International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys.13, No.1 (Apr-23)

http://dx.doi.org/10.12785/ijcds/130177

Deep Learning-based Analysis of Algerian Dialect Dataset Targeted Hate Speech, Offensive Language and Cyberbullying

Ahmed Cherif Mazari *[1] **and Hamza Kheddar** [2]

[1]*Mathematics and Computer Science Department, LSEA Laboratory, University of Médéa, Algeria*
[2]*Electrical Engineering Department, LSEA Laboratory, University of Médéa, Algeria*

Received 14 Jun. 2022, Revised 19 Dec. 2022, Accepted 6 Feb. 2023, Published 16 Apr. 2023

Abstract: Toxicity and hate speech on social media platforms can lead to cyber-crime, affecting social life on a personal and community level. Therefore, automatic toxicity and hateful content detection are necessary to enhance web content quality and fight against inappropriate speech spread through social media. This need is also a challenge when comments are posted and written in complex languages, such as Arabic, which is recognised for its difficulties and lack of resources. This paper introduces a new dataset for Algerian dialect toxic text detection, whereby we build an annotated multi-label dataset consisting of 14150 comments extracted from Facebook, YouTube and Twitter, and labelled as hate speech, offensive language and cyberbullying. To assess the practical utility of the created annotated dataset, several tests have been conducted using many classification models of traditional machine learning (ML), namely, Random Forest, Naïve Bayes, Linear Support Vector (SVC), Stochastic Gradient Descent (SGD) and Logistic Regression. Furthermore, several assessments have been conducted using Deep Learning (DL) models such as Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional-LSTM (Bi-LSTM) and Bidirectional-GRU (Bi-GRU). Experimental tests demonstrate the success of the Bi-GRU model, which achieved the highest results for DL classification, with 73.6% Accuracy and 75.8% F1-Score.

Keywords: Machine Learning, Deep Learning, Algerian Dialect, Cyberbullying Detection, Offensive Language Detection, Hate Speech Detection

1. INTRODUCTION

With the widespread adoption of Web2.0 in recent decades, individuals have become more engaged, whereby people around the world now have the ability to communicate and share their opinions freely and instantly by posting them on social networks. However, the Internet is not entirely safe; it can be a source of hateful and toxic content. Thus social media networks are grappling with how to ban these contents while maintaining free expression. Furthermore, hate speech can be sparked by the multiplicity of communities and individuals, their cultures, backgrounds, and beliefs. Therefore in each culture, the community has its own interpretations and acts differently depending on its culture.

Toxic language detection is a branch of study that examines human spoken and written language to determine whether it has aggressive, hateful, offensive, cyberbullying or not. Thus, numerous articles and studies have been published addressing language toxicity identification, and this study field is increasingly attracting the interest of researchers. Various automatic methods have been implemented and evaluated; among these methods are those that use traditional techniques and as well recent deep learning models. Traditional methods are mainly based on ML algorithms, such as Support Vector Machines (SVM) and Naive Bayes (NB) for classification. However, recent deep learning methods are based on vector representation and use artificial neural networks such as Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN).

The Arab world has seen a significant surge in the use of social network platforms. According to Arab social media research, in some countries, social media penetration has surpassed 90% of the population. As is well known, detecting toxicity and hate speech in Arabic is more difficult due to the language's complexity in terms of spelling, morphology and even worse for the spoken language named

*Corresponding author

E-mail: mazari.ahmedcherif@univ-medea.dz, kheddar.hamza@univ-medea.dz

by the dialect.

This paper concentrates on building the Algerian dialect dataset as well as its collection procedures. The spoken language of Algerian people, in which this imposes additional challenges such as the combination of Modern Standard Arabic (MSA), French and different terms derived from local and other languages; these latter conditions make existing state-of-the-art work inefficient to detect the toxicity and hate speech of the Algerian dialect on social media. Moreover, there is a significant deficit of resources, datasets, and suitable processing methods.

This study is devoted first, to building a new dataset targeted hate speech, offensive language and cyberbullying from various social network platforms, namely YouTube, Facebook and Twitter. Then, we apply several preprocessing tasks to tackle the related issues of dialect language. Finally, we test and investigate several machine learning techniques and deep learning models for toxicity detection by classification methods.

