machine learning & knowledge extraction

MDPI

Concept Paper

Detecting Arabic Cyberbullying Tweets Using Machine Learning

Alanoud Mohammed Alduailaj and Aymen Belghith *

College of Computing and Informatics, Saudi Electronic University, Abu Bakr Street, P.O. Box 93499, Riyadh 11673, Saudi Arabia
* Correspondence: a.belghith@seu.edu.sa or aymen.belghith@gmail.com

Abstract: The advancement of technology has paved the way for a new type of bullying, which often leads to negative stigma in the social setting. Cyberbullying is a cybercrime wherein one individual becomes the target of harassment and hatred. It has recently become more prevalent due to a rise in the usage of social media platforms, and, in some severe situations, it has even led to victims' suicides. In the literature, several cyberbullying detection methods are proposed, but they are mainly focused on word-based data and user account attributes. Furthermore, most of them are related to the English language. Meanwhile, only a few papers have studied cyberbullying detection in Arabic social media platforms. This paper, therefore, aims to use machine learning in the Arabic language for automatic cyberbullying detection. The proposed mechanism identifies cyberbullying using the Support Vector Machine (SVM) classifier algorithm by using a real dataset obtained from YouTube and Twitter to train and test the classifier. Moreover, we include the Farasa tool to overcome text limitations and improve the detection of bullying attacks.

Keywords: cyberbullying; classification; detection; machine learning (ML); Support Vector Machine (SVM); Arabic social media

Citation: Alduailaj, A.M.; Belghith, A. Detecting Arabic Cyberbullying Tweets Using Machine Learning. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 29–42. https://doi.org/10.3390/make5010003

Academic Editor: Francesco Buccafurri

Received: 3 October 2022
Revised: 31 December 2022
Accepted: 3 January 2023
Published: 5 January 2023

1. Introduction

Recently, the Internet and social media have become promising platforms for learning, sharing opinions, and exchanging ideas. Twitter is a popular interactive social networking platform that allows users to share their positive and negative comments. People now are digitally connected, regardless of time, location, and distance. Social network sites have become the trend, and the vast majority of people, especially teenagers, are eager to join and engage in online communities. The anonymity of social networks, where users often use pseudonyms rather than their actual names, has resulted in a huge number of online crimes, such as cyberbullying, making their activities more difficult to monitor. Cyberbullying is one of the most serious ethical issues on the Internet, and the number of people who have been victims of cyberbullying, especially teenagers, is disturbing. Cyberbullying is described as any offensive action taken by persons toward a victim via online media. In order to measure its prevalence, many studies have addressed cyberbullying, and findings have shown that cyberbullying is a prevalent issue among today's young people, with a growing number of victims [1]. Many cyberbullying detection mechanisms have been established to help in the monitoring and prevention of cyberbullying.

Researchers' work has grown in the area of cyberbullying detection. Despite its prevalence and negative effects within the Arabic culture, few studies have examined this type of attack in the Arabic language [2]. Moreover, the Arabic language has a rich morphological structure and a complex nature. These language characteristics, combined with the lack of focus on proposing detection mechanisms for Arabic cyberbullying, increase the challenge of proposing an efficient detection mechanism taking into account the Arabic context. Another challenge exists due to the variations in the environment surrounding users and their interactions; existing contributions in other languages are not relevant to the Arabic cultural context. In fact, there are several expressions that are not tolerated in

Arabic culture, while they are entirely acceptable in other cultures [2]. For example, the words "كلب", "Dog", and "حمار", "Donkey", are examples of domestic animals. However, it is not permissible to use these types of words in another context, such as describing an action or a person.

Recently, the automatic detection of cyberbullying has led to impressive improvements in the classification of cyberbullying, especially in the English language, such as in [3,4]. However, only a few pieces of research have been conducted using machine learning for Arabic cyberbullying on social networks, such as in [2,5]. Thus, we aim to improve the accuracy of Arabic cyberbullying detection and increase the performance and capabilities of Arabic Natural Language Processing (NLP).

The rest of this paper is organized as follows. Section 2 presents the background of cyberbullying, including the definition and machine learning approaches and, more specifically, the SVM classifier. Section 3 provides an overview of the existing cyberbullying detection mechanism. Section 4 describes the proposed model to detect cyberbullying comments on Arabic social media platforms, whereas Section 5 presents the experiments and obtained results. Finally, Section 6 contains the paper's conclusions.

2. Background

In this section, we present a brief background regarding cyberbullying and its impacts, machine learning approaches, and Natural Language Processing areas.

2.1. Definition of Cyberbullying

Cyberbullying does not have a single meaning, but it has been studied from multiple viewpoints in the literature and several meanings have been proposed. Cyberbullying is a type of harassment carried out and enabled by Information and Communication Technologies (ICT), including text-based data and messaging apps and various social media platforms [1]. Another definition describes cyberbullying as "an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself" [1].

A. Types of Cyberbullying

The different types of cyberbullying are described below:

- Scary: occurs when the poster sends scary messages [6]. The goal of this type is to intimidate the victim through fear.
- Flooding: in which the bully frequently sends the same statements, makes inflammatory comments, or presses the Enter key to prevent the victim from contributing to the conversation [6].
- Masquerade: involves the bully claiming to be a different person. This is intended to give the impression that the bully does not directly intimidate a victim [6].
- Flaming: a form of online combat where a bully sends emails or electronic messages, to an online group, that are dismissive or vulgar towards one or more victims, either privately or publicly [7].
- Trolling: includes intentionally posting comments that are in disagreement with others. The poster intends to elicit negative discussions or feelings [6].
- Harassment: a form of conversation wherein the bully constantly sends insulting messages to users [7].
- Denigration: also known as dissing. It occurs when an electronic bully promotes gossip or false claims about a person to damage the user's reputation or friendships [7].
- Outing: occurs when a user posts private or disruptive information about a victim on public social media. Furthermore, the relationship between the bully and the victim is generally narrow [7].
- Exclusion: involves the intentional removal of a person from a social community. This type of bullying usually occurs among young people and adolescents [6].

