Arabic hate speech detection system based on AraBERT

Palé Ollo Salomon
Higher Institute of Computer Science and Multimedia of Sfax
sfax, tunisia
ollosalomon@gmail.com

Zied kechaou
REGIM-lab: Research Groups on Intelligent Machines universty of sfax, national engineering school of sfax (ENIS),BP1173,3038,sfax,tunisia
Higher School of commerce of Sfax
sfax, tunisia zied.kechaou@ieee.org

Ali wali
REGIM-lab: Research Groups on Intelligent Machines universty of sfax, national engineering school of sfax (ENIS),BP1173,3038,sfax,tunisia
Higher Institute of Computer Science and Multimedia of Sfax
sfax, tunisia ali.wali@iee.org

2022 IEEE 21st International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC) | 978-1-6654-9084-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCICC57084.2022.10101577

Abstract -- Tunisia has entered a phase freedom of speech with access to social media since the Jasmine Revolution in 2011. Toxic contents such as abusive and hateful speeches have become omnipresent on Tunisian social media. Considering the side effects of these toxic contents on the psychology of users, it is necessary to detect them automatically. The dialect of Tunisian is underrepresented. As a consequence, there is not enough data set. In this paper, we present the data collection process with the aim of having a Tunisian reference dataset, to evaluate different models of hate speech and abuse detection. We also present our neural network model based on AraBERT. Our experimental results on our dataset shows that the AraBERT model performs better with an F1 score of 0.99.

Keywords: Tunisian dialect - Hate speech - Social media - Arabert

I. INTRODUCTION

Internet and social media offer to people a large range of benefits and opportunities to empower themselves in various ways. With this easy access to social media and freedom of speech, more than 8 million of Tunisians are using social media by 2022[1]. Once sensitive or even "taboo" topics such as religion and politics have become popular and widely discussed by Tunisians on social media. However, hate and abuse speech against individuals or groups of individuals have also increased. According to **[1]**, recent events such as the legalization of gender equality in inheritance, the appointment of a Jewish man as Tunisia's Minister of Tourism, and the murder of a sub-Saharan African student have triggered intense debates among Tunisians, among which most of them took place on social media networks, leading to a strong emergence of abusive/hate speech. This has created a need for tools to detect such contents online. While there is no clear distinction between hate speech (HS) and abusive speech (AS), the United Nations defines hate speech as « any type of verbal, written or behavioral communication that attacks or uses derogatory or discriminatory language in reference to a person or group on the basis of who they are. In other words, based on their religion, ethnicity, nationality, race, color, ancestry, gender or and other identity factor »[2]. Hate speech is therefore the set of ideas based on the superiority of someone according to a wide range of criteria (color, nationality, religion...) and incitement to discrimination based on the same criteria. Abusive speech is defined as any implicit or explicit insult or attack against other people, or any inappropriate language. Hate speech is known to be complex and ambiguous because it was not just a matter of identifying words. In **[2]**, the authors showed that the detection of hateful content is a difficult task compared to non-hateful content due to the lack of unique and discriminating linguistic features. Based on the definition of hate speech and abusive speech given, it is quite difficult to differentiate between hate speech and abusive speech which is usually subject to personal bias resulting in low inter-rater agreement **[3]**. The Arabic language is known to be difficult and ambiguous compared to English language **[4] [5] [6] [7]**, Arabic content on social networks **[8] [9] [10]** is noisy with different dialects, and most Arabic users do not care to use correct grammar or spelling. All of these factors make hate speech almost impossible to detect and identify using conventional features widely adopted in many language-based tasks. The Tunisian dialect called "Tounsi" or "Derja" is different from modern standard Arabic. It is a mixture of words and phrases of Amazigh, French, Turkish, Italian and other languages.

In the rest of this paper, we will present some related work in Section 2. In addition, we will represent our data collection and preparation process in section 3. Furthermore, our proposed models will be presented in section 4. As for section section 5, a brief discussion of the results. Finally, we will make a summary and perspectives for the future.