The rest of the paper is organised as follows: Section 2 presents related work on methods applied to toxicity and hate speech detection for the Algerian dialect. Section 3 details the contribution. Section 4 outlines the tests and findings. The conclusion is presented in section 5.

2. RELATED WORK

In the last years, using machine learning models, particularly deep learning, to Natural Language Processing (NLP) and toxicity and hate speech analysis tasks have attracted much attention from research teams. Several studies have looked into this phenomenon of toxicity detection in English text (e.g. cyberbullying, hate speech, abusive language, radicalisation detection, offensive language...). The work in [1] tackled the hate speech detection and the problem of offensive language using traditional ML algorithms, namely naïve Bayes, logistic regression, decision tree, random forest and linear SVMs, by which linear SVMs and logistic regression achieved the best in classifying offensive language, hate speech and clean vocabulary. The work of [2] presented the detection of hateful content in Tweets based on DL models, in which the authors conducted experiments with architectures of CNN and Long Short-Term Memory (LSTM) to learn semantic word embedding. The experimental results concluded that these architectures outperform existing methods. As well as the research of [3] that addressed the problem of discerning hateful content in social networks, studying the detection of aggressive language in tweeter content using DL models. The researchers trained and tested the LSTM model of RNN on a 16k tweets dataset; the obtained results achieved 93.2% for the F1-score performance metric. At COLING-2018, with the "Workshop on Trolling, Aggression and Cyberbullying" (TRAC-2018), the team [4] proposed the aggression identification study using DL and data augmentation technique, whereby they combined neural network algorithm with three logistic regression classifiers trained on n-grams (character and word). The team's work achieved around 60% F1-score on the English dataset, 38% on the Hindi Twitter dataset and 63% on the Hindi Facebook dataset. [5] presented a study for detecting, in various forms, violence and toxicity in comments, the researchers proposed a single model capsule for data augmentation to deal with implicitly, the results achieved 98.46% of ROC-AUC on the Kaggle dataset of toxic comments. Three architectures are implemented for detecting cyberbullying in social media in [6], incited by the reported success of neural network models, namely a CNN, a mixed LSTM-CNN and a hybrid LSTM-CNN-DNN (Deep Neural Network). Furthermore, the same work using DL is tested on the toxic comment Kaggle dataset by [7] to classify offensive comments on social networks. The approach proposed in [8] tackled cyberbullying identification on social networks using CNN founded on character level and shortcuts. The authors provided a new cyberbullying Chinese comments dataset, and the experiments were accomplished on both the Chinese and the English tweet datasets. The experimental results showed that this approach is suitable for cyberbullying detection tasks with a 74.3% F1-score.

Although recent advances have been made in the field of Arabic toxicity and hate speech, most resources in this area of study are either limited, particular, or not publicly available. [9] Analysed the automatic identification of violent language on Arabic Twitter social networks, using unigram and bigram counts. [10] addressed a new challenge in Arabic social media and introduced an unsupervised framework for detecting Arabic violence on Twitter. In the study of [11], the researchers created a new training dataset from Twitter consisting of 1690 posts by combining automatic and manual annotation of Arabic text, and then they tested machine learning models based on sentiment analysis to examine "violence and discrimination against women" content. [12] presented their system for identifying the offensive and hateful language in Arabic. This is performed by CNN, LSTM, Transfer Learning (TL) and Multi-Task Learning (MTL), which achieved a 73% F1-score in the hate speech detection task. Due to a lack of resources, [13] created a new open Arabic dataset annotated for irony identification referring to international affairs, football and social issues. It consists of 5,358 tweets written in MSA, dialectal Arabic and a mix of both. [14] suggested a work for Arabic hate speech detection against women, whereby the authors created a multilingual corpus from the YouTube platform and tested several ML and DL models, in which CNN outperformed other models. Furthermore, [15] demonstrated the effectiveness of RNN and CNN for the classification of Hirak-2019 (popular protest in Algeria during 2019) comments expressed in Algerian dialect and retrieved from social networks, the experiment was conducted on a corpus of 7800 comments. Even newer, [16], [17], addressed, in greater detail, offensive and hateful language detection and classification in Arabic Social Media. Moreover, [18] presented the impact of preprocessing tasks on detecting Arabic hateful and offensive language.