B. Impact of Cyberbullying

With the advances and widespread use of technology, many ethical issues have proliferated. Social networks have been used by a huge number of users, especially in the Arabic world and in the Middle East. Recent research realized in 2020 has shown very interesting results [8]. The percentage of young Arabs using social media has increased, from 25% in 2015 to 79% in 2020 [8]. In 2020, Egypt possessed the ninth largest national market of Facebook, with 44 million users, while Saudi Arabia had the eighth highest Twitter presence, reaching a high position in relation to the world. The same study showed that the users spending the most time on social networks are from the Middle East and Africa, with an average of three hours and a half per day. Unfortunately, this hyper-connectivity leads to several risks, including cyberbullying, which has become a concerning issue. A recent study performed between July 2019 and May 2020 in the U.S. showed that persons aged between 10 and 16 who are subject to cyberbullying or violence are exposed to suicidal thoughts at an increased rate of 50% [9]. In United States schools, research conducted by the Cyberbullying Research Group showed that more than 36% of students, aged between 12 and 17, were absent from school due to experiencing cyberbullying [10]. Cyberbullying can have negative effects on a victim's mental health and sense of self-worth. These cyberbullying statistics demonstrate the severity of bullying's detrimental effects and the urgent necessity to take decisive action to address it [11]. It is necessary for the target population's general health. Cyberbullied individuals experience social anxiety in 41% of cases, depression in 37%, and even suicidal thoughts in 26% of cases, as shown in Figure 1.

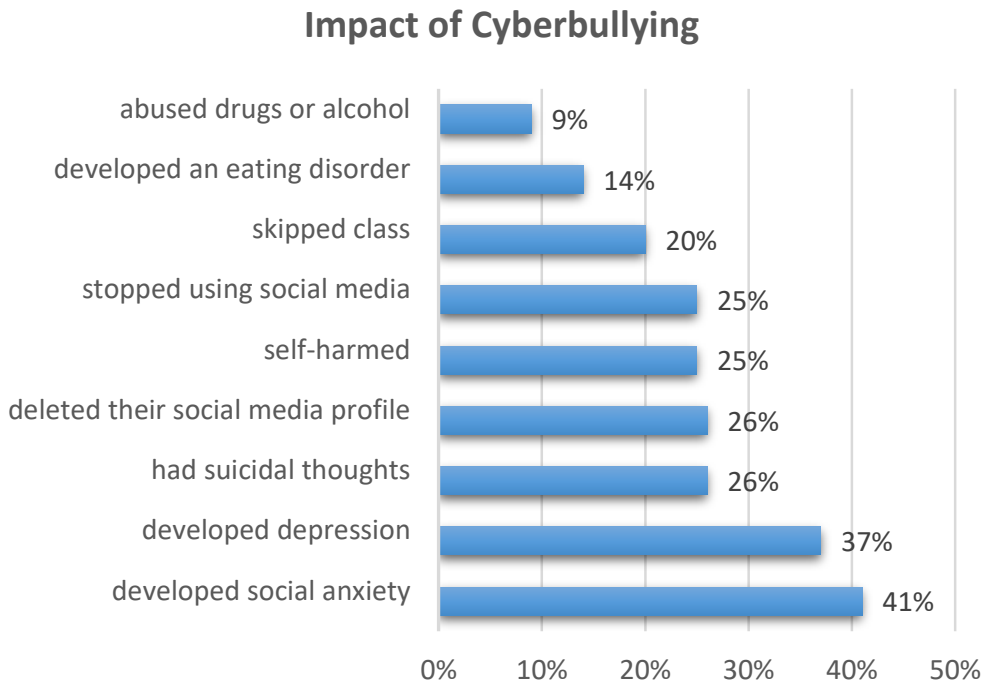

Figure 1. Examples of negative effects of cyberbullying.

On a local scale, a survey was conducted in Saudi Arabia in 2017 by the National Family Safety Program, entitled "*Relationship between Childhood Bullying and Anti-Social Behaviors among Adults*". Overall, 39% of the sample group, including 10156 participants, aged more than 18 years, indicated that they had been bullied at least once in their lives [12].

As a result, adverse cyberbullying consequences have been identified in Saudi Arabia. Therefore, his Royal Highness Prince Mohammed bin Salman bin Abdulaziz adopted, in 2020, child protection measures in cyberspace by launching projects to lead efforts related to developing the best initiatives, procedures, and practices to protect children from the growing number of cyber threats that target them while they are online [13].

C. Anti-Cyberbullying Legislation

If adequate proof is provided, the law may be helpful in reducing cyberbullying. A first step in addressing the issue is to collect evidence of cyberbullying, advising parents to collect evidence in the case of child cyberbullying. Parents should print out comments, photographs, and any other items that can be treated as direct evidence of cyberbullying or take screenshots. The proof should be reported to appropriate agencies, such as service providers, police, and civil rights officers. In Gulf Cooperation Council (GCC) countries, cyberbullying laws exist as well. For example, in the United Arab Emirates (UAE), uploading images of other people on online media without permission from the owner could result in a jail sentence of 6 months or a penalty under the UAE IT or Cybercrime Act of a half-million Dirhams (AED 500,000) [6].

