Improving Arabic Hate Speech Identification Using Online Machine Learning and Deep Learning Models

Hossam Elzayady, Mohamed S. Mohamed, Khaled Badran, and Gouda Salama

Abstract Due to the rising use of social media platforms on a global scale to interact and express thoughts freely, the spread of hate speech has become very noticeable on these platforms. Governments, organizations, and academic institutions have all spent substantially on discovering effective solutions to handle this issue. Numerous researches have been performed in several languages to find automated methods for identifying hate speech, but there has been minimal work done in Arabic. The findings of a performance evaluation of two machine learning models, namely the passive-aggressive classifier (PAC) and the Bidirectional Gated Recurrent Unit (Bi-GRU) augmented with an attention layer, are investigated in this work. Proposed models are developed and evaluated using a multi-platform Arabic hate speech dataset. We employ term frequency-inverse document frequency (TF-IDF) and Arabic word embeddings for feature extraction techniques after running a variety of pre-processing steps. The experimental results reveal that the two proposed models (PAC, Bi-GRU with attention layer) provide an accuracy of 98.4% and 99.1%, respectively, outperforming existing methods reported in the literature.

Keywords Arabic hate speech · Text mining · Online machine learning · Deep learning

1 Introduction

With growing of Internet use, the number of people using social networks (OSN) has also risen dramatically. OSN is now the most widely used and participative platform for expressing feelings, communicating, and transferring information [1,

H. Elzayady · M. S. Mohamed (✉) · K. Badran · G. Salama
Department of Computer Engineering, Military Technical College, Cairo, Egypt
e-mail: mohamedms@mtc.edu.eg

K. Badran
e-mail: khaledbadran@mtc.edu.eg

G. Salama
e-mail: gisalama@mtc.edu.eg

X.-S. Yang et al. (eds.), *Proceedings of Seventh International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 448,
https://doi.org/10.1007/978-981-19-1610-6_46

2]. As a result of the ease of social media platforms' accessibility and anonymity, this provides a fertile atmosphere for the dissemination of violent and damaging information because of the user's desire to dominate discussion and to share their beliefs or arguments [3]. Identifying hate speech on social media is a challenging task at the moment. Text written with the intention of injury, violence, or societal upheaval directed against a particular group is referred to as hate speech [4].

This form of behavior is both socially and psychologically detrimental to users, shaking their confidence in online social media [5]. Some nations and governments throughout the world have implemented laws to limit hate speech on social media platforms. Furthermore, a large number of organizations and firms are now required to assess hate speech on their platforms and take the necessary action (e.g., deletion) [6]. Hate speech on social media has been the subject of several studies that have developed a wide variety of approaches, concentrating on the English language, while there is a dearth of studies on Arabic language [7]. There are more than a billion people who speak Arabic as a first language, and it is the internet's fifth most popular language [8]. As a result of its morphological complexity and inherent ambiguity, handling Arabic language has proven to be difficult. Additionally, Arabic includes a huge number of dialects [9].

In this paper, our goal is to build two efficient models to detect Arabic hate speech. The first model is based on implemented online supervised learning classifier, namely the passive-aggressive classifier (PAC). PAC is generally used for large-scale learning. It is one of the few 'online learning algorithms'. Online machine learning techniques employ sequential input data, and the model is updated step by step. This method does not rely on pre-existing training data, as in traditional batch learning approaches. The second model is based on developed (BI-GRU), with an attention mechanism added to the network model, providing key words with a larger weight and non-key words with a lower weight, allowing important features to stand out more.

In the rest of this paper, related work is provided in Sect. 2. Section 3 explains the proposed methodology, including the dataset description, text preparation steps, feature extraction methods, and classification models. The experimental outcomes are discussed in Sect. 4. Finally, Sect. 5 illustrates the conclusion and future work.