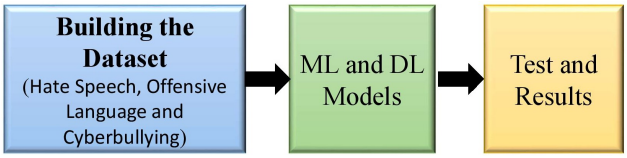

Figure 1. General diagram of the proposed approach

3. CONTRIBUTION

We present the broad diagram followed to design our proposed approach in Figure 1 below, as well as the details of each phase are shown in the following subsections. The work starts by collecting data from diverse social network platforms in order to build the dataset of Algerian dialect labelled in hate speech, offensive language and cyberbullying. Afterwards, we investigate and analyse this dataset by several algorithms based on ML and DL models.

A. *Building the dataset*

The procedures followed for the creation of the dataset are recommended by the literature review ([19], [20], [21], [22]). The protocol consists of the following steps:

1) *Data collecting*

In this step, we collect the textual data related to the Algerian dialect, which is written in Modern Arabic Standard (MSA), French, Dialect or/and in Arabizi (Algerian dialect transcripted in Latin script and Arabic numbers). Therefore, various popular Algeria media pages and posts such as Echourouk news TV, El Bilad TV and political pages hosted by Facebook, YouTube and Twitter platforms are used to retrieve textual comments and posts on twelve (12) different topics such as Politics, Religion, Price hikes, Legislative elections, Racism, Illegal immigration (Harraga), Prime Minister Jarrad, Drugs, Parliamentary politics, Moroccan politics, President's speech and misogyny. Whereby the collected data is related to hate speech, cyberbullying, and offensive and abusive language. The retrieving and collecting process of data is automatically achieved from the platforms via their packages and APIs as follows Facebook [1](graphAPI), [2]YouTube (Google API YouTube) and Twitter [3](Twitter API).

2) *Data filtering*

The retrieved textual data are filtered as follows:

- We manually delete all comments written in different languages other than French, Arabic and Algerian dialect.
- Retweets and repeated comments are also removed.
- Long comments that contain more than 128 words and short comments that contain under 10 words are also filtered.

[1]https://developers.facebook.com/docs/graph-api/
[2]https://developers.google.com/youtube/v3/
[3]https://developer.twitter.com/en/docs/twitter-api

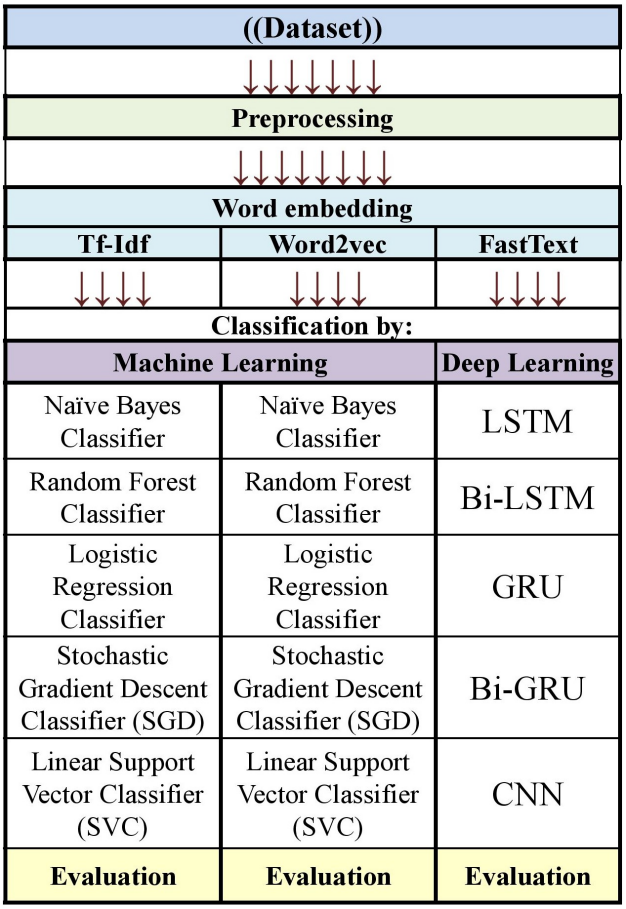

Figure 2. Steps and techniques of the proposed approach

Therefore, the obtained dataset consists of 14150 comments, in which 5579 comments from Facebook, 216 comments from Twitter and 8355 comments from YouTube. Table I below shows samples of collected comments labelled as hate speech, offensive language and/or cyberbullying.