[1] https://www.digital-discovery.tn/

[2] https://www.un.org/en/genocideprevention/

2022 IEEE 21th Int'l Conf. on Cognitive Informatics & Cognitive Computing (ICCI*CC'22)
Y. Wang, K.N. Plataniotis, B. Widrow, W. Pedrycz, W. Kinsner, P. Spachos & S. Kwong
978-1-6654-9084-9/22/$31.00 ©2022 IEEE

II. RELATED WORKS

Social media contains a wide range of toxic speech that affects many different targets. Examples of such language include hate speech and abusive language. Several works in the literature have been conducted to detect and locate these types of language. The work of **[11]** proposes an approach based on the CNN-BiLSTM architecture. This approach has achieved good results with respect to the F1 0.73 measure. Abuzayed et al adopted a "fast and simple" approach by which they study the effectiveness of 15 classical (e.g., SVM, decision tree..) and neural (e.g., CNN) learning models **[12]**, this approach shows that the best (neural) deep learning models outperform the classical models and their best classifier (which combines both CNN and RNN in a common architecture) obtained a macro-F1 score of 0.73. On datasets extracted from Twitter, the authors of **[13]** applied logistic regression, naive basis, Decision Trees, random forests and support vector machines as classification models for hate speech recognition. The dataset includes 24,802 tweets, the authors used a combination of unigram, bigram, trigram and sensitive scores as features. They also include tweet-specific features such as number of hashtags, user mentions, retweets and URLs. In the experiments, the best F1 macro score of 90% was obtained using logistic regression.

The objective of Hao Chen et al **[14]** was to explore automatic detection of abusive content using a variety of supervised machine learning techniques. They compared more traditional approaches with more recent neural network-based approaches, namely the SVM classifier and two deep neural classifiers: CNN and RNN. They also compared ngrams and word integrations for feature representation. They concluded that the use of word embeddings that have been pre-trained on the same data source as the next task is an advantage for the abusive content detection task. The SVM classifiers performed best on balanced datasets, with balance achieved through oversampling. The results of the full analysis of the ability of the different classifiers to handle class imbalance show that the deep learning models performed well on extremely unbalanced datasets, while the SVM was unable to identify the minority abusive content class.

Ahmad A. Al Sallab et al **[15]** used a deep learning architecture on the SA dialect dataset. They first experimented with a simple LSTM architecture on three dialectal SA datasets, with poor results. Then, they used a model that combines LSTM and CNN which led to better results. Finally, they obtained state-of-the-art results by combining the more elaborate BiLsTM and CNN models with more convolutional layers.

In the paper **[16]**,a dataset consisting of 4203 comments and divided into seven categories were used to train a deep recurrent neural network (DRN) model to classify and detect hate speech. These categories are religion, racism, anti-gender content, violent content, offensive content, insulting/intimidating content, normal positive and normal negative comments. The dataset was extensively preprocessed and labeled, and its features were extracted. The proposed RNN architecture, called DRNN-2, consisted of 10 layers with 32 batch sizes and 50 iterations for the classification task. Another model consisting of five hidden layers, called DRNN-1, was used only for binary classification. They achieved a recognition rate of 99.73% for the binary classification, 95.38% for the three Arabic comment classes, and 84.14% for the seven Arabic comment classes.

In the paper **[17]** the authors developed a high-quality textual corpus of Arabic hate and offense speech. They performed an extensive empirical analysis by evaluating a variety of feature selection methods in a supervised classification framework including machine and deep learning methods. They trained a large number of classifiers on 2-, 3-, and 6-class datasets with heterogeneous feature spaces. From their results, for traditional machine learning algorithms, SVM dominates Naive Bayes and Logistic Regression algorithms in all three tasks and all feature extraction methods. For binary classification, the combination of word (1-3) and character (1-5) ngrams achieved the highest accuracy with an F1-macro of 85.16%. For multi-class datasets, SVM achieved 73.11% and 66.86% using char-ngrams for the 3-class task and the hybrid of words and char-ngrams for the 6-class task, respectively. Their CNN+mBert deep neural network-based model outperformed all other learned models in all three prediction tasks with 87.05% for the 2-class task, 78.99% for the 3-class task and 75.51% for the 6-class task. The work of **[18]** is the first to address the problem of identifying religious hate speech on Arabic Twitter. In this work, they describe how they created the first publicly available Arabic dataset annotated for the task of hate speech detection. They then developed various classification models using lexicon-based, n-gram-based, and deep learning approaches. In their work, they present a detailed comparison of the performance of different approaches on a brand new dataset. They find that a (RNN) architecture with (GRU) and pre-trained word embeddings can adequately detect religious hate speech with 0.84 (AUROC).

