Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.048003

ARTICLE

A Machine Learning Approach to Cyberbullying Detection in Arabic Tweets

Dhiaa Musleh[1], Atta Rahman[1,*], Mohammed Abbas Alkherallah[1], Menhal Kamel Al-Bohassan[1], Mustafa Mohammed Alawami[1], Hayder Ali Alsebaa[1], Jawad Ali Alnemer[1], Ghazi Fayez Al-Mutairi[1], May Issa Aldossary[2], Dalal A. Aldowaihi[1] and Fahd Alhaidari[3]

[1]Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam, 31441, Saudi Arabia

[2]Department of Computer Information Systems, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam, 31441, Saudi Arabia

[3]Department of Networks and Communications, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam, 31441, Saudi Arabia

*Corresponding Author: Atta Rahman. Email: aaurrahman@iau.edu.sa

Received: 24 November 2023 Accepted: 28 May 2024 Published: 18 July 2024

ABSTRACT

With the rapid growth of internet usage, a new situation has been created that enables practicing bullying. Cyberbullying has increased over the past decade, and it has the same adverse effects as face-to-face bullying, like anger, sadness, anxiety, and fear. With the anonymity people get on the internet, they tend to be more aggressive and express their emotions freely without considering the effects, which can be a reason for the increase in cyberbullying and it is the main motive behind the current study. This study presents a thorough background of cyberbullying and the techniques used to collect, preprocess, and analyze the datasets. Moreover, a comprehensive review of the literature has been conducted to figure out research gaps and effective techniques and practices in cyberbullying detection in various languages, and it was deduced that there is significant room for improvement in the Arabic language. As a result, the current study focuses on the investigation of shortlisted machine learning algorithms in natural language processing (NLP) for the classification of Arabic datasets duly collected from Twitter (also known as X). In this regard, support vector machine (SVM), Naïve Bayes (NB), Random Forest (RF), Logistic regression (LR), Bootstrap aggregating (Bagging), Gradient Boosting (GBoost), Light Gradient Boosting Machine (LightGBM), Adaptive Boosting (AdaBoost), and eXtreme Gradient Boosting (XGBoost) were shortlisted and investigated due to their effectiveness in the similar problems. Finally, the scheme was evaluated by well-known performance measures like accuracy, precision, Recall, and F1-score. Consequently, XGBoost exhibited the best performance with 89.95% accuracy, which is promising compared to the state-of-the-art.

KEYWORDS

Supervised machine learning; ensemble learning; cyberbullying; Arabic tweets; NLP

1 Introduction

Cyberbullying is defined as a person or group using telecommunication and digital devices to threaten others via communication networks. It includes verbal abuse, harassment, and aggressive words. Cyberbullying may denigrate and unfairly criticize someone or impersonate others' identities. It has harmed many individuals worldwide, particularly teenagers, as it gets more accessible on widespread social networking sites [1]. Cyberbullying is the act of bullying on the internet, so there is no physical damage, but the victims express psychological harm. People will be rude behind the screen, and they are anonymous, so there is no liability for the bully. The number of people bullying the victim is enough to affect him/her with the words. Because in real life, one is driven by a limited number, but on the internet, it is up to thousands [2]. "One in five parents around the world say their child has experienced cyberbullying at least once in 2018" [3]. Cyberbullying is rising every day so that the numbers might become more extensive. It is, therefore, essential to prevent this by first detection, and then remedial actions may be taken employing cyber laws. Nowadays, most use Twitter (also known as X) to communicate and express ideas. Twitter has its own rules to stop the horrible act, whether it is cyberbullying or other things. However, the problem is that it needs to monitor 192 million [4] when you want to monitor up to millions of users and use the site 24 h a day and seven days a week. So, humans cannot do this monitoring manually. On the other hand, Machine Learning (ML) has become practical and can do the job without human intervention in real-time. Based on a survey conducted [4], several implications of cyberbullying have been noticed, as given in Fig. 1. The first Category is about those who experienced cyberbullying. The second Category is about its impact, and the third Category is about the type of cyberbullying that occurred to the individuals.

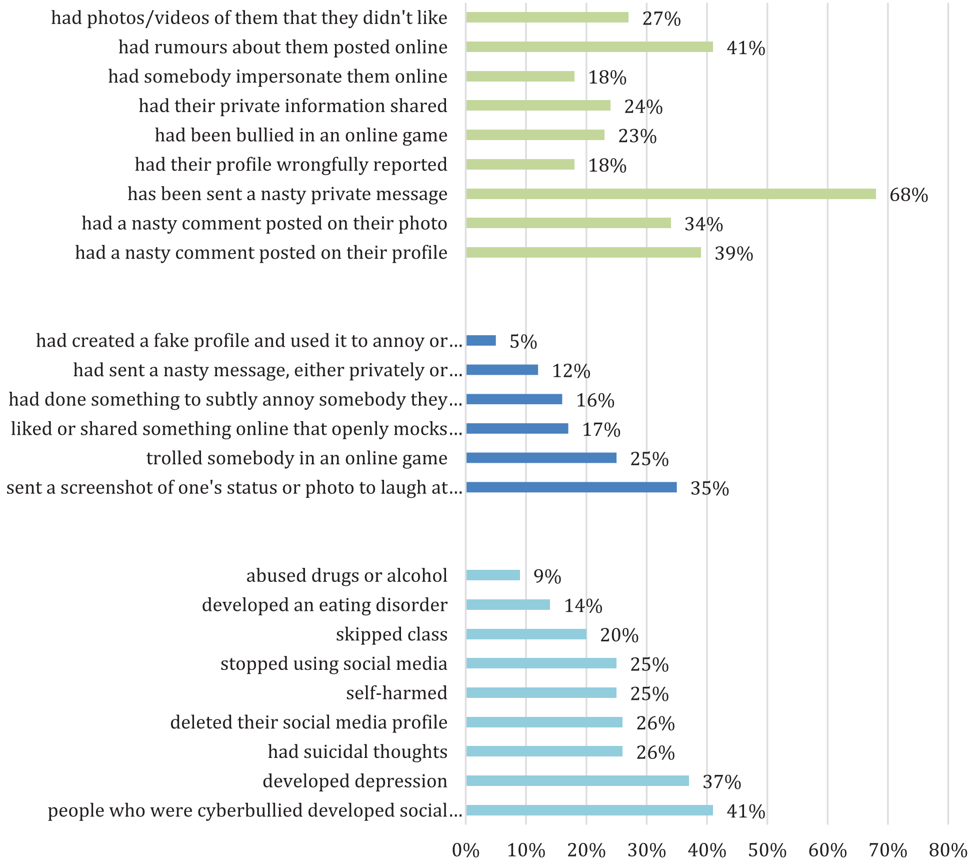

Figure 1: Cyberbullying statistics about its experience, impact, and type

The survey reveals the potential frequency, harm, and kind of impact social media users have, which is worrisome. There is clear evidence and motivation to conduct studies where cyberbullying should be adequately monitored, and appropriate measures should be taken by the authorities to prevent the consequences in society. It is one of the significant areas to target in Saudi Vision 2030, which aims to transform the individual's lifestyle and improve well-being. In the literature, this task has been frequently done for the English language; however, for the Arabic language, studies are limited, and more research is needed, especially considering the diverse dialects in the Arabic language. Arabic is one of the most popular languages, ranking sixth globally. There are 5.2% of Arabic users on the internet [5]. Many teenagers use social media applications like Twitter, Snapchat, Instagram, and TikTok. Because of bullies, teenagers may get depressed or fear facing the world. Bullying in the past was famous in schools and is now known as cyberbullying on the internet. Many teenagers stayed at home because of the Coronavirus disease 2019 (COVID-19), spending more time online, and due to cyberbullying, faced anxiety and other psychological issues [5]. So, the situation demands a system to detect cyberbullying, which can help prevent bullying. Motivated by that, in this paper, we propose to develop a model for preprocessing Arabic tweets and detecting cyberbullying using machine learning techniques. The proposed research aims to identify cyberbullying in Arabic tweets by automatically applying machine learning algorithms with the help of Arabic Natural Language Processing (ANLP) approaches. This research is motivated by relatively few studies identifying cyberbullying that have been conducted in the Arabic language. While an overwhelming increase in the number of Arabic speaking users has been witnessed over last few years due to extended support for the native and non-English languages over the social media platforms in general and twitter in particular. This is first of its kind study in the Kingdom of Saudi Arabia. The significant contributions and aims of the study that makes it prominent from others, are listed below:

- Collection and preparation of a standard dataset comprising Arabic tweets collected over the entire Arab region, especially Saudi Arabia, to comprehend diverse dialects of the language.
- A comprehensive review of related literature over the past decade for cyberbullying detection in general and particularly in Arabic language.
- Investigation of several machine learning approaches on the self-curated and secondary dataset from the literature.
- An improvement in the results compared to the state-of-the-art techniques in the literature.

