A hybrid approach based on personality traits for hate speech detection in Arabic social media

Hossam Elzayady, Mohamed S. Mohamed, Khaled M. Badran, Gouda I. Salama
Department of Computer Engineering, Military Technical College, Cairo, Egypt

Article Info

Article history:

Received Mar 13, 2022
Revised Jun 21, 2022
Accepted Jul 19, 2022

Keywords:

Deep learning
Hate speech
Machine learning
Personality traits
Text mining

ABSTRACT

In recent years, as social media has grown in popularity, people have gained the ability to freely share their views. However, this may lead to users' conflict and hostility, resulting in unattractive online environments. Hate speech relates to using expressions or phrases that are violent, offensive, or insulting to a minority of people. The number of Arab social media users is quickly rising, and this is being followed by an increase in the frequency of cyber hate speech in the area. Therefore, the automated detection of Arabic hate speech has become a major concern for many stakeholders. The intersection of personality learning and hate speech detection is a relatively less studied niche. We suggest a novel approach that is focused on extracting personality trait features and using these features to detect Arabic hate speech. The experimental results show that the proposed approach is superior in terms of the macro-F1 score by achieving 82.3% compared to previous work reported in the literature.

Corresponding Author:

Hossam Elzayady
Department of Computer Engineering, Military Technical College
Cairo, Egypt
Email: hossamelzaiade@gmail.com

1. INTRODUCTION

The importance of social media has increased as a consequence of the enormous evolution in internet users [1]. Users of social media do not have constraints on posting as much as they want, including critiques and reviews [2], [3]. The potential to conceal a person’s identity may be exploited by certain users, increasing the danger of technological misconduct [4], [5]. However, this kind of freedom when using social media comes with a plethora of concerns that have to be dealt with [6].

One of the most significant difficulties is scouting for hate speech on social media. Hate speech is “abusive communication that targets particular group characteristics, such as race, religion, or gender [7], [8]. Many social media firms are now required to assess hate speech on their sites and take necessary action (e.g., deletion). The frequency of hate speech makes social media sites unattractive to marketers [8]. Automated hate speech identification is critical owing to the amount of online content produced on social media platforms, making manual inspections almost impossible.

In the social media realm, many academics have investigated hate speech detection and suggested different methods to identify it, with particular emphasis on the English language [9], [10]. Detecting hate speech in Arabic media, on the other hand, is still a developing field. The morphological sophistication and lexical ambiguities of Arabic make working with it challenging. Another problem is that Arabic has a large variety of dialects within it [11], [12]. The psychology of hatred is one of the aspects that is straightly linked to the detection of hate speech but has generally been ignored in prior research. Though a significant number of studies have examined the correlation between personality and hate speech, current techniques for automated hate speech identification often ignore these research findings [13].

In this article, we propose a novel approach for detecting hate speech on Arabic social media, including two combined models. The first one is based on extracting personality trait features from hate speech. The second model identifies hate speech, including additional features derived from the first. Our approach reveals considerable contributions from a scientific perspective. As far as we know, we are the first to develop an automated method based on personality literature to identify Arabic hate speech. This actually offers a new direction of thinking about how personality affects the detection of hate speech.

The related work is presented in section 2. The suggested hybrid approach for detecting Arabic hate speech is introduced in section 3, along with the specifics of the two-stage process. Section 4 discusses the experimental setup and findings, while section 5 concludes and suggests directions for future work.

2. RELATED WORK

2.1. Personality detection from Arabic language

Salem *et al.* [14] proposed a machine learning method for predicting the personalities of Arabic users' Twitter accounts. They published a new Twitter-based personality traits dataset called (AraPersonality) for the Egyptian dialect. In order to gather and annotate the dataset, a questionnaire comprising many multiple-choice questions (MCQs) has been developed. They removed unnecessary data such as usernames and emails during the pre-processing and cleaning of user tweets gathered. The normalization and removing repeated letters processes were applied to ensure that all of the words remained consistent. They computed the term frequency-inverse document (TF-IDF) for each user individually. The four algorithms employed were k-nearest neighbors (KNN), decision trees (DT), support vector machines (SVM), and multinomial naïve Bayes (MNB). Binary representation has an average f1-score of 59%, whereas multiclass representation has an average F1-score of 33%.

Salim *et al.* [15] utilized the above-mentioned AraPersonality dataset. They investigated the impact of pre-processing and numerical features on the accuracy of personality prediction. Two preprocessing processes were applied. These processes consist of stemming and eliminating stop words. Along with the text feature, several numerical features with an absolute correlation value higher than 0.05 were employed. Compared with the baseline model in [14], the best binary model improved by 3.0%, while the best multiclass model improved by 6.7%.

2.2. Arabic hate speech detection

Abuzayed and Elsayed [16] examined the performance of hate speech detection in Arabic tweets using conventional and deep learning methods. They used the TF-IDF and word embedding features as feature extraction techniques. Multiple experiments were conducted with seven classical and eight neural learning models on a dataset of 8,000 tweets provided by (OSACT4 Workshop - SemEval 2020) to see which model performed the best. Basic text preparation was performed, including the removal of diacritics and letter normalization, as well as punctuation, foreign characters, and numbers were ignored. Results revealed that the traditional TF-IDF word representation outperformed word embedding using classical machine learning methods. The architecture based on combined convolutional neural network (CNN) and long short-term memory (LSTM) neural networks outperforms traditional machine learning methods by achieving a score of 73% macro-F1 on the dev set.