The authors of **[19]** proposed four different neural network architectures namely Convolutional Neural Network (CNN), Bidirectional Long-Term Short-Term Memory (Bi-LSTM), Bi-LSTM with attention mechanism and a combined CNN-LSTM architecture. These networks were trained and tested on a labeled dataset of Arabic YouTube comments. They applied Bayesian optimization techniques to tune the hyperparameters of these neural network models. After training and testing each network using 5-fold cross-

validation, the CNN-LSTM model achieved the highest recall (83.46%), followed by CNN (82.24%), Bi-LSTM with attention (81.51%), and Bi-LSTM (80.97%).

For the databases to be exploited, we find that in the work of **[20]** the authors have created a reference dataset called (L-HSAB) which includes 6000 tweets for the detection of hate speech and abusive language in Arabic and specifically in the Levantine dialect (Syrian and Lebanese). The proposed dataset is classified into three classes; normal, abusive and hateful. The authors of **[21]** constructed a dataset that contains 6600 tweets for the detection of Arabic religious hate speech.

III. COLLECTION AND PREPARATION OF DATA

We conducted our experiments with two datasets. The T-HSAB **[22]** set of 6039 tweets that contains three types of tweets (normal, abusive, hateful) and a dataset we collected that contains 10000 tweets. The collected corpus is a combination of the first corpus and a set collected by ourselves. T-HSAB which is the first Tunisian public dataset aiming to be a reference dataset for the automatic detection of Tunisian toxic content online. This corpus contains 6039 tweets grouped into abusive, hateful, normal. So we just take 4898 tweets between hateful and normal. In October 2020, using the extension for Google Sheets "Twitter Archiver", we collected 5102 Arabic tweets, the dataset is about any type of hate (religious, sexual, racial, etc.). We used this collection of tweets as a training and test dataset. We only included in the search rules unbiased terms referring to a religion name or people practicing that religion, gender, race, etc. Specifically, we did not use any descriptions used to insult people of a particular race or gender or religious affiliation. For example, when collecting tweets related to gender, we used the Arabic equivalent of the keywords: Tunisian woman (المرأة التونسية , (Tunisian girl)الطفلة التونسية...). Again, when collecting race-related tweets, we used the Arabic equivalent of the keywords: black(وصيف كحلوش،), white (أبيض...).

A. *Data Annotation*

In this work, we divide speech types into two categories:

• Hateful: it is a speech addressed to individuals, groups or moral entities, with the aim of provoking and harassing them. It is everything that is violent, unethical, racist, harmful behavior.

•Normal : it is a speech that is not Hateful. For example, a publication that contains news or praise or express an opinion politely.

The labeling was done manually by ourselves, where each publication is classified as normal, or hateful. The total number for each category is 6335 for Normal and 3665 for Hateful.