In Saudi Arabia, Anti-Cyber Crime decreed that anyone who harms another individual using digital platforms would be criminally liable and would be subjected to a jail sentence of not more than a year or would be expected to pay a penalty of up to SR 500,000, or both. Under the current anti-cybercrime law, an individual who invades the privacy of others by abusively using their phone camera or threatening, cyberstalking, or bullying others via social media can be severely punished [6].

2.2. Machine Learning

As is known, machine learning (ML) is a subfield of artificial intelligence (AI) that offers systems with the capability of learning and improving with the involvement of automation processes from previous experience and without having to be specifically programmed. ML can be explained as the capability of a system to learn for itself to make decisions instantly by applying a training dataset [14]. It is necessary for activities and tasks that are quite complex for a human, particularly working with codes, such as in the case of cyberbullying detection. Note that there are two ML approaches: supervised and unsupervised ML.

In supervised learning algorithms, the training dataset has class labels to build a model that can then be used to predict unlabeled data [14]. Decision Tree, Naïve Bayes (NB), K-Nearest Neighbors, and Support Vector Machine (SVM) are examples of classification algorithms that show reasonable accuracy and performance [14]. SVM is a binary classifier that assumes that data samples are clearly differentiated. It seeks to find the best hyperplane that maximizes the difference between the classes' margins [5]. Recall that our proposal, discussed in Section 4, will use SVM.

However, the unsupervised learning algorithms use unlabeled training datasets. The data do not include predicted classes and the system tries to build patterns between the data and group similar data together.

2.3. Natural Language Processing

Natural Learning Processing (NLP) is a field of computer science that intends to facilitate communication between machines and human beings. The main idea behind it is to create an automated environment to understand human language and the meanings of utterances. NLP is very significant as it has a major impact on our daily lives [15,16].

To the best of our knowledge, Arabic is the fourth most-used language on the Internet, with 400 million speakers in 22 different countries [15,16]. Classical Arabic is the language form used in literary texts, documents, and the Quran. Modern Standard Arabic originated from Classical Arabic, which is used for formal writing and conversations. Arabic includes many dialects used for informal writing and daily conversations. Dialects have no standard or codified form and are influenced by region-specific slang [15,16].

The Arabic language has a complex nature because of its rich morphology and different dialects. In addition, Arabic is a rich language written from right to left without capitalization, and each character's shape will change according to its position in the word. However, NLP applications are required to handle these complex problems, which are relevant to the nature and structure of the Arabic language [15,16].

3. Previous Work

In this section, we deliver a literature review of different classifications for diverse perspectives, which provides us with a clear vision of the researchers' latest findings and the success rate of each study in this area of research with their most accurate results in bullying classification.

Mouheb et al. [2] suggested a useful model for recognizing cyberbullying on Arabic platforms using the NB classifier. The dataset was extracted from the most popular social sites, such as Twitter and YouTube APIs, which included 26,000 Arabic comments. For the training dataset, the authors used the frequency of the most common words that dealt with cyberbullying and labeled the training dataset to measure the probabilities by the NB theoretical approach to run the classifier. The accuracy of the extracted Arabic comments obtained by the NB classification was 0.95.

Dalvi et al. [3] described a proposal to select a machine learning algorithm to detect and avoid cyber-security threats on Twitter. Both the SVM and NB classifier models were applied to train and test social media comments. The SVM and NB classifiers were effective in identifying true positive instances, with the accuracy of 71.25% and 52.70%, respectively.

Haidar et al. [5] introduced a multilingual cyberbullying identification method based on machine learning and Natural Language Processing methods. They collected datasets from Facebook and Twitter. The authors used n-gram attributes and tested a variety of classifiers, including NB and SVM. The authors showed that SVM obtained much higher precision. In a newer study, Haider et al. [7] expanded their previous research and introduced a model to identify and prevent bullying words in Arabic text by using ensemble machine learning.

Muneer et al. suggested an automated cyberbullying detection scheme based on the collection of a generic dataset of over 35,000 different tweets [17]. Furthermore, seven classifiers were used, including NB, SVM, AdaBoost (ADB), Stochastic Gradient Descent (SGD), Light Gradient Boosting Machine (LGBM), Logistic Regression (LR), and Random Forest (RF). The experimental results showed that LR provides the best results, with a median accuracy of around 90.57%. However, this study only focused on a few feature extraction methods.

Nandakumar et al. used the NB classifier algorithm and the SVM model applied to a Twitter dataset [4]. The terms were identified and filtered as part of the classification. Then, the classifier broke down each tweet into single words before calculating the probabilities and making a decision. The results showed that the NB classifier outperformed the SVM model in terms of precision. Note that these results were obtained in a text data classification context.

In [18], Mubarak and Darwish utilized the keyword list as a resource to construct a training dataset that they then applied to test and build an abusive word detector. The final result represents a character n-gram and applies a deep learning classifier to obtain a 90% F1 score.

In an attempt to increase the accuracy ratio of the detection of offensive words in Arabic, Alakrot et al. [19] provided a method to detect negative terms. The authors used an SVM classifier with n-gram features for the prediction, where they obtained an F1 measure of 0.82.

Following the previous studies, Mouheb et al. [2] used an NB classifier and NLP to clean and normalize text in the Arabic context; as a result, the accuracy of the extracted Arabic comments obtained by the NB classification was 0.95. In addition, Dalvi et al. [3] tried to measure the probabilities of retrieving tweets, and NB showed an accuracy percentage of 52.70, whereas the SVM resulted in 71.25% accuracy. Moreover, in [5], the authors used the SentiStrength mechanism on Arabic tweets with chosen classifications, and the authors found that SVM obtained 93.4% accuracy.

