ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 1, 783-790 (2021)

www.astesj.com

ASTESJ

ISSN: 2415-6698

# Text Mining Techniques for Cyberbullying Detection: State of the Art

Reem Bayari[*], Ameur Bensefia

*CISAM Division, Higher College of Technology, Abu Dhabi, 4102, UAE*

ARTICLE INFO

*Article history:*
*Received: 26 November, 2020*
*Accepted: 31 January, 2021*
*Online: 05 February, 2021*

*Keywords:*
*Natural Language Processing*
*Text Classification*
*Machine Learning*
*Deep Learning*
*Neural Network*

ABSTRACT

*The dramatic growth of social media during the last years has been associated with the emergence of a new bullying types. Platforms such as Facebook, Twitter, YouTube, and others are now privileged ways to disseminate all kinds of information. Indeed, communicating through social media without revealing the real identity has emerged an ideal atmosphere for cyberbullying, where people can pour out their hatred. Therefore, become very urgent to find automated methods to detect cyberbullying through text mining techniques. So, many researchers have recently investigated various approaches, and the number of scientific studies about this topic is growing very rapidly. Nonetheless, the methods are used to classify the phenomenon and evaluation methods are still under discussion. Subsequently, comparing the results between the studies and identifying their performance is still difficult. Therefore, the current systematic review has been conducted with the aim of survey the researches and studies that have been conducted so far by the research community in the topic of cyberbullying classification based on text language. In order to direct future studies on the topic to a more consistent and compatible perspective on recent works, we undertook a deep review of evaluation methods, features, dataset size, language, and dataset source of the latest research in this field. We made a choice to focus more on techniques that adopted neural networks and machine learning algorithms. After conducting systematic searches and applying the inclusion criteria, 16 different studies were included. It was found that the best accuracy was achieved when a deep learning approach is used particularly CNN approach. It was found also that, SVM is the most common classifier in both Arabic and Latin languages and outperformed the other classifiers. Also, the most widely used feature is N-Gram especially bigram and trigram. Furthermore, results show that Twitter is the main source for the collected datasets, and there are no unified datasets. There is also a shortage of studies in Arabic texts for cyberbullying identification in contrast with English texts.*

## 1. Introduction

Online social media is now a part of everyday life activity; without a digital footprint, it has become increasingly difficult to survive in this new age of digital media. Cyberbullying is defined as an electronic form of intentional harm and hate to someone and it's considered as a crime [1]. The work presented in [2] reports that cyberbullying had a major and long-term impact on the victims. Cyberbullying leaves both the abuser (predator) and the victim with mental and physical consequences. Different researchers [2], [3] have reported that victims attempted suicide due to many cyberbullying incidents, where they have been mentally abused by offensive and violent messages received from abusers. Numerous studies have shown that adolescents are the primary victims [3], [4]. Despite the regulations, presented in most of the countries, that protect and help bullying victims, there are still many people who suffer from this phenomenon. Indeed, if the victim or his family doesn't report the case of bullying, the victim will keep suffering, and the abuser will continue making other victims. Therefore, the early detection of bullying will help in finding an effective solution by protecting the targeted person and punishing the abuser. Measures to track and identify potentially harmful online behavior must therefore be implemented. Because of the large number of daily posts, and the huge amount of information that circulates through the different social media platforms, manual checking for all posts is just impossible.

[*] Corresponding Author: Reem Bayari, ralbayari@hct.ac.ae

https://dx.doi.org/10.25046/aj060187

Consequently, several studies focused on finding a way to autodetect the presence of cyberbullying quickly and effectively in order to avoid any serious consequences [5], [6].

In this paper, we present an exhaustive list of the most recent research dedicated to autodetecting cyberbullying by focusing mainly on machine learning, neuronal networks and deep learning techniques. We undertook a deep review of evaluation methods, features, dataset size, language, and dataset source of the latest research in this field. In the following section we give the reader a background of cyberbullying, followed by section 3 where we discuss and present the cyberbullying approaches based or website is denigrated.

To accomplish the primary objective of this study, we identified our research questions as follows:

- RQ1: Which dataset was mainly used for the classification of cyberbullying?
- RQ2: What was the size and language of the dataset?
- RQ3: What was the method of classification used? And which one has been used most?
- RQ4: What were the metrics of quality used?
- RQ5: Which approach proved the most effective?
- RQ6: Which features were the most frequently used with classifiers?