3) *Data annotation*

The data annotation is manually performed by three annotators, Algerian dialect native students, targeted in three labels: Hate Speech, Cyberbullying and Offensive Language, whereby each label is classified in 'Yes' or 'No'. The following Table II presents the number of comments for each label and each class.

B. *Dataset analysis by ML and DL models*

To prepare the built dataset for experimentation, first, the textual data must be prepared through the preprocessing NLP tasks. Then, these textual data are converted to input data by numerical vectors created by word embedding methods. Finally, we perform several tests on the dataset using various ML and DL algorithms and models. These different steps and techniques are illustrated in Figure 2.

TABLE I. Sample of dataset comments. HS, CB, and OL stand for hate speech, cyberbullying, and offensive language, respectively

Comment	Topic	HS	CB	OL	Source
امشي هزة تهزك تقول براح ازكيكك روح تملح على روحك	Price hikes	Yes	Yes	No	Facebook
لوكان يروحو قاع غاية روحو يا وجوه الذل والميزيرية	Illegal immigration (harraga)	Yes	Yes	Yes	YouTube
شكرا لك سيدي واستاذي نثق بكلامك لانك اهل اختصاص	Politics	No	No	No	YouTube
Oh my God !!!! . Une femme est une femme imbécile ☹ ☹ ☹ ce n'est pas une fraise. Et l'Afrique du Sud est développée, elle n'a pas besoin de tes fraises.	Misogyny	No	Yes	Yes	Facebook

TABLE II. Dataset composition. HS, CB, and OL stand for hate speech, cyberbullying, and offensive language, respectively

Label	HS		OL		CB	
Class	Yes	No	Yes	No	Yes	No
Number	7573	6577	5766	8384	4524	9626
Total	14150		14150		14150	

1) Preprocessing

For the preprocessing step of textual data used in the proposed approach, we follow [23], [24] strategies which include:

- Text segmentation into words.
- Keeping only the Arabic and French characters by using the [4]AlphabetDetector Api, deleting numbers, and removing unknown uni-codes characters and extra spaces.
- Substitution of URLs (≪ http ≫ or ≪ https ≫) by the < url > tag.
- Substitution of User mentions and email address by < user > tag.
- Substitution of emoticons by < emoticon > tag.
- Standardizing text words by removing repeated letters more than twice (e.g. Hellllllo to Hello, سلاااااام).
- Hashtags ≪ # ≫ composed of concatenated words are substituted by their separated word version.
- Lowercasing text.
- Keeping stopwords and punctuation marks to avoid destructing possible obfuscated words. Unlike [23], whereby [25] reported that punctuation and stopwords contribute to the text's meaning for the toxicity and hate speech detection tasks.

In addition, we carry out specific Arabic tasks such as letter normalisation:

- Unifying the letters that are written differently. Example { أ إ آ } are substituted by {ا}.
- Deleting all Arabic diacritics 'Tashkeel' like (fatha, damma, kasra, tashdid ... etc.).
- Removing the elongation as (اعـــــــــــــلام becomes اعلام) (in English media).

2) Word embedding

Machine learning models require numerical representations of input data for the training and testing phases. Among these representations are word embedding methods that are used to convert textual data, such as words, sentences, paragraphs, and documents, into digital vectors representing the implicit semantic relationships linking these textual elements. Moreover, one of the most effective unsupervised learning applications is embedding models, which have been extensively used in DL-based NLP algorithms.

In this work, beside using the Tf-Idf for traditional machine learning techniques, two different algorithms derived from the Word2vec model [26] are employed for word embedding, the Continuous Bag-Of-Words (CBOW) and the Skip-Gram (SG) for the ML algorithms. The second model is the FastText for the DL algorithms, developed by Facebook AI Research Laboratory [27]. We opted for FastText because this word embedding model is trained on textual data collected from the Facebook platform, which includes the Algerian dialect.