2 Related Work

Recently, there has been a dearth of research on Arabic natural language processing. The identification of online hate speech in an Arabic context has received little attention [10]. However, Al-Hassan and Al-Dossari [11] provided a research on text mining methodologies for dealing with hate speech in general, as well as issues for dealing with hate speech in the Arabic-speaking world. Husain and Uzuner [6] examined the most advanced natural language processing (NLP) approaches for Arabic offensive language identification, encompassing a wide range of topics such as hate speech, cyberbullying, pornography, and violent content. Haddad et al. [12]

constructed the first Arabic benchmark dataset in the Tunisian dialect known (T-HSAB). The dataset comprises 6,039 comments divided into three categories: hateful, abusive, and normal. Although they indicated that the comments were gathered from several platforms, they made no indication of which ones. In order to assess classification performance, classical machine learning classifiers used unigrams, bigrams, and trigrams were applied. All of the models were outperformed by the Naive Bayes (NB) model. Similarly, Mulki et al. [13] built a Twitter dataset for detecting hate speech and abusive language in the Levantine dialect named (L-HSAB), which seeks to prevent any hazardous words from being used automatically. Albadi et al. [14] introduced the first Arabic Twitter dataset to address the issue of religious hate, but they didn't come across any other kinds of hate speech. The dataset is used to train different classification models utilizing lexicon-based, ngrams-based, and deep-learning-based techniques. In terms of area under curve (AUC), gated recurrent unit (GRU) and pre-trained word embedding models excel over all other implemented models, earning a score of (84%). Elmadany et al. [15] used the publicly available (OSACT) dataset [16], in order to perform an Arabic hate speech detection task. Multiple M-BERT-based classifiers were employed with various fine-tuning settings. Macro F1 scores in this task didn't achieve remarkable progress comparable to those found in previous research that used more standard machine learning approaches. Hassan et al. [17] pre-processed the prior dataset (OSACT), for building a hybrid model of support vector machine (SVM) and deep neural networks for identifying abusive language. On the test set, the proposed model received an F1 score of 90.5%. Omr et al. [5] developed a binary system using 12 machine learning classifiers and two deep learning classifiers, presenting the first multi-platform dataset for Arabic hate speech identification. The RNN model had the greatest F1 score of 98.7%, with same accuracy, recall, and precision.

3 Methodology

The overall architecture of our approach is shown in detail in Fig. 1. Feature extraction techniques are applied to the dataset after it has been pre-processed using text mining techniques. Then PAC and BI-GRU models with an attention layer are applied for training. Finally, performance metrics are utilized for model evaluation.

3.1 Dataset Description

In our study, we have taken into account the first multi-platform dataset to identify hate speech in Arabic, which was gathered by [5]. Four social media networks contributed comments: Twitter, YouTube, Facebook, and Instagram. The dataset is well-balanced, unlike many others in previous work. There are a total of 10,000

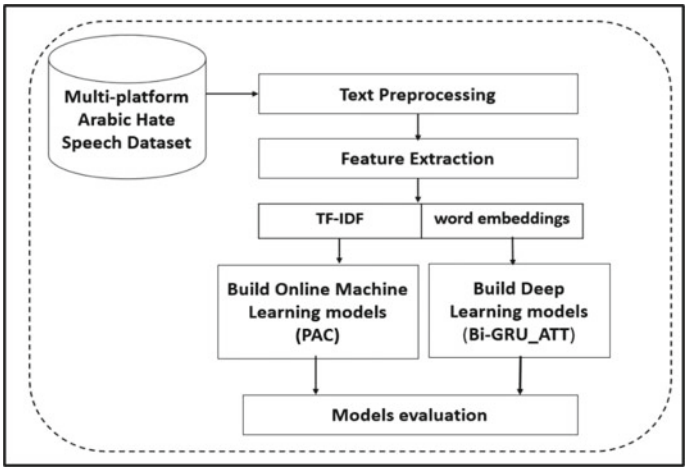

Fig. 1 Overview of methodology

Fig. 2 Word cloud of the multi-platform hate speech dataset

hateful comments, but there are also 10,000 non-hateful remarks. Figure 2 shows the world cloud of the utilized dataset.

3.2 Data Pre-processing

This is a vital step in data analysis since it eliminates data that is not strictly essential for the investigation. Pre-processing includes: deleting stop words, neglecting diacritics, discarding hashtags, eliminating punctuation, erasing links, remove empty lines, and normalizing Arabic letters as well as converting emoji and emoticons.