The rest of the paper is organized as follows: Section 2 provides the background of the study, and Section 3 is dedicated to related work. Section 4 contains the proposed methodology. Section 5 presents results and discussion, while Section 6 concludes the study.

2 Background

Cyberbullying is defined as any harm done constantly and intentionally through the internet. Cyberbullying has four core elements: harm, intent, repetition, and imbalance of power. Furthermore, to be considered cyberbullying, the bully must cause harm, and the damage must be caused willingly. Also, the harm must be repeated; a one-time insult will not be considered cyberbullying. In addition, the bully must have a higher power. For example, they are more prevalent [6]. Another study shows an increase in cyberbullying, as 56.1% of 351 users have admitted to being affected by cyberbullying. The study also demonstrates that the adverse effects of cyberbullying, fear, powerlessness, anger, anxiety, and sadness are the same as those of face-to-face bullying [7,8]. Since cyberbullying detection mainly relies on NLP algorithms, subsections focus on the commonly used steps to provide a background.

2.1 Natural Language Processing (NLP)

Humans use Natural language to communicate, whereas computers only analyze zeros and ones [8]. Technology today has advanced, so we can make computers study languages by transferring them to their understanding using NLP, which assists computers in understanding human languages. Nowadays, we use NLP a lot to communicate with computers and devices daily, for example, language translators. However, NLP is an old concept that was proposed in [9]. Then, in 1954, Georgetown University and IBM worked together to translate Russian sentences into English. It was an automatic system capable of high-quality translation of 250 words and six "grammar" criteria [10].

2.1.1 Tokenization

Tokenization is breaking down a phrase, sentence, paragraph, or even text document into smaller pieces like individual words or concepts. Tokens are the names given to each of these smaller units. Because Tokenization decreases word typographical variance, it is essential for the feature extraction and bag of words (BoW) procedures. Consequently, the words are transformed into features using a feature dictionary, vectors, or feature index. The feature's index is the frequency of (expression) in the vocabulary related to its usage [11]. Further, Tokenization can be used to count numbers and the words frequency.

2.1.2 Stemming

Stemming is the process of reducing a word to its word stem, which affixes to suffixes and prefixes or the word's roots, known as a lemma. The Arabic language is complicated compared to other languages in stemming techniques. Table 1 shows some examples of stemming by removing its prefixes and suffixes. It may also consider infixes.

Table 1: Prefix, suffix, and infix example

Word	Letter(s)	Stem	Type of affix
يشرب	ي	شرب	Prefix
لبيت	ل	بيت	Prefix
بالعلم	بال	علم	Prefix
ذهاب	ا	ذهب	Infix
شربت	ت	شرب	Suffix
بيته	ه	بيت	Suffix
علميا	يا	علم	Suffix

Kanan et al. [12] mentioned their experience using stemming to reduce the number of words in the text. By using the stem, all forms of a word (name, adverb, adjective, or character) are discarded by returning the word to its original root. The study focused on removing the prefixes since the authors believed it would make document classification more effective. They used three NLP preprocessing tools for detection: normalization, stop word removal and stemming. Then, they applied a set of machine learning algorithms, for classification: K-nearest neighbors (KNN), SVM, NB, RF, and J48. When compared with and without stemming, it was found that stemming increased accuracy and F1-score.

2.1.3 Removing Stop Words

Stop words are used as unnecessary filler words for the sentence; removing them is essential. For example, stop words in English are the 'A,' 'Is,' 'The,' and 'Are,' and there are many more. The study focuses on removing the Arabic stop words. These words have been prepared; some are in Table 2. For the Arabic Language, Abu El-Khair [13] defined a list that contains 56 Arabic stop words.

Table 2: Example of Arabic stop words

#	Word
1	انها
2	اثناء
3	اجل
4	في
5	احيانا
6	اذا
7	ايضا

2.1.4 N-Gram

N-gram is a series of letters that, when combined, will help understand their meaning. We use N-gram to spell sentences as every word, two words [14]. Following are examples of different N-grams: "My laptop" (2-grams), "I have kids" (3-grams), and "I am a student" (4-grams). There are three purposes for using N-gram: first, correcting spelling mistakes. For example, when "Los Angeles" is written as "Los Angeles," N-gram can predict and fix errors. Second, it can choose which words can be grouped. For example, it can indicate that "Los" and "Angeles" can be combined as "Los Angeles." Finally, it can assist in anticipating the next word to see the whole meaning if the user erases words like "I drink" and can predict the next word, "water or Juice" [15].

3 Related Work

This section will summarize cyberbullying-related studies by reviewing the preprocessing methods applied in each research, the classifiers used, the dataset provided, the features set, and the results.

3.1 Data Collection

This section presents the different approaches for data collection. First, a group of authors obtained their data primarily from Twitter and supplemented it with other sources. For example, Al-Ajlan et al. [16] received their dataset from Twitter utilizing the Twitter streaming API and querying with a bad word list, and the total number of tweets is 39,000. Haider et al. [17,18] gathered their dataset from Twitter and Facebook using two custom-built tools: a Twitter scraper written in PHP and a Facebook scraper written in Python. In addition, the tools were linked to the mango database server to save data. In [19], the data was gathered by AlHarbi et al. from Twitter through the Twitter API, Microsoft-FLOW, and YouTube comments and was then compiled into a single file comprising 100,327 tweets and comments. Mouheb et al. [20], a dataset containing 25,000 comments and tweets, was gathered from Twitter and YouTube using the YouTube Data API and the Twitter API, respectively. Haider et al. [21] gathered a data collection size of 34,890 from Twitter using a program created by the author team. Kanan et al. [12] gathered their dataset from Twitter using

RStudio and Stool, statistical and mathematical tools for extracting tweets. In addition, the dataset contains 19,650/6138 tweets. Phanomtip et al. [22] detected cyberbullying in the Twitter dataset, and the hate speech tweet dataset was extracted from a paper study, which contained 38,686 and 68,519 tweets, respectively. In [23], the dataset was collected by Banerjee et al. from Twitter, and the size is 69,874. Likewise, Bharti et al. [24] collected a mixed dataset of two public datasets gathered via Twitter API from Twitter. The dataset collected by Almutiry et al. [11] consisted of 17,748 tweets obtained from Twitter through the Twitter API and Arabi Tools. Lokhande et al. [25] also used Twitter as a dataset because it generates data daily. Almutairi et al. [26], after signing in to Twitter via the developer's app, created a new application. They generated access tokens and access token secrets, using Python programming language to access the Twitter API via an open-source library package called tweety. Jain et al. [27] utilized a combination of several datasets containing hate speech: the hate speech Twitter dataset, the hate speech language dataset with tweets, and the Wikipedia dataset. In [28], a Java application was created by Al-Mamun et al. to extract social media data. Twitter and Facebook were proposed for Bangla text. Using Facebook Graph API and Twitter Rest API, 1000 pieces were collected from Facebook, and 1400 Bangla public status updates were obtained from Twitter. Authors in [29] used Twitter API to collect data, then preprocessed it with the NL Toolkit (NLTK), and the term frequency–inverse document frequency (TF-IDF) vectorizer performed the extraction. Balakrishnan et al. [30], using the GamerGate hashtag, collected data from Twitter and used the cyberbullying dataset provided by authors in [31] gathered the information from three distinct social networks: Twitter, Formspring, and Wikipedia, each focusing on a different aspect of cyberbullying—racism and misogyny on Twitter, bullying on Formspring, and Wikipedia attacks.

The second group of authors collected data from sources other than Twitter in their fundamental research. Di-Capua et al.'s [32] dataset was published in public, and the data was taken from Formspring, me, and YouTube. Alakrot et al. [33] also used YouTube; their method was to select some YouTube channels, upload controversial videos about Arab celebrities, and collect comments from them. The data set collected about 15,050 comments. Several authors, like [2], who collected Formspring data, namely Me and MySpace, used Formspring. Reynolds et al. [34] also collected the data from Formspring. The authors selected random files from the 18,554 users and then used Amazon's Mechanical Turk service to determine the labels for the truth sets. Maral et al. [35] also used three datasets in their study: Formspring (a Q&A forum), Wikipedia talk pages (a collaborative knowledge repository), and Twitter (a microblogging platform). All these datasets are manually labeled and publicly available. In a study by Hani et al. [36], the Kaggle dataset was initially taken from a research paper from Formspring and contains 12,773 conversations. Kaggle also has been used by Srivastava et al. [37]. The last group of authors used multiple and different ways and datasets for their data collection. Rachid et al. [38] collected the dataset from Aljazeera.net (accessed on 10/05/2024), and the comments were selected using CrowdFlower.