Haddad *et al.* [17] provided the first Tunisian hate and abusive speech publicly available dataset called (T-HSAB) aimed at automating Tunisian hate content detection on the internet. Traditional machine learning classifiers like SVM and NB utilize unigrams, bigrams, and trigrams were applied. The NB model outperformed the SVM in all measures. F1 score 83.6%, recall 79.8%, precision 89.5%, and accuracy 87.9%.

Hassan *et al.* [18] applied several machine learning and deep learning techniques to detect offensive or hateful tweets. These included (CNN-BILSTM) and bagged SVM (SVM with bagged features). They performed basic preprocessing procedures to eliminate non-Arabic letters, diacritics, and punctuation, and to limit character repetition. A complicated ensemble classifier utilized SVM, bagged SVM, and CNN-BiLSTM was implemented. The ensemble classifier outperformed all other classifiers by using pre-trained Mazajak word embedding and CNN feature extraction. The highest accuracy was achieved with a score of 97.7%.

Aljarah *et al.* [6] used natural language processing (NLP) and machine learning to identify cyber hate speech in an Arabic context on Twitter. The data was suitable for analysis after removing non-Arabic characters, numbers, symbols, and punctuation. They investigated emotion features, TF vectors, bag of words (BoW) vectors, profile features, and TF-IDF vectors. A random forest (RF), Gaussian NB, and SVM were studied throughout the system design. The model's performance was evaluated using 10-fold cross-validation. The best performance was achieved by the random forest model, which included TF-IDF features as well as extra profile features. Djandji *et al.* [19] used the AraBERT model for offensive language recognition and hate speech detection and found it to be superior to single-task and multilabel methods.

2.3. Correlation of personality traits and hate speech

Earlier studies on personality-factor theories suggested that personalities were a major predictor of hate behavior [20]. The big 5 personality traits (openness, conscientiousness, extraversion, agreeableness, and neuroticism) have gotten a lot of attention [14], [20]. Lee and Ram [13] were particularly interested in studying the five personality factors and their association with hateful behavior. First openness (O), previous research has revealed conflicting or contradictory findings about the link between openness traits and hatred. Second, conscientiousness (C) has been linked to hate conduct in a negative way, according to prior studies. Third, previous research on the link between extraversion (E) and hate behavior has produced mixed findings. Fourth, generally, a negative relationship between agreeableness (A) and hatred has been shown in earlier studies. Finally, numerous studies have shown a positive relationship between neuroticism (N) and hate behavior. They developed a deep learning model based on personality for identifying hate speech online. Two real-world examples were used to verify the proposed model's accuracy. The findings indicate that the proposed model outperformed current baselines, including one developed by Google.

In summary, these associations between personality characteristics and hate speech may be observed. To the best of our knowledge from existing literature, this work is the first to infer personality features from the aforementioned AraPersonality dataset, and then combine those features with the Arabic hate speech dataset to get even better results.

3. PROPOSED HYBRID MODEL ARCHITECTURE FOR ARABIC HATE SPEECH DETECTION

Detailed explanations of the framework model are presented in this section. Figure 1 depicts the suggested hate speech detection system block diagram. The proposed model consists of two phases. The functions of each phase will be discussed in more detail in the following subsections.

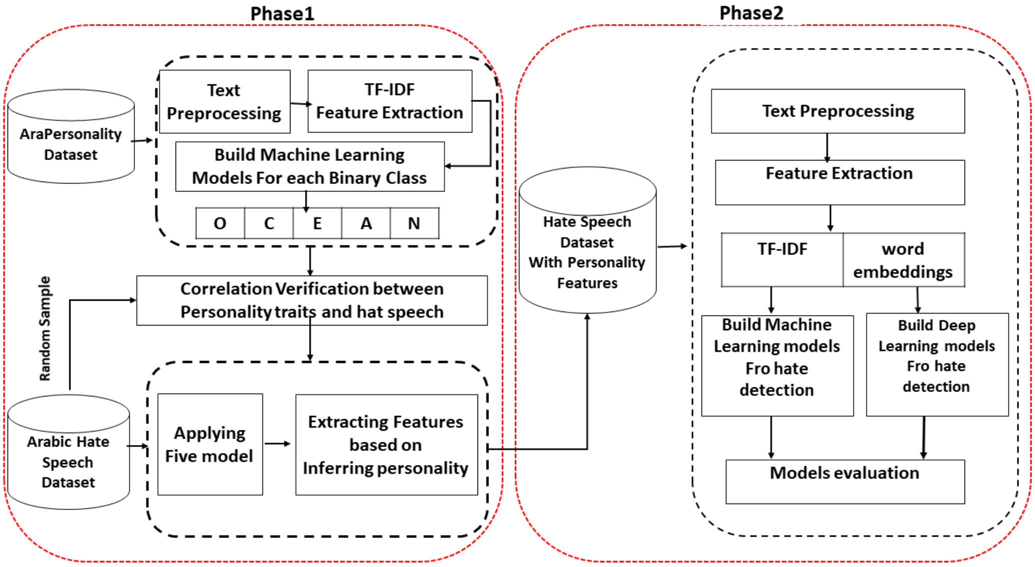
Phase1
Phase2
AraPersonality Dataset
Text Preprocessing
TF-IDF Feature Extraction
Build Machine Learning Models For each Binary Class
O
C
E
A
N
Correlation Verification between Personality traits and hat speech
Random Sample
Arabic Hate Speech Dataset
Applying Five model
Extracting Features based on Inferring personality
Hate Speech Dataset With Personality Features
Text Preprocessing
Feature Extraction
TF-IDF
word embeddings
Build Machine Learning models Fro hate detection
Build Deep Learning models Fro hate detection
Models evaluation

