Computer Systems Science & Engineering
DOI:10.32604/csse.2022.020023
Article

Tech Science Press

Consensus-Based Ensemble Model for Arabic Cyberbullying Detection

Asma A. Alhashmi* **and Abdulbasit A. Darem**

Northern Border University, Arar, 9280, Saudi Arabia
*Corresponding Author: Asma A. Alhashmi. Email: asma.alhashmi@nbu.edu.sa
Received: 06 May 2021; Accepted: 30 June 2021

Abstract: Due to the proliferation of internet-enabled smartphones, many people, particularly young people in Arabic society, have widely adopted social media platforms as a primary means of communication, interaction and friendship making. The technological advances in smartphones and communication have enabled young people to keep in touch and form huge social networks from all over the world. However, such networks expose young people to cyberbullying and offensive content that puts their safety and emotional well-being at serious risk. Although, many solutions have been proposed to automatically detect cyberbullying, most of the existing solutions have been designed for English speaking consumers. The morphologically rich languages-such as the Arabic language-lead to data sparsity problems. Thus, render solutions developed for another language are ineffective once applied to the Arabic language content. To this end, this study focuses on improving the efficacy of the existing cyberbullying detection models for Arabic content by designing and developing a Consensus-based Ensemble Cyberbullying Detection Model. A diverse set of heterogeneous classifiers from the traditional machine and deep learning technique have been trained using Arabic cyberbullying labeled dataset collected from five different platforms. The outputs of the selected classifiers are combined using consensus-based decision-making in which the F1-Score of each classifier was used to rank the classifiers. Then, the Sigmoid function, which can reproduce human-like decision making, is used to infer the final decision. The outcomes show the efficacy of the proposed model comparing to the other studied classifiers. The overall improvement gained by the proposed model reaches 1.3% comparing with the best trained classifier. Besides its effectiveness for Arabic language content, the proposed model can be generalized to improve cyberbullying detection in other languages.

Keywords: Consensus; cyberbullying detection; arabic language; offensive contents; ensemble learning; deep learning

1 Introduction

The vast proliferation of social networks has a substantial impact on individuals and communities. People become more connected and individually wired to the huge networks of relatives, friends, followers, and other people in common. Social media bring people to get to know each other more closely, and improve individual communication skills, language, writing, communication with someone's ideas, talents, and experiences. However, some negative users misuse this digital world by spreading offensive content to embarrass or harm other users. Such offensive content is known as cyberbullying. Cyberbullying can be described as the use of digital means, such as smart electronic devices that are connected to the internet, to post content intending to harm or embarrass others [1,2]. Cyberbullying can cause depression, emotional and physical stress, destruction of self-esteem, and social isolation among many others. According to Cyberbullying Research Center 2019, 36% of children aged between 12–17 years old in the US are cyberbullied monthly [3]. In a study published by Florida Atlantic University, 70% of cyberbullying cases are underreported [4]. Google trend search shows that "cyberbullying" has received significant attention recently. The National Center for Health Statistics (NCHS) found that the increases in cyberbullying contributed to youth suicides. People under 25 years old who were exposed to cyberbullying are at risk twice as compared to other people [4,5]. Although most cyberbullying is attributed to social media, there are many other platforms where cyberbullying takes place, such as online gaming, cell phone services, websites, and other sharing platforms [6]. Moreover, cyberbullying can come in different formats, such as texts, images, videos, and audios. However, text-based cyberbullying is the most popular form used by misbehaving users. Due to the enormous amount, variety, and velocity of data and information generated by users, manual detection and/or relying on user reports are not an efficient or effective way to overthrow cyberbullying [2,7–12]. Accordingly, automatic cyberbullying detection is important to defend the cyber realm and protect communities and individuals.

Automatic detection of cyberbullying became the subject of many studies over the past few years [1,4,5,9,13,14]. However, cyberbullying has dramatically increased recently [2,15]. Different types of cyberbullying were investigated and many tools and techniques were used to analyze the text-content and detect the cyberbullying phenomenon [13]. Natural language processing (NLP) techniques were commonly used to process the text data to extract useful patterns for detecting cyberbullying [16]. Natural language-based data, such as text, usually unstructured. Thus, NLP techniques and algorithms were used to convert the text-based features to structured features suitable for machine learning algorithms. However, cyberbullying detection faces many challenges that need to be deeply investigated and solved. One of these challenges is that most of the existing cyberbullying detection systems were designed for the English language. Cyberbullying incidents were also reported in Arab countries. For example, 60% of youth in Arabic Gulf countries reported being victims of cyberbullying [17]. Arabic language, which is spoken by 330 million native people, has unique characteristics different from English and other languages [18]. The morphology of the Arabic language is rich with a high degree of affixation, roots in word stems, and interspersed vowel patterns [19]. The Arabic language can be divided into modern standard language, classical language, and informal language. Each of these types is used in different circumstances in the daily life of Arabic society [20,21]. For instance, Classical Arabic is used for religion-related content, such as prayers and speeches, while Modern Standard Arabic (MSA) is used for education, writing, and news reporting. Meanwhile, the informal language variety is the daily spoken language by people with their family and friends. Most of the social media and online communication is conducted using the informal Arabic variety. Each type of the Arabic language has its own lexicon that is different. People usually mix between those three types during informal writing and speaking [20,21]. These unique characteristics of the Arabic language render solutions developed for other languages ineffective once applied to the Arabic language.

Despite the effect of cyberbullying and offensive speech in social media on the Arabic society and the ineffectiveness of the solutions designed for other languages, few research studies were conducted focusing on cyberbullying detection in the Arabic content [22–25]. Most of the existing works directly adopt models designed for other languages, particularly English. However, due to the Arabic language's distinct characteristics, such models are ineffective and the models' detection accuracy is low. That is, due to the rich morphology of the Arabic language, the use of small dataset size used to construct the detection models leads to data sparseness problems which degrade the detection accuracy. Collecting huge datasets for Arabic cyberbullying is not a trivial task and needs substantial effort, time and cost. Alternatively, there is a need to improve the detection accuracy without increasing the size of the datasets.