## 2. Research Method

The main stages of the review methodology (research strategy, quality evaluation criteria) were outlined in this section, as in Figure 1.

### 2.1. Data Sources

In January 2020, we started this study, and we included most studies from 2016 to 2019 and some studies before that. We used the following databases: IEEE, Science Direct, ACM.

### 2.2. Inclusion and exclusion criteria

The inclusion and exclusion criteria were developed in the selection criteria stage to ensure that the research included in this study was valuable and relevant and would lead us to our main objective.

*Inclusion criteria*

- Papers that are purely for the classification of cyberbullying in texts.
- Papers that used neural networks / machine learning in the classification.
- Papers that explained the model and its performance measures.
- Papers which mentioned the size and language of the dataset.

*Exclusion criteria*

- Papers not for text classification
- Papers that didn't mention the findings' accuracy.
- Paper had not been published in journal or conference
- Paper did not use neural networks / machine learning in the classification.
- Paper that didn’t mention the size and language of the dataset.

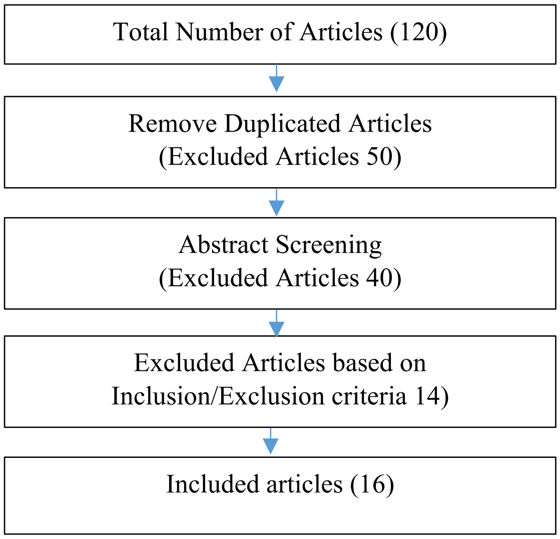
Total Number of Articles (120)
Remove Duplicated Articles (Excluded Articles 50)
Abstract Screening (Excluded Articles 40)
Excluded Articles based on Inclusion/Exclusion criteria 14)
Included articles (16)

Figure 1: Data Sources

### 2.3. Search strategy

We used the bellow keywords to collect all the previous studies:

- “Cyberbullying” and “classification” and “neural networks” and “text”
- “Cyberbullying” and “classification” and “deep learning” and “text”
- “Cyberbullying” and “detection” and “deep learning” and “text”
- “Cyberbullying” and “classification” and “machine learning”
- “Cyberbullying” and “categorization” and “text”
- “Cyberbullying” and “classification” and “text”

### 2.4. Quality assessment evaluation

In this part, we designed quality assessment questions to make a checklist for the research and ensure that it would satisfy the aim of this systematic review.

- Q1: Was the corpus (size and language) identified and described well?
- Q2: Were the text classification approach described clearly?
- Q3: Was the performance of the method identified clearly?

## 3. Cyberbullying Background

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities.

For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

Cyberbullying is defined as the use of communication technology and information such as messages, photographs or videos in order to spread aggressive actions with the intention of harming others [1]. Unlike, bullying, cyberbullying does not require the presence of the victim in the same place or near the bully's place [7].Therefore, it differs from traditional bullying that depends on direct abuse towards victims who could be children, adolescents, or women through physical aggression and intentional, visible behavior [8]. Due to the development of technology and the increase of using smart devices, cyberbullying has become more common and represent a real problem, nothing seems to be able to stop. This is because it is done via the Internet, which involves unknown distances and sources, that allow users to speak without restrictions and it is easy to repeat the aggressive actions at any time and could spread rapidly. Cyberbullying it's not a new problem, in 2005, [2], indicated that the correlation between bullying and psychological symptoms is a reality. These symptoms may compromise risk factors involved in psychopathology. The authors indicate that bullying causes violence, delinquency, depression, anxiety, self-destructive, identity and suicidal issues, and that such symptoms could lead to psychopathology. Cyberbullying's effects are profound and could have major and long-term effects compared to traditional bullying, especially for teenagers who present the largest proportion of victims. According to statistics of [3], [4], several victims of cyberbullying tried to commit suicide because of the, degrading, and violent texts that abusers sent to them .