Figure 1. The proposed Arabic hate speech detection framework

3.1. Inferring personality features from text (Phase 1).

This phase aims to infer personality trait features from Arabic text. Many studies gathered and addressed personality trait datasets in English [21]–[24]. After our literature survey, we found just one dataset was gathered in Arabic, specifically, for Egyptian dialect [14]. This phase includes three main processes: the first primary aim is to develop five machine learning models for each of the five-character traits (O, C, E, A, N). This process involves data cleaning and pre-processing, extracting features from AraPersonality dataset, training machine learning classifiers, and then evaluating these classifiers to determine which the best is.

3.1.1. AraPersonality dataset

The dataset contains the profiles of 92 Egyptian Twitter users, as well as their Twitter feeds and their personality ratings based on the five major personality traits. Each user has an average of around 3,200 tweets [14], [15]. It is noteworthy that this dataset has been published on the Kaggle website by the authors.

3.1.2. Data preprocessing for AraPersonality

Data preparation is a critical step in data analysis since it eliminates any data that is deemed superfluous. Preprocessing includes:

- Letters normalization: we normalize *alif* (أ،آ،إ to ا), *alif maqsura* (ي،ئ to ى).
- Redundancy removal: we reduce the length of some of the letters that appear several times, for example, if a user writes (اهلاااااااااا) he/she meant (اهلا).
- Stemming: we eliminate meaningless additions from a word in order to restore it to its original form with the same meaning intact.
- Emoji and emoticon conversion: we applied changing emoji and emoticons into Arabic textual labels that explain the content of them.
- Keep only Arabic language: we used alphabet-detector python library to remove any other language letters except Arabic.

In addition to the previous processes, we also performed the following steps: remove stop words, ignore diacritics, disregard hashtags, eliminate punctuation, delete links, and remove empty lines.

3.1.3. Feature extraction

We use in this phase TF-IDF feature extraction technique. Today, 83% of text-based recommender systems in digital libraries employ TF-IDF as a term weighting method [25]. The usefulness of a word in the corpus is measured numerically, calculated by (1). Where TF is the number of times a specific word (*t*) appears in the document. Using (2), IDF determines if a term (*t*) is frequent or rare in all documents (*n*) in order to know its importance. The document frequency *(d)* is the number of documents (*d*) that include the term (*t*) [26], [27].

$$tf\text{–}idf = tf\ (t) \times\ idf(d,t) \tag{1}$$

$$idf(d,t) = log\frac{N}{df(d,t)} + 1 \tag{2}$$

3.1.4. Baseline models

The baseline models contain a variety of supervised machine learning algorithms, which will be utilized to show how each has an influence on the classifier. These algorithms are MNB, SVM, DT, and KNN [28]–[30]. All algorithms are trained in this phase using the TF-IDF technique, which is previously described.

3.2. Personality-based model for Arabic hate speech detection (Phase 2)

During this phase, we utilize the hate speech dataset that has been modified. Basic text preparation is applied to get our dataset ready for feature extraction process. Both TF-IDF and word embedding techniques were used for the feature extraction. Several experiments were carried out using a variety of classical and neural learning models for identifying Arabic hate speech on Twitter.

3.2.1. Arabic hate speech dataset

The shared task of (OSACT) in LREC 2020 offers (SemEval 2020 Arabic offensive language dataset, subtask A, B). Each tweet includes two labels: one indicating if it is offensive or not, and the other indicating whether it is hated speech or not. The dataset is published into two parts, one for training and the other for development. There are 6,839 tweets in the training dataset, including 1,371 offensive tweets and 350 hate speech tweets. There are 1,000 tweets in the development dataset, including 179 offensive tweets and 44 hate speech tweets [16], [27].

3.2.2. Data preprocessing for Arabic hate speech

To prepare our dataset for feature extraction, we do the following text preparation steps. First, we have taken off any punctuation, strange letters, numbers (including user mentions), and diacritics. Second, the normalization of Arabic characters was accomplished by augmenting the pre-trained word embedding model (AraVec2.0).

3.2.3. Text representation (non-contextual representations)

We use TF-IDF which was earlier explained, and word embeddings as our primary feature extraction methods since they are easy and problem-independent. The capability to transform words into a real vector of dimensions is provided by word embedding [31]. This enables the exploration and discovery of any word similarity [32]. It is accomplished by figuring out how words are represented as vectors by looking at the sentences in which they appear [33]. We employed the pre-trained AraVec2.0 for Arabic word embedding architecture [34]. Considering that we deal with tweets, the pre-trained Skip Gram 300D-embeddings are utilized for this study.