To this end, this study focuses on improving the efficacy of the existing cyberbullying detection models for Arabic content by designing and developing a Consensus-Based Ensemble Cyberbullying Detection Model. To achieve this aim, four phase methodology is adopted. In the first phase, multiple datasets that contain cyberbullying were collected and combined. The second phase is the data preprocessing which includes normalization, noise cleaning, stemming, tokenization, and building corpus using NLP techniques. The third phase is the features extraction and representation phase in which unstructured text features will be converted to structured and numerical representation of the words using Term-Frequency and Invers Documents Frequency (TF-IDF). Then, multiple machine learning as well as Sequential Deep Learning (SDL) algorithms are trained using a dataset collected from different platforms. The best set of these trained classifiers has been used for constructing the proposed ensemble model. The F1-Score, a harmonic means of recall and precision, is used to rank the classifier outputs. These scored outputs are aggregated to make consensus-based decision making where each classifier contributes to the final decision. Finally, the aggregated value is loaded to the sigmoid function to resemble human-based decision-making. The results show that the proposed model is more effective than the other trained classifiers.

The rest of the paper is laid out as follows. The related work is presented in Section 2. The detailed methodology of this study is described in Section 3. Section 4 presents the results of the experiments and discussion. Section 5 outlines the conclusion and future work.

2 Related Work

In general, there are a few studies that have been done in the detection and analysis of cyberbullying, particularly for the Arabic language [17,22–28]. Ana Kovacevic [7] in her survey paper, studied the use of web content mining to detect cyberbullying. This paper examines various ideas for making the internet a safer place by employing a web content mining strategy for detecting and monitoring cyberbullying. Another survey was presented by Haidar et al. [17] for the techniques used for cyberbullying detection in Arabic content. There are many machine learning techniques that were used for designing the detection models. For example, Nearest Neighbor Estimators (NNE), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), and Deep Learning (DL). NLP techniques were also used for feature representations [29]. NLP techniques allow computers to comprehend human speech in its raw and unprocessed state. Most of the analyzed data used are from social media, such as Twitter [2,23,26,30], Facebook [7,23], Formspring [23], WhatsApp, Vine, Instagram [2,31], Ask.fm, YouTube [2,23] and Packet.

Nahar et al. [32] proposed a weighting scheme of feature selections as an efficient method for detecting cyberbullying messages from social media. It provided a graph model for extracting the cyberbullying network which was then used to use rating algorithms to classify the most violent predators and victims. Parime et al. [33] suggested a data mining and machine learning method to detect and prevent cyberbullying by detecting the existence or absence of cyberbullying using a dataset from popular social networking websites. It also goes into the social aspects of cyberbullying and how they can be addressed, as well as how the issue can be tackled in general. Using the fuzzy logic method, Sheeba et al. [34] suggested low-frequency keyword extraction with emotion classification and cyberbullying detection. Keywords are extracted from transcripts using the fuzzy logic method which defined three features for words: frequency estimation, noun extraction (using the Qtag tool), and clustering the words using the

C-means algorithm. It also uses fuzzy logic to detect explicit and implicit word expressions and also detect cyberbullying terms from transcripts.

Reynolds et al. [35] suggested a machine learning solution to detect cyberbullying. They developed rules to automatically detect cyberbullying content using a language-based (machine learning) system for identifying patterns used by bullies and victims. In a small sample of form spring results, they successfully classify 78.5 percent of posts that involved cyberbullying by tracking the percentage of curse and insult words inside the message or post. The study of Bosse et al. [36] focused on a Normative Agent System to avoid cyberbullying. A system of normative agents physically present in a simulated society is used in this approach. To detect norm breaches, the agents use a belief–desire-intention (BDI) model and a variety of tactics, including insult and following. They attempt to improve user behavior by using encouragement and punishment. The application has been introduced in a simulated world for children called Club Time Machine. According to the findings, agents have the ability to reduce the number of norm abuses in the long run. Hosseinmardi et al. [31] research focused on comparing popular Ask.fm and Instagram users to understand cyberbullying better. This study looked at users who used frequently two famous online social networking sites, Instagram and Ask.fm. The negativity and positivity of word use in posts by common users of these two social networks were analyzed. They looked at the posting behavior of famous user profiles and looked at how it correlated with negativity. However, all these works proposed for English languages don't meet the unique characteristics of the Arabic language. In semi-anonymous social networks, Mazari [37] attempted to explain cyberbullying activity. His research mostly focused on analyzing negative user activity on the Ask.fm social network, which was based on several incidents of cyberbullying that have resulted in suicidal behavior. They looked at the frequency of derogatory terms in Ask.fm "question-answers" profiles as well as a social network of "question-answers" and "likes". They also looked at the characteristics of users who cut in social networks.

There are few studies on cyberbullying detection in the Arabic language. Al-Ajlan et al. [30] abstracted a suggestion for automatic extraction of the features to design a cyberbullying detection model for the Arabic language using deep learning techniques. On a Twitter-based dataset, a Convolutional Neural Network (CNN) was trained. However, such an approach has not been evaluated and validated. Using NB techniques, a model for detecting cyberbullying in Arabic texts was proposed by Mouheb et al. [22]. The dataset used for the training and testing was collected from Twitter and Youtube. The reported accuracy is 92.5% out of 672 cyberbullying comments. However, the dataset used for the testing is relatively small (less than 10% as it is obtained from the confusion matrix of the testing data set). It contains obvious keywords that indicate cyberbullying. Mouheb et al. [26] proposed a design system for real-time cyberbullying detection in the Twitter stream. However, the system was designed based on a collection of keywords extracted from the dataset and weightage mechanisms. However, such a heuristic approach will result in high false alarms due to the lack of consideration of the semantic expression in the dataset.