Furthermore, the comments were classified into three categories: Obscene (533 comments), offensive (25,506 comments), and clean (5653); the data by Husain in [39] was provided by the shared task of the fourth workshop on Open-Source Arabic Corpora and Corpora Processing tools (OSACT) in Language Resource and Evaluation Conference (LREC) 2020. Furthermore, the data was released into three different data parts: training dataset (1000 tweets), development dataset (1367 tweets), and testing dataset (5468). In [40], the dataset used by Alfageh et al. was taken from a publicly available dataset that contains 15,000 comments from YouTube. In [41], the dataset had 2218 sessions provided by a research report collected from Instagram using snowball sampling. Yin et al. [42] used data from three datasets: Kongregate, Slashdot, and MySpace. Then, they manually labeled a randomly selected subset of the threads from each labeled dataset. A study in [43] provided a semi-synthetic, flexible, and scalable dataset to mitigate shortcomings of current cyberbullying detection datasets. The dataset covers various traits such as hostility, reprise, aim to hurt, and issues among peers. Özel et al. [44] manually collected around 900 messages from Twitter and Instagram. However, only half of them contained cyberbullying terms. Dadvar et al. [45] collected cyberbullying related data from MySpace and Fundacion Barcelona Media. Similarly, Buan et al. [46] used the third version of the bullying traces dataset that was created by Professor Zhu at the University of Wisconsin-Madison. Authors in [47] collected their data from Perverted-Justice (PJ), an American organization investigates, identifies, and publicizes the conduct of adults who solicit online sexual conversations with adults posing as minors.

Based on the brief review of cyberbullying detection datasets, it is apparent that most of the data has been collected, processed, and annotated manually by the researchers. Most of the data sources are comprised of English language while for Arabic language datasets are limited. Moreover, Twitter and Instagram are the most famous targeted social media.

3.2 Review of Machine Learning Algorithms in NLP

3.2.1 Support Vector Machine (SVM)

SVM is a fast and dependable classification algorithm. Supervised machine learning has been shown to give excellent and accurate performance results. This section summarizes the primary use of the SVM algorithm for word classification, focusing on cyberbullying detection.

Phanomtip et al. [22] start by filtering the tweets by removing every unimportant information like tags (@), URLs, and locations. Then, they used embedding methods: TF-IDF and document to vector (Doc2Vec) to extract the vector from the filtering tweet and feed it into the linear SVM for the classification. With a large dataset of 68 K tweets, the results state that the experiments for the linear SVM gave excellent accuracy in detecting cyberbullying, with 91% and 86% accuracy using TF-IDF and Doc2Vec, respectively. For a small dataset, Hani et al. [36] used 1.6 K posts and applied the TF-IDF method to extract the vector for the linear SVM. The result was also reasonable accuracy, more than 89%. Buan et al. in [46] followed the same approach by applying the linear SVM to distinguish between bullying and other texts. The idea was to weigh each term by giving it a gram scale. The accuracy of a dataset consisting of more than 14 K tweets was about 77%. Likewise, studies in [47,48] used the same feature and the SVM to reduce the number of parts by weighing their importance to the class attribute.

3.2.2 Naïve Bayes (NB)

NB technique is a simple text categorization algorithm. It is a probabilistic approach for each attribute in each class set. It has been effectively used for various issues and applications but excels in NLP. The studies using NB as their main algorithm are from Rachid et al. in [38], Mouheb et al. in [20], Kanan et al. in [12], Haider et al. in [18], Al-Mamun et al. in [28] and Agrawal et al. in [31]. They used the NB model and collected their data set from Twitter but slightly differed in the preprocessing step and methodology. First, Rachid et al. [38] eliminated Arabic and English punctuation, HTML codecs, numbers and symbols, and words of size one. In addition, diacritics and normalization of Arabic text are all part of the preprocessing stages. In addition, the models used are NBM machine learning and Bag of a Word, which resulted in 87% Presisions, 35% Recall, and 50% F1-score. Second, Mouheb et al. [20], the preprocessing steps include removing Arabic diacritics, Arabic determiner, and normalizations. Furthermore, using the NB model resulted in 95% accuracy. Third, Kanan et al. in [12], the preprocessing steps removed diacritics, non-Arabic literals, symbols, normalization, stop words, stemming, and 91% accuracy. Fourth, Al-Mamun et al. [28] investigated the NB approach for cyberbullying detection in users' posts in Bangla text. The NB scored an F1-score equal to

69% and achieved an accuracy of 60.98%. In English language text, however, the F1-score was 39%, and the accuracy was 40.98%. Finally, Agrawal et al. in [31] investigated the NB approach with character N-grams as a feature selection method. Consequently, the F1-score for the NB was 35.9% for bullying detection, 68.6% for racism on the Formspring dataset, 64.7% for sexism on the Twitter dataset, and 66.5% for attack-related text in Wikipedia. For word unigrams, NB F1-scores were 0.025 for bullying in Formspring, 0.617 for racism, 0.635 for sexism on Twitter, and 0.659 for the attack on Wikipedia. Alfageh et al. in [40] and Özel et al. in [44] utilized the NBM. To begin with, the data set was obtained from YouTube, and the data is balanced; the preprocessing process includes the removal of all non-Arabic literals, symbols punctuation, URLs, hashtags and Arabic diacritics, normalization, and stemming. The model employed is NBM, which has an accuracy of 78.4%. Consequently, Özel et al. [44], in their research on detecting cyberbullying for the Turkish language NBM, were the most successful in their experiments for running time and accuracy. SVM followed them in second place. On the other hand, even though instance-based K-nearest neighbors (IBK) have no training, their running duration is lengthy. J48, on the other hand, has a relatively short testing time but a pervasive training time. As a result, NBM was the top performer. On the other hand, the studies from Bharti et al. [24] performed the worst among all the classifiers and NB algorithms. The accuracy was 66.86%, with an F1-score of 68.18% in cyberbullying detection. Further, the decree showed an accuracy of 90.59% and an F1-score of 92.79%.

3.2.3 SVM and NB

In their study, authors [12] discussed supervised machine learning techniques in Arabic social media content for detecting cyberbullying. They rely on the classifying process using several classifiers: SVM and NB. They start by stating their dataset of 14 K tweets collected from Twitter using RStudio and Rtool and 2 K posts from Facebook. Afterward, they used the Waikato Environment for Knowledge Analysis (WEKA) Toolkit to implement preprocessing tools like Stop-Word-Removal and Stemming. They conclude that SVM gives a better result when the size of the datasets increases, with an F-measure of more than 90%. At the same time, NB gave 10% fewer results than SVM, with an F-measure of less than 85%. Haider et al. [18] used the same approach with a massive dataset of 126K posts collected from the same sources (Twitter and Facebook). The results also favored SVM with an F-measure of more than 92%, while NB achieved an F-measure of 90%. Rachid et al. [38] provided a dataset of 32 K collected from Aljazeera.net, considering if it is a balanced or unbalanced dataset. They applied two feature extraction methods: TD-IDF and BoW. The result shows that SVM is better than NB regardless of whether the dataset is balanced or unbalanced. SVM gave a higher detection accuracy when applying the TD-IDF method, while NB used the Bow method. Haider et al. [21] provided almost the exact dataset size as Rachid et al., but they collected it from Twitter. Also, they applied two different feature extraction methods, Boosting and Bagging. They conclude that SVM beat NB by about 0.4% accuracy higher for both ways. Atoum [48] proposed a simulated annealing (SA) model for the detection process. He used SVM and NB as supervised machine-learning classification tools for this model. He provides a dataset of 5K tweets collected from Twitter API. The experiment results indicated that SVM classifiers have outperformed NB classifiers as they achieved an average accuracy value of more than 90%, while NB gained about 81%. Bharti et al. [24] followed the same approach but with a large dataset of 57 K tweets. The results were similar: SVM accuracy was 91.48%, and NB was 91.35%. From the review, we can conclude that SVM is a promising supervised ML model for classification regardless of the dataset size. While NB significantly lacks accuracy when the dataset decreases [49]. Such as, Dalvi et al. [29] used a minimal dataset; the results were about 71% for the SVM model and 52% for the NB model.