In fact, the SVM classifier performs well in different mechanisms and languages. Hence, the results reveal that for Arabic bullying detection, SVM outperforms NB on most tested datasets and achieves the best result. Overall, these studies highlight the fact that

bullying detection has not been investigated deeply and that it is relatively a new research area. In view of all that has been mentioned so far, in this paper, we aim to address this issue and optimize the detection of cyberbullying in Arabic with high accuracy. This framework has shown moderate accuracy results; however, there is still room for improvement as it requires further research on the treatment of Arabic content and the use of weighted equations or word frequency analysis to classify cyberbullying comments according to their strength, which will enhance the performance measures for the Arabic language.

4. Proposed System and Methodology

In our methodology, we have developed a machine learning model to detect Arabic cyberbullying tweets using the supervised classifier SVM, as shown in Figure 2. The methodology involves the following steps: (1) data collection, (2) pre-processing, (3) extraction of different scenarios, (4) feature extraction, (5) classification (SVM), and (6) evaluation metrics.

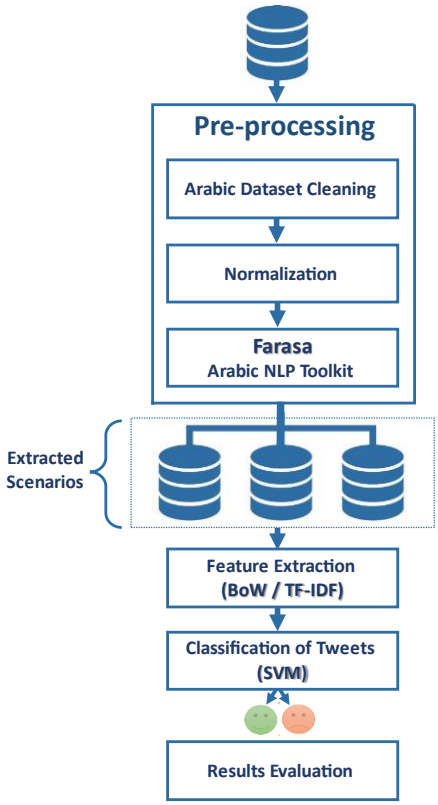

Figure 2. Cyberbullying detection in Arabic using SVM.

A. Data Collection

In order to prepare our model for training, the dataset was collected on 5 February 2021 from the Twitter and YouTube APIs, which included 30,000 Arabic comments (kaggle.com/datasets/alanoudaldealij/arabic-cyberbullying-tweets) (accessed on 18 December 2022). The training dataset was labeled as bullying and non-bullying posts depending on the most common and frequent bullying keywords in Arabic society, which were already collected manually from the frequent words in the posts. After this, we extracted an extensive list of bullying words to be used for the classification of the bullying and non-bullying classes.

B. Data Pre-Processing

The Natural Language Toolkit (NLTK) is useful for data pre-processing and tokenizing words to split a large sample of text into individual words. Then, before feeding the data into the model to be built, it must be cleaned and normalized as follows.

- **Data Cleaning**: This can be performed by removing URLs, hashtags, "@" mentions, numbers, non-Arabic words, and any irrelevant parts of the collected tweets.

In addition, each word is refined through a series of steps. The two proposed methods for refinement are

- Deleting letter repetitions of words;
- Removing stopwords to return the words as text in case they do not belong to the stopword list.

However, the Arabic language has a complex nature and structure and many grammatical forms; each word can be available in different patterns but convey the same meaning, which is mainly due to the rich morphology of Arabic. To handle this issue, we proceeded as follows.

- **Normalization:** During this process, the words are normalized by eliminating any potential letter misinterpretations. Then, we substitute some Arabic letters with their official form, due to the common misspelling of some words—for example, changing "أ", "آ", and "إ" to "ا".
- **Farasa (https://farasa-api.qcri.org/lemmatization/) (accessed on 15 February 2021) Arabic NLP Toolkit:** This is an interesting text processing toolkit for Arabic text. It outperforms many segmenters related to Arabic, such as Stanford and MADAMIRA [20]. Additionally, Farasa is built on ranks, as per the SVM analysis with linear-based kernels, proposing a variety of lexicons, features, and vocabulary to rank possible word segmentations. Furthermore, Farasa offers different features, such as segmentation, stemming, spellchecking, a Named Entity Recognizer (NER), Part of Speech tagging (POS tagging), and diacritization. In our model, we use stemming and segmentation features as follows.
  1. **Farasa Stemming**: Farasa provides data stemming, because words have different structures, especially in the Arabic language (e.g., "حمار", "حمارة", "يالحمار", and "حمير"). To correct this flaw, stemming is done before classification. The method of stemming is the reduction of inflected terms to their root format. The number of attributes is reduced when stemming is used. Suffixes, infixes, and prefixes are all removed. For example, the word "حمارة" contains the prefix "ة" at the end, the word "يالحمار" contains the suffix "يا" at the start of the term, and the word "حمير" contains infix "ي" in the middle of the term. All of the terms are grouped together by using stemming under one keyword "حمار", which is the root of these words.
  2. **Farasa Segmentation:** Farasa provides a data segmentation technique by splitting a sentence into smaller sections, known as segments. It is commonly used in the field of text processing. Each section has its own meaning. For example, applying segmentation to sentence "مجنونة وحمارة مافي مثلها" results in "حمار+ة ما+في مثل+ها مجنون+ة و+". Thus, this tool could increase the detection of bullying words.