To sum up, cyberbullying detection has received increasing attention from researchers recently. However, most of the existing cyberbullying detection methods have been designed for the English language. Arabic language, which has different and unique characteristics than other languages, has not received adequate research attention. Most of the models designed for detecting Arabic cyberbullying were directly adopted from those designed for the English language. Because of the rich morphology of the Arabic language, which includes a high degree of interspersed vowel patterns, roots in word stems, and affixation, most of the existing solutions suffer in terms of data sparsity and cause low detection accuracy. Some solutions used deep learning to extract the hidden features and skip the feature selection. However, deep learning performance depends on the size of the labeled data available for the training. That is, most of the Arabic datasets used to train existing solutions are relatively small and do not guarantee an adequate solution to the data sparsity problem, since collecting a large labeled dataset for Arabic cyberbullying is not a trivial task. Therefore, the goal of this research is to design a consensus-based model that can improve the accuracy of Arabic cyberbullying detection by ensemble of a diverse set of machine learning classifiers. The data sparsity problem was solved by a set of diverse classifiers. The results of the classifiers are inferred using a human decision-making mechanism, namely the sigmoid function. A comprehensive review of the proposed model is presented in the next section.

3 Proposed Consensus-Based Detection Model

The majority of the current cyberbullying detection models are trained using an English language-based textual dataset. There are few cyberbullying detection models proposed for other languages. Arabic language which is different in its syntax and semantics comparing to other languages, is yet to be tested in-depth by researchers. The performance of the model constructed using the English language are ineffective for Arabic text due to its language variants, implicit sentiments, and grammar complexity. The proposed model in this study consists of four phases: dataset collection, dataset preprocessing, features extraction and representation, and model construction. The proposed model is illustrated in Fig. 1. A detailed explanation of each phase is presented in the following sections.

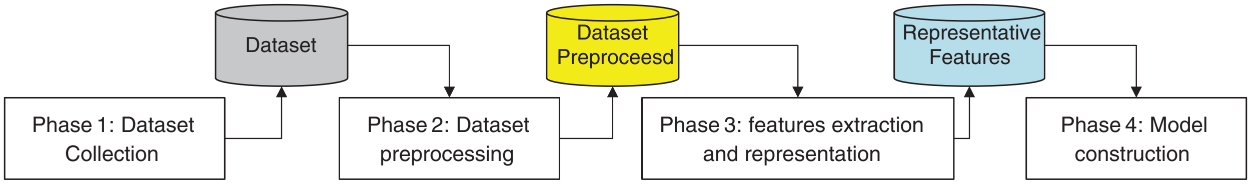
Dataset
Dataset Preproceesd
Representative Features
Phase 1: Dataset Collection
Phase 2: Dataset preprocessing
Phase 3: features extraction and representation
Phase 4: Model construction

Figure 1: Methodology adopted for constructing the proposed model

3.1 Dataset Collection Phase

Due to the lack of availability of Arabic Dataset for cyberbullying detection, the dataset from other languages has been translated to Arabic language in two stages. In the first stage, the google translation of API was used to generate the first draft of the translation. Google translate is the most accurate and common online translation used for many languages [38,39]. In 2016, the google translation library has been redesigned using the Neural Machine Translation model (NTM) [38,39]. NTM raised the translation accuracy from 63% using Statistical Machine Translation (STM) model methods to 72% on average nearing human-based translation quality which is 77% and English-Arabic translation can reach 85% accuracy [40]. In the second stage, the dataset is corrected manually using human-based translation. The polarity of the words in cyberbullying is modified according to the spoken Arabic language. Similarly, the implicit sentiment has been corrected as well.

3.2 Pre-Processing Phase

In the preprocessing phase, the text transformed into a more digestible form to help machine learning algorithms perform better. The preprocessing phase is illustrated in Fig. 2. The preprocessing phase consists of tokenization, normalization, stemming, and lemmatization [17,19,23,25]. In the tokenization stage, the text is split into smaller parts (usually words) called tokens. Each token can be used as a feature in the next stage. The normalization includes removing the irrelevant or unnecessary parts of the text, such as numbers, hyperlinks, special characters, words in Latin, symbols, dates, punctuations, stop words, sparse terms, and white spaces. Such text parts and letters work as noise that destructs the classification accuracy. Then, the stemming is done to convert words to their original source. This is usually done by removing the vowels in Arabic words. After normalization and stemming, the lemmatization process is conducted. In the lemmatization stage, the suffixes and prefixes of the words are removed to reduce the similar words that have different forms and thus represent them more accurately.

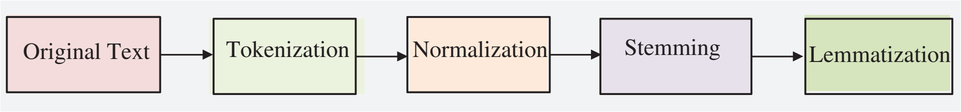

Figure 2: Text preprocessing

3.3 Feature Representation Phase

In this phase, the textual features in the dataset were converted to a numerical representation. The information retrieval statical technique used in this study is called Terms-Frequency and Inverse Document Frequency (TF-IDF) [41]. Each instance in the dataset is described by a vector called features vector. The feature vectors contain the TF-IDF value of each token in the dataset. The length of the feature vector depends on the number of unique words obtained from the dataset. The TF-IDF value of a word reflects how important that word is for the classification in the corpus. For example, offensive or abusive words will be found more in the cyberbullying annotated instances. Let $tf_{i,j}$ be the number of occurrences of the word i in the instance j, the df_i be the total number of instances the word i occurred in, and N be the total number of instances. Then, the TF-IDF wight $w_{i,j}$ of word i in the instance j can be measured in Eq. (1) in the following way:

$$\mathrm{w_{i,j}} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{1}$$

While df consider the terms that are equally important, the term df is used to suppress the weight of the terms that occur on all the documents while scaling up the weight of the rare terms.