3.2.4. Baseline models

In this phase, both machine learning and deep learning models are implemented. RF, extra trees, DT, SVM, Gradient Boosting, XGBoost, and logistic regression (LR) were among the traditional machine learning models that we investigated. For deep learning, both recurrent neural networks (RNNs) and CNN were evaluated. We experimented with a variety of RNN designs, including LSTM, Bidirectional long short term memory (BI-LSTM), and a gated recurrent unit (GRU) [35]–[37]. We also experimented with using a hybrid of CNN and RNN models. According to prior research, in NLP tasks like sentiment analysis, the architecture that combines CNN and RNN excels better than CNN or RNN alone [38]. Figure 2 depicts the (CNN-RNN) model architecture. The joint model uses LSTM, BI-LSTM, and GRU as types of RNN.

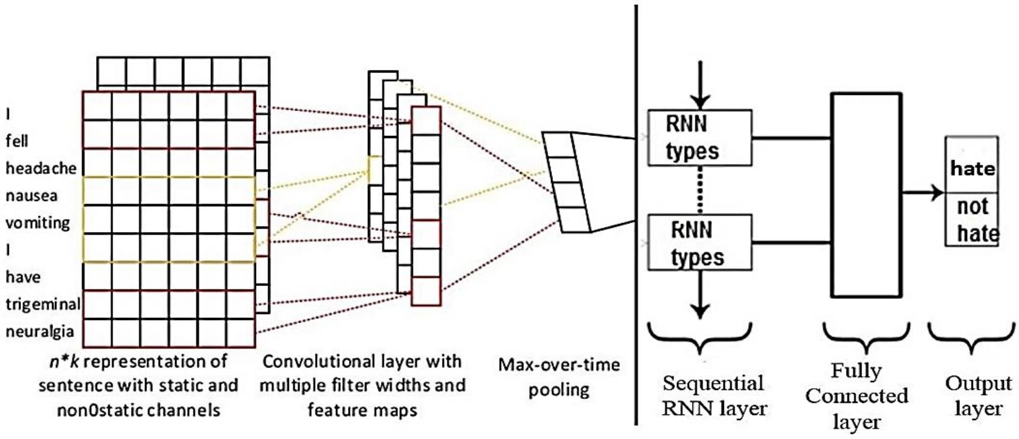

Figure 2. Joint CNN and RNN models architecture

3.2.6. Our AraBERT based proposed model (contextualized representations)

AraBERT model is the foundation of our proposed model. AraBERT is an Arabic pre-training BERT transformer model, pretrained on a large Arabic corpus that excelled in a variety of Arabic natural language processing applications [39]. Our suggested model is based on augmenting AraBERT with personality trait features.

We used AraBERTv1, which is effective for Arabic because it is based on Farasa Segmenter, to decompose the sentences into pieces (tokenization). Moreover, at the beginning of each sentence, the special [CLS] symbol is added, followed by the token [SEP] to separate sentences and inserted at the end. We used in our study the pretrained "bert-base-arabertv1" Arabic embedding with 768 hidden dimensions, 12 attention heads and encoder layers, and 110 M parameters.

4. EXPERIMENTAL SETUP AND DISCUSSION

The findings and evaluation of the implemented models for each phase that were previously presented are discussed in this section. In the first phase, two experiments are conducted. Experiment A involves five models for each binary class for inferring personality-based features from the text. Experiment B verifies the correlation between offensive language and personality characteristics. The models used in [16], [19] were the subject of our second phase investigation. We have used the identical models that were proposed in [16]. During this phase, experiment C was carried out to compare the findings after adding personality features to the original dataset. Basically, all experiments are run using Google Colab Pro. These libraries were used to create the experiment: NumPy, Pandas, Re, Alphabet Detector, as well as Sklearn and Keras.

4.1. Experiment A: inferring personalities from text

The AraPersonality dataset, which contains 3,200 categorized tweets for each user, is used to test our proposed models and compare them to the approach provided in [15]. The histogram for the binary representation of the dataset is shown in Figure 3, where the x-axis indicates the name of a trait and the y-axis reflects the number of samples in the dataset, with the two bars representing the number of samples with scores of 0 and 1 in that trait. The dataset is filtered and cleaned utilizing intensive preprocessing techniques. Only text features are used; user tweets are converted into a bigram vector that shows the frequency of each word in the tweets using TF-IDF.

Table 1 shows the resultant F-measure of our proposed models for each binary class of dataset and the model in [15] using the four classifiers. The bolded text indicates the class maximum value for our proposed models. An asterisk is put beside the best value of each class in the baseline models. It is worth noting that the best results of classes (O, E, A, and N) produced using our approach exceed those obtained using [15], whereas the top outcome in class (C) is the same for both approaches. This superiority in most of the results is most likely due to the use of more steps in preprocessing phase. Preprocessing improves model performance by reducing dimensionality and eliminating unnecessary content. Steps are demonstrated in detail in (3.1.2). The model that performed the best in each class was chosen to infer personal traits in subsequent experiments.

