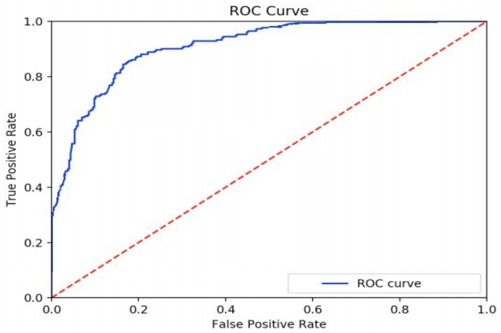

Figure 2: Receiver Operating Characteristic Curve

Below is a comparison of obtained results by previous research discussed in the established solutions section. F1-score is the most representative metric and was reported by all papers.

A more detailed analysis of the results of Table 1 must depend on the implementation details mentioned in the Established Solutions section. Half the papers used for comparison in Table 1 did not report the accuracy, which is not an issue as F1-score is a better metric to compare across different datasets that are unbalanced. It must be noted that (Ousidhoum *et al.*, 2019) was included as the dataset produced in the paper was incorporated into our dataset, yet their predicted labels broke down the offensive texts into one of five sub-categories, which explains the lower F1-score. Notably, Mulki *et al.* (2019) produced the best reported F1-score, yet that is due to the fact that it only focused on a single dialect (Levantine), and is further supported by the usage of less advanced techniques, where Naive Bayes produced the best results. The most logical comparison is with Mohaouchane *et al.* (2019) and Mubarak *et al*. (2020), with the former training on the dataset by Alakrot *et al*. (2018), and the latter using a similar approach to ours.As compared with (Mohaouchane *et al.*, 2019), their best results were obtained using a CNN model with Aravec word embedding, with a CNN+LSTM model coming a close second at 83.65%. However, the dataset was still less diverse than the one in this implementation, coming from a single social media platform (YouTube) and with the dataset consisting of a low amount of Levantine and Egyptian dialect instances relative to ones from the Gulf (Alakrot *et al*., 2018), which is bound to affect the results. When it comes to Mubarak *et al.* (2020), their usage of multilingual BERT which under performs relative to a pre-training specific to Arabic, coupled with a significantly smaller dataset (10,000) explains the relative increase in F1-score, with the implementation in Mubarak *et al.* (2020) achieving a 79.7% score. Despite this hypothesized challenge, the BERT-base model with AraBERT pre-training was able to outperform other implementations. This can be attributed to BERT's ability to “understand” contextualized information, making it less sensitive to specific words and more capable of identifying them in their appropriate context.

Table 1: Reported Metrics Comparison

	Best Reported Metrics	
Paper	**Accuracy**	***F*1-Score**
This Implementation	85	**84.88**
(Ousidhoum *et al.*, 2019)*	-	56
(Albadi *et al.*, 2018)	79	78
(Mulki *et al.*, 2019)**	90.3	**89.6**
(Mubarak *et al.*, 2017)	-	60
(Mohaouchane *et al.*, 2019)	87.84	84.05
(Mubarak *et al.*, 2020)	-	79.7
(Chowdhary *et al.*, 2020)	74	68
(Abuzayed and Elsayed, 2020)	-	73

Some of the models included in the comparison are added as a baseline, since these included the datasets used to construct the one used in this implementation. Specifically those found in references (Ousidhoum *et al.*, 2019; Albadi *et al.*, 2018; Mulki *et al.*, 2019; and Mubarak *et al.*, 2017).

5. Conclusion

In this paper, we tackle the problem of automatically detecting offensive language on multi-dialect Arabic social media. We built a larger, aggregated dataset from previously labeled datasets with some modifications to maintain data cohesion, and made sure to check methodology of labeling so that the dataset remains valid. The performance of the BERT-base model (Devlin *et al.*, 2018) using AraBERT's pretrained weights as a fine-tunable starting point (Antoun *et al.*, 2020). Pre-processing and segmentation were applied using the Farasa segmentation process by Abdelali *et al.* (2016). We were able to achieve an F1-score of 84.88% on a more representative, yet more difficult dataset. This is due to the BERT models ability for recall and context identification, which has been a breakthrough in Natural Language Processing (NLP) tasks. The AraBERT pretraining enabled us to use that model with positive results. A larger focus was placed on Mohaouchane *et al.* (2019) and Mubarak *et al*. (2020) in analysis part of Section IV as these were the only implementations to attempt a deep learning approach with a comparable dataset for a similar objective, which points to a lack of sufficient research in this important topic.

For future work, we would recommend incorporating other metadata from social media posts such as likes, replies or otherwise into the set of features that would further introduce context to the model, which we believe would increase the model accuracy.

References

Abdelali, A., Darwish, K., Durrani, N. and Mubarak, H. (2016). Farasa: A Fast and Furious Segmenter for Arabic. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations,11-16. DOI: 110.18653 / v1 / N16-3003

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, 109-114.

Alakrot, A., Murray, L. and Nikolov, N. S. (2018). Dataset Construction for the Detection of Anti-Social Behaviour in Online Communication in Arabic. *Procedia Computer Science*, 142, 174-181.

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), 69-76.

Antoun, W., Baly, F. and Hajj, H. (2020). AraBERT: Transformer-Based Model for Arabic Language Understanding. arXiv preprint arXiv:2003.00104.

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, 6203-6212.

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.

Mohaouchane, H., Mourhir, A. and Nikolov, N. S. (2019). Detecting Offensive Language on Arabic Social Media using Deep Learning. In 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), *IEEE*, 466-471.

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, 52-56.

Mubarak, H., Rashed, A., Darwish, K., Samih, Y., andAbdelali, A. (2020). Arabic Offensive Language on Twitter: Analysis and Experiments. arXiv preprint arXiv:2004.02192.

Mulki, H., Haddad, H., Ali, C. B. 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, 111-118.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y. and Yeung, D. Y. (2019). Multilingual and Multi-aspect Hate Speech Analysis. arXiv preprint arXiv:1908.11049.

Salem, F. (2017). The Arab Social Media Report 2017: Social Media and the Internet of Things: Towards Data-Driven Policymaking in the Arab World. MBR School of Government, Dubai.

Soliman, A. B., Eissa, K. and El-Beltagy, S. R. (2017). Aravec: A set of Arabic Word Embedding Models for Use in Arabic NLP. *Procedia Computer Science*, 117, 256-265.

Cite this article as: Ahmed Fahmy. (2022). Detecting Offensive Language in Multi-Dialectal Arabic Social Media. *International Journal of Data Science and Big Data Analytics*, 2(1), 20-25. doi: 10.51483/ IJDSBDA.2.1.2022.20-25.