3) Classification

For the classification phase, the dataset is divided into two parts; training and testing, representing 70% and 30% of the dataset size, respectively. Then, different ML and DL algorithms and methods are experimented with using the Accuracy and F1-score metrics to evaluate the performances.

- **Machine Learning:** For traditional machine learning methods, we use both the TF-IDF technique and Word2Vec embedding for the five classification algorithms such as Random Forest (RF) classifier, Multinomial Naïve Bayes classifier, LinearSVC Support Vector classifier (kernel = 'rbf'

[4]https://pypi.org/project/alphabet-detector/

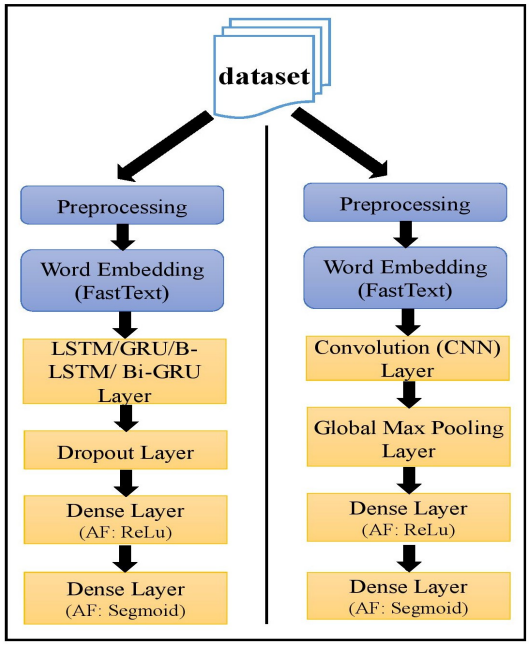

Figure 3. LSTM/GRU/B-LSTM/Bi-GRU and CNN Architectures

and C='1.0'), Stochastic Gradient Descent (SGD) classifier (loss='log' and penalty='l2') and Logistic Regression (LR) classifier (c=0.5).

- **Deep Learning:** For deep learning methods, we use the following architecture for each algorithm, LSTM, GRU, Bi-LSTM or Bi-GRU. The embedding matrix weights are computed using the FastText models, then the (LSTM/GRU/Bi-LSTM/Bi-GRU) layer that scans the feature map. To reduce the over-fitting of training, we add the dropout layer (with a probability = 0.5). The results are then sent into a single feed-forward layer (fully-connected) using the activation function ReLu, and the layer output is fed into a Sigmoid layer to find the output classes. We employ Adam optimiser with ten epochs, as shown the Figure 3.
  For the CNN model, we use the convolution (CNN) layer to scan the feature map returned by word embedding, then the global max pooling layer is used for the output before the first dense layer with the activation function ReLu and the first dense layer with the activation function Sigmoid as presented in Figure 3.

4. TEST AND RESULTS

For the experimental part, the packages of Scikit-learn for ML and Keras (with TensorFlow) for DL on the Google Colab Python environment are used. Thus, the test results by traditional ML classifiers are presented in Table III and Figure 4.