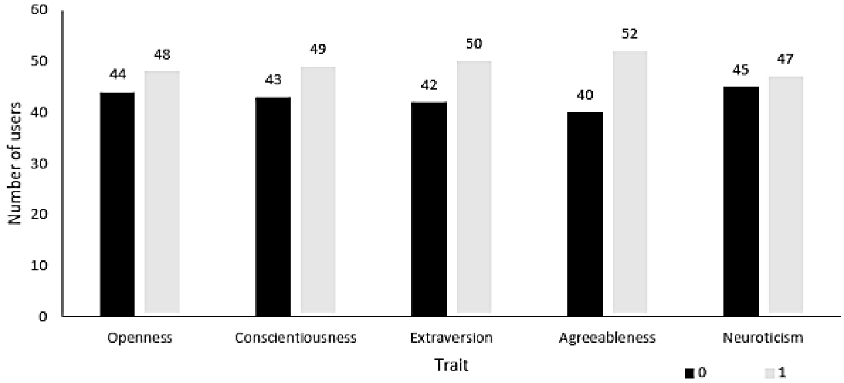

Figure 3. Histogram of AraPersonality dataset using binary representation

Table 1. F1-scores in binary representation

Model	Proposed models					Baseline models in [15]				
	O	C	E	A	N	O	C	E	A	N
DT	.61	.52	.58	.66	.63	.59 *	.48	.62*	.65	.63
KNN	.52	**.53**	**.74**	**.72**	**.69**	.50	.53*	.62*	.71*	.59
MNB	**.71**	.35	.31	.59	.62	.49	.25	.30	.56	.49
SVM	.57	.38	.50	.60	.66	.55	.31	.43	.54	.61 *

4.2. Experiment B: verifying correlation between offensive language and personality

We used the models developed in a previous experiment to verify the correlation between offensive language and personal traits. Because of its superiority in results, the KNN algorithm is used for classes (C, E, A, and N). On the other hand, MNB is selected for class O. For this experiment, we randomly selected a sample of 340 tweets from the Arabic hate speech dataset's development set. The chosen sample is balanced, 170 is offensive and 170 is not offensive.

Figure 4 reveals the performance measures of each implemented model. It is obvious that model (N) exceeds all other models in terms of precision, reaching 58%. On the other hand, the models (A, C) outperform the rest of the models in terms of recall, achieving 95% and 81%, respectively. These results demonstrate the ability of model (N) to recognize offensive speech due to its high precision rate. Due to their high recall rate, the two models (A, C) are capable of detecting non-offensive speech. According to the findings, we demonstrated that there are correlations between personality traits and hate speech. As a result, we have a strong desire to extract personality trait features from texts employing earlier models and add them during the hate speech classification process, aiming to achieve better performance.

4.3. Experiment C: Arabic hate speech detection based on personality features

At the moment, after confirming the correlation in the prior experiment, we decided to add personality features to the Arabic hate speech dataset by applying the five models to the whole dataset. For each tweet, we now have five binary features in addition to the text feature. In order to evaluate how effective these additional features are on the dataset.

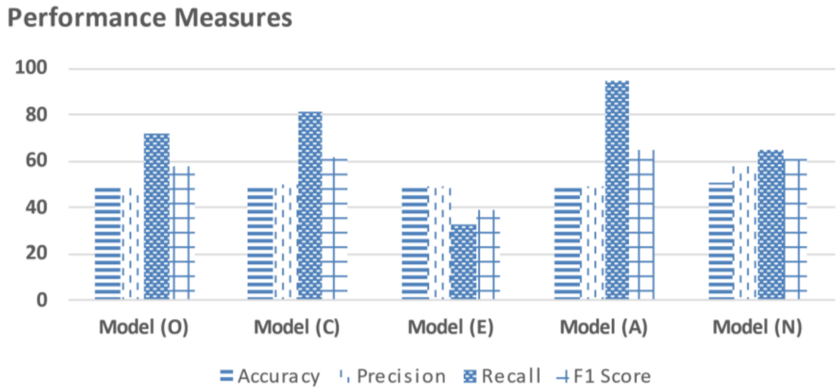

Figure 4. Performance evaluation of proposed personality model using hate speech sample

We used the same approach as Abuzayed and Elsayed [16] and compared the results after incorporating the personality trait binary features with the textual features. They only participated in subtask B of the whole dataset, which determines whether or not a tweet includes hate speech. Our models' overall performance is measured using a macro-averaged F1-score.

First, we used TF-IDF and binary features to test the seven conventional machine learning models described in section 3.2.4. From Figure 5, it is obvious that our suggested approach, based on adding personality trait features, outperforms the approach compared when using XGB, extra trees, DT, and gradient boosting machine learning algorithms. While both approaches give equal results when using SVM. The comparative approach outperformed our approach when using RF and LR algorithms. Overall, our proposed approach achieved (2%) better result than the best score achieved by comparative approach.

Second, we used pre-trained word embeddings (AraVec 2.0) and binary features to test seven deep-learning models. The first three are RNN models, namely LSTM, BLSTM, and GRU. The fourth is CNN. The next three combined CNN and different types of RNN. The parameters selected for deep learning models are shown in Table 2.

