Classifying toxicity in the Arabic Moroccan dialect on Instagram: a machine and deep learning approach

Rabia Rachidi[1], Mohamed Amine Ouassil[2], Mouaad Errami[2],
Bouchaib Cherradi[1,2,3], Soufiane Hamida[2,4], Hassan Silkan[1]
[1]Department of Computer Science, Faculty of Science, Chouaib Doukkali University, El Jadida, Morocco
[2]EEIS Laboratory, ENSET of Mohammedia, Hassan II University of Casablanca, Mohammedia, Morocco
[3]Department of Computer Science, Provincial Section of El Jadida, CRMEF Casablanca-Settat, El Jadida, Morocco
[4]Department of Electrical Engineering and Intelligent Systems, University Mohamed V, SupMTI of Rabat, Rabat, Morocco

Article Info

Article history:

Received Feb 12, 2023
Revised Mar 16, 2023
Accepted Mar 23, 2023

Keywords:

Cyberbullying
Deep learning
Machine learning
Moroccan dialect
Social media
Toxicity

ABSTRACT

People crave interaction and connection with other people. Therefore, social media became the center of society’s life. Among the brightest social media platforms nowadays with a massive number of daily users there is Instagram, which is due to its distinctive features. The excessive revealing of personal life has put users in the spots of getting bullied and harassed and getting toxic revues from other users. Numerous studies have targeted social media to fight its harmful side effects. Nevertheless, most of the datasets that were already available were in English, the Arabic Moroccan dialect ones were not. In this work, the Arabic Moroccan dialect dataset has been extracted from the Instagram platform. Furthermore, feature extraction techniques have been applied to the collected dataset to increase classification accuracy. Afterward, we developed models using machine learning and deep learning algorithms to detect and classify toxicity. For the models’ evaluation, we have used the most used metrics: accuracy, precision, F1-score, and recall. The experimental results gave modest scores of around 70% to 83%. These results imply that the models need improvement due to the lack of available datasets and the preprocessing libraries to handle the Moroccan dialect of Arabic.

Corresponding Author:

Bouchaib Cherradi
Department of Computer Science, Faculty of Science, Chouaib Doukkali University
El Jadida, Morocco
Email: bouchaib.cherradi@gmail.com

1. INTRODUCTION

Social media sites are quite popular and used by millions of users daily. It is considered an interactive technology that facilitates the creation and sharing of information, ideas, interests, and much more [1]. Social media as a marketing strategy has helped many businesses and brands to grow and promote themselves [2]. Additionally, it has assisted people in finding more effective ways to connect and communicate with one another. Social media gained a lot of popularity these recent years due to its user-friendly features [3]. Social media platforms such as Twitter, Facebook, Instagram, and others have given people the opportunity to connect across the world, which make the world at our fingertips whenever we want to. The most targeted age group to these platforms is youth due to lots of reasons, and the most common one is maintaining friendships. Social media tools offer social interaction and easy creation of content by users which makes it currently an integral part of many young people’s lives [3]. Unfortunately, many users got addicted to social media platforms where they binge all day scrolling down their phones, only to find out they’ve lost track of time and end up procrastinating and not getting their important stuff or job done. Overuse and exhaustion of social media are widespread phenomena that have been harmful to people's health and productivity [4]. Social media platforms by being so powerful and with such a massive reach cannot be perfect and all good, its privileges could not be denied but it has severe side effects that can't be neglected [5], [6]. Thousands of individuals had experienced major issues with their mental health, emotional instabilities, and time wastage as a result of social networking [7]. Although numerous studies have targeted social media to take down its damaging side effects, a greater number of the researchers used English datasets, and only a few used Arabic datasets [8]. Thus, most of these researchers use only the standard Arabic in formal speech like newspapers because Arabic dialects can be challenging and quite tricky. Arabic is an ancient Semitic language used as an official language in 22 Arab countries, it is spoken by many Muslims around the world [9]. Arabic is ranked fourth among Internet languages after English, Chinese, and Spanish, according to Internet World Stats [10]. Arabic speakers make up over 185 million online users or 4.8% of all Internet users. Nearly 300 million people are native Arabic speakers [11]. There seem to be three basic types of Arabic: Arabic dialect (AD), modern standard Arabic (MSA), and classical Arabic (CA).

Classical Arabic and modern standard Arabic forms are used in all arab countries, while the Arabic dialect refers to dialectal varieties used locally in each country, in daily conversations. Classical Arabic is the oldest form of Arabic that exists in the coran, ancient works, and religious texts. However, Standard Arabic is the simplified and standardized form of classical Arabic with some grammatical changes. It is used in official oral or written communications, the administration, and the educational world [12]. MSA and AD can be found written in two ways: in Arabic characters or Latin letters and numerals [13].

In 2021, the researchers Lepe-Faúndez *et al.* [8] did a study about detecting cyberbullying in the Spanish language in tweets using hybrid model for detecting. The evaluation uses three different Spanish language corpora: the chilean, Mexican, and Chilean-Mexican corpora. The approach used machine learning (ML) algorithms including SVM, NB, and RF. Their work got good results between 70% and 89%. Wu *et al.* [9], the scientists did a study based on a hierarchical squashing-attention network to classify the degree and seriousness of cyberbullying incidents. Their research sought to create a dataset of Chinese-language cyberbullying incidents classified as mild, medium, or serious as well as create a new squashing-attention technique. They achieved mediocre accuracy, ranging from 40% to 60%. The field of text analysis has been the focus of eminent researchers. Most of them have used English datasets due to their availability everywhere on websites. Numerous researchers targeted English datasets using different approaches. Another study has been conducted by Abbasi *et al.* [14] about deep learning (DL) for religious, race, and ethnicity toxic comments detection and classification and got great results that reached 90% using RNN, LSTM, and BiLSTM. Toxicity is a common problem that's why researchers from diverse backgrounds have worked on this issue. The researchers Nguyen *et al.* [5] did toxic speech detection for social media in the Vietnamese language using PhoBERT, they got medium results up to 70%. Another study has been applied to the georgian language by the scientist Lin and Ali [6] using NB, SVM, CNN, and RNN, they have achieved high results of up to 80%.