A. Extraction of Different Scenarios

After the pre-processing step, we extracted six different scenarios from our dataset, which are as follows.

1. **Cleaned_ normalize without Farasa tool:** This is a pre-processed dataset obtained by removing URLs, hashtags, "@" mentions, numbers, and non-Arabic words, and also deleting letter repetitions of the word as well as the single characters. Moreover, we apply normalization to this scenario.
2. **Stemmed_ normalize with FarasaStemmer tool:** This is a dataset cleaned with FarasaStemmer via the reduction of inflected terms to their root forms.
3. **Segmented_normalize with FarasaSegmenter tool:** This is a dataset cleaned with FarasaSegmenter by segmenting sentences into individual words.
4. **Cleaned_normalize_no_stopwords:** This is a dataset cleaned by removing stopwords.

5. **Stemmed_ normalize_no_stopwords**: This is a dataset cleaned with FarasaStemmer by removing stopwords.
6. **Segmented_normalize_no_stopwords**: This is a dataset cleaned with FarasaSegmenter by removing stopwords.

Note that the total count of stopwords is 243, so our model can often be improved if we do not take these words into account. After this, we insert these scenarios into an SVM classifier, to define which scenario gives the best result and the highest accuracy.

B. Feature Extraction

In this process, the defined model converts the data into another format in such a way that the machine learning algorithms can be applied. To obtain the most significant features from the given dataset, the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer and Bag of Words (BoW) algorithms are the ideal choices. The most significant features of data are mainly extracted and arranged as a featured list. TF-IDF is the most common weight to restore data based on the text mining approach. This weight is a specific technique to assess the values of the collected words through statistical analysis in a corpus of documents [21]. We used various TF-IDF analyzers in this study, including Unigram and Bigram.

Typically, the normalized Term Frequency (TF) is usually used to calculate the weight. It is equal to the frequency of a term (t) that appears in a document (d), divided by the total number of words in the document.

$$TF_{t,d} = \frac{n_{t,d}}{\sum_k n_{t,d}} \tag{1}$$

Inverse Document Frequency (IDF) is computed as the log of the ratio of number of documents divided by the total number of documents that include the term (t). The weight of rare terms is defined by the IDF among all documents.

$$IDF(wt) = \log(\frac{N}{df_t}) \tag{2}$$

Lastly, to compute the TF-IDF weight of the term (wt) for all the terms in the corpus, we simply multiply the TF by IDF as follows:

$$TF - IDF(wt) = TF_{t,d} \times \log(\frac{N}{df_t}) \tag{3}$$

BoW is a representation that counts how many times each word appears in the text to generate fixed-length vectors. Each tweet is provided as data input, and the number of repetitions of the term individually in the tweet is calculated, resulting in a presentation based on the numerical value of the term, known as a vector [22].

A. Classification

After feature extraction, we divide the collected dataset into training and testing sets at random in a 70:30 ratio. The testing dataset is used to validate the model in the final stage. The training data will be inserted into the SVM classifier for training the model. SVM is a machine learning algorithm for supervised classification. SVM can be used for classification as well as regression [3]. The probabilities for each group are also determined by using LinearSVC [3].

LinearSVC is similar to Support Vector Classifier (SVC) with the "linear"-based kernel parameter, but it is implemented as lib-linear instead of libsvm and provides flexibility in penalties and miss functions, as well as the ability to handle massive quantities of data [3]. This class also works on both sparse and dense data, and multiclass support is carried out by involving the on-vs-the-rest tactic. In addition, LinearSVC is another faster implementation of SVC for the case of a linear kernel.

B. Evaluation Metrics

In order to evaluate the performance of our proposal, we assess the results of classification [3]. The accuracy (A), recall (R), F1 score (F), and precision (P) are also measured by the following formulas:

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$F = 2 \cdot \frac{P \cdot R}{P + R} \tag{6}$$

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where:

- **True positives (TP)** refer to correct classification as positive instances.
- **True negatives (TN)** refer to correct classification as negative instances.
- **False positives (FP)** refer to incorrect classification where the result is the predicted class "yes" but the actual class is "no".
- **False negatives (FN)** refer to incorrect classification where the output is the predicted class "no" but the exact class is "yes".

5. Results

Based on the experiments' results, we show that the SVM model with the Farasa NLTK achieves better results in detecting Arabic cyberbullying tweets compared to the NB classifier. As a baseline, we compared the performance in different experiments with several scenarios that involve the Farasa NLP Toolkit via the SVM classifier, which is considered an effective classifier for classification and regression tasks for text mining. However, the conducted experiments' analysis indicated that the recently built SVM with the Farasa Stemmer tool obtained good state-of-the-art results on the Arabic cyberbullying tweets that were tested. In order to analyze the results and draw conclusions from the most accurate models related to the different scenarios, it is necessary to consider all the results from all the experiments and analyze them.

Firstly, to achieve our experiments, we applied different testing sizes to determine which is the suitable size to provide the best results for our model. Then, the results of using the SVM algorithm with the TF-IDF vectorizer are compared with those using the different testing sizes of our dataset to obtain the best accuracy and SVM using the recall, precision, and F1 score parameters. As shown in Table 1, the Stemmed_no_stopwords scenario provides accuracy of 95.742% when the testing ratio is 0.3, which is very close to the Segmented_no_stopwords scenario with accuracy of 95.516%. However, when the testing ratio is 0.7, the results become less accurate. Note that a ratio of 70:30 provides the best results.

Table 1. Summary of accuracy results with different testing ratios.