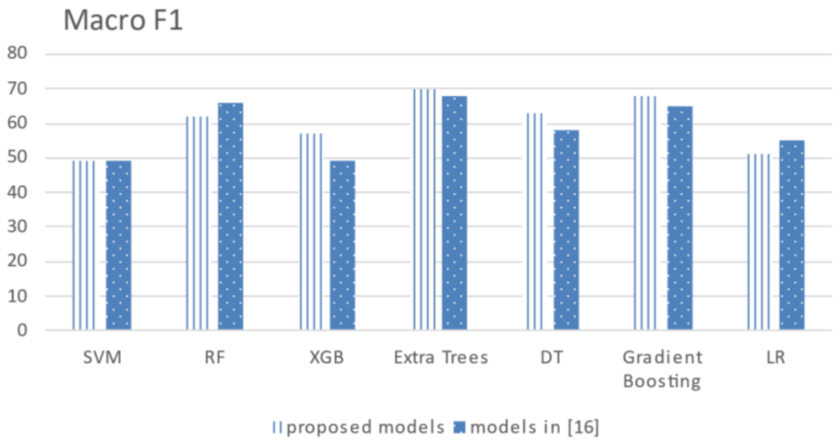

Figure 5. Performance evaluation of our proposed Arabic hate speech detection model using ML compared with models in [16]

Table 2. Parameters selected for deep learning models

Hyper parameter	Value
Embedding Dimension	300
Filters	25
Convolution function	ReLU
Kernel Size	5
Number of hidden units	16
Dropout	0.5
Optimizer	Adam

Finally, we implemented our proposed AraBert model augmented by personality features. In this experiment, we depend on Hugging Face library. We fine-tune AraBERT to be used for classification. This is done by adding a fully connected layer and a SoftMax layer. We applied the following setting: maximum length=128, patch size=16, epoch=3, epsilon=1e-8, and learning rate=2e-5. The classifier was developed using the encoder's pooled output together with a basic feed-forward neural network (FFNN) layer. The PyTorch-Transformers library was used to conduct the experiments.

Figure 6 shows the effectiveness of our proposed approach versus the comparative approach. Based on the results, we have several interesting observations. First, the results show that our proposed approach is clearly superior to the comparative approach when using all deep learning models. Second, the three combined models of CNN, and RNN types, CNN-LSTM, CNN-BILSTM, and CNN-GRU showed the highest performance, surpassing all other deep neural networks as well as traditional machine learning models. Last observation, the combined (CNN-LSTM) model achieved the best score of 77%, exceeding the compared model, which achieved a score of 73%. The superiority of these joint models may be due to the following reasons: CNN can be used to retrieve higher-level word feature sequences and RNNs types can catch long-term correlations across window feature sequences, respectively. Table 3 compares the performance of our suggested AraBERT model to the comparative method in [19]. Our suggested model, which was enhanced with personality trait features, earned the highest performance score of 82.3%, exceeding all outperforming all classical machine learning models and deep learning models used in [16].

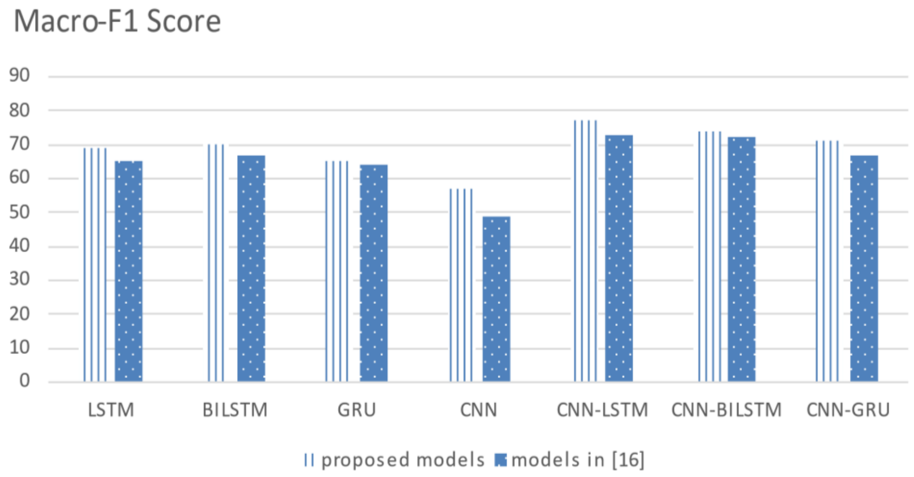

Figure 6. Performance evaluation of our proposed Arabic hate speech detection using DL compared with models in [16]

Table 3. Performance evaluation of our AraBERT model compared by model in [19]

Model	Macro-F1
AraBERT in [19]	80.6%
Our proposed AraBERT model	82.3%

5. CONCLUSION

In this paper, a personality-based framework for identifying Arabic hate speech is implemented. To our knowledge, this is the first research to utilize psychology theories to build a computational hate speech detection system in Arabic. Three experiments were conducted. First experimental outcomes confirm the high effectiveness of proposed models by applying intensive preprocessing steps. The results from the second experiment verify the theoretical psychological correlation between each personality trait and its association with hate speech. The third experiment shows the effectiveness of our proposed hate speech detection technique, which combines personality traits with textual features. For upcoming plans, there are several interesting directions. First, we will extend our proposed framework to include multi-personality trait features rather than binary; second, we will investigate sampling methods in greater depth to address the issue of imbalanced data; and third, the proposed approach may be improved for future goals in Arabic hate speech classification using multi-task learning approach.

REFERENCES

[1] K. Cortis and B. Davis, “Over a decade of social opinion mining: a systematic review,” *Artificial Intelligence Review*, Springer Netherlands, vol. 54, no. 7, pp. 4873–4965, 2021.