However, only a few researchers counted by fingers showed interest in Arabic dataset. One of the top six most spoken languages in the world. There are around 300 million persons who are native Arabic speakers [10]. Despite the huge growth of the available online Arabic researches, there is a lack of work in this area. Some research returned that to the difficulty of the process of classifying Arabic text than classifying other languages like English and other European languages [11]. However, due to some structural problems with the language itself as well as a lack of tools available to aid the researchers, Arabic language classification was unable to keep up with the pace [12]. The most recent Arabic studies on toxicity have been done by the scientist Al-Bayari and Abdallah [15] in 2022 using the most popular social media platform: Instagram; where he got medium results up to 69%. In 2022 as well, another study has been made by Shannag *et al.* [16] about the construction and evaluation of multi-dialects of Arabic cyberbullying corpus using ML algorithms such as SVM and RF that achieved a high accuracy from 72.4% to 82.7%.

In this study, we aim to address the issue of toxic comments on social media, specifically on the popular platform Instagram. This is a pressing concern as toxic comments can harm individuals and harm the online community as a whole [14]. Our approach to addressing this issue is by developing methods to detect toxic comments in the content posted on Instagram. The main contribution of this research is the creation of a new dataset from scratch specifically for the purpose of toxic comment detection on Instagram. This is different from previous studies that have relied on datasets collected from Twitter, as the characteristics and patterns of toxic comments may differ between platforms. By collecting and organizing our own dataset from Instagram, we can ensure that it is well-suited for our research and the task of detecting toxic comments on this platform. This research is significant because it provides new insights into the issue of toxic comments on social media and proposes novel methods to address it. By using Instagram as our source of data, we can shed light on a commonly overlooked aspect of this problem and contribute to the development of more effective approaches to combating toxic comments online. The remainder of this paper is organized as follows: section 2 focused on the architecture of the proposed system. The materials and method are described in section 3. section 4 provides the results and discussion. Eventually, section 5 sums up this paper.

2. ARCHITECTURE OF THE PROPOSED DETECTION SYSTEM

The methodology of this experimentation involves a five-step process for analyzing and understanding data collected from Instagram using Selenium. Each step is carefully planned and executed to achieve the ultimate goal of providing insights and predictions based on the data. Figure 1 summarizes the steps involved in the methodology and provides a visual representation of the process. The goal of this experimentation is to provide insights and predictions based on the data collected from Instagram using Selenium, and to develop models that can be used for future analysis and predictions.

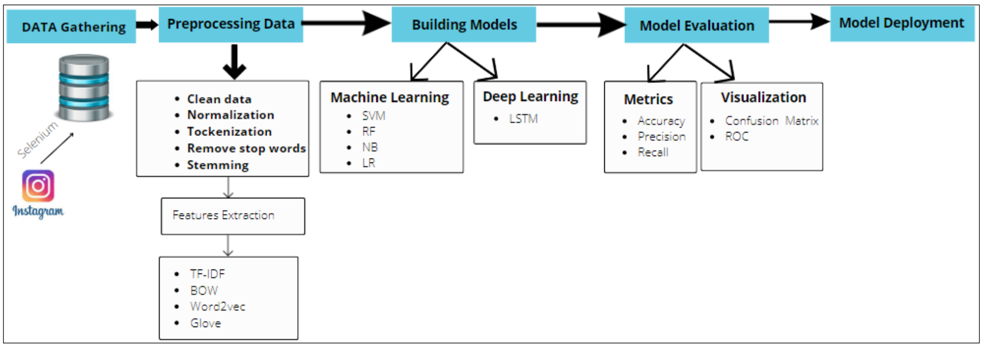


Figure 1. The method steps

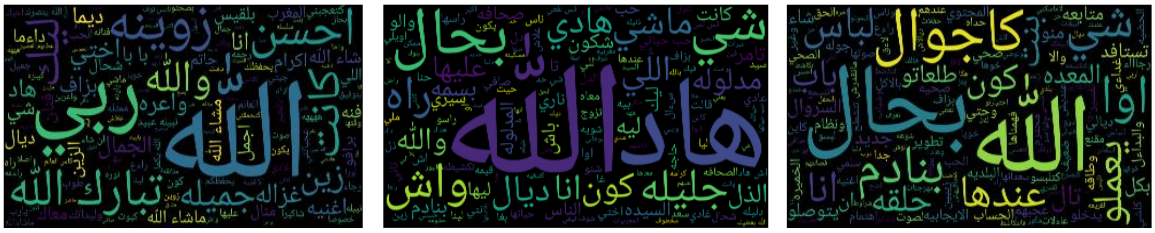
The first step of the method is data collection. This step involves using Selenium, an automation tool for web browsing, to access the Instagram website and extract the desired data. Selenium's functions and libraries are utilized to automate the process of collecting data, ensuring that the data is collected in a consistent and reliable manner. The second step is data cleaning and preprocessing. This step is crucial as it ensures that the data collected is relevant, accurate, and of a high quality. The data is checked for missing or incorrect information and any errors are corrected. The data is also transformed into a format that is suitable for analysis. This step helps to improve the accuracy and reliability of the results obtained in later steps. The third step is model building. This step involves selecting the appropriate algorithms and training the models. The models are fine-tuned to optimize their performance, which is essential for accurate predictions. The models are designed to provide insights and make predictions based on the data collected from Instagram. The fourth step is model evaluation. The performance of the models is evaluated using metrics such as accuracy, precision, recall, and F1-score. This step helps to determine the effectiveness of the models in accurately predicting the desired outcomes. The results of the evaluation are used to further refine and improve the models. The final step of the method is model deployment. The models are integrated into existing systems and made accessible to users through applications such as mobile apps or websites. This step is important as it allows the models to be put into practical use and provides valuable insights and predictions to users.