3.2.4 Random Forest

A random forest (RF) comprises many discrete decision trees that perform with a forest. Each tree in the RF generates a class prediction, and the class with the most votes becomes the model's prediction. Their performance is strict to match other algorithms because the trees shield each other from their faults. While some trees may be incorrect, many others will be correct. They can also handle various feature types, including binary, categorical, and numerical. During review RF was used in 7 out of 36 research. There is only one paper used alone. The others used it with other classifiers to compare RF with the others. Al-Ajlan et al. [16] used RF to detect cyberbullying because of its prominence in this field. The model was based on the user's personality and is determined by the Big Five and Dark Triad models. Psychopathy 91.8% had the best RF performance in Dark Triad while the others followed by a small margin of 2%. The Big Five had RF higher than the rest, with almost 91% Neuroticism, Extraversion, and Agreeableness, followed by 90%. So, when the lower RF was 90%, their research shows that the Big Five and Dark Triad models were proven to enhance RF in cyberbullying detection considerably. In three of seven of the research, Husain [39], Kanan, et al. [12], and Agrawal et al. [31], RF was the second-highest performing model. The first research is Husain [39], which detects offensive Arabic language from a dataset of tweets using machine learning models in groups (AdaBoost, Bagging, and RF). Count and TF-IDF were utilized for feature extraction. Their result showed that bagging at 88% had the best F1-score performance, while the RF was the second F1-score by 87%. The second research, Kanan et al. [12], aims to identify such harmful written posts; they proposed using Machine Learning. They used a variety of classifier algorithms (KNN, SVM, NB, RF, and Decision Trees (also known as J48). The first method involved testing each classifier with all the preprocessing stages; RF had a higher F-measure of 94.49%. The second method involved testing the classifiers without stemming; RF had the first F1-score of 94.4%. The third method was to try the classifiers without removing stop words. RF had the best F-measure of 93.4%. The fourth method was to compare the performance of the classifiers with all preprocessing stages on Facebook data only, followed by Twitter records, where SVM had the best F-measure of 94.4%, then RF followed by 91.4% F-measure on Twitter. While on Facebook, SVM had the first F-measure by 91.7l%, while RF was the second by 94.1% F-measure. Remaining studies had RF as the third-best F-measure Rachid et al. [38], Bharti et al. [24], and Jain et al. [27]. Whereas Jain et al. [27] used SVM, Logistic Regression (LR), RF, and Multi-Layered Perceptron (MLP) for classification. Three feature extraction methods were employed: the first technique was the BoW model, second with TF-IDF and last is word to vector (word2vec). The greatest F-measure value for the Twitter dataset is obtained as 93.4% when utilizing BoW and LR. The second F-measure was RF, with 93.4%. The TF-IDF's best F-measure was SVC at 93.9%, followed by LR at 93.6%, and then the third F-measure was RF at 93.3%. The best F1-score in word2vec was MLP at 92.2%; RF was the second F-measure at 92.2%. The best F-measure for the Wikipedia dataset was 83.7 % when TF-IDF and SVC were employed after LR at 82.5%, then RF at 81.8%. Then, BoW was the second-best F-measure. Their best was LR at 82.9%; then RF was the second with 81.8% F-measure. Word2Vec had the lowest F-measures. Their highest was MLP, at 82.5%. At the same time, the lowest F-measure was RF at 77.6%. Then, Bharti et al. [24] extracted features using BoW. They then utilized several classification methods: RF, decision tree, NB, XGBoost, SVM, and logistic regression. The best F-measure was LR, with about 94.06%. Then, SVM by 91.99%. The third is RF 91.8%. The last research was Rachid et al. [38], who gathered their dataset from the Arabic news channel Aljazeera website. The dataset is 32 K comments. Their best machine learning models had 85% accuracy; they got that with three different classifiers, RF, XGBoost and SVM with TF-IDF and N-gram.

3.2.5 SVM vs. RF

Some work has been done using SVM and RF techniques. Rachid et al. [38] used massive and imbalance Arabic datasets collected from the Aljazeera.net website. After they implemented the SVM algorithm, the F-measure was 98%, and with the RF algorithm, it was 80%. Bharti et al. [24] also used the same dataset feature, but the dataset was collected from Twitter in English. The result was 93.3% for SVM and 93% for RF. Jain et al. [27] have a dataset like that of Bharti et al. [24], and they got 93.9% for SVM and 93.3% for RF. All three researchers used TF-IDF for feature extraction and found that SVM is better than RF. On the other hand, Kanan et al. [12] and Husain [39] used large and imbalance datasets in the Arabic language. They conclude that RF is better than SVM. SVM is better when the dataset is large enough, more than 20 K, and RF will be better when the dataset is less than 20 K. So, it shows that there will be no difference if the dataset is an Asian foreign language [50,51].

3.2.6 Logic Regression (LR)

LR is a popular machine learning algorithm that uses the logistic function to classify data. Moreover, it is best used if the classification only holds two categories: bullying and non-bullying. However, it can be used in multiclass classification if one runs the algorithm multiple times using the one-versus-all technique. In most related work, logistic regression showed a higher result than other algorithms. Machine learning (ML) has been used as a deep learning (DL) baseline. Bharti et al. [24] used LR as one of the ML models, and LR was the best overall, with an accuracy of 92%. Rachid et al. [38] have also used LR as a baseline firstly with a BoW on a dataset with more CB (cyberbullying) than NCB (non-cyber bullying) and had an F1-score of 30%, which is a very low due to the imbalance. The second time, it was used on another imbalance dataset with more NCB than CB to make the ratio realistic; the F1-score of the second time was 53%, which is a clear improvement. The third time, LR was used with character level TF-IDF N-gram on a balanced dataset, and it got an F1-score of 84%. Agrawal et al. [31] used LR with both character N-grams and word unigrams, and the highest F1-score was 72% and 76%, respectively. In addition to being used as a baseline for DL, LR has been used for ensemble machine learning. Husain in [39] got an F1-score of 81% for LR, exceeding the decision tree and being 1% lower than SVM. Haidar et al. [21] had a different purpose for LR as they used it as one of five single learners incorporated for stacking ensemble machine learning. Jain et al. [27] have experimented with three feature selection methods: BoW, TF-IDF, and word2vec on two data sets from Twitter and Wikipedia. They got the highest accuracy of 92% when LR was combined with BoW. Alfageh et al. [40] have also used TF-IDF with LR, but the F1-score was lower by 1.8% than when it was used with count vectorization, which had an F1-score of 78.6%.

3.2.7 Other Classifiers

Authors in [16] proposed a Convolution Neural Network (CNN) for cyberbullying detection by adopting profound learning principles instead of machine learning. They implemented their system by applying four steps, starting with embedding the texts as a numerical representation. Then, convolving the input vectors to detect the features and compress the output of the convolution process to smaller matrices so that only significant and transparent parts are considered. The last step includes the dense layer, which will feed all the outputs of the previous layers to all NN's neurons. They provided a dataset of 39K tweets collected using Twitter streaming API for the experiment. The results of an automated cyberbullying detector with no human involvement were excellent as the accuracy was more than 95% of the detection process. Banerjee et al. [23] applied the same method with a larger dataset consisting of 69K tweets, and the results were also outstanding with more than 93% accuracy. Srivastava et al. [37] performed different models of deep learning algorithms in detecting insults in social commentary, and the most important of them are Gated Recurrent Units (GRU), Long short-term memory (LSTM), and Bidirectional LSTM (BLSTM). They first applied data preprocessing to percolate the comments, including text cleaning, Tokenization, stemming, lemmatization, and stop-word removal. After that, the filtered data is passed to the deep learning algorithms for prediction. The results were very similar, as BLSTM outperformed the others by 82.18% accuracy, while GRU and LSTM achieved 81.46% and 80.86%, respectively. In comparing CNN with GRU, LSTM, and BLSTM, Benaissa et al. [38] provided a dataset of 32 K Arabic comments collected from Aljazeera.net for testing. From the results, it seems CNN surpasses the other models by 1% F1-score. Also, combined, they achieved an average of 84% F1-score in a balanced dataset. The KNN is a supervised ML algorithm that may be used for both classification and regression tasks [50]. However, it is mainly employed for categorization and prediction. It is defined by two characteristics: a non-parametric learning algorithm and a lazy learning algorithm. Haidar et al. [21] when they compare KNN by F1-measure to other classifiers. They had that KNN is the lowest in bagging, 90.4%, and in Boosting, 89.8%. Also, Kanan et al. [12] found that KNN has the lowest F1-score, 78.2%, and 61.9%, in Facebook and Twitter datasets. Al-Mamun et al. [28] choose those classifiers (i.e., NB, SVM, Decision Tree, and KNN). In their results, KNN was observed better than NB.

Based on the related work, it is evident that cyberbullying detection in Arabic language is among the hottest areas of research that need significant attention [52,53]. More studies are needed to investigate the overwhelming amount of data being produced by the social media platforms by abundant of users especially in Arabic language. Investigation of smart techniques could potentially help identify this adverse effect of social media precisely to prevent undesirable incidents. In this regard, nine well-known machine learning algorithms have been shortlisted based on their effectiveness in the similar problems observed in the literature. Table 3 presents a summary of the reviewed literature.