### *3.1. Cyberbullying Categories*

There are several categories of cyberbullying as stated in [9], [10].

- Flooding: Consists of the bully giving the same one regularly comments, nonsense comments, or even by clicking on the enter key, in order not to let the victim contribute to the conversation.
- Masquerade: includes logging in to bully in a forum, chat room, or software using the account name of another person in order to bully a victim or tarnish the image of the victim.
- Flaming (bashing): involves two or more users attacking each other on a personal level. The conversation consists of a heated, short lived argument, and there is bullying language in all of the users' posts.
- Trolling (baiting): is intentionally posting opinion agreeing with other posts in an emotionally loaded thread in order to provoke a war, even though the comments do not actually reflect the actual opinion of the poster.
- Harassment: resembles conventional bullying most closely with the assumed relationship of the bully and victim. This form of cyberbullying involves sending the victim repeatedly abusive messages over extended periods of time.
- Cyberstalking and cyberthreats: involve sending messages that include harm attacks, which are threatening or extremely aggressive, or include extortion.
- Denigration: this kind of bullying includes Writing obscene, negative, or false rumors of someone to others or sharing them on a public forum or chat room, or website.
- Outing: this kind of bullying includes posting confidential, personal, or embarrassing information in a public chat room or forum. This type of bullying is close to denigration but requires the abuser and the victim to have a close relationship with each other either online or in person.
- Exclusion: this type of cyberbullying occurred most frequently in chat rooms or conversations between young people and adolescents by ignoring the victim.

### *3.2. Cyberbullying prevention methods and limitations*

Due to the extreme worldwide prevalence of cyberbullying and it's direct link with many negative psychological symptoms, researchers have studied the relationship of cyberbullying with several factors in order to help in cyberbullying prevention. The work presented in [11], found that supplying sympathy and strengthening the relationship between caregiver and adolescents affect positively in cyberbullying prevention. On the other hand , the work that represented in [12], reports that the improving of awareness on cyberbullying issues played an important role in cyberbullying prevention. In addition, some countries considered the cyberbullying as a crime. Therefore, they put the laws in dealing with anyone doing such a crime as well as encouraged people to report any case of bullying, as in UAE. This is all to try to prevent children and teens from engaging in cyberbullying as well as to assist cyber victims to deal with the adverse effects of cyberbullying. Although, methods and tools continue to enhanced in cyberbullying detection , the access restrictions on high-quality data limit the applicability of state-of-the-art techniques. Consequently, much of the recent research uses small, heterogeneous datasets, without a thorough evaluation of applicability [13].

### *3.3. Cyberbullying automatic detection methods*

The effective solution to detect bullying in the posts over social media is to build machine-based automated systems. It's for categorizing information and producing reports where cyberbullying is detected so that with fewer losses all recorded incidents can be quickly sorted out and addressed. Different methods are used to detect cyberbullying such as machine learning techniques, Natural Language Processing (NLP), and Deep Learning (DL). Examples include in [14], several NLP models such as Bag of Words (BoW), Latent Semantic Analysis (LSA) and Latent Dirichlet Allotment (LDA) used to detect bullying in Social Networks. On the other hand , some of the autodetection methods were based on word embedding, which expands a list of predefined offensive terms and assigns different weights to obtain bullying and latent characteristics [15].

## 4. Cyberbullying Detection Approaches Language-based

In this section, we present the most relevant works conducted in cyberbullying detection. We organized them based on the selected language, either Arabic or Latin, and both the features and the classification used in both cases.

### *4.1. Detection in Arabic language*

Alakrot et al in [6], investigated of using word-level and N-gram features with common pre-processing methods, including extra normalization effect on the efficiency of a trained SVM classifier .This is to detect offensive comments. Authors created dataset of 15,050 comments by collecting comments about the famous Arabic people from the YouTube platform. The dataset is available for public.As a result the stemming with pre-processing enhanced the detection of offensive language in casual Arabic text. In addition, the use of N-gram features increased the classifier efficiency. Despite the combination of stemming and N-gram features showed a negative impact on precision and recall ,pre-processing with stemming and N-grams (1-5) achieved the best performance as highlighted in Table 1.