[2] S. Rathore, P. K. Sharma, V. Loia, Y.-S. Jeong, and J. H. Park, “Social network security: issues, challenges, threats, and solutions,” *Information Sciences*, vol. 421, pp. 43–69, Dec. 2017, doi: 10.1016/j.ins.2017.08.063.

[3] M. O. Hegazi, Y. Al-Dossari, A. Al-Yahy, A. Al-Sumari, and A. Hilal, “Preprocessing Arabic text on social media,” *Heliyon*, vol. 7, no. 2, Feb. 2021, doi: 10.1016/j.heliyon.2021.e06191.

[4] A. Al-Hassan and H. Al-Dossari, “Detection of hate speech in Arabic tweets using deep learning,” *Multimedia Systems*, no. 0123456789, 2021, doi: 10.1007/s00530-020-00742-w.

[5] F. Husain and O. Uzuner, “A survey of offensive language detection for the Arabic language,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 1, pp. 1–44, Apr. 2021, doi: 10.1145/3421504.

[6] I. Aljarah *et al.*, “Intelligent detection of hate speech in Arabic social network: a machine learning approach,” *Journal of Information Science*, vol. 47, no. 4, pp. 483–501, Aug. 2021, doi: 10.1177/0165551520917651.

[7] O. Istaiteh, R. Al-Omoush, and S. Tedmori, “Racist and sexist hate speech detection: literature review,” in *International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, Oct. 2020, pp. 95–99, doi: 10.1109/IDSTA50958.2020.9264052.

[8] M. S. Jahan and M. Oussalah, “A systematic review of hate speech automatic detection using natural language processing,” *arXiv:2106.00742*, May 2021, doi: 10.48550/arXiv.2106.00742.

[9] G. Kovács, P. Alonso, and R. Saini, “Challenges of hate speech detection in social media,” *SN Computer Science*, vol. 2, no. 2, Apr. 2021, doi: 10.1007/s42979-021-00457-3.

[10] L. Yuan, T. Wang, G. Ferraro, H. Suominen, and M.-A. Rizoiu, “Transfer learning for hate speech detection in social media,” *arxiv.org/abs/1906.03829*, Jun. 2019, doi: 10.48550/arXiv.1906.03829.

[11] M. Alruily, “Classification of Arabic tweets: a review,” *Electronics*, vol. 10, no. 10, 2021, doi: 10.3390/electronics10101143.

[12] S. M. Abdou and A. M. Moussa, “Arabic speech recognition: challenges and state of the art,” in *Computational Linguistics, Speech and Image Processing for Arabic Language*, World Scientific, 2018, pp. 1–27.

[13] K. Lee and S. Ram, “PERSONA: personality-based deep learning for detecting hate speech,” *International Conference on Information Systems, ICIS 2020 - Making Digital Inclusive: Blending the Local and the Global*, 2021.

[14] M. S. Salem, S. S. Ismail, and M. Aref, “Personality traits for Egyptian Twitter users dataset,” in *Proceedings of the 2019 8th International Conference on Software and Information Engineering*, Apr. 2019, pp. 206–211, doi: 10.1145/3328833.3328851.

[15] M. Salim, S. Saad, and M. Aref, “Preprocessing the Egyptian Arabic dialect for personality traits prediction,” *International Journal of Intelligent Computing and Information Sciences*, vol. 19, no. 1, pp. 1–12, Jun. 2019, doi: 10.21608/ijicis.2019.62603.

[16] A. Abuzayed and T. Elsayed, “Quick and simple approach for detecting hate speech in Arabic tweets,” *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, no. May, pp. 109–114, 2020.

[17] H. Haddad, H. Mulki, and A. Oueslati, “T-HSAB: a Tunisian hate speech and abusive dataset,” *Communications in Computer and Information Science*, vol. 1108, pp. 251–263, 2019, doi: 10.1007/978-3-030-32959-4_18.

[18] S. Hassan, Y. Samih, H. Mubarak, A. Abdelali, A. Rashed, and S. A. Chowdhury, “ALT submission for OSACT shared task on offensive language detection,” in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, 2020, pp. 61–65.

[19] M. Djandji, F. Baly, W. Antoun, and H. Hajj, “Multi-task learning using Arabert for offensive language detection,” in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, 2020, pp. 97–101.

[20] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding, “Peer to peer hate: hate speech instigators and their targets,” in *12th International AAAI Conference on Web and Social Media (ICWSM 2018)*, Apr. 2018, pp. 52–61.

[21] A. Aref, R. H. Al Mahmoud, K. Taha, and M. Al-Sharif, “Hate speech detection of Arabic shorttext,” in *9th International Conference on Information Technology Convergence and Services (ITCSE)*, 2020, pp. 81–94.

[22] V. Varshney, A. Varshney, T. Ahmad, and A. M. Khan, “Recognising personality traits using social media,” in *IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, Sep. 2017, pp. 2876–2881, doi: 10.1109/ICPCSI.2017.8392248.

[23] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio, “TwitPersonality: computing personality traits from tweets using word embeddings and supervised learning,” *Information*, vol. 9, no. 5, May 2018, doi: 10.3390/info9050127.