The following table shows an example of each category :

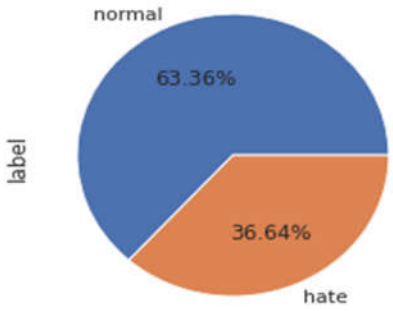

Figure 1: Distribution of data

TABLE 1: EXAMPLE OF ANNOTATION

Label	Example
Hateful	الموت يستحقون كلهم ، فاسدة حكومة إنها
Normal	الله اولاد انتم

B. *Data pre-processing*

Texts extracted from Twitter sites, are full of noise. Tweets typically contain URLs, punctuation, symbols and tags such as @, RT and <>, which we can now safely remove from the corpus. We discarded or normalized them without affecting the classification task. We performed the following preprocessing operations.:

- Normalization Alef : {} ٱ, آ, إ, أ→ا
- Normalization Alef Maqsoura : ى→ي
- Normalization of Ta Marbouta : ة→ه
- Normalization of hashtags : by removing the underscores and the # symbol.
- Removal of "stop-words": or we can call them empty words, these are words that have no real meaning and do not carry any meaning. Since they are very common and widely used, they cannot characterize, in the lexical sense, a text in relation to another. The Arabic language contains 750 stop-words.
- Removal of diacritics, html punctuation marks

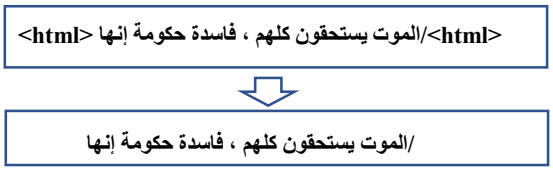

Figure 2: Remove html punctuation marks

- Removal emojis or replace them with their equivalent, non-Arabic characters and single letter words.

Figure 3: Remove emoji

- Managing elongated words: One approach to managing word lengthening is to only remove all repeated characters if the repetition was three or more with retention of a record along each tweet of the number of lengthened words it contained, which can serve as a predictive function for sentiment analysis.

IV. NEURAL NETWORK MODELS

In this section, we briefly describe the feature representation of the evaluated models and then describe the architectures of these models.

A. Representation of the characteristics

In this step, we transform the textual data into a representation that can be used for the task we want to accomplish. There are different ways to represent textual information, in our implementation we use word embeddings. Word embeddings are a dense vector representation of words. We use Bert **[23]** for word embeddings.

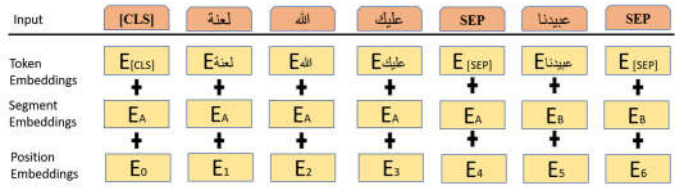

Figure 4: Incorporating words with Bert

```
array([[2.0000e+00, 7.2850e+03, 2.4766e+04, ..., 0.0000e+00, 0.0000e+00,
        0.0000e+00],
       [2.0000e+00, 1.2689e+04, 3.8990e+03, ..., 0.0000e+00, 0.0000e+00,
        0.0000e+00],
       [2.0000e+00, 3.6710e+03, 4.2450e+04, ..., 0.0000e+00, 0.0000e+00,
        0.0000e+00],
       ...,
       [2.0000e+00, 7.1070e+03, 2.0100e+02, ..., 0.0000e+00, 0.0000e+00,
        0.0000e+00],
       [2.0000e+00, 2.3330e+04, 2.0890e+03, ..., 0.0000e+00, 0.0000e+00,
        0.0000e+00],
       [2.0000e+00, 2.3836e+04, 1.2430e+03, ..., 0.0000e+00, 0.0000e+00,
        0.0000e+00]])
```

Figure 5: Overview of the result of the words embedded with Bert

BERT is the acronym for "Bidirectional encoder representations from transformers". It is a so-called pre-training language model based on neural networks. It is a bidirectional NLP model because it reads editorial content in its entirety and in both directions. BERT uses Transformer, an attention mechanism that learns the contextual relationships between words (or sub-words) in a text. In its vanilla form, Transformer comprises two distinct mechanisms - an encoder that reads the text input and a decoder that produces a prediction for the task. Since the goal of BERT is to generate a language model, only the encoder mechanism is needed.