3.4 Model Construction Phase

The datasets were divided into two categories: training part and testing part. The training subsets are used to construct the classifier, while the testing subsets are used to evaluate the classifier's results. Several machine learning algorithms were investigated in the comparison, such as SDL, RF, SVM, ANN, XGBoost (XGB). Then, a heterogeneous set of Ensemble Classifiers were combined and the decision is taken. Each classifier contributed by scaling the degree of cyberbullying in consensus-based decision-making. Fig. 3 shows the proposed classification model of using machine-learning algorithm.

The outputs of the classifiers are combined for consensus decision-making as follows. Let s_i be the output of the classifier i and w_i be the weight given to the classifier based on its accuracy, the consensus value z is calculated in Eq. (2) where n is the total number of classifiers.

$$\mathrm{z} = \sum_{i=1}^{n} \frac{w_i s_i}{n} \tag{2}$$

The classification decision $d_{(x)}$ of sample x is calculated using the sigmoid in Eq. (3) where e is the natural exponential function.

$$\mathrm{d_{(x)}} = \frac{1}{1 + e^{-z}} \tag{3}$$

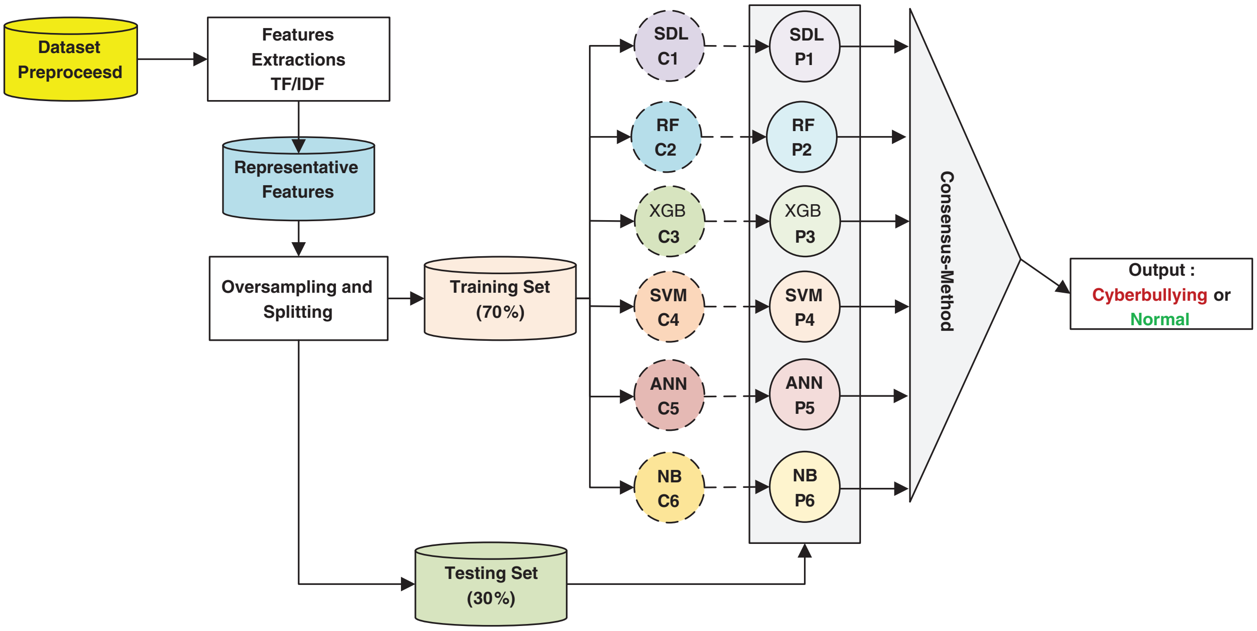

Figure 3: The proposed consensus-based ensemble cyberbullying detection model

Fig. 4 below depicts the pseudocode of the algorithm used for the online operation of the proposed consensus-based ensemble cyberbullying detection model. The sigmoid function was chosen because it is quite similar to a process of real-world thinking when a human is making a decision. It also adds the fuzziness aspect to the conventional linear process. This ensures linear growth in the middle while truncating the curve at the extremities.

```
Pseudocode: Online Operation - Consensus-Based Ensemble Model for Arabic Cyberbullying Detection
Input: $Post\ p, TF/IDF\ Dictionary\ D, Set\ of\ trained\ model\ H,\ Classifiers'\ Weights\ W$
Output: $Post\ Class\ c_p$
Step 1: Pre-processing
    $Normalization$: remove_non_ascii, to_lowercase, remove_punctuation, replace_numbers, remove_stopwords,
    $Teknonization\ and\ Steming$: word_split, stem_words
    $Lemmatization$: lemmatize_verbs
Step 2: Features Representation and Mapping
    $\forall\ d_i\ \epsilon\ p\ do$
        Get TF/IDF score $w_{i,j}\ from\ the\ TF/IDF\ Dictionary\ D$
        Map $d_i$ $to$ the features vector $v_p \leftarrow d_i$
Step 3: Ensemble-based Classification
    $\forall\ h_i\ \epsilon\ H\ do$
        $s_i \longleftarrow h_i(v_p)$ : $s_i$ is the classification score from classifier $h_i$
        $S \xleftarrow{append} s_i$
Step 4: consensus-Based decision-making
    $z \longleftarrow \sum_{i=1}^{n} \frac{w_i s_i}{n}$ : $w_i$ is the weight of $h_i$
    $c_p \longleftarrow \frac{1}{1+e^{-z}}$
    return $c_p$ the binary class
```

Figure 4: The pseudocode

3.5 Performance Analysis

3.5.1 Datasets

Although there is no common standard dataset for cyberbullying in social media, such as Twitter, Facebook, and YouTube [42]. Formspring is a commonly used source of the cyberbullying datasets (https://www.chatcoder.com/data.html). For the Arabic language, few works were proposed. Most of these works use private datasets which were created by the authors and not available for the public. Besides, most of these datasets are relatively small datasets and may not be representative. This is because the Arabic language has a rich morphology such that a small dataset leads to data sparsity problem [18]. Therefore, five datasets were combined and used in this research work to train and evaluate the proposed model. Each dataset was extracted from different platforms, namely Twitter, WhatsApp, Vine, Instagram, and Packet. The details of each dataset are shown in Tab. 1. In total, 23462 samples have been considered in the experiments among which 17122 are normal samples and 6340 are cyberbullying samples.