Table 3: Summary of literature review

Ref.	Classifier	Language	Dataset source and size	Feature extraction	Metric	Strength	Weakness
[11]	SVM	Arabic	Twitter API 17,748 tweets	Light Stemmer, Arabic Stemmer Khoja, TF-IDF	Accuracy 85.49%	N/A	Small dataset
[12]	KNN, SVM, NB, RF, J48	Arabic	Twitter API, 4000 tweets FB 2138 posts		N/A	Multiple experiments	N/A
[19]	Lexicon PMI Entropy Chi-square	Arabic	Twitter API, YouTube, Microsoft Flow. 100K+	The content of the tweet		Original dataset, lexicon approach.	Imbalance data with stop words
[22]	SVM	English	Twitter API 67K tweets	TF-IDF, and Doc2Vec	74%	Large dataset	Presumed sentence correction
[24]	NB, RF, J48 XGBoost, SVM, LR	English	From old papers 57,787 tweets	BoW	Accuracy 92.60 Precision 98.73 F1-score 94.20 Recall 92.07	6-way classification Big dataset	Imbalanced dataset
[25]	SVM, CNN Keras library	English	Twitter API		N/A	N/A	N/A
[26]	SVM	Arabic	Twitter API 8154 tweets	TF-IDF	Accuracy 82%, Precision 81%, F1-score 82%	Resampling was used to balancing	Minimal pre-processing

(Continued)

Table 3 (continued)

Ref.	Classifier	Language	Dataset source and size	Feature extraction	Metric	Strength	Weakness
[27]	SVM, LR, RF, MLP	English	Twitter & Wikipedia dataset 35,787 tweets 40K comments	BoW TF-IDF word2vec	Accuracy 92.1% Precision 95.9% Recall 92.7% f-measure 93.9%	Large dataset 4-way classifier	Imbalanced dataset
[33]	NB, CNB, LR	Arabic	15,000 YouTube comments	TF-IDF Vectorizations	F1-score 78.6%	Large dataset and features	Minimal pre-processing
[35]	CNN LSTM BLSTM SSWE	English	16K tweets 10K comments on Wikipedia 12K Q&A	N/A	N/A	Many classifiers	Data augmentation was missing
[37]	DL, GRU, LSTM, BLSTM, RNN	English	Kaggle dataset	N/A	Accuracy 82.12%	4-way classification	Imbalance dataset Accuracy is low
[38]	Deep Learning	Arabic	Aljazeera.net 32K comments	BoW N-grams, OOV embeddings	F-score 84%	Hybrid methods.	N/A
[39]	SVM, LR, J48, RF, Bagging, AdaBoost	Arabic	Twitter API 7835 tweets		Count features, TF-IDF	Emoticons and emojis turned into text	The results representation is not clear
[46]	SVM, LSTM, GRU	English	Bullying traces data set. 7321 tweets	N/A	Precision: 82.56% Recall: 86.42% Accuracy: 86.90%	N/A	Old and small dataset
[48]	SVM, NB Chi-square	English	Twitter API 5628 tweets	N/A	SVM, 4-gram 92.02%	N/A	Small dataset
[50]	MARBERT and BERT	Arabic	Around 24K tweets	N/A	F1-score 75%	Spam detection	Needs improvement
[51]	Machine Learning, SVM, NB	Arabic	30K tweets and comments from Twitter/YouTube	TF-IDF BoW	Accuracy SVM: 95.742% NB: 70.942%	Good accuracy	Tweets mixed with comments
[52]	LSTM Bi-LSTM	Arabic	10K tweets from Twitter	N/A	Accuracy 88% for LSTM and BiLSTM	Three datasets used for analysis	Short and imbalance dataset
[53]	Machine Learning	Arabic	Twitter with 4140 tweets	AraBERT, TF-IDF	Accuracy 89%	Three datasets used	Limited instances

4 Methodology

This section covers the steps involved in the proposed study that can be considered as the theoretical framework of the study: Dataset collection, dataset preprocessing, features extraction, generating the models, and evaluation metrics. Different phases of the proposed methodology are shown in Fig. 2.

4.1 Dataset Collection and Labeling

The dataset was collected from Arabic Twitter using the API, and it consists of about 9K tweets from diverse users from the Middle East, especially Saudi Arabia. It provides a linguistic variety in terms of diverse dialects used in the Arab world, not just one standard dialect such as modern standard Arabic (MSA), that is mainly used in similar studies conducted in Saudi Arabia. The reason behind the MSA is that most of the urban areas Arabic speaking users utilize MSA during their conversations [50,51]. While in the rural and semi-urban areas, nonstandard, slang and local diverse dialects are mainly observed during the data collection process. The collected tweets dataset was manually labeled. In this regard, it was divided into two labels or classes: Bullying and non-bullying for binary classification. Within the Bullying class, tweets contained at least one bullying word (considering the standard Arabic dictionary), and in the non-bullying class, tweets are entirely free of them.

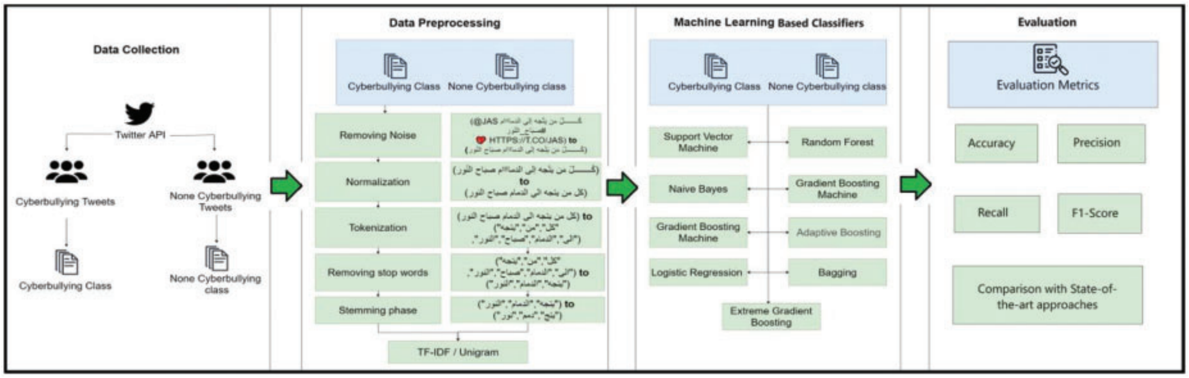

Figure 2: Different phases of the proposed methodology

4.2 *Dataset Preprocessing*

Preprocessing is a critical stage in a machine learning model since it cleans and prepares the dataset so that it may be used to train the classifier. In our case, the tweets are written in various dialects rather than traditional Arabic. Therefore, we have applied NLP approaches to cope with various issues posed by tweets in the Arabic language. It was applied as follows:

4.2.1 Data Cleaning, Handling Missing Values and Outliers

Before any work is done on the dataset, it needs to be cleaned from the inconsistent data by deleting the duplicate and broken (inconsistent) Tweets. The missing values and outliers were handled by the standard NLP preprocessing methods such as imputation [8]. Oversampling technique has been used to balance the collected dataset [8]. Nonetheless, no outliers have been detected in the dataset.

4.2.2 Normalization

The dataset was cleaned and normalized into a uniform text. This process was implemented by Python programming language using "pyarabic" libraries and regular expressions. Here we remove the Hashtags, Usernames, Numbers, URLs, English Letters, Special Characters, Quotations Marks, Brackets, Emojis, Repeated Letters, Tashkeel (like: "أَكَلَ مُحَمَّد تُفَّاحَة" to "أكل محمد تفاحة"), and Tatweel (like: "كتـــــــــــــــــــــــــــــــاب" to "كتاب") [8].

4.2.3 Stop-Word Removal

Stop-words are meaningless terms that do not aid in the analysis. We developed a stop-word dictionary by collecting it from "countwordsfree.com" (accessed on 10/05/2024). Example stop-words: "امس", "يناير", "انت", "انا", etc.

4.2.4 Tokenization

When punctuation and white space marks are encountered, Tokenization is used to split the sentences into tokens. For example, “أكل محمد تفاحة” tokenized into “تفاحة”, “محمد”, “أكل”. The tokenize class from the “pyarabic.araby” package was used.

4.2.5 Stemming

Stemming is a technique for returning a word to its root by removing suffixes and prefixes. For example: “كتاب، كاتب، يكتب” converted into “كتب”. ISRI Stemmer was utilized in this study and was imported from “nltk.stem.isri” library.

4.3 Features Extraction

Focusing on finding the best ML techniques to detect cyberbullying within Arabic tweets, we tested the best model features that gave the highest performance accuracy. N-gram extracted the required features with unigram range and TF-IDF methods. Also, nine supervised ML models were used to train the dataset, including SVM, NB, RF, LR, LightGBM, CatBoost, XGBoost, AdaBoost, and Bagging.

4.3.1 Term Frequency–Inverse Document Frequency (TF-IDF)

TF-IDF is most used for text classification and feature extraction. Term frequency is the number of times a word appears within a document, as given in Eq. (1). The Inverse Document Frequency returns how common or rare a word is in the entire record set, as given in Eq. (2). So, if the word is ubiquitous and appears in many records, this number will be 0. Otherwise, it will be 1, as given in Eq. (3).

$$\mathrm{TF}\,(\mathrm{d}) = \log(1 + \mathrm{freq}(\mathrm{d})) \tag{1}$$

$$\mathrm{IDF}\,(\mathrm{t}) = \log\left(\frac{\mathrm{n}}{\mathrm{DF}\,(\mathrm{t})}\right) + 1 \tag{2}$$

$$\mathrm{TF} - \mathrm{IDF}\,(\mathrm{t}, \mathrm{d}) = \mathrm{TF}\,(\mathrm{t}, \mathrm{d}) \times \mathrm{IDF}(\mathrm{t}) \tag{3}$$

4.3.2 Sequence of N Words (N-Gram)

N-grams extract the sequence of N words. In our case, we used Unigram with range (min: 1, max: 1). For example, the sentence “محمد أكل التفاحة” should be divided into {‘التفاحة’, ‘أكل’, ‘محمد’}.