Scenario	0.3	0.5	0.7
Cleaned	94.766%	94.297%	93.956%
Stemmed	95.593%	94.907%	94.508%
Segmented	95.417%	94.817%	94.363%
Cleaned _no_stopwords	94.875%	94.37%	93.939%
Stemmed_no_stopwords	95.742%	94.915%	94.555%
Segmented_no_stopwords	95.516%	94.809%	94.427%

After selecting a ratio of 70:30, we have two categories with which to analyze our dataset using the SVM classifier, and then we stem it down with the TF-IDF and BoW

techniques. The BoW vectorizer is a convenient tool to extract significant features from documents. BoW transforms text-based data into matrix format by evaluating the frequency of words within the given document file.

The results obtained show that Stemmed_no_stopwords provides the best result with the SVM model using the TF-IDF vectorizer, with 95.742% accuracy. We also note that SVM with the BoW vectorizer shows interesting performance, with accuracy of 95.661%. As shown in Table 2, the best accuracy is provided by the Stemmed_no_stopwords scenario when using NLTK. In fact, the stemming tool reduces the inflected terms to their root forms. Suffixes, infixes, and prefixes are all removed to increase the detection of bullying words. Therefore, we can conclude that Farasa has increased the accuracy, outperforming the state-of-the-art for Arabic data stemming, because the words have different structures, especially in the Arabic language. Meanwhile, the model achieves better results when removing stopwords.

Table 2. Summary of best accuracy results of extracted scenarios.

Best Scenario Based on Accuracy {High to Low}	TF-IDF Ngram_Range (1, 2)	BoW Ngram_Range (1, 2)
Stemmed {no_ stopwords}	95.742%	95.661%
Stemmed	95.593%	95.227%
Segmented {no_ stopwords}	95.516%	95.186%
Segmented	95.417%	94.956%
Cleaned {no_ stopwords}	94.875%	94.644%
Cleaned	94.766%	94.522%

The confusion matrix for the previous experiments using the SVM classifier with the TF-IDF vectorizer displays the number of bullying ("1") and normal ("0") predictions, which can be compared to the actual number in the dataset. We show that the Stemmed_no_stopwords scenario is the most effective at rating true negative and true positive results. Indeed, approximately 7.28% of tweets were accurately categorized as cyberbullying out of 537 bullying tweets, while approximately 88.46% of tweets were not categorized accurately as cyberbullying out of 6524 non-bullying tweets, as shown in Table 3. However, the Stemmed_no_stopwords scenario is more effective at rating true negative and true positive results in the SVM model.

Table 3. The confusion matrix of the SVM classifier.

Scenario	TN	TP	FN	FP
Cleaned	6517	472	306	80
	88.37%	6.40%	4.15%	1.08%
Stemmed	6531	519	259	66
	88.56%	7.04%	3.51%	0.89%
Segmented	6513	524	254	84
	88.31%	7.11%	3.44%	1.14%
Cleaned {no_ stopwords}	6522	475	303	75
	88.43%	6.44%	4.11%	1.02%
Stemmed {no_ stopwords}	6524	537	241	73
	88.46%	7.28%	3.27%	0.99%
Segmented {no_ stopwords}	6514	523	255	83
	88.33%	7.09%	3.46%	1.13%

Recall that the standard confusion matrix can be applied to determine the recall and precision, as well as obtain an F1 score for the extracted scenarios. Table 4 demonstrates the performance of this approach in terms of these characteristics. Depending on the scenario, Stemmed_no_stopwords using the Farasa Stemmer achieves the highest accuracy, indicating that it is best suited for Arabic cyberbullying detection. Figure 3 shows the accuracy of SVM with the TF-IDF vectorizer for all scenarios.

Table 4. Performance of extracted scenarios.

Scenario	Class	Precision	Recall	F1 Score
Cleaned	No (0)	0.96	0.99	0.97
	Yes (1)	0.86	0.61	0.71
	Overall	0.91	0.80	0.84
Stemmed	No (0)	0.96	0.99	0.98
	Yes (1)	0.89	0.67	0.76
	Overall	0.92	0.83	0.87
Segmented	No (0)	0.96	0.99	0.97
	Yes (1)	0.86	0.67	0.76
	Overall	0.91	0.83	0.87
Cleaned {no_ stopwords}	No (0)	0.96	0.99	0.97
	Yes (1)	0.86	0.61	0.72
	Overall	0.91	0.80	0.84
Stemmed {no_ stopwords}	No (0)	0.96	0.99	0.98
	Yes (1)	0.88	0.69	0.77
	Overall	0.92	0.84	0.88
Segmented {no_ stopwords}	No (0)	0.96	0.99	0.97
	Yes (1)	0.86	0.67	0.76
	Overall	0.91	0.83	0.87

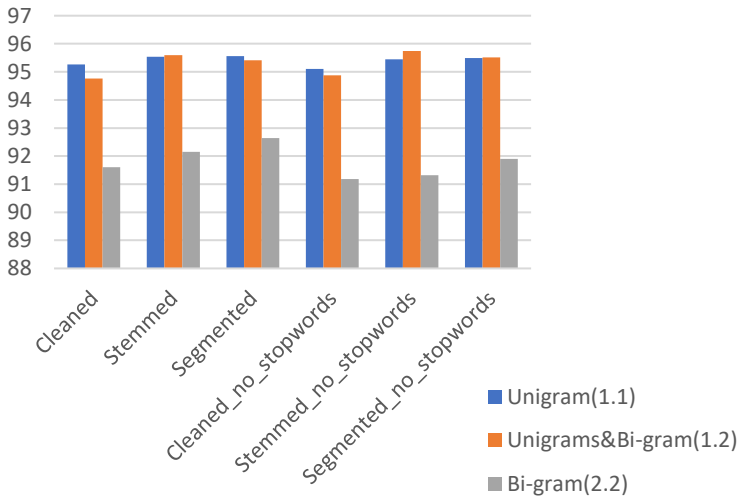

Figure 3. Accuracy results of extracted scenarios by SVM classifier.