Table 1: List of the used datasets

Dataset names	Total samples	Normal (Non-cyberbullying)	Cyberbullying
Twitter	13471	11501	1970
WhatsApp	1281	1028	253
Vine	1332	666	666
Instagram	6097	2899	3198
Packet	1281	1028	253
Total	23462	17122	6340

3.5.2 Experimental Setups

Due to the class imbalance between the normal and cyberbullying samples, the cyberbullying samples have been oversampled using Synthetic Minority Oversampling Technique (SMOTE). Thus, the total number of samples was 34,244 samples. The datasets were split into two subsets: training (70%) and testing (30%). Python programming with libraries, such as *nltk* library for natural language processing, scikit-learn library for conventional machine learning classifiers, imbalance for resampling of unbalanced data, and *Keras* library for deep learning among other libraries were used for the dataset preprocessing, oversampling, feature extraction, representation, and model training and testing. Six machine learning classifiers were used for the comparison with the proposed model, namely Random Forest (RF), Artificial Neural Network (ANN), NB, SVM, XGB and the SDL. As shown in Fig. 3, six classifiers that have the topmost performance were chosen to construct the proposed ensemble model.

3.5.3 Performance Measures

Four performance measures that are commonly used in the literature [10–12,24] were used to evaluate the proposed cyberbullying detection model. The first measure is classification accuracy (Acc) illustrated in Eq. (4) which is the ratio between the true classified samples and the total samples in the dataset. The second performance measure is precision (Pre) illustrated in Eq. (5) which is the ratio between the true classified normal samples and the total number of samples that are classified normal. The third performance measure is recall (Rec) illustrated in Eq. (6) which is the ratio between the true classified normal samples and the total number of normal samples. The fourth performance measure is the F1-measure ($F1$) illustrated in Eq. (7) which is the harmonic mean between recall and precision. The following formulates depict how each measure can be calculated.

$$\mathrm{Acc} = \frac{Number\ \ of\ \ True\ \ Classified\ \ Samples}{Total\ \ Number\ \ of\ \ Samples} \tag{4}$$

$$\mathrm{Pre} = \frac{Number\ \ of\ \ True\ \ Classified\ \ Normal\ \ Samples}{Total\ \ Number\ \ of\ \ Samples\ \ Classified\ \ Normal} \tag{5}$$

$$\mathrm{Rec} = \frac{Number\ \ of\ \ True\ \ Classified\ \ Normal\ \ Samples}{Total\ \ Number\ \ of\ \ Normal\ \ Samples} \tag{6}$$

$$\mathrm{F-Measure} = \frac{2\ \times\ Pre\ \times Rec}{Pre + Rec} \tag{7}$$

4 Results and Discussions

The results acquired from the proposed model are compared to the results of other machine learning classifiers. The performance in terms of precision, accuracy, F1-measure, and recall shown in Tab. 2 and Fig. 5. The combined dataset was used to train the proposed model and the six classifiers used for the comparison. Then, the proposed model was constructed from these classifiers by combining the output of these classifiers using weighted average as discussed in Eq. (3).

Table 2: Performance analysis

	Accuracy	Precision	Recall	F1-Measure
ANN	86.32	86.73	85.91	86.32
XGB	85.51	86.30	85.46	85.87
NB	73.23	90.86	67.72	77.60
SVM	87.17	88.31	86.73	87.52
RF	87.24	87.37	87.67	87.52
SDL	83.03	77.90	86.62	82.03
The proposed model	88.54	89.24	88.41	88.82