3. MATERIALS AND METHOD

3.1. Dataset collection and preprocessing

Data collection is the process of gathering information about a certain topic. Social media platforms, such as Instagram, provide vast amounts of data for researchers to explore due to the large number of active users. Instagram has experienced rapid growth and has the most monthly active users compared to other platforms such as Facebook and Twitter. Instagram scraping involves automatically collecting publicly available data from Instagram users, including both images and text. This makes Instagram a valuable source of data for researchers. Web scraping is the process of gathering data from web pages using various tools and technologies. Selenium is a popular open-source web-based automation tool that is well-suited for web scraping [17]. It offers a variety of functions to navigate through web pages and fetch the required content, making it a powerful tool for data extraction. Additionally, Selenium has a user-friendly interface, making it easy to develop and run tests efficiently [18].

In the process of gathering the dataset for this project, we used Selenium. After the long process of collecting the dataset, we gathered it in an excel file under 3 columns: user name, comment, and time/date. Then added the column classification that contains the comments' categories: positive, toxic, or neutral. The first step in cleaning a dataset involves three key actions. These include removing non-Arabic text, deleting emojis, and removing punctuation and special characters. The removal of non-Arabic text ensures that only relevant data is included in the analysis. Emojis are removed as they are not recognizable by natural language processing (NLP) in the text preprocessing stage. Removing punctuation and special characters helps to standardize the data and treat each text equally, making it easier to analyze. After this phase, we went to preprocessing the dataset by deleting stop words normalization, and stemming. After classifying the dataset, we did the words' cloud for each category, the Figure 2 shows these words' cloud. Figure 2(a) represents the positive words cloud, while Figure 2(b) indicates the toxic words cloud, and Figure 2(c) presents the neutral words cloud.



(a) (b) (c)

Figure 2. Words samples: (a) positive words cloud, (b) toxic words cloud, and (c) neutral words cloud

3.2. Features extraction

3.2.1. The term frequency-inverse document frequency (TF-IDF)

A statistical technique that is frequently employed in information retrieval and natural language processing. To reflect the fact that some words are used more frequently than others overall, the tf-idf value increases according to the number of times a word appears in the document and therefore is offset by the number of documents in the corpus that contain the term [19], [20].

3.2.2. Bag of words

The bag-of-words model is considered a simplifying representation applied in natural language processing and information retrieval (IR). Bag of words (BOW) describes the occurrence of words within a document. Every time we use an algorithm in NLP, it operates on numbers. Our text cannot be entered into the algorithm directly. As a result, the text is preprocessed using the bag of words model, which creates a bag of words from it and keeps track of how many times the most common words are used overall. Bag of words and TF-IDF differ significantly in that the former merely counts the frequency of words (TF) while the latter also includes some sort of inverse document frequency (IDF).

3.2.3. Word2Vec

It is a natural language processing technique. Word2vec is a learned representation of text where words with a similar representation have the same meaning. This approach of representing words and documents is considered the breakthrough key of DL in challenging NLP problems.

3.2.4. GloVe

GloVe stands for global vectors for word representation. It is a technique for obtaining vector representations for words. With GloVe, any word in a corpus of text can be intuitively converted into a position in a high-dimensional space. It can be used to identify connections between words, such as synonyms, connections between brands and products, zip codes, and locations.

3.3. Machine learning algorithms

3.3.1. Support vector machines

Support vector machines (SVMs) are supervised learning algorithms used for classification and regression. Its primary goal is to locate a hyperplane in an N-dimensional space that divides the datasets into classes. SVM is considered to be a learning method based on the theory of statistical learning. It can produce accurate classification outcomes without the need for lots of training data [21]. Although SVM works well with the default value, its results could be significantly improved by optimization parameters [22]. Support vector machine draws a hyperplane or set of hyperplanes in a space with high dimensions, to divide the data points into classes [23]. A good separation of the dataset points is fulfilled by the hyperplane with the biggest distance to the nearest training data point of any class which is referred to as the margin. The classifier's generalization error decreases as the margin increases [24]. This algorithm has two kernel types either linear or non-linear. Polynomial, Gaussian, and Sigmoid kernels are examples of common non-linear kernels.

3.3.2. Naïve bayes

Naïve bayes (NBs) belong to the ML classification algorithms using Bayes' theorem as their foundation. This theorem describes, based on prior knowledge of conditions that could be related to an event, the probability of this event. It is based on probability models that incorporate strong independence assumptions [25]. These models share a common principle. It is a very well-known sentiment analysis algorithm. This algorithm predicts the tag of text and determines the probability for each tag of the given text and the output is the one with the highest probability [26].

3.3.3. Random forest

Random forest (RF) is a ML algorithm, that performs supervised learning used for both classification and regression. Where each random forest is composed of multiple decision trees that work together as a group to produce one prediction. This algorithm relies on the decision trees, basically it relies on the majority vote of these trees i.e., the standard combination method for random forest ensembles [27]. It is simply an assemblage of decision trees whose results are grouped into one final result with the highest number of votes. A RF algorithm should have many trees between 64-128 trees [28]. Three different NB models can be used: Bernoulli for binary feature vectors, multinomial for discrete counts, and gaussian for classification.