Authors in [16], authors represented the first study in utilizing deep learning in Arabic cyberbullying detection .They utilized the same dataset in [17], with little changes. Changes include removing all hyperlinks, un-Arabic characters and emoticons. The dataset was tokenized into words to remove all unneeded characters before building the model's layers. Word Embedding was created after that. The dataset is divided into 80% for training and 20% for testing, then they trained a Feed Forward Neural Network FFNN. Authors considered the final decision of the accuracy  93.33 percent with validation accuracy 94.27, for the seven-layer network. Although it achieved an accuracy 94.56% with the three hidden layers.

In [17],  it is suggested a system for detecting cyberbullying in English and Arabic text. The only features included in the first stage were text (content of tweet) and language (English, Arabic). In the second stage, authors used an affective tweets package, specifically the TweetToSentiStrength Feature Vector filter. SentiStrength used for weighting the tweets, (2 to 5) for positive feelings and (-2 to -5) for negative feelings ,(1 , -1) for representing neutral feelings. The English lexicon files were used by SentiStrength where subsequently replaced by custom-built Arabic files including weighted profane words .Haidar et al, built two datasets. First one was obtained from Facebook which reached 0.98GB of size in order to verify the system. Second one was collected from twitter to train and test the system. The Arabic dataset was collected from different dialects Lebanon, Syria, Gulf Area and Egypt mainly, it contains 35273 unique tweets after removing all duplicates. Authors show that there is a difference between the “yes” precisions between the two classifiers. In terms of precision, SVM was much higher with  “yes” class. However, the overall system precision was 93.4 for SVM, 90.1 for NB.

A study by [18], utilized a collection of predefined obscene words as seeds words to collect another list from a large set of 175 million tweets . These were used to create a list of obscene words to detect offensive language and hate speech. Authors then generated new list from 3,430 words by performing Log Odd Ratio (LOR) method on unigrams and bigrams features. Authors evaluated the detection of offensive language by using five methods which are (seeds Words (SW),SW+LOR (unigram),SW+ LOR (bigram), LOR (unigram), LOR (bigram). The highest precision achieved from using the list generated by LOR(unigram).Authors have made  the dataset public for research as well as the list of obscene words and hashtags.

Table 1 provides a comparative summary between the different approaches in Arabic cyberbullying detection discussed above.

### 4.2. Detection in Latin language

In [19], authors used sentiment analysis to detect instances of bullying in the social network using Naïve Bays classifier (NB). Authors worked on a balanced dataset consisting of 5000 English tweets. Authors collected messages that contained one of these words “Gay,” “Homo,” “Dike,” and “Queer”. For training data, queer word used to classify the positive tweets while the presence of any of these terms "Gay," "Homo," "Dike” used to classify the negative tweets. As a result, NB classifier achieved accuracy 67.3%.

Table 1: Comparative summary between the different approaches in Arabic cyberbullying detection

| Ref | Dataset Size | Language | Platform | Performance | | | Approach | Features |
|---|---|---|---|---|---|---|---|---|
| | | | | Prec | Rec | F1 | | |
| [6] | 15,050 comments | Arabic | YouTube | 88% | 77% | 82% | SVM | N-grams (1-5), word-level |
| | | | | 83% | 80% | 81% | | |
| [16] | 4.913 records | Arabic | Twitter | Accuracy: 93.33 with Validation accuracy: 94.27 | | | FFNN | Unspecified |
| | 30.890 records | | | | | | | |
| [17] | Arabic 35273 | Arabic, English | Twitter, Facebook | 90.1 | 90.9 | 90.5 | NB | Text Language, SentiStrength Lexicon |
| | English 91431 | | | 93.4 | 94.1 | 92.7 | SVM | |
| [18] | 288 words and phrases. 127 Hashtag, 3.430 word | Arabic | Twitter, Al-Jazeera | 98% | 41% | 58% | SW, SW+LOR (unigram), SW+LOR (bigram), LOR (unigram), LOR (bigram) | Predefined list, Unigram, Bigram |