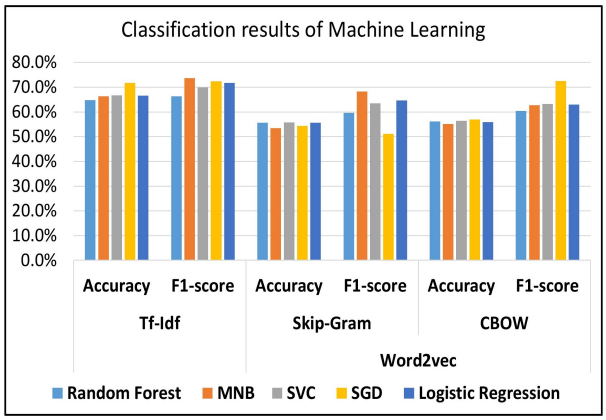

Figure 4. Graphical results of machine learning classification

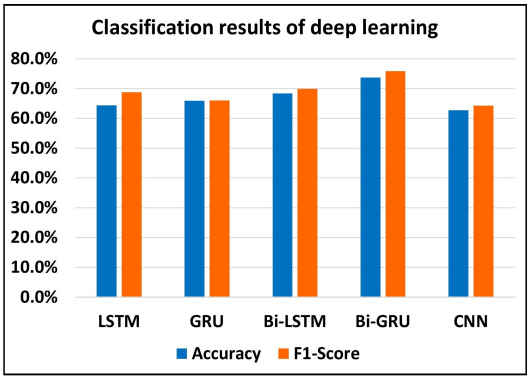

Figure 5. Graphical results of deep learning classification

Discussion 1

- We notice that the Tf-Idf is the most adapted technique compared to the word2vec embedding for traditional machine learning classifier, whereby the best results of F1-score 73.6% is achieved by the MNB algorithm, and Accuracy of 71.6% is achieved by the SGD algorithm.
- When using word2vec embedding, we notice that the Accuracy values are less than the values of F1-Score, and this is because the dataset is a multi-label. Moreover, the best result is achieved in CBOW embedding, obtained by the SVC 72.4% of F1-score and 56.9% of Accuracy.

The test results by traditional ML classifiers are depicted in Table IV and Figure 5.

Discussion 2

- We also notice that the F1-Score rate is higher than the Accuracy values, which is also due to the multi-label kind of the dataset.
- The best result is obtained by Bi-GRU, which

TABLE III. Results of ML classifiers

ML Classifier	Word embedding Tf-Idf		Word embedding (Word2vec)			
			Skip-Gram		CBOW	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Random forest	64.7%	66.2%	55.5%	59.5%	56.1%	60.3%
Multinomial NB	66.3%	**73.6%**	53.4%	**68.2%**	55.1%	62.7%
SVC	66.6%	69.9%	**55.7%**	63.4%	56.3%	63.1%
SGD	**71.6%**	72.3%	54.3%	51.0%	**56.9%**	**72.4%**
Logistic Regression	66.5%	71.7%	55.6%	64.6%	55.8%	62.9%

TABLE IV. Results of DL classifiers

DL Classifier	Word embedding (FastText)	
	Accuracy	F1-score
LSTM	64.3%	68.7%
GRU	65.8%	65.9%
Bi-LSTM	68.3%	69.8%
Bi-GRU	**73.6%**	**75.8%**
CNN	62.7%	64.2%

achieved 75.8% of F1-Score and 73.6% of Accuracy. The GRU model is a variant of RNN, and the Bidirectional GRU model is a Bidirectional version of RNN. It is based on the concept of transmitting the input sequence in two directions: as-is and backward to two recurrent layers that simultaneously conduct inverse training. By gathering information from previous (i.e., forward pass) and future (i.e., backward pass) contexts, a more contextualised representation of the input is produced. Therefore, recurrent algorithms like the bidirectional GRU are better suited for processing sequential data such as texts and dialects.

- Algorithms derived from RNN such as LSTM, B-LSTM, GRU and Bi-GRU achieve more performance than the CNN algorithm. This proves that RNN models are better suited for text analysing.

- All classification results from DL models exceeded results from traditional ML algorithms. The performances of DL models are better than ML models since they have the advantage of generating their own features and being independent of feature extraction techniques. More precisely, DL models- are better than ML models-based text classification algorithms in efficiency and accuracy ([28], [29], [30]). In addition, the FastText model of word embedding used as input to our models is more suitable for DL since it was trained by Facebook on dialect texts.

In summary, we notice that using Bi-GRU applied on the built dataset achieved the best performance compared to other state-of-the art research using their models and algorithms but experienced own their datasets, as shown in Table V.

5. CONCLUSION AND FUTURE WORK

In this article, we presented a completed and validated study to deal with Algerian dialect toxicity speech analysis using machine learning and deep learning models. First, we built a new [5]Algerian dialect dataset that targeted hate speech, offensive language and cyberbullying consisting of 14150 comments collected from Facebook, YouTube and Twitter. Afterwards, adapted preprocessing steps were implemented before the filtering and annotation stage. Then different ML classifiers (Multinomial Naïve Bayes, Random Forest, Logistic Regression, Stochastic Gradient Descent (SGD) and Linear Support Vector (SVC)) furthermore five DL models (LSTM, GRU, Bi-LSTM, Bi-GRU and CNN) were implemented and tested.