Finally, to guarantee that only Arabic-based letters remain when the process is completed, we utilize the alphabet-detector Python package.

3.3 Feature Extraction

We used term frequency-inverse document (TF-IDF) and word embeddings as our major feature extraction techniques since they are straightforward and problem-independent. First, TF-IDF calculates the relevance of a word to a document in a set of documents [2, 18]. As a result, this technique of operation distinguishes between common and significant words. Second, the most widely used distributed representation of terms is word embeddings. This makes it possible to investigate and identify any word similarity [7]. For Arabic word embedding architectures, we used the pre-trained AraVec2.0 [2].

3.4 Classification Models

Supervised online learning and deep neural networks are used as classification models in our experiments, as stated in the following subsections.

Passive-aggressive classifier. PAC is a notable classifier in online learning algorithms. If the classification produces the desired outcome, this algorithm remains inactive. However, it gets aggressive if the categorization produces an inaccurate result. It does not converge, in contrast to the majority of other algorithms [19]. The key premise of this algorithm is that it observes data, learns from it, and then discards it without retaining it. A classification upgrade is accomplished by solving a restricted optimization problem: The new classification should be as close to the previous one as feasible, with at least a unit margin on the most recent cases [19, 20]. In the face of noise, forcing a unit margin might be excessively aggressive. The passive-aggressive classifier takes a matrix of TF-IDF features as input. As a result, a model is constructed that is trained on the data from the training set and then applied to the test set to assess the classification's performance.

Bidirectional Gated Recurrent Unit with Attention. Two control gates, a reset gate and an update gate, are included in the GRU neural network [21]. Bi-GRU is a sequence processing model made up of two GRUs. One takes information in a forward direction, whereas the other takes it backwards [22, 23]. Text categorization using the Bi-GRU approach relies on associations between words. Instead than using keyword significance in selecting a text's categorization, they evaluate all words equally. By augmenting BI-GRU with an attention mechanism, it is possible to learn which words are more critical to the categorization by giving these keywords a larger weight. Results in a variety of text categorization tasks have been demonstrated to be improved by using this mechanism [23, 24].

4 Experimental Analysis and Results

The results and assessment of the implemented models are presented in this section. All tests are done in Google Colab Pro by using: NumPy, pandas, re, Alphabet Detector, Sklearn, and Keras packages. The results are determined in accordance with the accuracy, precision, recall, and F1 score values.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \tag{1}$$

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{false positive}} \tag{2}$$

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{false negative}} \tag{3}$$

$$F1 - \text{score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{4}$$

In order to evaluate our suggested models and compare it with approach proposed in [5], experiments are carried out using a multi-platform Arabic hate speech dataset with a total of 20,000 categorized comments, as described in (3.1). Pre-processing techniques which described in (3.2) are applied for getting rid of any noise from the dataset. Training and test sets are generated from the dataset. The training phase consumes 80% of the data, whereas the testing process consumes 20% of the data (Table 1).

We suggested two models, the first of which is based on PAC algorithms and was trained using the TF-IDF technique. The second model is built on BI-GRU and has an attention mechanism; the model is trained using pre-learned word embeddings (AraVec 2.0). Table 2 displays the parameters used in BI-GRU with attention model.

The results acquired by all the algorithms for the various performance measures are shown in Table 2. According to Table 2, it is obvious that deep learning performs a little better than online machine learning, and BI-GRU with attention is the best

Table 1 Tuned values of the hyperparameters

Hyperparameter	Value
Embedding dimension	300
Loss function	Categorical_crossentropy
Bidirectional GRUs unit	64
Optimizer	Adam
Batch size	128
Dropout	0.5
Number of epochs	10

Table 2 Models' evaluation performances

Model	Accuracy	Precision	Recall	F1_score
PAC (%)	98.4	98.51	98	98.42
BI-GRU with attention (%)	99.1	99.2	99.1	99.1