Table 2: Comparative summary between the different approaches in Latin cyberbullying detection

| Ref | Dataset Size | Language | Platform | Performance | | | | Approach | Features |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Precession | Accuracy | Recall | F1 | | |
| [19] | 5000 tweets | English | Twitter | | 67.3% | | | NB | Predefined list |
| [20] | 1000 tweets 3.045 posts 20.921 questions and answers. | English | Twitter, FormSpring.me, YouTube | 60-81% | 69-73% | 26-94% | 4-74% | GHSOM | Syntactic Semantic Sentiment Social |
| | | | | 60% | -- | 40% | -- | C4.5 | |
| | | | | - | 67 % | -- | -- | NB | |
| | | | | --- | -- | 67% | -- | SVM | |
| [21] | 1608 conversations | English | FormSpring.me | 89.6% | 77.65 | 91.1% | 89.8%. | SVM | bigram, trigram,4-gram |
| [22] | 900 messages written | Turkish | Twitter, Instagram | - | - | - | 54% | J48 | Undefined |
| | | | | - | - | - | 81% | NBM | |
| | | | | - | - | - | 84% | IBK | |
| | | | | - | - | - | 64% | SVM | |
| [23] | 14,509 tweets | English | Twitter | -- | 78% | -- | -- | SVM | word skip-grams, and Brown clusters |
| [24] | 25k tweets | English | Twitter | 91% | -- | 90% | 90% | logistic regression with L2 regularization | bigram, unigram, and trigram, syntactic structure sentiment lexicon, number of characters, words, and syllables in each tweet. |
| [15] | 1762 tweets | English | Twitter | 76.8 | | 79.4 | 78.0 | SVM | BOF, Latent Semantic Bullying |

The authors in [20] , used an unsupervised approach for detecting cyberbully traces over social platforms. This is by utilizing Growing Hierarchical Self-Organizing Map. The suggested model is based on machine learning (decision tree C4.5, SVM, NB) as well as techniques derived from NLP (Natural Language Processing) such as semantic, syntactic features of textual sentences. Authors in this work tested the proposed model in three different datasets and platforms. The performance range for each classifier is covered in table 2.

The authors in [21], adopted a supervised approach for cyberbullying detection. Authors used different machine learning classifiers, TFIDF and sentiment analysis algorithms for features extraction. The classifications were evaluated on different n-gram language models. Authors found out, neural network with 3-grams achieved higher accuracy 92.8% compared to SVM with 4- grams that achieved 90.3%. Furthermore, NN exceeded other classifiers on the same dataset in another work. The dataset obtained from Kaggle (Formspring.me) and consists of 1608 English instance conversations, labeled under two classes (Cyberbullying, non-Cyberbullying). Each class consists of 804 instances. The performance average rate for each classifier is highlighted in table 2.

A study by [22], authors adopted a supervised approach to the identification of bullying and harassment in posted messages in Turkish language. Authors used information gain and chi-square methods for features selection. Authors used the same labeled dataset. It was collected from Kaggle (Instagram and Twitter). Authors calculated the accuracy and running time for many machines learning classifiers, including SVM, Decision Tree (C4.5), Naïve Bayes Multinomial, and K Nearest Neighbors (KNN). Authors compared the accuracy of classifiers under different conditions, NBM classifier was the most efficient classifier before features selection is implemented, while IBk achieved the most efficient when 500 features were selected. As shown in table two.

A study by [23] , authors used different methods of text classification to differentiate between hate ,profanity expressions, and other texts. In determining the baseline, authors used standard lexical characteristics and a linear SVM classifier. Authors applied a linear SVM classifier on three groups of features extracted surface n-grams, word skip-grams, and Brown clusters. The best accuracy (78%) achieved when authors used the character 4-gram model.

In [24], authors used a hate speech lexicon to collect tweets that containing hate speech keywords. It's used then to label a sample of these tweets under three categories (hate, offensive, normal) speech. Authors trained a multiclass classifier to distinguish these different categories. Authors used bigram, unigram, and trigram features as well as features for the number of characters, words, and syllables in each tweet. In addition, authors included binary and count indicators for hashtags, mentions, retweets, and URLs. Author then tested a variety of models; each model was tested by using 5-fold cross-validation. Authors found that the Logistic Regression and Linear SVM model outperformed other models. For the final model, logistic regression with L2 Regularization were used for the final model. The final model then trained by using the whole dataset to predict the label for each tweet. The best performance is highlighted in table 2.

In [15], authors introduced a novel learning method for cyberbullying recognition called Embedding Enhanced Bag-of-Words (EEBOW).EEBoW mixes BoW (Bag of Words) features, latent semantic features and bullying features .Bullying features are derived from word embedding, capturing the semantic details behind words. Authors reported that the EBoW model outperforms other comparable models, including Semantic-enhanced BOW (SEBOW), BoW, LDA(Latent Dirchilet Allocation), LSA(Latent Semantic Analysis) and BOW . The performance of best model (BoW) is highlighted in table two.