As an ending, the best results are reached for ML classification using Tf-Idf by the SGD classifier: 71.6% of Accuracy, and the MNB classifier: 73.6% of F1-score. Regarding the word2vec embedding, the SGD algorithm outperforms others by 56.9% of Accuracy and 72.4% of F1-Score.

Concerning DL classification using FastText embedding, the Bi-GRU algorithm produces the highest results with 73.6% Accuracy and 75.8% F1-Score. Finally, the experimentation revealed that the classification results obtained using DL models beat those obtained using traditional ML algorithms, most likely because to the fact that textual data is better analysed by RNN algorithms such as LSTM, GRU, Bi-LSTM and Bi-GRU.

In future work, we intend to create a large dialectal dataset about the spoken language of Arab countries. Moreover, we plan to study the field by recent Transfer Learning NLP techniques using BERT, GPT-2 or GPT-3 models and perform more evaluations to determine which embeddings and algorithms are best suited for dialect toxicity detection tasks.

ACKNOWLEDGEMENT

Special thanks to annotators: students at the MI Dpt / University of Médéa, Algeria (Ms C.N., B.K.S, B.N).

[5]https://sourceforge.net/projects/alg-dialect-toxicity-speech/

TABLE V. Comparison with existing work

Work	Dataset	Model	Accuracy	F1-score
Our work	AlgD 14150 Comments	**FastText +Bi-GRU**	**73.6%**	**75.8%**
		Tf-IDF+NB	66.3%	73.6%
		Tf-IDF+SGD	71.6%	72.3%
		W2V(CBOW)+SGD	56.9%	72.4%
HS detection on Twitter using TL [19]	Urdu Hate speech 10526 Tweets	FastText + BiGRU	72%	67%
Detection of HS in Arabic tweets using DL [20]	11 000 Arabic Tweets	SVM	-	65%
		GRU	-	70%
		CNN+LSTM	-	73%
		LSTM/ CNN+GRU	-	72%
A DL Framework for Automatic Detection of HS Embedded in Arabic Tweets [21]	9833 Arabic Tweets For multi-class classification	w2v(SG)+CNN	73%	-
		w2v(SG)+CNN-LSTM	70%	-
		w2v(SG)+BiLSTM-CNN	73%	-

REFERENCES

[1] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, 2017, pp. 512–515.

[2] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th international conference on World Wide Web companion*, 2017, pp. 759–760.

[3] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Effective hate-speech detection in twitter data using recurrent neural networks," *Applied Intelligence*, vol. 48, no. 12, pp. 4730–4742, 2018.

[4] J. Risch and R. Krestel, "Aggression identification using deep learning and data augmentation," in *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, 2018, pp. 150–158.

[5] S. Srivastava, P. Khurana, and V. Tewari, "Identifying aggression and toxicity in comments using capsule network," in *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, 2018, pp. 98–105.

[6] H. Rosa, D. Matos, R. Ribeiro, L. Coheur, and J. P. Carvalho, "A "deeper" look at detecting cyberbullying in social networks," in *2018 international joint conference on neural networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[7] M. Anand and R. Eswari, "Classification of abusive comments in social media using deep learning," in *2019 3rd international conference on computing methodologies and communication (ICCMC)*. IEEE, 2019, pp. 974–977.

[8] N. Lu, G. Wu, Z. Zhang, Y. Zheng, Y. Ren, and K.-K. R. Choo, "Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts," *Concurrency and Computation: Practice and Experience*, vol. 32, no. 23, p. e5627, 2020.

[9] H. Mubarak, K. Darwish, and W. Magdy, "Abusive language detection on arabic social media," in *Proceedings of the first workshop on abusive language online*, 2017, pp. 52–56.

[10] K. E. Abdelfatah, G. Terejanu, A. A. Alhelbawy *et al.*, "Unsupervised detection of violent content in arabic social media," *Computer Science & Information Technology (CS & IT)*, vol. 7, 2017.

[11] O. El Ansari, Z. Jihad, and M. Hajar, "A dataset to support sexist content detection in arabic text," in *International Conference on Image and Signal Processing*. Springer, 2020, pp. 130–137.

[12] I. A. Farha and W. Magdy, "Multitask learning for arabic offensive language and hate-speech 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. 86–90.