4.4 Model Generation and Evaluation

This study implemented nine supervisor machine learning models using Python, Sklearn, Catboost, LightGPM, and XGBoost libraries to determine the best classification model. The accuracy as given in Eq. (4), precision as given in Eq. (5), recall as given in Eq. (6), and F1-score as given in Eq. (7) will be used to evaluate the classifiers in this research. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) will be determined using these equations [49].

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{6}$$

$$\text{F1-score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{7}$$

5 Results and Discussion

This section presents the results of all classifiers in detail and demonstrates various effects of Unigram and TF-IDF on text classification for each model. After preprocessing the dataset and extracting the features, the dataset will be fed to the classifiers to determine whether a tweet is cyberbullying. To better evaluate the proposed models, two sets of experiments were conducted in this study, the first with the collected dataset and the second with an existing dataset. It is worth noting that no overfitting was observed during the analyses. The following sections summarize the results of the two experiments.

5.1 Experimental Results with the Collected Dataset

The models were built using a diverse dataset collected and curated in this experiment. The dataset was collected through Twitter API, consisting of about 9K tweets. Nine classifiers were used in these experiments, which are: SVM, RF, LR, AdaBoost, CatBoost, LightGBM, Bagging, XGBoost, and NB. The outcomes of each classifier's model evaluation predictions are shown in Table 4.

Table 4: Experimental results with the collected dataset

Classifier	Feature extraction	Accuracy	Recall	Precision	F1-score
SVM	TF-IDF	87.93	83.33	85.57	84.44
	Unigram	86.41	83.51	82.18	82.84
RF	TF-IDF	89.17	80.99	**90.44**	85.46
	Unigram	89.13	82.34	89.17	85.62
NB	TF-IDF	83.43	75.14	81.29	78.09
	Unigram	83.43	75.14	81.29	78.09
LR	TF-IDF	88.14	78.92	89.66	83.95
	Unigram	87.93	82.07	86.51	84.23
CatBoost	TF-IDF	89.84	84.05	89.45	86.67
	Unigram	89.06	82.97	88.47	85.63
LightGBM	TF-IDF	89.63	84.41	88.65	86.45
	Unigram	89.49	83.24	89.28	86.15
AdaBoost	TF-IDF	88.21	81.44	87.68	84.45
	Unigram	88.42	80.9	88.65	84.6
XGBoost	TF-IDF	**89.95**	84.23	89.56	**88.82**
	Unigram	89.17	82.7	88.95	85.71
Bagging	TF-IDF	89.27	86.67	86.12	86.39
	Unigram	88.39	**88.02**	83.36	85.63

Table 4 shows that XGBoost has the highest accuracy and F1-score, as it reaches 89.95% and 86.82%, respectively, using TF-IDF as a feature extraction approach. Also, RF achieved 90.44% precision using TF-IDF, considered the highest score in the table. The highest score for the Recall goes for the Bagging classifier, as it obtained 88.02% using Unigram as feature extraction. The TF-IDF approach enhances the classifier's accuracy more than the N-grams, as shown in Fig. 3. To achieve better performance, it is best to use the TF-IDF rather than N-gram for extracting the features from text.

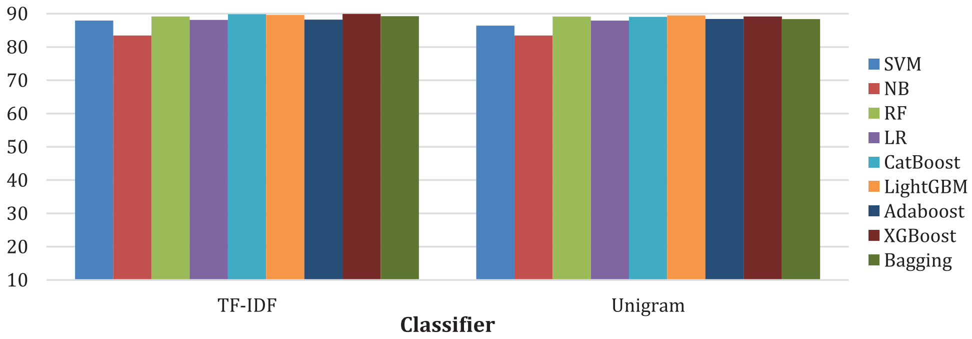

Figure 3: The impact of feature extraction on text classification

5.2 *Experimental Results with an Existing Dataset*

To compare the proposed models with existing research, we investigated the proposed methodology using the dataset used by [33]. This dataset contains 15 K tweets; 39% are labeled as cyberbullying. Their work passed the dataset through preprocessing steps: Tokenization, filtering, normalization, and stemming and then they used N-gram and SVM. In the proposed work, we improved accuracy as follows. First, the dataset was imbalanced, so we added more tweets to make it a balanced dataset. Second, the dataset goes through many steps of preprocessing techniques: removing repeated tweets, tokenization, removing noise (numbers, English letters, emojis), normalization, removing stop words, and stemming. After that, we applied TF-IDF and N-gram for feature extraction and SVM for classification, following the same steps. It is apparent from Table 4 that the proposed approach achieves higher values of evaluation metrics than [33] with the same classifier. Furthermore, we applied LR, to investigate the classification result. TF-IDF with SVM has the highest results, followed by N-gram with LR and N-gram with SVM. The comparison is given in Table 5. It is also apparent that the proposed algorithms are also promising in terms of scalability. The possible reason behind these outcomes is TF-IDF that have proven to be instrumental in NLP overall. On contrary, SVM has been investigated with N-gram, but it did not provide a similar performance. The same is evident in Table 6 where TF-IDF has been combined with XGBoost. As far as the computational complexity of the proposed models is concerned, during training ensemble models take longer than the traditional models in terms of convergence rate. It is obvious due to the nature of the ensemble algorithms where the consensus takes more time.

5.3 *Qualitative Comparison with State-of-the-Art*

To compare the scheme with the state-of-the-art, we have selected some recent studies with three common aspects (Language (Arabic), social media (Twitter), and Data collection region). However, their dataset was different than the one proposed. In this regard, a comparison was made with three recent approaches [52–54] from 2023. The proposed scheme outperforms the schemes given in [52] with

NB, [53], and [54] in terms of accuracy by 19%, 1.95%, and 0.95%, respectively, while using XGBoost with TF-IDF. However, the scheme in [52] outperformed the proposed scheme regarding accuracy and precision using the SVM classifier. Nonetheless, the authors mentioned in [52] that the results are with a segmented scenario and not generalized for their whole dataset. In terms of precision, the proposed scheme using RF with N-gram performs better than [53–54]. Regarding Recall and F1-score, the proposed scheme is better than [53] with a litter margin. Nonetheless, the scheme in [54] marginally performs better than the proposed scheme regarding Recall and F1-score. The comparison with state-of-the-art is presented in Table 6.

Table 5: Comparison with an existing dataset

Approach	Feature extraction	Classifier	Accuracy	Precision	Recall	F1-score
Alakrot et al. (2018) [33]	N-gram	SVM	85%	81%	78%	80%
Proposed approach	N-gram	SVM	86.01%	88.67%	82.27%	85.35%
	TF-IDF	SVM	86.55%	90.20%	81.74%	85.76%
	N-gram	LR	86.42%	89.27%	82.5%	85.75%

Table 6: Comparison with state-of-the-art

Approach	Feature extraction	Classifier	Accuracy	Precision	Recall	F1-score
Alduailaj et al. (2023) [51]	TF-IDF	SVM	95.742%	92%	84%	88%
	BoW	NB	70.942%			
Alzaqebah et al. (2023) [52]	None	LSTM	88%	88%	88%	88%
Mursi et al. (2023) [53]	TF-IDF	MLP	89%	88%	90%	89%
Proposed	TF-IDF	XGBoost	89.95%	90.44%	88.02%	88.82%
	N-gram	RF				
	Unigram	Bagging				

5.4 Discussion

The proposed scheme addresses the cyberbullying detection problem from Arabic tweets. It is from the hottest areas of research in behavioral studies and modelling of online users [54]. The dataset in this regard has been collected from diverse Arab regions, annotated, and preprocessed before applying a broad spectrum of approaches in machine learning. The analyses are made with the same dataset and a secondary dataset from the literature. The proposed scheme was promising in both cases. Wholistically, it was observed that in terms of feature extraction methods, the schemes involving TF-IDF performed better than other feature selection methods such as BoW, and N-gram. Moreover, the preprocessing techniques used in the studies with the diverse dialects used in the Arabic language tweets make a difference. In contrast to schemes in [33,52–54], the proposed scheme is comparable in all metrics. The dataset was collected by means of Twitter APIs and fused prior to investigation, and it does not contain any users' personal information of any type. The scheme can easily be generalized and/or extended to other social media platforms such as YouTube and Facebook comments. The findings of the study can be interpreted as: the models can be used in social media platforms to combat cyberbullying by identifying potential users and contents containing such offensive language and taking strict actions like reporting and blocking the accounts to prevent the emotional and phycological damage to the victims of cyberbullying.