Table 2 provides a comparative summary between the different approaches in Latin cybe4rbullying detection as discussed above

## 5. Deep-Learning in Cyberbullying detection

Aauthors in [25] , proposed a novel algorithm to detect cyberbullying . The proposed algorithm is based on a convolution neural network (CNN) with semantics features by utilizing word embedding. It is to eliminate the needs for features extraction process. CNN-CB model consists of four layers: embedding, convolutional, max pooling and dense. Authors then applied the algorithm on dataset consist of about 39,000 English tweets .Authors then compared the accuracy result with the SVM classifier .Authors reported that the CNN-CB algorithm outperforms classical machine learning with accuracy 95 percent as shown in Table 3.

In [26], it is presented a novel approach to optimize Twitter cyberbullying detection based on deep learning (OCDD).The proposed approach eliminates the features extraction and selection phases . It's to preserve the semantics of words by replacing the tweet by a set of word vectors. Authors then fed it to a convolutional neural network (CNN) for classification phase along with metaheuristic optimization a algorithm for parameter tuning. This is to find the optimal or near optimal values.

In [27], it is proposed an aggregate approach for the two deep learning models. The first one is character-level convolutional neural network (CNN). It's for capturing the low-level syntactic knowledge from the character series. The second is word-level (long-term recurrent convolutional networks) LRCN. It's to capture semantic high-level information from word sequence, complementing the CNN model. Authors used dataset contains in total 8,815 comments from Kaggle. Authors reported that the hybrid model's sensitivity and accuracy are 0.5932 and 0.7081, respectively. Also, the aggregated approach is significantly enhanced the performance as well as outperformed other machine learning methods in cyberbullying detection.

In [28], it is proposed a novel pronunciation-based neural network (PCNN) for cyberbullying detection .The proposed model is to overcome the misclassification that produced from using misspelled words. Authors phoneme text codes as interface for a coevolution neural network. This technique is to correct spelling errors which did not change the pronunciation. Authors then fed it to CNF in order to detect cyberbullying. Authors compared the performance of models using two datasets, collected from Twitter (1313 tweets) and Formspring.me (13,000 messages). Authors also solved the problem of datasets balance with different techniques in order to compare the result between balanced and imbalanced datasets. Authors compared PCNN performance with previous work and found PCNN outperform the other methods. Authors reported that PCNN performed better when it is applied on the Twitter dataset than Formspring.me dataset. In addition, PCNN and CNN Random model performed better than CNN with pre-trained.

## 6. Discussion

This section will present the results according to the language and method in three separate sections , following the same order of study that followed in section four.

### *6.1. Cyberbullying detection in Arabic language.*

For Cyberbullying detection in Arabic language, the findings indicate that, SVM classifier is the most used classifier in the classification of Arabic text [6], [17]. Also, most of studies used Twitter platform as a source to collect the datasets[16]-[18]. For the surveyed papers, there was no unified dataset, and the maximum size of the dataset is 35273 tweets built by the writers in a study [17],and this size is small compared to the English dataset available. In addition, the results show that most common investigated feature is N-Grams particularly unigram and bigram [6], [18] . The common used method for measuring the accuracy are same to other languages (Precesion,Recall,F1). Furthermore, the highest accuracy achieved when authors used SVM classifier with Language SentiStrength Lexicon feature in [17] followed by NB classifier that is also investigated in the same study.Generally, it was found that deep learning algorithms have not been researched with the Arabic language as much as in English.

### *6.2. Cyberbullying detection in Latin language*

The results of this analysis concluded that out of seven papers analyzed, the SVM classifier was tested five times, followed by the

NB classifier that was examined three times. Also, the most of the datasets are collected from Twitter platform , there is still no unified dataset and the maximum size of the dataset is 25k tweets in [24]. In addition, the results show that most common investigated feature is Bigram, Trigram. The commonly used methods for measuring the accuracy are same to other languages (Precession, Recall, F1).Also, the highest accuracy is archived by using SVM classifier that represented with bigram, trigram,4-gram features in [21].