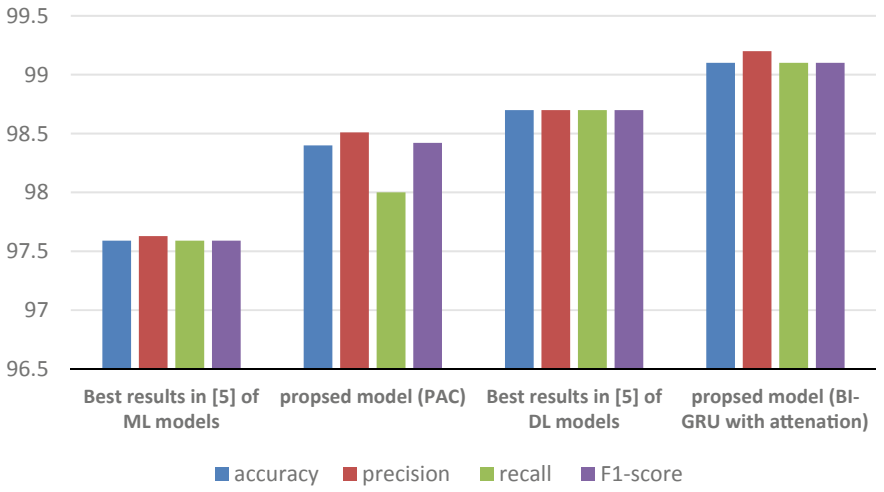

Fig. 3 Performance evaluation of our proposed Arabic hate speech detection models using ML compared with models in [5]

architecture for classifying Arabic hate speech in online social networks with an accuracy of 99.1%, F1 score of 99.1%, recall of 99.1%, and precision equal to 99.2%.

The effectiveness of our proposed methodology in comparison to the comparative methodology in [5] is shown in Fig. 3. In light of the findings, we have discovered some interesting observations. First, the results show that our proposed model BI-GRU with attention is clearly superior to the online machine learning PAC and comparative approach including traditional machine learning and recurrent neural network models. Second, our proposed model based on online machine learning algorithm PAC exhibited the best performance, outperforming all classical machine learning models used in [5]. Finally, it can be deduced that our suggested model Arabic hate speech BI-GRU with attention performed the best when compared to the other described in the related work section.

5 Conclusions

In this study, we identify hate speech in Arabic social media using the first Arabic hate speech dataset gathered from several platforms. We suggest two effective models

using the online machine learning algorithms PAC and Bi-GRU augmented by an attention layer. Several data preparation and text representation techniques have been conducted. The results indicate a notable improvement in the accuracy of the online machine learning classifier PAC compared with conventional machine learning algorithms. The results also showed the effectiveness of the Bi-GRU with attention model and its superiority over all models used in classifying hate speech in the Arabic language. For upcoming plans, we plan to assess the effects of various contextualized word embedding techniques (e.g., BERT, GPT, GPT-2, and Elmo) on hate speech models. Another area of future work is to look into recognizing other types of harmful information on social media, such as video or audio with hateful speech.

Acknowledgements The authors would like to thank **Dr. Ahmed Omar**, From Computer Science Department, Faculty of Science, Minia University, for his support in providing us with the dataset.