[24] G. Farnadi *et al.*, “Computational personality recognition in social media,” *User Modeling and User-Adapted Interaction*, vol. 26, no. 2–3, pp. 109–142, Jun. 2016, doi: 10.1007/s11257-016-9171-0.

[25] H. Elzayady, M. S. Mohamed, K. Badran, and G. Salama, “Improving Arabic hate speech identification using online machine learning and deep learning models,” in *Proceedings of Seventh International Congress on Information and Communication Technology*, 2023, pp. 533–541, doi: 10.1007/978-981-19-1610-6_46.

[26] H. Elzayady, K. M. Badran, and G. I. Salama, “Sentiment analysis on Twitter data using Apache spark framework,” in *13th International Conference on Computer Engineering and Systems (ICCES)*, Dec. 2018, pp. 171–176, doi: 10.1109/ICCES.2018.8639195.

[27] F. Husain, “OSACT4 shared task on offensive language detection: intensive preprocessing-based approach,” in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, 2020, pp. 53–60.

[28] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: a survey,” *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.

[29] P. Yang and Y. Chen, “A survey on sentiment analysis by using machine learning methods,” in *IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Dec. 2017, pp. 117–121, doi: 10.1109/ITNEC.2017.8284920.

[30] K. M. Alomari, H. M. ElSherif, and K. Shaalan, “Arabic tweets sentimental analysis using machine learning,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10350, 2017, pp. 602–610.

[31] D. Sagheer and F. Sukkar, “Arabic sentences classification via deep learning,” *International Journal of Computer Applications*, vol. 182, no. 5, pp. 40–46, Jul. 2018, doi: 10.5120/ijca2018917555.

[32] H. Elzayady, K. M. Badran, and G. I. Salama, “Arabic opinion mining using combined CNN - LSTM models,” *International Journal of Intelligent Systems and Applications*, vol. 12, no. 4, pp. 25–36, Aug. 2020, doi: 10.5815/ijisa.2020.04.03.

[33] A. Alwehaibi and K. Roy, “Comparison of pre-trained word vectors for Arabic text classification using deep learning approach,” in *17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2018, pp. 1471–1474, doi: 10.1109/ICMLA.2018.00239.

[34] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, “AraVec: a set of Arabic word embedding models for use in Arabic NLP,” *Procedia Computer Science*, vol. 117, pp. 256–265, 2017, doi: 10.1016/j.procs.2017.10.117.

[35] B. Roshanfekr, S. Khadivi, and M. Rahmati, “Sentiment analysis using deep learning on Persian texts,” in *Iranian Conference on Electrical Engineering (ICEE)*, May 2017, pp. 1503–1508, doi: 10.1109/IranianCEE.2017.7985281.

[36] P. Vateekul and T. Koomsubha, “A study of sentiment analysis using deep learning techniques on Thai Twitter data,” in *13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Jul. 2016, pp. 1–6, doi: 10.1109/JCSSE.2016.7748849.

[37] K. Baktha and B. K. Tripathy, “Investigation of recurrent neural networks in the field of sentiment analysis,” in *2017 International Conference on Communication and Signal Processing (ICCSP)*, Apr. 2017, pp. 2047–2050, doi: 10.1109/ICCSP.2017.8286763.

[38] A. Hassan and A. Mahmood, “Convolutional recurrent deep learning model for sentence classification,” *IEEE Access*, vol. 6, pp. 13949–13957, 2018, doi: 10.1109/ACCESS.2018.2814818.

[39] W. Antoun, F. Baly, and H. Hajj, “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*, Marseille, France. European Language Resource Association. Feb. 2020, pages 9–15.

BIOGRAPHIES OF AUTHORS

Hossam Elzayady is a Ph.D. candidate at the Department of computer engineering, received a bachelor’s degree in computer engineering and a Master of Science degree from the MTC, Cairo, Egypt, in 2005 and 2018, respectively. His research interests are in artificial intelligence, data science, and machine learning. He can be contacted at hossamelzaiade@gmail.com.

Mohamed S. Mohamed received his M.S. degree (2011) and B.S (2004) in Computer Engineering from Military Technical College, Egypt (MTC). He also received his Ph.D. degree (2018) in electrical and computer engineering from University of Idaho, USA (UI). He is currently a faculty member at the electrical and computer engineering department at MTC. His research focuses on cyber security, malicious act, and variabilities issues related to connected vehicles, survivable systems, and networks. He can be contacted at mohamedms@mtc.edu.eg.

Khaled M. Badran received a bachelor’s degree in computer engineering and a Master of Science degree from the MTC, Cairo, Egypt, in 1995 and 2000, respectively. He also received a Ph.D. degree in electrical and computer engineering from Sheffield University, UK, in 2009. He is currently a faculty member of the Department of Computer Engineering, MTC. His research interests are in artificial intelligence, data mining, semantic web, and database security. He can be contacted at khaledbadran@mtc.edu.eg.

Gouda I. Salama received his Bachelor of Engineering and Master of Engineering from MTC, Cairo, Egypt, in 1988 and 1994, respectively. As well, he received a Ph.D. degree in electrical and computer engineering from Virginia Tech. University, USA in 1999. He is currently a faculty member with the Department of Computer Engineering, MTC. His research interests are in image and video processing, pattern recognition, and information security. He can be contacted at gisalama@mtc.edu.eg.