Furthermore, the three variants of TF-IDF unigram, bigram, and trigram for their relevance as a classification function showed that the TF-IDF n-gram range of (1, 2) performed better than TF-IDF unigram/bigram and obtained the best accuracy with the SVM classifier, as summarized in Figure 4. However, these results may be inconclusive as larger datasets can affect the results, and, in this case, it is necessary to check whether bigram and unigram perform better on these larger datasets in future work.

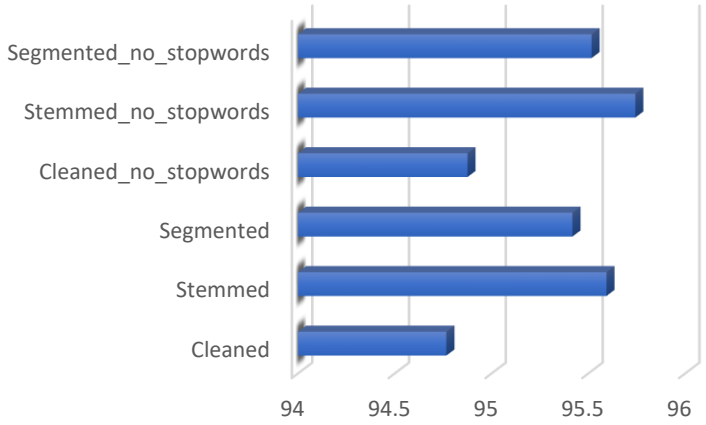

Figure 4. Performance evaluation of SVM with different TF-IDF analyzers.

Accuracy can be further improved by using balanced data, where the training samples for positive and negative comments are nearly the same. Comparison of the performance of similar models shows that the SVM model achieves better performance and a higher resolution with TF-IDF (unigram and bigram) range (1, 2), where Stemmed_no_stopwords scored a total of 95.742% for accuracy. As a result, we noticed that the stemming tool converts all different structures of words into their root forms and this helps the classifier to identify the bullying words and assign them to the correct class. Moreover, the model achieved better results when these stopwords were not taken into account, to reduce obfuscation in Arabic sentences.

As a synthesis, the results show that SVM outperforms NB in terms of classification accuracy. In this study, recall that we have classified Arabic cyberbullying tweets using two machine learning techniques: NB and the SVM classifier. Then, we have shown how the SVM classifier provides the best results in predicting cyberbullying comments with the TF-IDF vectorizer. From previous works, the highest accuracy was obtained by Mouheb et al. [2], who used the NB classifier and obtained 0.95; as compared to these results, we enhanced the performance with our mechanism by including the Farasa Stemmer, and the score is 95.742% for accuracy.

The accuracy of both the NB and SVM classifiers for different scenarios is shown in Table 5. SVM provides the best accuracy with a percentage of 95.742%, while NB obtains accuracy of 70.942% in the Segmented scenario. Note that this result is the exact opposite of the SVM classifier, with Stemmed_no_stopwords having the highest results. Therefore, a combined approach of classification can be the best solution when we need to consider different contexts (scenarios). Moreover, it is necessary to check whether the results can be improved with additional data processing and classifier modifications. Finally, training the model with more data over larger batches should be considered.

Table 5. Classifiers' accuracy for extracted scenarios.

Scenario	SVM	NB
Cleaned	94.766%	70.915%
Stemmed	95.593%	70.807%
Segmented	95.417%	**70.942%**
Cleaned {no_ stopwords}	94.875%	69.79%
Stemmed {no_ stopwords}	**95.742%**	68.773%
Segmented {no_ stopwords}	95.516%	69.79%

6. Conclusions

In our work, we trained the SVM model with a large-scale Arabic dataset containing approximately 30,000 comments. We then tested the SVM model on a different Twitter dataset due to the popularity of this application as a platform for collecting text data in order to classify cyberbullying comments. We showed that the performance of SVM with the TF-IDF vectorizer using Farasa NLTK achieved the best cyberbullying classification. The results obtained were then compared to those of the NB classifier with different parameters of ngram range and with additional feature extraction, such as BoW. Recall that BoW counts how many times each word appears in the text to generate fixed-length vectors by using CountVectorizer. The results showed that SVM still outperformed NB in detecting cyberbullying content, with a percentage of 95.742%. The high accuracy of our model will help to protect users from the actions of social network bullies.

For future work, we recommend testing our model with millions or even more daily posts on social media applications. Additionally, the model could be embedded in texting applications, so that users could increase their awareness of bullying in social networks and delete any bullying comments automatically. Moreover, we can integrate a new version with law enforcement and the "كلنا أمن" application to track and solve serious cases of bullying, particularly those that result in suicide or death.

Further future goals include training our model using deep learning methods rather than machine learning, and the results from the two training methods can be compared.