References

1. Cortis K, Davis B (2020) Over a decade of social opinion mining. Springer, Netherlands. https://doi.org/10.1007/s10462-021-10030-2
2. Aljarah I, Habib M, Hijazi N, Faris H, Qaddoura R, Hammo B, Abushariah M, Alfawareh M (2020) Intelligent detection of hate speech in Arabic social network: a machine learning approach. J Inf Sci. https://doi.org/10.1177/0165551520917651
3. Jahan MS, Oussalah M (2021) A systematic review of Hate Speech automatic detection using Natural Language Processing (2021)
4. Salminen J, Hopf M, Chowdhury SA, Jung S, gyo, Almerekhi H, Jansen BJ (2020) Developing an online hate classifier for multiple social media platforms. Human-centric Comput Inf Sci 10:1–34. https://doi.org/10.1186/s13673-019-0205-6
5. Omar A, Mahmoud TM (2020) Comparative performance of machine learning and deep learning algorithms for Arabic Hate Speech detection in OSNs comparative performance of machine learning and deep learning algorithms for Arabic Hate Speech detection in OSNs. Springer International Publishing. https://doi.org/10.1007/978-3-030-44289-7
6. Husain F, Uzuner O (2021) A survey of offensive language detection for the Arabic Language. ACM Trans Asian Low-Resour Lang Inf Process 20:1–44. https://doi.org/10.1145/3421504
7. Abuzayed A, 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, pp 109–114
8. Al-Hassan A, Al-Dossari H (2021) Detection of hate speech in Arabic tweets using deep learning. Multim Syst. https://doi.org/10.1007/s00530-020-00742-w
9. Hegazi MO, Al-Dossari Y, Al-Yahy A, Al-Sumari A, Hilal A (2021) Preprocessing Arabic text on social media. Heliyon. 7:e06191. https://doi.org/10.1016/j.heliyon.2021.e06191
10. Faris H, Aljarah I, Habib M, Castillo PA (2020) Hate speech detection using word embedding and deep learning in the Arabic Language Context Hate Speech detection using word embedding and deep learning in the Arabic Language context. https://doi.org/10.5220/0008954004530460
11. Al-Hassan A, Al-Dossari H (2019) Detection of hate speech in social networks: a survey on multilingual corpus, pp 83–100. https://doi.org/10.5121/csit.2019.90208
12. Haddad H, Mulki H, Oueslati A (2019) T-HSAB: A Tunisian Hate Speech and abusive dataset. Commun Comput Inf Sci 1108:251–263. https://doi.org/10.1007/978-3-030-32959-4_18

13. Mulki H, Haddad H, Bechikh Ali C, Alshabani H (2019) L-HSAB: a Levantine Twitter dataset for hate speech and abusive language, pp 111–118. https://doi.org/10.18653/v1/w19-3512
14. Albadi N, Kurdi M, Mishra S (2018) Are they our brothers? Analysis and detection of religious hate speech in the Arabic Twittersphere. In: Proceedings of the 2018 IEEE/ACM international conference on advances in social networks analysis and mining, ASONAM 2018, pp 69–76. https://doi.org/10.1109/ASONAM.2018.8508247
15. Elmadany A, Zhang C, Abdul-Mageed M, Hashemi A (2020) Leveraging affective bidirectional transformers for offensive language detection, pp 102–108
16. Mubarak H, Darwish K, Magdy W, Elsayed T, Al-Khalifa H (2020) Overview of {OSACT}4 {A}rabic 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
17. Hassan S, Samih Y, Mubarak H, Abdelali A, Rashed A, Chowdhury S (2020) ALT submission for OSACT shared task on offensive language detection, pp 61–65
18. Elzayady H, Badran KM, Salama GI (2019) Sentiment analysis on Twitter Data using Apache spark framework. In: Proceedings—2018 13th international conference on computer engineering and systems, ICCES 2018, pp 171–176. https://doi.org/10.1109/ICCES.2018.8639195
19. Gupta S, Meel P Passive-aggressive classifier. Springer, Singapore. https://doi.org/10.1007/978-981-15-7345-3
20. Nagashri K, Sangeetha J Passive-aggressive classifier and other machine learning algorithms. Springer, Singapore. https://doi.org/10.1007/978-981-33-6987-0
21. Li P, Luo A, Liu J, Wang Y, Zhu J, Deng Y, Zhang J (2020) Bidirectional gated recurrent unit neural network for Chinese address element segmentation. ISPRS Int J Geo-Information 9. https://doi.org/10.3390/ijgi9110635
22. Tay NC, Tee C, Ong TS, Teh PS (2019) Abnormal Behavior recognition using CNN-LSTM with attention mechanism. In: 2019 IEEE international conference on electrical, control and instrumentation engineering, ICECIE 2019—proceedings. https://doi.org/10.1109/ICECIE47765.2019.8974824
23. Haddad B, Orabe Z, Al-Abood A, Ghneim N (2020) {A}rabic offensive language detection with attention-based deep neural networks. In: Proceedings of the 4th workshop on open-source Arabic Corpora and processing tools, with a shared task on offensive language detection, pp 76–81
24. Mohaouchane H, Mourhir A, Nikolov NS (2019) Detecting offensive language on Arabic social media using deep learning. In: 2019 6th International conference on social networks analysis, management and security, SNAMS 2019, pp 466–471. https://doi.org/10.1109/SNAMS.2019.8931839