3.3.4. Logistic regression

Logistic regression (LR) is one of the most well-known ML algorithms, within the category of supervised learning technique. This algorithm predicts the outcome of a categorical variable, which implies that the result has to be a categorical or discrete value, in which it can either be true or false, 0 or 1, and or Yes or No. In addition, logistic regression gives the pass to the use of discrete or categorical predictors and provides the ability to adjust for multiple predictors. Due to this, logistic regression is particularly helpful for the analysis of observational data when adjustment is required to reduce the potential bias resulting from variations in the groups being compared [29]. LR model after calculating the weights of the input computes a sum of the input features and then calculates the logistic of the result.

3.4. Deep learning algorithms

The deep learning (DL) algorithms are a subfield of machine learning that are inspired by the structure and function of the human brain. They use artificial neural networks, which consist of multiple interconnected processing nodes, to analyze complex patterns and relationships in large amounts of data. These algorithms are capable of learning from the input data without relying on explicit rules or programmed instructions. The learning process is accomplished through the adjustment of network weights and biases based on the input data, which allows the network to progressively improve its performance. Deep learning algorithms are used in a variety of applications, including image and speech recognition, natural language processing, and autonomous systems. They have proven to be particularly effective in solving problems where traditional machine learning techniques have fallen short. However, these algorithms require large amounts of data and computational resources to train, which can make their implementation challenging. Despite this, the ongoing advancement in computing power and the increasing availability of large datasets make Deep Learning a rapidly growing area of research and development in the field of artificial intelligence. Figure 3 illustrates LSTM memory cell structure.

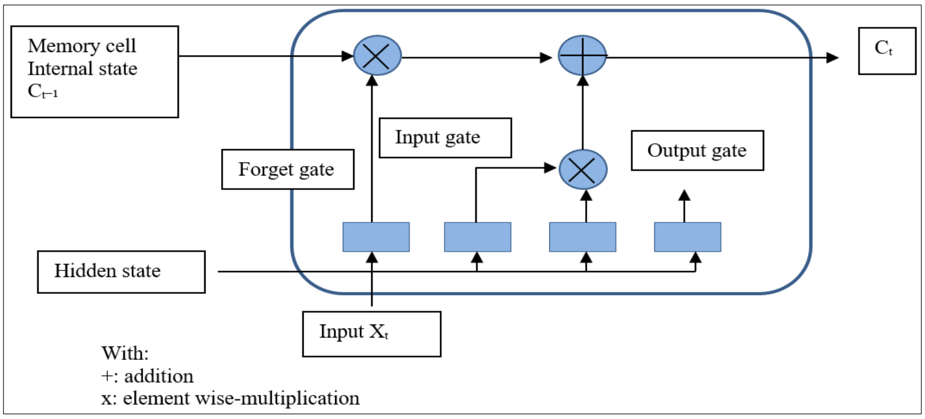


Figure 3. LSTM memory cell

Long short-term memory networks, or LSTMs, are a subset of recurrent neural networks. It is regarded as a strong kind of RNN. An LSTM performs numerous math operations, instead of just feeding its results into the following section of the network. Algorithm 1 represents the operation of an LSTM network.

Algorithm 1. LSTM algorithm

```
1: Input sequence: The LSTM model takes an input sequence, which could be a sequence of text,
   audio, or numerical data.
2: Input preprocessing: The input sequence is preprocessed (vectorization, normalization...)
3: Gates: The input gate, forget gate, and output gate are the three gates that the LSTM
   model uses to regulate the information flow through the network. The LSTM model can
   selectively recall or forget data from earlier time steps thanks to these gates.
4: Cell state: A cell state is used by the LSTM model to store and update data over time.
   The state of the cell serves as a memory that can store relevant information about the
   input sequence.
5: Hidden state: In order to extract the pertinent data from the input sequence at each time
   step, the LSTM model also employs a hidden state. Based on the input, the previous hidden
   state, and the cell state, the hidden state is updated.
6: Output: At each time step, the LSTM model generates an output, which could be a scalar
   number, a probability distribution, or a sequence of values. Based on the updated hidden
   state and cell state at that time step, the output is generated.
```

LSTM layers allow the model to retain information over longer periods, unlike the traditional RNNs. The input data is put into the first layer, and then the output of that layer is put into the second layer, and so on. Each layer processes the input data and generates a new representation, which is then passed to the next layer. After the preprocessing of the text and getting its word vectors of it, we input them in the convolutional network layer as features. The hidden layers intermediate layers between the input and output layers and the place where all the computation is done. The output layer produces the result for given inputs. Due to their capacity to understand long-term connections between data time steps, LSTMs are used to learn, process, and classify sequential data. Sentiment analysis, language modeling, speech recognition, and video analysis are some of the most popular LSTM applications. Figure 4 shows an example of LSTM layers.

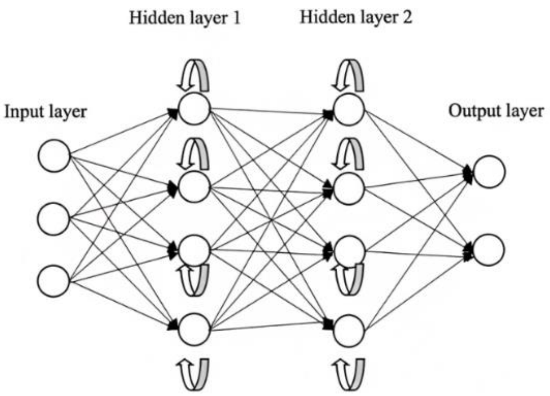


Figure 4. LSM layers

4. RESULTS AND DISCUSSION

4.1. Performance measures

Performance measures are statistical indicators used to evaluate the effectiveness and accuracy of machine learning models. Two commonly used performance measures are the confusion matrix and metrics. The confusion matrix is a table that summarizes the number of correct and incorrect predictions made by a model in binary classification tasks. It provides information on four key metrics: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). These metrics can be used to calculate several important measures of model performance such as precision, recall, F1-score, and accuracy. Precision measures the proportion of positive predictions that are actually positive, recall measures the proportion of actual positive cases that are correctly identified, F1-score is the harmonic mean of precision and recall, and accuracy measures the overall accuracy of a model in making predictions. These performance measures are important in choosing the best model for a given problem and for fine-tuning it for optimal performance.