[13] I. Abbes, W. Zaghouani, O. El-Hardlo, and F. Ashour, "Daict: A dialectal arabic irony corpus extracted from twitter," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 6265–6271.

[14] I. Guellil, A. Adeel, F. Azouaou, M. Boubred, Y. Houichi, and A. A. Moumna, "Sexism detection: The first corpus in algerian dialect with a code-switching in arabic/french and english," *arXiv preprint arXiv:2104.01443*, 2021.

[15] A. C. Mazari and A. Djeffal, "Deep learning-based sentiment analysis of algerian dialect during hirak 2019," in *2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*. IEEE, 2021, pp. 233–236.

[16] B. AlKhamissi and M. Diab, "Meta ai at arabic hate speech 2022: Multitask learning with self-correction for hate speech classification," *arXiv preprint arXiv:2205.07960*, 2022.

[17] Z. Boulouard, M. Ouaissa, and M. Ouaissa, "Machine learning for hate speech detection in arabic social media," in *Computational Intelligence in Recent Communication Networks*. Springer, 2022, pp. 147–162.

[18] F. Husain and O. Uzuner, "Investigating the effect of preprocessing arabic text on offensive language and hate speech detection," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 4, pp. 1–20, 2022.

[19] R. Ali, U. Farooq, U. Arshad, W. Shahzad, and M. O. Beg, "Hate speech detection on twitter using transfer learning," *Computer Speech & Language*, vol. 74, p. 101365, 2022.

[20] A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in arabic tweets using deep learning," *Multimedia Systems*, pp. 1–12, 2021.

[21] R. Duwairi, A. Hayajneh, and M. Quwaider, "A deep learning framework for automatic detection of hate speech embedded in arabic tweets," *Arabian Journal for Science and Engineering*, vol. 46, no. 4, pp. 4001–4014, 2021.

[22] H. Mubarak, A. Rashed, K. Darwish, Y. Samih, and A. Abdelali, "Arabic offensive language on twitter: Analysis and experiments," *arXiv preprint arXiv:2004.02192*, 2020.

[23] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A bert-based transfer learning approach for hate speech detection in online social media," in *International Conference on Complex Networks and Their Applications*. Springer, 2019, pp. 928–940.

[24] H. Sohn and H. Lee, "Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations," in *2019 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2019, pp. 551–559.

[25] H. H. Saeed, K. Shahzad, and F. Kamiran, "Overlapping toxic sentiment classification using deep neural architectures," in *2018 IEEE international conference on data mining workshops (ICDMW)*. IEEE, 2018, pp. 1361–1366.

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[27] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[28] B. Behera, G. Kumaravelan, and P. Kumar, "Performance evaluation of deep learning algorithms in biomedical document classification," in *2019 11th international conference on advanced computing (ICoAC)*. IEEE, 2019, pp. 220–224.

[29] M. JayaSree and L. K. Rao, "A deep insight into deep learning architectures, algorithms and applications," in *2022 International Conference on Electronics and Renewable Systems (ICEARS)*. IEEE, 2022, pp. 1134–1142.

[30] W. Wei, L. Xiaolin, and L. Yao, "News text classification based on attention mechanism," in *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*. IEEE, 2020, pp. 462–466.

Ahmed Cherif Mazari Associate Professor at the university of Médéa and Researcher at the Laboratory of Advanced Electronic Systems (LSEA), Médéa (Algeria). Doctorate degree in Computer science from Biskra University (Algeria), Magister degree in NLP from Algiers University and Engineer degree in Computer science from ESI (Algiers-Algeria). His research interests are in Arabic Natural Language Processing (ANLP), Hate Speech detection, Sentiment Analysis, Information Retrieval and Deep Learning.

Hamza Kheddar Associate Professor at University of Medea and Researcher at the Laboratory of Advanced Electronic Systems (LSEA), Medea (Algeria). He obtained a Ph.D degree in Telecommunication from USTHB University (Algeria), a Magister degree in electronics (Spoken communication) from USTHB University, and an Engineer degree in telecommunication from ENST-TIC (Oran-Algeria). His research interests include, but are not limited to: speech steganography, digital watermarking, data hiding, speech processing, intrusion detection, biometrics, deep learning, 5G coding, and information security.