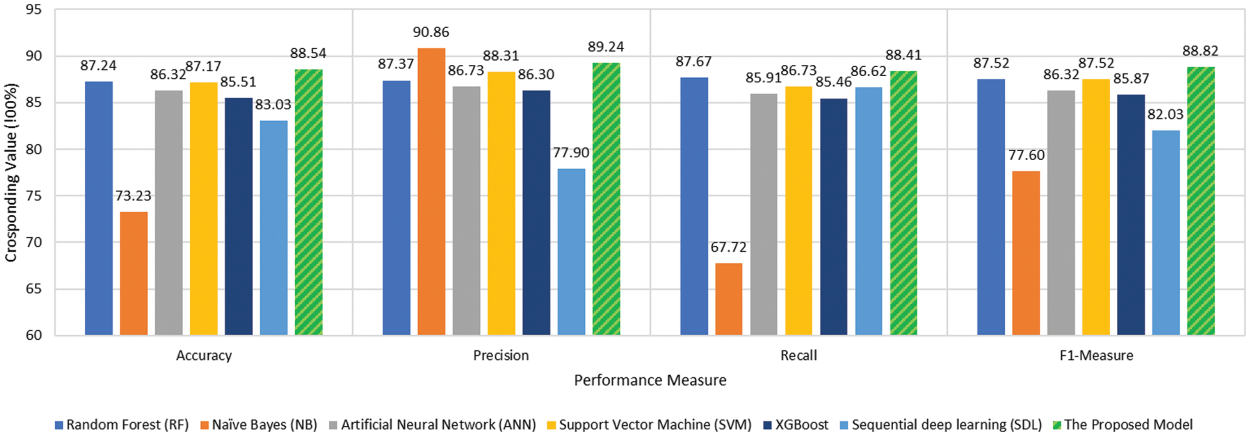

Figure 5: Performance comparison

The proposed model achieved the best accuracy among all the tested classifiers. The ensemble learning-based algorithm RF classifier makes better predictions comparing to the other tested classifier due to the use of ensemble learning and the majority voting scheme as decision making. However, comparing with the proposed model, the use of heterogeneous classifiers along with consensus-based decision making is more effective than the homogeneous classifiers used by the RF algorithm. The harmonic mean F-measure better describes the performance of the tested classifiers due to its ability to convey the balance between precision and recall. The proposed model achieved the best tradeoff between the precision and recall 88.82% comparing to the other tested classifiers while deep learning achieved the worst tradeoff between the precision and recall in terms F1-Measure. This is due to the sparsity problem of the Arabic language in which the scale of the training data affects the performance of the classifier. The deep learning algorithm needs a large number of training samples of Arabic language which is lacked by the current cyberbullying detection studies. As can be noticed in Fig. 5 and Tab. 2, the proposed consensus-based ensemble model achieved the highest performance compared to the other tested models.

Tab. 3 lists the percentage of the improvement gained compared with the tested classifiers. The overall improvement gained by the proposed model reach 1.3% comparing with the RF and SVM which are the best-tested classifiers.

Table 3: Improvement ratio

	Accuracy (%)	Precision (%)	Recall (%)	F1-measure (%)
ANN	2.22	2.51	2.5	2.5
XGB	3.03	2.94	2.95	2.95
NB	15.31	−1.62	20.69	11.22
SVM	1.37	0.93	1.68	1.3
RF	1.3	1.87	0.74	1.3
SDL	5.51	11.34	1.79	6.79

Figs. 6 and 7 depict the detailed performance measures used to get deep insights into the effectiveness of the proposed model. As can be seen in Fig. 6, the proposed model achieved the best and highest tradeoff between True Negative Rate (TNR) and True Positive Rate (TPR) among all the tested algorithms. Although the Naïve Bayes classifier achieved the highest TPR value between the tested models and the lowest TNR value among the tested models. In terms of False Negative Rate (FNR) and False Positive Rate (FPR), the proposed model achieved the lowest and the best reduction of both FNR and FPR.

The ensemble learning with the set of homogenous classifiers, such as the case of the RF algorithm. RF achieved better detection performance than the other tested classifiers. However, the ensemble of heterogenous classifiers used by the proposed consensus-based model outperforms the RF algorithm. Therefore, a diverse set of heterogeneous classifiers with consensus-based decision-making can decrease the impact of the data sparsity problem and enhance the detection accuracy of Arabic language-based cyberbullying. This concept can be extended to other languages as well as other domains to improve the classification performance.