$$Specificity=\frac{TN}{TN+FP} \tag{1}$$

$$Sensivity=\frac{TP}{FN+TP} \quad (2)$$

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Precision=\frac{TP}{TP+FP} \quad (4)$$

$$F1\text{-}score=2*\frac{Precision* Sensivity}{Precision+Sensivity} \quad (5)$$

4.2. Training results

Dataset contains 3,242 samples, 80% for training, and the other 20% for testing. For each class of the classification, there are almost 1,000 instances. To import and implement the proposed DL model, we used the Tensorflow 2.9 library in this study. We constructed the model in Python and relied on the Jupyter notebook and google colab platforms for training the ML and DL models, and evaluating the models' performances. Before training the models, we used grid search for tuning the parameters. SVM has some hyper-parameters like C or gamma values. Instead of trying all combinations and seeing what parameters work best, we used grid search for better results and time-saving. When it comes to models' execution time, SVM takes a lot of time for the training and the other models consume less time compared to it.

4.3. Testing results

We adopt LSTM as the basis for this experiment alongside SVM RF LR and NB. The results after building the models are shown in Table 1. The highest score was on the side of ML algorithms, especially the one used with tf-idf.

Table 1. Results summary

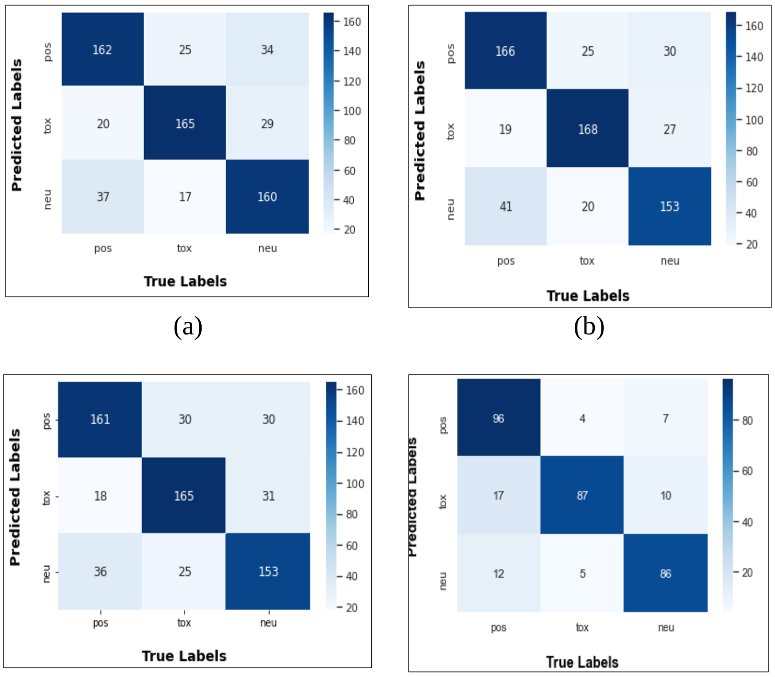
Models	TF-IDF (%)			BAG OF WORDS			WORD2VEC
	Uni-gram	1g+2g	1g+2g +3g	Uni-gram	1g+2g	1g+2g+3g	(%)
SVM	75.04	75.04	73.81	72.27	71.96	71.96	-
RF	71.8	70.42	69.65	69.49	64.87	70.57	-
LR	75.5	74.58	73.81	73.96	72.73	72.88	-
NB	69.95	70.57	71.03	69.18	70.11	70.11	-
LSTM		-			-		83.64

We used metrics such as accuracy, precision, F1-score, and recall for evaluating the performances of models. It is observed that the SVM model performs the best, with an accuracy of 75.04% and an F1 of 78.38%, which shows strength compared with its competitors of the other models. On the other hand, model NB got the lowest accuracy with 71.03% and an F1 of 76.95%. Table 2 shows the used metrics like accuracy, precision, f1-score, and recall.

Table 2. Models' performance summary

Models	Accuracy (%)	Precision (%)	F1-score (%)	Recall (%)
SVM	75.04	79.71	78.38	77.10
RF	71.8	81.45	74.01	83.17
NB	71.03	76.66	76.95	80.37
LR	75.5	79.80	78.67	77.57
LSTM	83.64	83.59	83.49	83.51

These are the confusion matrix that we obtained and they summarize the performance of the classification algorithms. The following matrixes belong to the algorithms with the higher scores in each category. The confusion matrixes are shown in the Figure 5. Figure 5(a) presents the confusion matrix of SVM using tf-idf uni-gram method, while Figure 5(b) shows the confusion matrix of SVM using tf-idf 1g+2g method, and Figure 5(c) indicates the confusion matrix of SVM using tf-idf 1g+3g method. Figure 5(d) shows the confusion matrix of LSTM. The results of the experimentation indicate that LSTM performed the best among all the algorithms tested. For the ML algorithms: The results showed that SVM outperformed other algorithms. SVM achieved the best results, followed by NB in second place, logistic regression in third place, and RF in last place. This shows the effectiveness of LSTM over the other ML algorithms in analyzing and predicting based on the data collected from Instagram.