B. Proposed model architectures

We used the AraBERT model on two classification tasks (Binary and multi-class). The AraBERT model is based on Bert.

➢ AraBERT

AraBERT **[24]** is the BERT pre-trained specifically for Arabic language with the aim of achieving the same success as BERT for English language. The performance of AraBERT is compared to Google's multilingual BERT and other state-of-the-art approaches.

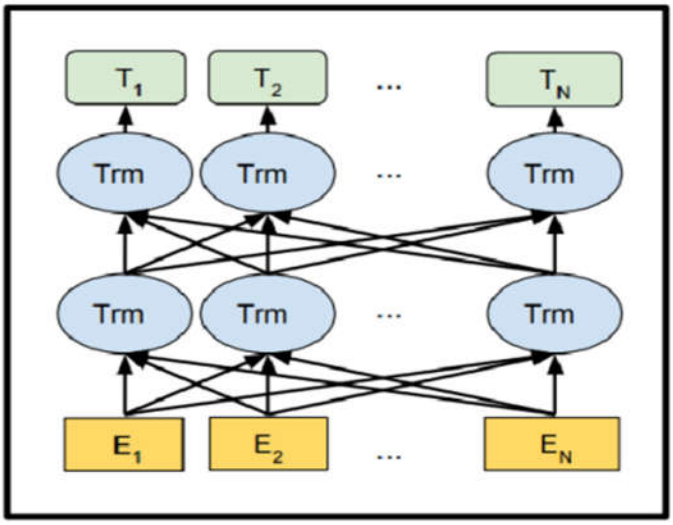

Figure 6: Architecture of the BERT model

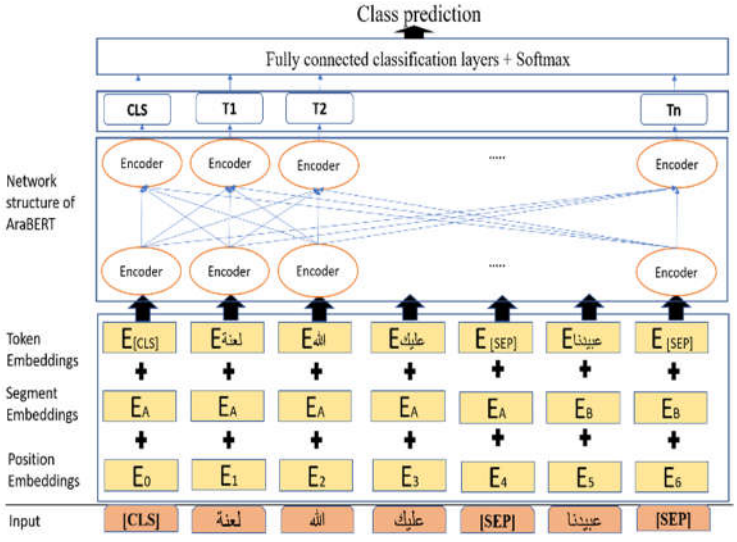

Figure 7: Architecture of AraBERT

In this experiment, we refined the cased 02 version of the pre-trained model by adding a binary classification layer to classify tweets as hate or normal. For multi-class classification with AraBERT, we added a multi-class classification layer (three classes) to classify tweets as hate, abusive or normal. To prepare our data before running it through the model, we followed the same steps as described in Section 3.2. In addition to these steps, we removed html artifacts, replaced emojis with their equivalents. We also replaced email URLs with special tokens. Finally, we applied the farasa segmentation.

V. EXPERIENCES

We conducted experiments to evaluate the proposed models for the Arabic hate speech detection tasks. We performed a binary classification and a multi-class classification task. For the binary classification, the tweets are classified into two classes: hate and normal. For the multi-class classification, tweets are classified into hate, abusive or normal. In this section, we describe the datasets used in our experiments and the experiment setup, including model implementation, hyperparameter fitting, baselines, and evaluation measures.