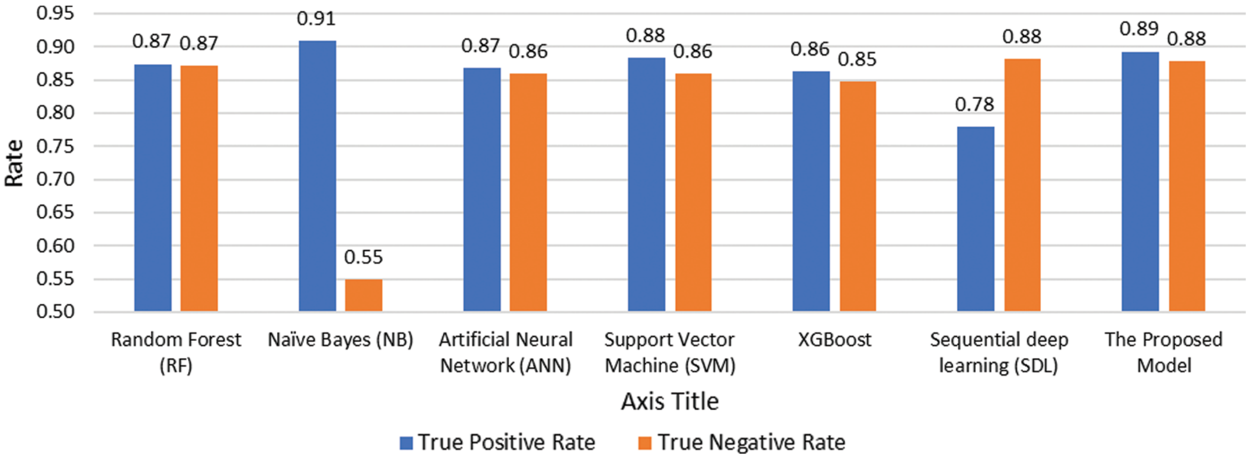

Figure 6: Performance comparison of TNR and TPR

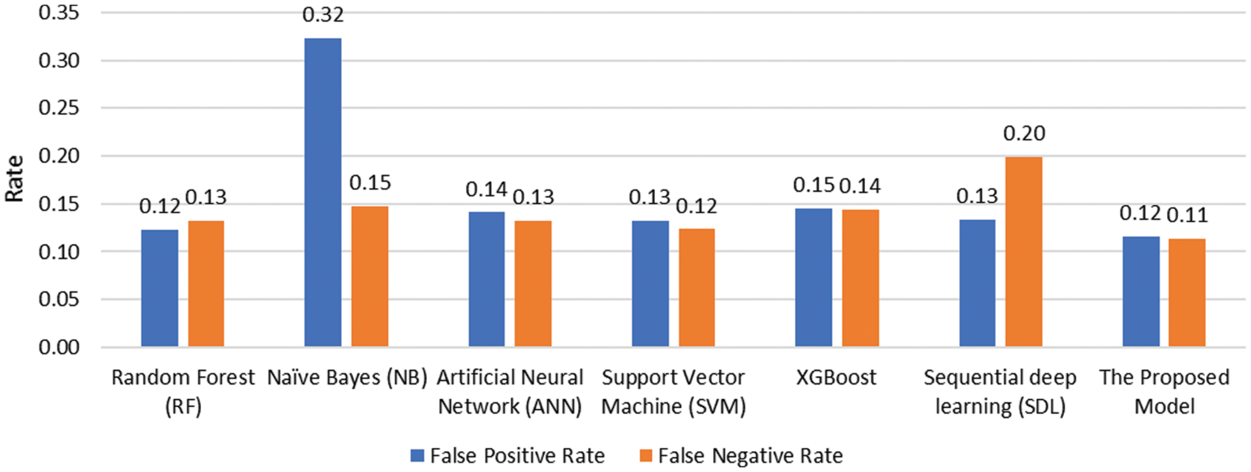

Figure 7: Performance comparison of FNR and FPR

5 Conclusion and Future Work

In this study, a consensus-based ensemble cyberbullying detection model is proposed for detecting Arabic cyberbullying and offensive speech. The proposed model consists of four main components: text preprocessing, features extraction and representation, ensemble-based classification, and consensus-based decision making. Most of the existing cyberbullying detection models were designed for the English language or languages with relatively similar characteristics like Latin-based languages. Most of these models don't meet the Arabic language's unique characteristics, leading to misrepresentation of cyberbullying in Arabic data. The proposed model in this study improves the accuracy of the existing technology through the power of diversity and consensus decision-making. Six machine learning classifiers were trained and the output of these classifiers were combined using a weighted average decision-making system to achieve the consensus instead of the majority voting. The proposed model achieved the best performance for cyberbullying detection comparing to the other models.

In this study, the focus directed to the model designed, features extraction, representation, and selection will be explored in-depth in our future studies. The dataset used in the study which contains the modern standard Arabic language is translated from other languages. Due to complexity of Arabic culture, there are some expressions in Arabic language which are considered offensive speech whereas such expressions may be considered normal in other language. Thus, a labeled dataset that considered regional variations for Arabic cyberbullying is important for improving the performance in future research.

Acknowledgement: The authors gratefully acknowledge the approval and the support of this research study by Grant No. SAT-2018-3-9F-7871 from the Deanship of Scientific Research at Northern Border University, Arar, K.S.A. Moreover, assistance provided by Dr. Fuad A. Ghaleb was greatly appreciated.

Funding Statement: This research was funded by Deanship of Scientific Research at Northern Border University, Grant No. SAT-2018-3-9F-7871.

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

References

[1] S. Hinduja and J. Patchin, “Bullying, cyberbullying and suicide,” *Archives of Suicide Research*, vol. 14, no. 3, pp. 206–221, 2010.

[2] V. Balakrishnan, S. Khan and H. Arabnia, “Improving cyberbullying detection using twitter users’ psychological features and machine learning,” *Computer Security*, vol. 90, no. 101710, pp. 101710, 2020.

[3] J. Patchin, “2019 cyberbullying data,” Cyberbullying.org. [Online]. Available: https://cyberbullying.org/2019-cyberbullying-data. [Accessed: 05-May-2021].

[4] Florida Atlantic University, “Nationwide teen bullying and cyberbullying study reveals significant issues impacting youth,” *Science Daily*, 21-Feb-2017. Retrieved July 5, 2021 from www.sciencedaily.com/releases/2017/02/170221102036.htm.

[5] C. Nixon, “Current perspectives: The impact of cyberbullying on adolescent health,” *Adolescent Health Medicine and Therapeutics*, vol. 5, pp. 143–158, 2014.

[6] L. McInroy and F. Mishna, “Cyberbullying on online gaming platforms for children and youth,” *Child and Adolescent Social Work Journal*, vol. 34, pp. 597–607, 2017.

[7] N. Aliyu, A. Dogo, F. Ajibade and T. Abdurauf, “Analysis of cyber bullying on facebook using text mining,” *Journal Application Artificial Intelligent*, vol. 1, no. 1, pp. 1–12, 2020.

[8] i. Riadi, “Detection of cyberbullying on social media using data mining techniques,” *International Journal of Computer Science and Information Security*, vol. 15, no. 3, pp. 244–250, 2017.

[9] A. A. Nuaimi, “Effectiveness of cyberbullying prevention strategies in the UAE,” in *ICT Analysis and Applications*, Singapore: Springer, pp. 731–739, 2021.

[10] E. Ates, E. Bostanci and M. Guzel, “Comparative performance of machine learning algorithms in cyberbullying detection: Using turkish language preprocessing techniques,” arXiv preprint arXiv: 2101.12718, 2021.

[11] C. Iwendi, G. Srivastava, S. Khan and P. Maddikunta, “Cyberbullying detection solutions based on deep learning architectures,” *Multimedia Systems*, vol. 5, pp. 1–14, 2020.

[12] A. Ali and A. Syed, “Cyberbullying detection using machine learning,” *Pakistan Journal of Engineering and Technology*, vol. 3, no. 2, pp. 45–50, 2020.

[13] M. Vyawahare and M. Chatterjee, “Taxonomy of cyberbullying detection and prediction techniques in online social networks,” In: Jain, L., Tsihrintzis, G., Balas, V., Sharma, D. (Eds.), *Data Communication and Networks. Advances in Intelligent Systems and Computing*, Singapore: Springer, vol. 1049, pp. 21–37, 2020.

[14] W. Li, “A design approach for automated prevention of cyberbullying using language features on social media,” in *2019 5th Int. Conf. on Information Management*, Cambridge, UK: University of Cambridge, 2019.

[15] S. Gordon, “Beware parents and educators cyberbullying increasing during pandemic,” Verywellfamily.com. [Online]. Available: https://www.verywellfamily.com/cyberbullying-increasing-during-global-pandemic-4845901. [Accessed: 05-May-2021].