### *6.3. Deep-Learning in Cyberbullying detection*

After the deep systematic review, the results show that the CNN is the most common used method in the classification of cyberbullying. Also, CNN method had been investigated five times in different studies under different conditions . The results of review show also, that the most of datasets are collected from Twitter platform as well as there is no unified dataset and the maximum size of the dataset is 39,000 tweets, this is small to investigate the deep learning techniques. In addition for deep learning algorithms, the widely used approach for measuring the accuracy is the same for machine learning algorithms (Precession, Recall, F1). In addition, the highest accuracy is archived in [28] study when CNN is used, and outperformed the machine learning algorithms.

To conclude the result for both languages, the best accuracy was achieved when deep learning approaches were used particularly when CNN is applied in [28]. Also, deep learning algorithms were utilized more in the classification of cyberbullying in the English text, while machine learning used more with the Arabic text. Moreover, SVM is most common classifier in both Arabic and Latin languages and outperformed the other classifiers, it was examined seven times . Furthermore, N-Gram is the most widely used function for both Arabic and Latin languages with classifiers. The Twitter platform also primarily provides the origins of most of the datasets followed by FormSpring.me .

## 7. Conclusion

In this paper, we conducted an in-depth analysis of 16 studies on automatic cyberbullying detection methods based on text language. we undertook a deep review of evaluation methods, features, dataset size, language, and dataset source of the latest research in this field.We focused only on techniques that adopted neural network and machine learning algorithms. This is to direct future studies on this topic to a more consistent and compatible perspective on recent works, and to provide a practical and effective implementation for future systems. It was found that the best accuracy was achieved when a deep learning approach is used especially when CNN used. It was found also that, SVM is the most common classifier in both Arabic and Latin languages, and outperformed the other classifiers.Also, the most widely used feature is N-Gram especially bigram, trigram. In addition, Twitter is the main source for the collected datasets. Furthermore, there is no unified data sets. Although, cyberbullying prevention methods were adopted, largely, but most of the literature work aimed to enhance the detection by adding a new feature, as a number of features increased the process of features selection and extraction become more complicated. On the other hand, most of the work is done to find automated English language solutions, although each language actually has different structures and rules. In addition, there is no standard dataset and list of bad words to be counted as being used in the cyberbullying detection efforts. Finding an effective solution to detect cyberbullying helps a lot in protecting the targeted person.

## Conflict of Interest

The authors declare no conflict of interest.

## References

[1] J.W. Patchin, S. Hinduja, "Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying," Youth Violence and Juvenile Justice, **4**(2), 148–169, 2006, doi:10.1177/1541204006286288.

[2] T. Ivarsson, A.G. Broberg, T. Arvidsson, C. Gillberg, "Bullying in adolescence: Psychiatric problems in victims and bullies as measured by the Youth Self Report (YSR) and the Depression Self-Rating Scale (DSRS)," Nordic Journal of Psychiatry, **59**(5), 365–373, 2005, doi:10.1080/08039480500227816.

[3] P.W. Agatston, R. Kowalski, S. Limber, "Students' Perspectives on Cyber Bullying," Journal of Adolescent Health, **41**(6 SUPPL.), 59–60, 2007, doi:10.1016/j.jadohealth.2007.09.003.

[4] T. Beran, L.I. Qing, "Cyber-harassment: A study of a new method for an old behavior," Journal of Educational Computing Research, **32**(3), 265–277, 2005, doi:10.2190/8YQM-B04H-PG4D-BLLH.

[5] E.A. Abozinadah, A. V. Mbaziira, J.H.J. Jones, "Detection of Abusive Accounts with Arabic Tweets," International Journal of Knowledge Engineering-IACSIT, **1**(2), 113–119, 2015, doi:10.7763/ijke.2015.v1.19.

[6] A. Alakrot, L. Murray, N.S. Nikolov, "Towards Accurate Detection of Offensive Language in Online Communication in Arabic," Procedia Computer Science, **142**, 315–320, 2018, doi:10.1016/j.procs.2018.10.491.

[7] J.D. Marx, "Healthy communities: What have we learned and where do we go from here?," Social Sciences, **5**(3), 2016, doi:10.3390/socsci5030044.

[8] M.A. Campbell, "Cyber Bullying: An Old Problem in a New Guise?," Australian Journal of Guidance and Counselling, **15**(1), 68–76, 2005, doi:10.1375/ajgc.15.1.68.