*A. **Dataset***

For binary classification, we used our dataset of 10,000 tweets that we created as described in Section 3. We divided the data into training data (80%) or 8000 tweets, validation data (10%) or 1000 tweets, and test data (10%) or 1000 tweets. We used the 80% for training the model, and 10% for validation to adjust the model weights. The test data (10%) was used to test the model to see its performance on never seen data. For multi-class classification, we used the T-HSAB dataset of 6039 tweets. The data was divided into training, validation and test sets. The training set contained 80% (4831 tweets) of the dataset, the validation and test set each contained 10% or 604 tweets per set.

*B. **Experimental parameters***

We implemented all neural network models using the Keras library with TensorFlow as the backend. We ran the experiments on Google Collaboratory, which provides a free Jupyter notebook environment with GPU gas pedal. The binary model architecture was experimented with binary cross-entropy as the loss function, "Adam" as the optimizer, Relu as the hidden layer activation function and the Sigmoid as the output layer activation function. The batch size is 16, we tried to change it to 32 or 64, but the Google Collaboratory environment sputtered several times. For the learning rate, we tested several values {1e-5, 2e-5,3e-5} and got the best result with 1e-5 over 14 epochs. As a metric, we used the macro averages of precision (P), recall (R), accuracy, F1 measures. Since our dataset is relatively unbalanced and given the serious consequences of not detecting hateful tweets, we presented our results using precision (P), recall (R) and F1 measures. The multi-class AraBERT model was experimented with categorical cross-entropy, used for multi-class classification, "Adam" as an optimizer, Relu as the hidden layer activation function and Softmax as the output layer activation function. We tested several values of the learning rate {1e-5, 2e-5,3e-5,4e-5} and obtained the best result with 3e-5 over 12 epochs.

VI. RESULTS OF THE EXPERIMENTS AND DISCUSSION

In this section, we discuss the results of the series of experiments.

TABLE 2: RESULTS OF THE EVALUATION OF THE TESTED MODELS

Class	Algorithms	P(%)	R(%)	F1(%)
2	AraBERT	0,9987	0,9987	0,9987
3	AraBERT	0.9955	0.9955	0.9955

We observe the results obtained by our models in terms of precision, recall, and F1 measure. These results were expected since the AraBERT model has proven to be very powerful, obtaining state-of-the-art results in many natural language processing tasks such as sentiment analysis, question answering.

As stated by Antoun **[24]**, AraBERT achieved state-of-the-art performance in sentiment analysis, named entity recognition and question answering tasks. and question answering tasks, which is confirmed by the results we obtained for sentiment analysis. The effectiveness of Arabert on the Arabic (Tunisian) dialect is due to the fact that it is pre-trained on a large corpus and to the Transformer architecture on which it is based. The attention layers of the Transformer allowed our models to efficiently select the relevant input series for the output class, and thus increase the accuracy score **[25]**.

VII. CONCLUSION

In this paper, we have addressed the problem of detecting abusive and hate speech. We developed a high-quality textual corpus of hateful and abusive speech for the Tunisian dialect and labeled it into two classes: normal and hateful. We evaluated the AraBERT model for binary and multi-class classification. The results obtained met our expectation which was to show the effectiveness of the proposed models in the task of hate speech detection. The binary model performed well with an F1 measure of 0.9987. The AraBERT model of multi-class classification also obtained a good F1 result of 0.9955.

ACKNOWLEDGMENT

The research leading to these results has received funding from the Ministry of Higher Education and Scientific Research of Tunisia under the grant agreement number 01-02 20PEJC.

RÉFÉRENCES

[1] H.Haddad, H.Mulki and A.Oueslati., "T-HSAB: A tunisian hate speech and abusive dataset," Conférence internationale sur le traitement de la langue Arabe, 2019.

[2] Z. Zhang, L. Luo, "Hate speech detection: A solved problem? The challenging case of long tail on Twitter," Semantic Web,vol.10, pp. 925-945, 2019.