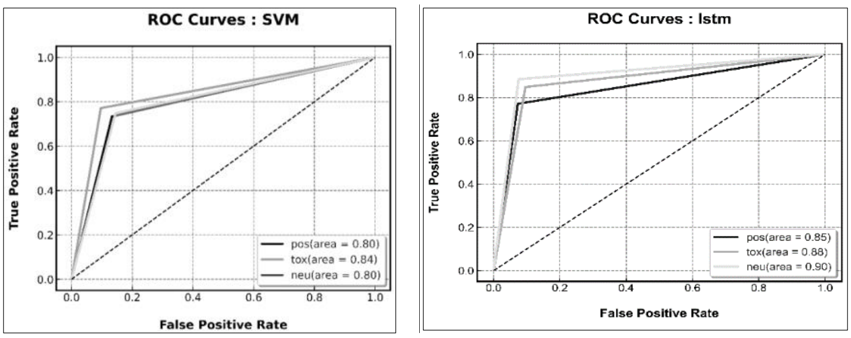


(a) (b)

(c) (d)

Figure 5. The classifications provided by the four studied models. (a) SVM tf-idf unigram, (b) SVM tf-idf 1g+2g, (c) SVM tf-idf 1g+3g, and (d) LSTM

The performance of the algorithms was evaluated using receiver operating characteristic (ROC) curves and the area under the curve (AUC) measure. ROC is a probability curve that gives a measure of separability and shows how well the model is able to distinguish between classes. The higher the AUC, the better the model is at making predictions. Classifiers with curves closer to the top-left corner indicate better performance. Figure 6 shows the ROC curves: the Figure 6.a indicated the ROC curve of SVM and the Figure 6.b shows the ROC curve of LSTM. These figures provide a visual representation of the performance of the algorithms and indicate that LSTM outperforms SVM in terms of separating the classes and making predictions.



(a) (b)

Figure 6. ROC curves of the studied classifiers: (a) SVM and (b) LSM

4.4. Discussion

In this study, we have worked on social media cyberbullying in the Moroccan dialect. The dataset that we have used we collected it from scratch using the famous platform Instagram. The texts in this dataset have been classified into 3 subclasses: positive, toxic, or neutral. We could not find datasets in Arabic or Moroccan dialect for this project, the existing ones are mostly in English. Moreover, each category is recognized as follows: 0 for toxic, 1 for positive, and 2 for neutral. In the first step, after organizing the dataset, we applied a sequence of preprocessing operations to the dataset. These operations include removing emojis and Latin characters, deleting stop words, and removing punctuation and alphanumeric characters. Just after we used the word embedding on the texts to convert them into vectors. In the proposed system architecture, we presented two ways to represent words to assess the effect of feature extraction on the proposed model. In approach 1, we used tf-idf with the three grams. The second method we used BOW. Then we applied the algorithms for the two methods. Overall, the results that we got out of this experiment were medium due to some difficulties with the Moroccan dialect and the dataset size. Table 3 summarizes the accuracy of articles reviewed in related work that used standard languages.

Table 3. Related work accuracies of standard languages summary

Ref.	Language	Techniques/algorithms						
		SVM (%)	*RF (%)*	*NB (%)*	*LR (%)*	*LSTM (%)*	*BERT (%)*	*CNN (%)*
[2]	Arabic standard	68.33	-	68.33	-	77.95	-	-
[4]	English	-	-	-	-	-	98	-
[5]	Vietnamese	78.0	79.10	-	79.91	80.00	-	-
[6]	Georgian	81.6	-	81.6	-	-	-	-
[14]	English	-	-	-	-	95.81	-	-
[1]	English	-	-	-	-	91	-	93
[30]	Hindi, Marathi	-	-	-	85.18	77.77	-	-
[8]	Spanish language: Chilean, Mexican, and Chilean-Mexican corpora	86.90	87.17	75.30	-	-	-	-
[9]	Chinese	54.50	46.40	-	-	-	57.4	60

In this paper, we used the Arabic language and precisely the Moroccan dialect. Table 4 shows the accuracy of the articles that used dialects as well. Haidar *et al.* [31], used tweets database summed up to 4.93 GB, in Lebanese, Syrian, Gulf Area, and Egyptian dialects to detect and stop cyberbullying in Arab countries.

Table 4. Related work accuracies of Arabic dialects summary

Ref.	Language	Techniques/algorithms			
		SVM (%)	RF (%)	NB (%)	LR (%)
[15]	Arabic dialects (Jordanian, Egyptian, and Iraqi)	69	67	66	66
[31]	Arabic (Lebanese, Syrian, Gulf, and Egyptian)	93.4	-	90.1	-
[32]	Arabic (Gulf)	-	-	95.9	-
[16]	Arabic (Egyptian, Gulf, and Levantine)	80.27	80	-	82.7

5. CONCLUSION AND PERSPECTIVES

In this work, we have used ML algorithms SVM, NB, and RF. We used DL algorithms as well like RNN-LSTM which gave medium results. We trained and tested the compiled corpus on four ML models including SVM, NB, RF, and LR, alongside RNN-LSTM. The experiments showed that the LSTM model outperformed the ML models with an accuracy rate of 83.64% and an F1-score rate of 83.49%. However, the results still need enhancement to perform better. The modest results we obtained are due to the difficulty of the Moroccan language and its preprocessing plus the other tools. In future work, we suggest making the model better suited to dealing with the Arabic language and especially the Moroccan dialect to improve results. Furthermore, we can use other DL models as the Arabic version of the BERT model (AraBERT). Moreover, hybrid models where we can combine SVM with another algorithm might increase the scores.

REFERENCES

[1] K. R. Purba, D. Asirvatham, and R. K. Murugesan, “A study on the methods to identify and classify cyberbullying in social media,” in *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*, 2018, pp. 1–6, doi: 10.1109/ICACCAF.2018.8776758.