Author Contributions: Conceptualization, A.M.A. and A.B.; methodology, A.M.A. and A.B.; software, A.M.A.; validation, A.B.; formal analysis, A.M.A. and A.B.; resources, A.M.A.; writing—original draft preparation, A.M.A.; writing—review and editing, A.B.; visualization, A.M.A. and A.B.; supervision, A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data set used in this paper is available on: kaggle.com/datasets/alanoudaldealij/arabic-cyberbullying-tweets (accessed on 15 January 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Al-Ajlan, M.A.; Ykhlef, M. Optimized Twitter Cyberbullying Detection based on Deep Learning. In Proceedings of the 2018 21st Saudi Computer Society National Computer Conference, Four Seasons, Saudi Arabia, 25–26 April 2018; Available online: https://ieeexplore.ieee.org/abstract/document/8593146 (accessed on 15 January 2021). [CrossRef]
2. Mouheb, D.; Albarghash, R.; Mowakeh, M.F.; Aghbari, Z.A.; Kamel, I. Detection of Arabic Cyberbullying on Social Networks Using Machine Learning. In Proceedings of the 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, 3–7 November 2019. [CrossRef]
3. Dalvi, R.R.; Baliram Chavan, S.; Halbe, A. Detecting A Twitter Cyberbullying Using Machine Learning. In Proceedings of the 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 13–15 May 2020; Available online: https://ieeexplore.ieee.org/document/9120893 (accessed on 23 January 2021). [CrossRef]
4. Nandakumar, V. Cyberbullying Revelation In Twitter Data Using Naïve Bayes Classifier Algorithm. *Int. J. Adv. Res. Comput. Sci.* **2018**, *9*, 510–513. [CrossRef]
5. Haidar, B.; Chamoun, M.; Serhrouchni, A. Multilingual Cyberbullying Detection System: Detecting Cyberbullying in Arabic Content. In Proceedings of the 2017 1st Cyber Security in Networking Conference (CSNet), Rio de Janeiro, Brazil, 18–20 October 2017. [CrossRef]
6. Alduailej, A.H.; Khan, M.B. The challenge of cyberbullying and its automatic detection in Arabic text. In Proceedings of the 2017 International Conference on Computer and Applications (ICCA), Dubai, United Arab Emirates, 6–7 September 2017; Available online: https://ieeexplore.ieee.org/abstract/document/8079791/ (accessed on 2 February 2021). [CrossRef]
7. Haidar, B.; Chamoun, M.; Serhrouchni, A. Arabic Cyberbullying Detection: Enhancing Performance by Using Ensemble Machine Learning. In Proceedings of the 2019 International Conference on Internet of Things (ithings) and IEEE Green Computing and Communications (Greencom) and IEEE Cyber, Physical and Social Computing (Cpscom) and IEEE Smart Data (Smartdata), Atlanta, GA, USA, 14–17 July 2019. [CrossRef]
8. Radcliffe, D.; Abuhmaid, H. How the Middle East Used Social Media in 2020. Ninth Annual Series of Pablications, SSRN Electron, April 2021. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3826011 (accessed on 2 February 2021).

9. Sumner, S.A.; Ferguson, B.; Bason, B.; Dink, J.; Yard, E.; Hertz, M.; Hilkert, B.; Holland, K.; Mercado-Crespo, M.; Tang, S.; et al. Association of Online Risk Factors with Subsequent Youth Suicide-Related Behaviors in the US. *JAMA Netw. Open* **2021**, *4*, e2125860. [CrossRef]
10. Brown, V.; Clery, E.; Ferguson, C. Estimating the Prevalence of Young People Absent from School due to Bullying. *Natl. Cent. Soc. Res.* **2011**, *1*, 1–61.
11. Label, D. Cyberbullying Statistics. Available online: https://www.ditchthelabel.org/cyber-bullying-statistics-what-they-tell-us (accessed on 7 July 2021).
12. National Project for Cyberbullying Control. Available online: https://nfsp.org.sa/ar/Pages/default.aspx (accessed on 20 April 2021).
13. NCA. Governor of Riyadh Region Announces Two Global Cybersecurity Initiatives, Sponsored by Crown Prince Mohammed Bin Salman. Available online: https://nca.gov.sa/news?item=53 (accessed on 18 May 2021).
14. Kulkarni, P. *Reinforcement and Systemic Machine Learning for Decision Making*; Wiley-IEEE Press: Hoboken, NJ, USA, 2012.
15. Shaalan, K.; Siddiqui, S.; Alkhatib, M.; Abdel Monem, A. Challenges in Arabic Natural Language Processing. In *Computational Linguistics, Speech and Image Processing for Arabic Language*; WSPC: Singpore, 2018; pp. 59–83. [CrossRef]
16. Alduailej, A.; Alothaim, A. AraXLNet: Pre-Trained Language Model for Sentiment Analysis of Arabic. *J. Big Data* **2022**, *9*, 1–21. [CrossRef]
17. Muneer, A.; Fati, S.M. A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. *Future Internet* **2020**, *12*, 187. [CrossRef]
18. Mubarak, H.; Darwish, K. *Arabic Offensive Language Classification on Twitter*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 269–276.
19. Alakrot, A.; Murray, L.; Nikolov, N.S. Towards Accurate Detection of Offensive Language in Online Communication in Arabic. *Procedia Comput. Sci.* **2018**, *142*, 315–320. [CrossRef]
20. Abdelali, A.; Darwish, K.; Durrani, N.; Mubarak, H. Farasa: A Fast and Furious Segmenter for Arabic. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, San Diego, CA, USA, 12–17 June 2016. [CrossRef]
21. Parisi, A. *Hands-on Artificial Intelligence for Cybersecurity: Implement Smart AI Systems for Preventing Cyber Attacks and Detecting Threats and Network Anomalies*; Packt Publishing Ltd.: Birmingham, UK, 2019.
22. Alam, K.S.; Bhowmik, S.; Prosun, P.R.K. Cyberbullying Detection: An Ensemble Based Machine Learning Approach. In Proceedings of the 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 4–6 February 2021; pp. 710–715. [CrossRef]