[3] Z. Waseem, T. Davidson, D. Warmsley and I. Weber, "Understanding Abuse: A Typology of Abusive Language Detection Subtasks," Association for Computational Linguistics, pp. 78-84, 2017.

[4] Z. Kechaou, M. B. Ammar and A. M. Alimi, "A new linguistic approach to sentiment automatic processing," IEEE ICCI, pp. 265-272, 2010.

[5] Z. Kechaou, M. B. Ammar and A. M. Alimi, "Improving e-learning with sentiment analysis of users' opinions," In 2011 IEEE global engineering education conference (EDUCON), pp. 1032-1038, 2011.

[6] Z. Kechaou, A. Wali , M. B. Ammar, H. Karray and A. M. Alimi, "A novel system for video news' sentiment analysis," Journal of Systems and Information Technology, 2013.

[7] Z. Kechaou, M. B. Ammar and A. M. Alimi, "A multi-agent based system for sentiment analysis of user-generated content," International Journal on Artificial Intelligence Tools, 2013.

[8] Z. Kchaou, S. Kanoun, "Arabic stemming with two dictionaries," 2008 International Conference on Innovations in Information Technology, pp. 688-691, 2008.

[9] Z. Kechaou, S. Kanoun, "A new-arabic-text classification system using a Hidden Markov Model," International Journal of Knowledge-based and Intelligent Engineering Systems, pp. 201-210, 2014.

[10] M. Abbes, Z. Kechaou and A. M. Alimi, "Enhanced deep learning models for sentiment analysis in arab social media," International Conference on Neural Information Processing, pp. 667-676, 2017.

[11] I. A. Farha, W. Magdy, "An online Arabic sentiment analyser," Association for Computational Linguistics, p. 192–198, 2019.

[12] A. Abuzayed, T. Elsayed, "Quick and Simple Approach for Detecting Hate Speech in Arabic Tweets," European Language Resource Association, p. 109–114, 2020.

[13] T. Davidson, D. Warmsley, M. Macy, I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," Conf. Web Soc. Media, p. 512–515, 2017.

[14] H. Che, S. McKeever and S. J. Delany, "A Comparison of Classical Versus Deep Learning Techniques for Abusive Content Detection on Social Media Sites," Computer Science, vol 11185, p. 117–133, 2018.

[15] K. A. Kwaik, M. Saad, C. Stergios, S. Dobnik, "LSTM-CNN Deep Learning model for Sentiment Analysis of Dialectal Arabic," Communications in Computer and Information Science, p. 108–121, 2019.

[16] F. Y. Al Anezi, "Arabic Hate Speech Detection Using Deep Recurrent Neural Networks," MDPI AG, p. 6010 , 2022.

[17] S. Alsafari, S. Sadaoui and M. Mouhoub, "Hate and offensive speech detection on Arabic social media," Online Social Networks and Media, pp. 1-15, 2020.

[18] N. Albadi, M. Kurdi, S. Mishra, "Are They Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere," In the proceedings of The 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2018.

[19] H. Mohaouchane, A. Mourhir and N. S. Nikolov, "Detecting Offensive Language on Arabic Social Media Using Deep Learning," 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2019.

[20] H. Mulki, H. Haddad, C. B. Ali, H. Alshabani, "L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language," Association for Computational Linguistics, pp. 111–118, 2019.

[21] N. Albadi, M. Kurdi, S. Mishra, "Are They Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere," In the proceedings of The 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2018.

[22] H. Hatem, M. Hala and O. Asma, "T-HSAB: A Tunisian Hate Speech and Abusive Dataset," SPRINGER INTERNATIONAL PUBLISHING AG, 2019.

[23] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Association for Computational Linguistics, p. 4171–4186, 2019.

[24] W. Antoun, F. Baly and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," European Language Resource Association, p. 9–15, 2020.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez and L. Kaiser, "Attention Is All You Need," Proceedings of the 31st International Conference on Neural Information Processing Systems, p. 6000-6010, 2017.