[16] M. Moreno, A. Gower, H. Brittain and T. Vaillancourt, "Applying natural language processing to evaluate news media coverage of bullying and cyberbullying," *Prevention Science*, vol. 20, no. 8, pp. 1274–1283, 2019.

[17] B. Haidar, M. Chamoun and A. Serhrouchni, "Multilingual cyberbullying detection system: Detecting cyberbullying in arabic content," in *2017 1st Cyber Security in Networking Conf.*, Rio de Janeiro, Brazil, 2017.

[18] M. Biltawi, A. Awajan and S. Tedmori, "Arabic reading comprehension benchmarks created semiautomatically," in *2020 21st Int. Arab Conf. on Information Technology*, Zarqa University, Jordan, 2020.

[19] K. Meftouh, M. Laskri and K. Smaïli, "Modeling arabic language using statistical methods," *Arabian Journal for Science and Engineering*, vol. 35, no. 2C, pp. 69–82, 2010.

[20] N. Abdulla, N. Ahmed, M. Shehab and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in *2013 IEEE Jordan Conf. on Applied Electrical Engineering and Computing Technologies*, Amman, Jordan, 2013.

[21] M. Abdul-Mageed, M. Diab and M. Korayem, "Subjectivity and sentiment analysis of modern standard arabic," in *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, 2011.

[22] D. Mouheb, R. Albarghash, M. Mowakeh, Z. Aghbari and I. Kamel, "Detection of arabic cyberbullying on social networks using machine learning," in *2019 IEEE/ACS 16th Int. Conf. on Computer Systems and Applications*, Abu Dhabi, United Arab Emirate, 2019.

[23] 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.

[24] B. Rachid, H. Azza and H. Ben Ghezala, "Classification of cyberbullying text in arabic," in *2020 Int. Joint Conf. on Neural Networks*, Glasgow, United Kingdom, 2020.

[25] B. Haidar, M. Chamoun and A. Serhrouchni, "Arabic cyberbullying detection: Enhancing performance by using ensemble machine learning," in *2019 Int. Conf. on Internet of Things and IEEE Green Computing and Communications and IEEE Cyber, Physical and Social Computing and IEEE Smart Data*, Atlanta, GA, USA, 2019.

[26] D. Mouheb, M. Abushamleh, Z. Aghbari and I. Kamel, "Real-time detection of cyberbullying in arabic twitter streams," in *2019 10th IFIP Int. Conf. on New Technologies, Mobility and Security*, Canary Island Spain, 2019.

[27] B. Haidar, M. Chamoun and A. Serhrouchni, "Arabic cyberbullying detection: Using deep learning," in *2018 7th Int. Conf. on Computer and Communication Engineering*, Kuala Lumpur, Malaysia, 2018.

[28] D. Mouheb, R. Ismail, S. Qaraghuli, Z. Aghbari and I. Kamel, "Detection of offensive messages in arabic social media communications," in *2018 Int. Conf. on Innovations in Information Technology*, AL AIN, UAE, 2018.

[29] N. Indurkhya and F. Damerau, "Handbook of natural language processing," *Computational Linguistics*, vol. 37, no. 2, pp. 395–397, 2011.

[30] M. Al-Ajlan and M. Ykhlef, "Optimized twitter cyberbullying detection based on deep learning," in *2018 21st Saudi Computer Society National Computer Conf.*, Saudi Arabia, 2018.

[31] H. Hosseinmardi, S. Li, Z. Yang, Q. Lv, R. Rafiq *et al.*, "A comparison of common users across instagram and ask.fm to better understand cyberbullying," in *2014 IEEE Fourth Int. Conf. on Big Data and Cloud Computing*, Sydney, Australia, pp. 355–362, 2014.

[32] V. Nahar, X. Li and C. Pang, "An effective approach for cyberbullying detection," *Communications in Information Science and Management Engineering*, vol. 3, no. 5, pp. 238, 2013.

[33] S. Parime and V. Suri, "Cyberbullying detection and prevention: Data mining and psychological perspective," in *2014 Int. Conf. on Circuits, Power and Computing Technologies*, Nagercoil, India, 2014.

[34] J. Sheeba and K. Vivekanandan, "Low frequency keyword extraction with sentiment classification and cyberbully detection using fuzzy logic technique," in *2013 IEEE Int. Conf. on Computational Intelligence and Computing Research*, Tamilnadu, India, 2013.

[35] K. Reynolds, A. Kontostathis and L. Edwards, "Using machine learning to detect cyberbullying," in *2011 10th Int. Conf. on Machine Learning and Applications and Workshops*, Honolulu Hawaii USA, 2011.

[36] T. Bosse and S. Stam, “A normative agent system to prevent cyberbullying,” in *2011 IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology*, Lyon, France, 2011.

[37] A. Mazari, “Cyber-bullying taxonomies: Definition, forms, consequences and mitigation strategies,” in *2013 5th Int. Conf. on Computer Science and Information Technology*, Amman, Jordan, 2013.

[38] D. Lunić, N. Stanišić, A. Njeguš and I. Đerić, “Google translate accuracy evaluation,” in *Proc. of the Int. Scientific Conf.-Sinteza*, Republic of Slovenia, 2020.

[39] S. Tsai, “Using google translate in EFL drafts: A preliminary investigation,” *Computer Assisted Language Learning*, vol. 32, no. 5–6, pp. 510–526, 2019.

[40] S. Mohammad, M. Salameh and S. Kiritchenko, “Sentiment lexicons for arabic social media,” in *Proc. of the Tenth Int. Conf. on Language Resources and Evaluation*, Republic of Serbia, pp. 33–37, 2016.

[41] F. Sebastiani, “Machine learning in automated text categorization,” *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[42] F. Elsafoury, “Cyberbullying datasets,” Mendeley.com. [Online]. Available: https://data.mendeley.com/datasets/jf4pzyvnpj/1, [Accessed: 04-Summer-2021].