5.5 Limitations of the Study and Future Work

As far as the limitation of the study is concerned, the dataset is limited in terms of the number of instances, though it contains diversified tweet samples collected from diverse middle eastern regions with various dialects. Those regions include Iraq, UAE, Oman, Kuwait, Qatar, Jordan, and other Arabic speaking nations. The scheme is robust against the dialectic variations and capable of handling them effectively as evident in the results and discussion section. Nonetheless, the major chunk of dataset was collected from Saudi Arabia's major cities like Riyadh, Dammam, and Jeddah where modern standard Arabic (MSA) dialect is usually evident in social media. That could be a potential bias in the dataset. Such issues can be overcome by using data augmentation techniques by adding more data with a uniform distribution and balanced sampling techniques. Moreover, accuracy and other metrics can be improved by using advanced data preprocessing techniques with different feature extraction methods [55], deep learning and ensemble learning approaches. Additionally, the encoders/transformers like Bidirectional Encoder Representations from Transformers (BERT) with their Arabic counter parts namely ARBERT (Arabic BERT) and MARBERT (Modern Standard Arabic with BERT) can also be investigated. MARBERT is a large-scale pre-trained masked language model focused on both Dialectal Arabic (DA) and MSA [50].

6 Conclusions

Detection of cyberbullying is getting more difficult as internet users have too many ways of bullying without being identified. Cyberbullying can threaten individuals and cause the victims to commit suicide or go into depression, so its detection is necessary. Several studies have been conducted in the literature but mainly in English, while only a few studies exist in Arabic. In this study, we have proposed and developed Machine Learning models for Cyberbullying Detection from Arabic tweets with diverse dialects. We improved the proposed model significantly using feature extraction methods. The dataset achieved high results using XGBoost, Bagging, and RF classifiers, with XGBoost getting the highest accuracy of them all at 89.95%, using TF-IDF as feature extraction. In the future, we intend to enhance the dataset by data augmentation techniques. Moreover, in future, we aim to improve the accuracy of the detection method by applying hybrid models, investigating diverse datasets, and more than two feature extraction methods for the NLP, which will help further fine-tune the models.

Acknowledgement: The authors like to acknowledge CCSIT, IAU, Dammam for using the resources.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Conceptualization, Dhiaa Musleh; Data curation, Minhal AlBo-Hassan, Jawad Alnemer and Ghazy Al-Mutairi; Formal analysis, Atta Rahman, Hayder Alsebaa and Dalal Aldowaihi; Funding acquisition, May Aldossary and Fahd Alhaidari; Investigation, Dalal Aldowaihi and Fahd Alhaidari; Methodology, Dhiaa Musleh, Mohammed Alkherallah and Hayder Alsebaa; Project administration, Atta Rahman; Resources, Jawad Alnemer; Software, Mohammed Alkherallah,

Minhal Al-Bo-Hassan, Mustafa Alawami and Jawad Alnemer; Supervision, Dhiaa Musleh and Dalal Aldo-waihi; Validation, Mustafa Alawami, Hayder Alsebaa and May Aldossary; Visualization, Ghazy Al-Mutairi; Writing–original draft, Mohammed Alkherallah, Minhal AlBo-Hassan, Mustafa Alawami and Ghazy Al-Mutairi; Writing–review & editing, Atta Rahman, May Aldossary and Fahd Alhaidari. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data can be requested from the corresponding author.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] W. N. Hamiza Wan Ali, M. Mohd, and F. Fauzi, "Cyberbullying detection: An overview," in *2018 Cyber Resilience Conf.*, Putrajaya, Malaysia, 2018, pp. 1–3. doi: 10.1109/CR.2018.8626869.

[2] B. Sri Nandhini and J. I. Sheeba, "Online social network bullying detection using intelligence technique," *Procedia Comput. Sci.*, vol. 45, no. 1, pp. 485–492, 2015. doi: 10.1016/j.procs.2015.03.085.

[3] M. Iqbal, "Twitter revenue and usage statistics (2021) Business of App," Jul. 5, 2021. Accessed: Sep. 14, 2021. [Online]. Available: https://www.businessofapps.com/data/twitter-statistics/

[4] D. Label, "Cyberbullying statistics," Accessed: Jan. 16, 2021. [Online]. Available: https://www.ditchthelabel.org/cyber-bullying-statistics-what-they-tell-us

[5] "Most common languages used on the internet as of January 2020, by share of internet users," Jun. 2020. Accessed: Sep. 14, 2021. [Online]. Available: https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/

[6] J. Patchin and S. Hinduja, "Measuring cyberbullying: Implications for research," *Aggress. Violent Behav.*, vol. 23, no. 4, pp. 69–74, Jul.–Aug. 2015. doi: 10.1016/j.avb.2015.05.013.

[7] D. N. Hoff and S. N. Mitchell, "Cyberbullying: Causes, effects, and remedies," *J. Educ. Adm.*, vol. 47, no. 5, pp. 652–665, Aug. 2009. doi: 10.1108/09578230910981107.

[8] A. Alqarni and A. Rahman, "Arabic tweets-based sentiment analysis to investigate the impact of COVID-19 in KSA: A deep learning approach," *Big Data Cogn. Comput.*, vol. 7, no. 1, pp. 16, 2023. doi: 10.3390/bdcc7010016.

[9] W. J. Hutchins, "The georgetown-IBM experiment demonstrated in January 1954," in *6th Conf. Assoc. Mach. Translat. Americas*, Washington DC, USA, 2004.

[10] S. A. Mandal, "Evolution of machine translation," *Towards Data Science*, Accessed: Jan. 4, 2023. [Online]. Available: https://towardsdatascience.com/evolution-of-machine-translation-5524f1c88b25

[11] S. Almutiry and M. Abdel Fattah, "Arabic cyberbullying detection using Arabic sentiment," *Egyptian J. Lang. Eng.*, vol. 8, no. 1, pp. 39–50, Apr. 2021. doi: 10.21608/ejle.2021.50240.1017.

[12] T. Kanan, A. Aldaaja, and B. Hawashin, "Cyber-bullying and cyber-harassment detection using supervised machine learning techniques in Arabic social media contents," *J. Int. Technol.*, vol. 21, no. 5, pp. 1409–1421, Sep. 2020.

[13] I. Abu El-Khair, "Effects of stop words elimination on Arabic information retrieval," *Int. J. Comput. & Inform. Sci.*, vol. 4, no. 3, pp. 110–133, 2006.

[14] S. L. Aouragh, A. Yousfi, S. Laaroussi, H. Gueddah, and M. Nejja, "A new estimate of the N-gram language model," *Procedia Comput. Sci.*, vol. 189, no. 1, pp. 211–215, 2021. doi: 10.1016/j.procs.2021.05.111.

[15] S. Srinidhi, "Understanding Word N-grams and N-gram probability in natural language processing," *Towards Data Science*, Accessed: Apr. 23, 2021. [Online]. Available: https://towardsdatascience.com/understanding-word-n-grams-and-n-gram-probability-in-natural-language-processing-9d9eef0fa058

[16] M. A. Al-Ajlan and M. Ykhlef, "Deep learning algorithm for cyberbullying detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 9, pp. 199–205, 2018. doi: 10.14569/IJACSA.2018.090927.

[17] B. Haidar, M. Chamoun, and A. Serhrouchni, “Arabic cyberbullying detection: Using deep learning,” in *2018 7th Int. Conf. Comput. Commun. Eng.*, Kuala Lumpur, Malaysia, 2018, pp. 284–289. doi: 10.1109/ICCCE.2018.8539303.

[18] B. Haidar, M. Chamoun, and A. Serhrouchni, “A multilingual system for cyberbullying detection: Arabic content detection using machine learning,” *Adv. Sci. Technol. Eng. Syst. J.*, vol. 2, no. 6, pp. 275–284, 2017. doi: 10.25046/aj020634.

[19] M. S. AlHarbi, B. Y. AlHarbi, N. J. AlZahrani, M. M. Alsheail, J. F. Alshobaili and D. M. Ibrahim, “Automatic cyber bullying detection in Arabic social media,” *Int. J. Eng. Res. Technol.*, vol. 12, no. 12, pp. 2330–2335, 2019.

[20] 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 Int. Conf. Comput. Syst. Appl.*, Abu Dhabi, UAE, 2019, pp. 1–5. doi: 10.1109/AICCSA47632.2019.9035276.

[21] B. Haidar, M. Chamoun, and A. Serhrouchni, “Arabic cyberbullying detection: Enhancing performance by using ensemble machine learning,” in *Proc. IEEE Joint ithings, GreenCom, CPSCom) and SmartData*, Atlanta, GA, USA, 2019, pp. 323–327. doi: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00074.