[9] S. Nadali, M.A.A. Murad, N.M. Sharef, A. Mustapha, S. Shojaee, "A review of cyberbullying detection: An overview," International Conference on Intelligent Systems Design and Applications, ISDA, 325–330, 2014, doi:10.1109/ISDA.2013.6920758.

[10] Willard, "Parent Guide to Cyberbullying and Cyberthreats," 1–14, 2014.

[11] R.P. Ang, D.H. Goh, "Cyberbullying among adolescents: The role of affective and cognitive empathy, and gender," Child Psychiatry and Human Development, **41**(4), 387–397, 2010, doi:10.1007/s10578-010-0176-3.

[12] W. Cassidy, K. Brown, M. Jackson, "'Under the radar': Educators and cyberbullying in schools," School Psychology International, **33**(5), 520–532, 2012, doi:10.1177/0143034312445245.

[13] C. Emmery, B. Verhoeven, G. De Pauw, G. Jacobs, C. Van Hee, E. Lefever, B. Desmet, V. Hoste, W. Daelemans, "Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity," Language Resources and Evaluation, 2020, doi:10.1007/s10579-020-09509-1.

[14] J.M. Xu, K.S. Jun, X. Zhu, A. Bellmore, "Learning from bullying traces in social media," NAACL HLT 2012 - 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 656–666, 2012.

[15] R. Zhao, A. Zhou, K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," ACM International Conference Proceeding Series, **04-07-Janu**, 1–6, 2016, doi:10.1145/2833312.2849567.

[16] B. Haidar, M. Chamoun, A. Serhrouchni, "Arabic Cyberbullying Detection: Using Deep Learning," Proceedings of the 2018 7th International Conference on Computer and Communication Engineering, ICCCE 2018, 284–289, 2018, doi:10.1109/ICCCE.2018.8539303.

[17] B. Haidar, M. Chamoun, A. Serhrouchni, "A multilingual system for cyberbullying detection: Arabic content detection using machine learning," Advances in Science, Technology and Engineering Systems, **2**(6), 275–284, 2017, doi:10.25046/aj020634.

[18] H. Mubarak, K. Darwish, W. Magdy, "Abusive Language Detection on Arabic Social Media," 52–56, 2017, doi:10.18653/v1/w17-3008.

[19] H. Sanchez, "Twitter Bullying Detection," Homo, 2011.

[20] M. Di Capua, E. Di Nardo, A. Petrosino, "Unsupervised cyber bullying

detection in social networks," Proceedings - International Conference on Pattern Recognition, **0**, 432–437, 2016, doi:10.1109/ICPR.2016.7899672.

[21] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, A. Mohammed, "Social media cyberbullying detection using machine learning," International Journal of Advanced Computer Science and Applications, **10**(5), 703–707, 2019, doi:10.14569/ijacsa.2019.0100587.

[22] S.A. Özel, S. Akdemir, E. Saraç, H. Aksu, "Detection of cyberbullying on social media messages in Turkish," 2nd International Conference on Computer Science and Engineering, UBMK 2017, 366–370, 2017, doi:10.1109/UBMK.2017.8093411.

[23] S. Malmasi, M. Zampieri, "Detecting hate speech in social media," International Conference Recent Advances in Natural Language Processing, RANLP, **2017-Septe**, 467–472, 2017, doi:10.26615/978-954-452-049-6-062.

[24] T. Davidson, D. Warmsley, M. Macy, I. Weber, "Automated hate speech detection and the problem of offensive language," Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017, 512–515, 2017.

[25] M.A. Al-Ajlan, M. Ykhlef, "Deep learning algorithm for cyberbullying detection," International Journal of Advanced Computer Science and Applications, **9**(9), 199–205, 2018, doi:10.14569/ijacsa.2018.090927.

[26] M.A. Al-ajlan, "Optimized Twitter Cyberbullying Detection based on Deep Learning," 2018 21st Saudi Computer Society National Computer Conference (NCC), 1–5, 2018.

[27] S. Bu, S. Cho, A Hybrid Deep Learning System of CNN and LRCN to Detect Cyberbullying from SNS Comments, Springer International Publishing, 2018, doi:10.1007/978-3-319-92639-1.

[28] X. Zhang, J. Tong, N. Vishwamitra, E. Whittaker, J.P. Mazer, R. Kowalski, H. Hu, F. Luo, E. Dillon, "Based Convolutional Neural Network," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 740–745, 2016, doi:10.1109/ICMLA.2016.0132.