[2] H. Li, W. Mao, and H. Liu, “Toxic comment detection and classification,” in *CS299 Machine Learning*, 2019.

[3] M. Vichare, S. Thorat, C. S. Uberoi, S. Khedekar, and S. Jaikar, “Toxic comment analysis for online learning,” in *2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, IEEE, 2021, pp. 130–135, doi: 10.1109/ACCESS51619.2021.9563344.

[4] H. Fan *et al.*, “Social media toxicity classification using deep learning: Real-world application uk brexit,” *Electronics*, vol. 10, no. 11, p. 1332, Jun. 2021, doi: 10.3390/electronics10111332.

[5] L. T. Nguyen, K. Van Nguyen, and N. L. T. Nguyen, “Constructive and toxic speech detection for open-domain social media comments in Vietnamese,” in *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, 2021, pp. 572–583, doi: 10.1007/978-3-030-79457-6_49.

[6] N. Lashkarashvili and M. Tsintsadze, “Toxicity detection in online Georgian discussions,” *International Journal of Information Management Data Insights*, vol. 2, no. 1, p. 100062, Apr. 2022, doi: 10.1016/j.jjimei.2022.100062.

[7] G. Song, D. Huang, and Z. Xiao, “A study of multilingual toxic text detection approaches under imbalanced sample distribution,” *Information*, vol. 12, no. 5, p. 205, May 2021, doi: 10.3390/info12050205.

[8] M. Lepe-Faúndez, A. Segura-Navarrete, C. Vidal-Castro, C. Martínez-Araneda, and C. Rubio-Manzano, "Detecting aggressiveness in tweets: a hybrid model for detecting cyberbullying in the Spanish language," *Applied Sciences*, vol. 11, no. 22, Art. no. 22, Jan. 2021, doi: 10.3390/app112210706.

[9] J. L. Wu and C. Y. Tang, "Classifying the severity of cyberbullying incidents by using a hierarchical squashing-attention network," *Applied Sciences*, vol. 12, no. 7, Art. no. 7, Jan. 2022, doi: 10.3390/app12073502.

[10] H. Mubarak, K. Darwish, and W. Magdy, "Abusive language detection on arabic social media," in *Proceedings of the First Workshop on Abusive Language Online*, Vancouver, BC, Canada: Association for Computational Linguistics, 2017, pp. 52–56. doi: 10.18653/v1/W17-3008.

[11] R. Mamoun and M. A. Ahmed, "A comparative study on different types of approaches to the arabic text classification," in *Proceedings of the 1st International Conference of Recent Trends in Information and*, 2014.

[12] H. M. Al Chalabi, S. K. Ray, and K. Shaalan, "Question classification for Arabic question answering systems," in *2015 International Conference on Information and Communication Technology Research (ICTRC)*, May 2015, pp. 310–313. doi: 10.1109/ICTRC.2015.7156484.

[13] B. Haidar, M. Chamoun, and A. Serhrouchni, "Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content," in *2017 1st Cyber Security in Networking Conference (CSNet)*, Oct. 2017, pp. 1–8. doi: 10.1109/CSNET.2017.8242005.

[14] A. Abbasi, A. R. Javed, F. Iqbal, N. Kryvinska, and Z. Jalil, "Deep learning for religious and continent-based toxic content detection and classification," *Scientific Reports*, vol. 12, no. 1, pp. 1–12, 2022.

[15] R. Al-Bayari and S. Abdallah, "Instagram-based benchmark dataset for cyberbullying detection in Arabic text," *Data*, vol. 7, no. 7, p. 83, Jun. 2022, doi: 10.3390/data7070083.

[16] F. Shannag, B. H. Hammo, and H. Faris, "The design, construction and evaluation of annotated Arabic cyberbullying corpus," *Education and Information Technologies*, vol. 27, no. 8, pp. 10977–11023, Sep. 2022, doi: 10.1007/s10639-022-11056-x.

[17] K. U. Manjari, S. Rousha, D. Sumanth, and J. S. Devi, "Extractive text summarization from web pages using selenium and TF-IDF algorithm," in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, Tirunelveli, India: IEEE, Jun. 2020, pp. 648–652. doi: 10.1109/ICOEI48184.2020.9142938.

[18] S. Hamida, B. Cherradi, O. El Gannour, O. Terrada, A. Raihani, and H. Ouajji, "New database of french computer science words handwritten vocabulary," in *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*, Jul. 2021, pp. 1–5. doi: 10.1109/ICOTEN52080.2021.9493438.

[19] S. Hamida, O. El Gannour, B. Cherradi, H. Ouajji, and A. Raihani, "Efficient feature descriptor selection for improved Arabic handwritten words recognition," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 5, p. 5304, Oct. 2022, doi: 10.11591/ijece.v12i5.pp5304-5312

[20] S. Hamida, O. El Gannour, B. Cherradi, H. Ouajji, and A. Raihani, "Handwritten computer science words vocabulary recognition using concatenated convolutional neural networks," *Multimed Tools Appl*, Nov. 2022, doi: 10.1007/s11042-022-14105-2.

[21] J. Ye, X. Cheng, J. Zhu, L. Feng, and L. Song, "A DDoS attack detection method based on SVM in Software Defined Network," *Security and Communication Networks*, vol. 2018, pp. 1–8, 2018, doi: 10.1155/2018/9804061.

[22] I. Syarif, A. Prugel-Bennett, and G. Wills, "SVM parameter optimization using grid search and genetic algorithm to improve classification performance," *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)*, vol. 14, no. 4, p. 1502, Dec. 2016, doi: 10.12928/telkomnika.v14i4.3956.

[23] S. Learn, *1.4. Support Vector Machines-Scikit-learn 0.23. 2 Documentation*. 2019.