[22] A. Phanomtip, T. Sueb-in, and S. Vittayakorn, “Cyberbullying detection on tweets,” in *2021 18th Int. Conf. Elect. Eng./Electron., Comput., Telecommun. Inf. Technol.*, Chiang Mai, Thailand, 2021, pp. 295–298. doi: 10.1109/ECTI-CON51831.2021.9454848.

[23] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, “Detection of cyberbullying using deep neural network,” in *Proc. ICACCS*, Coimbatore, India, 2019, pp. 604–607. doi: 10.1109/ICACCS.2019.8728378.

[24] S. Bharti, A. K. Yadav, M. Kumar, and D. Yadav, “Cyberbullying detection from tweets using deep learning,” *Kybernetes*, vol. 51, no. 9, pp. 1–13, 2021.

[25] M. Lokhande, A. Suryawanshi, N. Kaulgud, R. Joshi, and P. Ingle, “Detecting cyber bullying on twitter using machine learning techniques,” *J. Critical Rev.*, vol. 7, no. 19, pp. 1090–1094, 2020.

[26] A. R. Almutairi and M. A. Al-Hagery, “Cyberbullying detection by sentiment analysis of tweets’ contents written in Arabic in Saudi Arabia society,” *Int. J. Comput. Sci. Netw. Secur.*, vol. 21, no. 3, pp. 112–119, 2021.

[27] V. Jain, V. Kumar, V. Pal, and D. K. Vishwakarma, “Detection of cyberbullying on social media using machine learning,” in *2021 5th Int. Conf. Comput. Methodologies Commun.*, Erode, India, 2021, pp. 1091–1096. doi: 10.1109/ICCMC51019.2021.9418254.

[28] A. Al-Mamun and S. Akhter, “Social media bullying detection using machine learning on Bangla text,” in *2018 10th Int. Conf. Elect. Comput. Eng.*, Dhaka, Bangladesh, 2018, pp. 385–388. doi: 10.1109/ICECE.2018.8636797.

[29] R. R. Dalvi, S. B. Chavan, and A. Halbe, “Detecting twitter cyberbullying using machine learning,” in *2020 4th Int. Conf. Intell. Comput. Control Syst.*, Madurai, India, May 2020, pp. 297–301. doi: 10.1109/ICICCS48265.2020.9120893.

[30] V. Balakrishnan, S. Khana, T. Fernandez, and H. R. Arabnia, “Cyberbullying detection on twitter using big five and dark triad features,” *Pers. Individ. Dif.*, vol. 141, no. 15, pp. 252–257, Apr. 2019. doi: 10.1016/j.paid.2019.01.024.

[31] S. Agrawal and A. Awekar, “Deep learning for detecting cyberbullying across multiple social media platforms,” in *Proc. ECIR 2018, Adv. Inform. Retrieval*, Grenoble, France, 2018, pp. 141–153.

[32] M. di Capua, E. di Nardo, and A. Petrosino, “Unsupervised cyber bullying detection in social networks,” in *2016 23rd Int. Conf. Pattern Recognit.*, Cancun, Mexico, 2016, pp. 432–437. doi: 10.1109/ICPR.2016.7899672.

[33] A. Alakrot, L. Murray, and N. Nikolov, “Towards accurate detection of offensive language in online communication in Arabic,” *Procedia Comput. Sci.*, vol. 142, no. 1, pp. 315–320, 2018. doi: 10.1016/j.procs.2018.10.491.

[34] K. Reynolds, A. Kontostathis, and L. Edwards, “Using machine learning to detect cyberbullying,” in *Proc. 10th Int. Conf. ML and Appl. and Workshops*, Honolulu, HI, USA, 2011, pp. 241–244.

[35] D. Maral and K. Eckert, "Cyberbullying detection in social networks using deep learning based models; a reproducibility study," in *2016 23rd Int. Conf. Pattern Recognit.*, Bratislava, Slovakia, 2020, pp. 245–255. doi: 10.1109/ICPR.2016.7899672.

[36] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer and A. Mohammed, "Social media cyberbullying detection using machine learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 5, pp. 703–707, 2019. doi: 10.14569/issn.2156-5570.

[37] G. Srivastava, S. Khan, C. Iwendi, and P. K. Reddy Maddikunta, "Cyberbullying detection solutions based on deep learning," *Multimed. Syst.*, vol. 29, no. 1, pp. 1839–1852, 2023. doi: 10.1007/s00530-020-00701-5.

[38] B. A. Rachid, H. Azza, and H. H. Ben Ghezala, "Classification of cyberbullying text in Arabic," in *2020 Int. Joint Conf. on Neural Netw. (IJCNN)* , Glasgow, UK, 2020, pp. 1–7. doi: 10.1109/IJCNN48605.2020.9206643.

[39] F. Husain, "Arabic offensive language detection using machine learning and ensemble," arXiv:2005.08946, 2020.

[40] D. Alfageh and T. Alsubait, "Comparison of machine learning techniques for cyberbullying," *Int. J. Comput. Sci. Netw. Secur.*, vol. 21, no. 1, pp. 1–5, 2021.

[41] I. Nazar, D. S. Zois, and M. Yao, "A hierarchical approach for timely cyberbullying detection," in *2019 IEEE Data Sci. Workshop (DSW)*, Minneapolis, MN, USA, 2019, pp. 190–195. doi: 10.1109/DSW.2019.8755598.

[42] D. Yin, Z. Xue, and L. Hong, "Detection of harassment on Web 2.0," in *Proc. Content Anal. Web Workshop at Www*, Madrid, Spain, Apr. 21, 2009, pp. 1–6.

[43] N. Ejaz, F. Razi, and S. Choudhury, "Towards comprehensive cyberbullying detection: A dataset incorporating aggressive texts, repetition, peerness, and intent to harm," *Comput. Human Behav.*, vol. 153, no. 3, pp. 108123, 2024. doi: 10.1016/j.chb.2023.108123.

[44] S. A. Özel, S. Akdemir, H. Aksu, and E. Saraç, "Detection of cyberbullying on social media messages in Turkish," in *2017 Int. Conf. Comput. Sci. Eng. (UBMK)*, Antalya, Turkey, 2017, pp. 366–370. doi: 10.1109/UBMK.2017.8093411.

[45] M. Dadvar and F. deJong, "Cyberbullying detection: A step toward a safer internet yard," in *WWW '12 Companion: Proc. 21st Int. Conf. on World Wide Web*, Lyon France, Apr. 2012, pp. 121. doi: 10.1145/2187980.2187995.

[46] T. A. Buan, S. Steria, and R. Ramachandra, "Automated cyberbullying detection in social media using an SVM activated stacked convolution LSTM network," in *ICCDA '20: Proc. 2020 4th International Conf. Comput. Data Anal.*, CA, USA, Mar. 2020, pp. 170–174.

[47] N. Potha and M. Maragoudakis, "Cyberbullying detection using time series modeling," in *Proc. IEEE Int. Conf. Data Mining Workshop*, Shenzhen, China, 2014, pp. 373–382.

[48] J. O. Atoum, "Cyberbullying detection through sentiment analysis," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Las Vegas, NV, USA, 2020, pp. 292–297.

[49] D. A. Musleh *et al.*, "Arabic sentiment analysis of YouTube comments: NLP-Based machine learning approaches for content evaluation," *Big Data Cogn. Comput.*, vol. 7, no. 127, pp. 1–15, 2023. doi: 10.3390/bdcc7030127.

[50] A. Alotaibi *et al.*, "Spam and sentiment detection in Arabic tweets using MarBert model," *Math. Model. Eng. Prob.*, vol. 9, no. 6, pp. 1574–1582, 2022. doi: 10.18280/mmep.090617.

[51] A. M. Alduailaj and A. Belghith, "Detecting Arabic cyberbullying tweets using machine learning," *Mach Learn. Knowl. Extr.*, vol. 5, no. 1, pp. 29–42, 2023. doi: 10.3390/make5010003.

[52] M. Alzaqebah *et al.*, "Cyberbullying detection framework for short and imbalanced Arabic datasets," *J. King Saud Univ.– Comput. Inform. Sci.*, vol. 35, no. 8, pp. 101652, 2023. doi: 10.1016/j.jksuci.2023.101652.

[53] K. T. Mursi and A. M. Almalki, "ArCyb: A robust machine-learning model for Arabic cyberbullying tweets in Saudi Arabia," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 9, pp. 1059–1067, 2023. doi: 10.14569/IJACSA.2023.01409110.

[54] Atta-ur-Rahman, S. Dash, A. K. Luhach, N. Chilamkurti, S. Baek and Y. Nam, "A neuro-fuzzy approach for user behaviour classification and prediction," *J. Cloud Comput.*, vol. 8, no. 1, pp. 1–15, 2019. doi: 10.1186/s13677-019-0144-9.

[55] V. Balakrisnan and M. Kaity, "Cyberbullying detection and machine learning: A systematic literature review," *Artif. Intell. Rev.*, vol. 56, no. Suppl 1, pp. 1375–1416, 2023. doi: 10.1007/s10462-023-10553-w.