[24] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[25] G. N. Kiomourtzis, L. V. Kouros, V. Magoula, and N. Koursioumpas, "Collision avoidance system implementation based on EEG frequency Analysis using ML Techniques," M.Sc. Thesis, National and Kapodistrian University of Athens, Greece, 2020.

[26] V. Vangara, S. P. Vangara, and K. Thirupathur, "Opinion mining classification using naive bayes algorithm," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 9, no. 5, pp. 495–498, 2020.

[27] P. Bonissone, J. M. Cadenas, M. C. Garrido, and R. A. Díaz-Valladares, "A fuzzy random forest," *International Journal of Approximate Reasoning*, vol. 51, no. 7, pp. 729–747, 2010.

[28] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?" in *international workshop on machine learning and data mining in pattern recognition*, Springer, 2012, pp. 154–168.

[29] D. W. Hosmer, *Lemeshow S. Applied logistic regression*. USA: John Wiley and Sons, 2000.

[30] R. Pawar and R. R. Raje, "Multilingual Cyberbullying Detection System," in *2019 IEEE International Conference on Electro Information Technology (EIT)*, 2019, pp. 040-044, doi: 10.1109/EIT.2019.8833846.

[31] B. Haidar, C. Maroun, and A. Serhrouchni, "A multilingual system for cyberbullying detection: Arabic content detection using machine learning," *Advances in Science, Technology and Engineering Systems Journal*, vol. 2, pp. 275–284, Dec. 2017, doi: 10.25046/aj020634.

[32] D. Mouheb, R. Albarghash, M. F. Mowakeh, Z. A. Aghbari, and I. Kamel, "Detection of Arabic cyberbullying on social networks using machine learning," in *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, Nov. 2019, pp. 1–5. doi: 10.1109/AICCSA47632.2019.9035276.

BIOGRAPHIES OF AUTHORS

Rabia Rachidi received a master's degree in business intelligence and big data analytics (BIBDA) in 2022, from the Faculty of Science Chouaib Doukkali university in El Jadida, Morocco. Currently she is a Ph.D. student at Chouaib Doukkali university. Her research interests are primarily in AI, machine learning, deep learning, and NLP. In addition, she is a Ph.D. candidate in the optimization research emerging systems networks and imaging laboratory (LAROSERI) at the faculty of science El Jadida (Chouaib Doukkali University), Morocco. She can be contacted at email: rachidirabia99@gmail.com.

Mohamed Amine Ouassil received the B.Sc. in mathematics and computer science engineering in 2009 from the Faculty of Science and Technology of Beni Mellal, Morocco, and the M.Sc. degree in data science and big data in 2021 from High National School for Computer Science and Systems Analysis (ENSIAS) of Rabat, Morocco. He is currently working as guidance counsellor at National Ministry of Education Morocco. He is also a Ph.D. candidate at electrical engineering and intelligent systems (EEIS) laboratory in ENSET Mohammedia, Hassan II University of Casablanca (UH2C). His research interests reside in the fields of machine learning, artificial intelligence and natural language processing. He can be contacted at email: ouassil.amine@gmail.com.

Mouaad Errami was born in 1998 at Taroudant, Morocco. He received his engineering degree in data engineering and data science from the National Institute of statistics and applied economics of Rabat in 2020. Currently, he is working as a systems engineer at Rabat. He is also a Ph.D. candidate at electrical engineering and intelligent systems (EEIS) laboratory in ENSET Mohammedia, Hassan II University of Casablanca (UH2C). Highly interested in the world of machine learning and artificial intelligence, his articles delve heavily into the realm of NLP such as detecting fake news and sentiment analysis; is also interested in developing this domain when it comes to Arabic language. He can be contacted at email: mouaad.errami@gmail.com.

Bouchaib Cherradi was born in 1970 at El Jadida, Morocco. He received the B.S. degree in Electronics in 1990 and the M.S. degree in applied electronics in 1994 from the ENSET Institute of Mohammedia, Morocco. He received the DESA diploma in Instrumentation of Measure and Control (IMC) from Chouaib Doukkali University at El Jadida in 2004. He received his Ph.D. in Electronics and Image processing from the faculty of science and technology, mohammedia. Dr. Cherradi works as an associate professor in CRMEF-El Jadida. In addition, he is associate researcher member of electrical engineering and intelligent systems (EEIS) laboratory in ENSET of Mohammedia, Hassan II University of casablanca (UH2C), and LaROSERI Laboratory on leave from the faculty of science, El Jadida (Chouaib Doukali University), Morocco. He is a supervisor of several Ph.D. students. He can be contacted at email: bouchaib.cherradi@gmail.com.

Soufiane Hamida is a 29-year-old researcher from Rabat, Morocco, is highly knowledgeable in the field of machine learning methodologies for pattern recognition, as evidenced by his Ph.D. degree. He further honed his skills and expertise in the field of educational technology by obtaining his master’s degree from the higher normal school of tetouan, Abdelmalek Essaadi University in 2017. Currently, he is actively contributing to the advancement of research at the electrical engineering and intelligent systems research laboratory at Hassan II University of Casablanca, Morocco. Furthermore, he is making significant efforts towards furthering research at the GENIUS Laboratory at SupMTI in Rabat, Morocco. He can be contacted at email: hamida.93s@gmail.com.

Hassan Silkan received the Ph.D. in computer sciences from Sidi Mohamed Ben Abdellah University, FSDM, Morocco. Currently, he is a professor in Chouaib Doukkali University, department of computer science, faculty of sciences El Jadida, Morocco. His research areas are shape representation, similarity search, content-based image retrieval, database indexing, machine learning and deep learning. He can be contacted at email: silkan.h@ucd.ac.ma.