ASM Sc. J., 19, 2024

https://doi.org/10.32802/asmscj.2023.1448

Popular Cyberbullying Detection Methods on Social Media

Z. Lamghari[1,2*]

[1]*Department of Computer Science, Faculty of Sciences, Mohammed V University, Rabat, 10000, Morocco.*

[2]*Laboratory of Sciences, Engineering, and Management (LSEM), High School of Technology (EST), Sidi Mohamed Ben Abdellah University, Immouzer Road, Fez, 2427, Morocco.*

In contemporary times, the utilisation of social networks entails inherent risks, particularly on platforms such as Instagram, Twitter, and TikTok, which facilitate anonymous usage and the creation of fictitious profiles. The cloak of anonymity has fostered an ideal environment for cyberbullying, a form of harassment that unfolds in the virtual realm accessed by children and youth through information and communication technologies like the Internet and cell phones. Cyberbullying is pervasive across numerous countries globally. Consequently, a multitude of studies have delved into diverse solutions aimed at preventing cyberbullying, including various approaches and techniques for identifying and detecting it through the analysis of text language. This systematic review seeks to encompass these scientific studies that primarily focus on the classification of cyberbullying based on the language employed in online communication.

Keywords: cyberbullying; text mining; bullying victims; artificial intelligence; performance analysis

I. INTRODUCTION

The act of cyberbullying can also occur alongside harassment in school settings, such as in classrooms, playgrounds, and other areas. Harassment often involves the creation of online conversations, groups, or pages on social media platforms, specifically targeting an individual. Internet users then use this dedicated space to insult and demean the person (Perera & Fernando, 2021), providing an outlet for their anger. Like other forms of harassment, cyberbullying can have various consequences for the victim, including problems at school, such as dropping out or absenteeism, as well as impacts on their self-esteem, social skills, mental health, and even suicidal thoughts or behaviours. It's not just limited to teenagers, as even adults, including teachers, can also fall victim to cyberbullying (Jaber *et al.*, 2022). In fact, cyberbullying can even affect children as young as elementary school age. Studies have shown that teenagers are the most commonly targeted group for cyberbullying (Alotaibi & Mukred, 2022).

Despite the existence of laws in many countries aimed at protecting and assisting victims of bullying (Urbaniak *et al.*, 2022), bullying continues to be a persistent problem for many individuals. In cases where the victim or their parents do not report the bullying incident, the victim may continue to suffer while the perpetrator seeks out new victims. Therefore, identifying instances of bullying is crucial in order to implement effective remedies that protect the victim and hold the abuser accountable. However, manual monitoring of all posts on social media platforms is impractical due to the sheer volume of comments and information exchanged. As a result, several studies have focused on developing methods to promptly and accurately detect cyberbullying in order to mitigate its harmful effects (Refaee, 2021; Alakrot *et al.*, 2018; Marx, 2016; Jhaver *et al.*, 2022),

*Corresponding author's e-mail: zineb.lamghari@usmba.ac.ma

including the establishment of indicators to monitor and recognise potentially dangerous online behaviour.

We have compiled the latest publications on automatic detection of cyberbullying, with a particular focus on pattern recognition, neural networks, and deep learning techniques. Our systematic review aims to survey scientific studies that have addressed the classification of cyberbullying based on text language. The upcoming sections are organised as follows: Section 2 describes our methodology, Section 3 provides the theoretical background related to our study field, Section 4 presents the most recent research on cyberbullying detection, Section 5 documents techniques that utilise advanced Deep Learning algorithms for detecting cyberbullying on social media data, and finally, Section 6 summarises the entire paper.

To achieve the primary objective of this study, we developed research questions. The first question investigates the most commonly used dataset for cyberbullying classification, while the second question explores the popular amount and lexicon defined in existing datasets. The third question focuses on the most commonly used categorisation methods for addressing cyberbullying issues. The fourth question examines the quality metrics employed in existing data. The fifth question investigates the most successful techniques used for cyberbullying detection, and finally, the sixth question delves into the characteristics of the most commonly used classifiers.

II. RESEARCH METHODOLOGY

As shown in Figure 1, this section outlines the main steps of our review approach, which include performance assessment metrics and research methodology. Our study commenced in January 2022 and encompassed studies published from 2016 to 2021, as well as earlier studies, obtained from ACM, Science Direct, and IEEE databases. The articles selected underwent a rigorous criteria-based evaluation process to ensure their relevance and contribution to our objective. Our inclusion criteria consisted of: 1) Scientific papers with a focus on text-based cyberbullying classification; 2) Scientific papers utilising neural networks or machine learning for classification; 3) Scientific papers providing details on the model and its performance indicators; and 4) Scientific papers citing the dataset's size and language. Conversely, the exclusion criteria were: 1) Scientific papers not suitable for text classification; 2) Scientific papers lacking mention of findings' veracity; 3) Papers not published in a journal or presented at a conference; 4) Papers that did not utilise neural networks or machine learning for classification; and 5) Scientific papers failing to mention the dataset's size or language.

We gathered all prior studies using the following keywords combined with the “Cyberbullying” keyword:

- “classification”.
- “Neural networks”.
- “text”.
- “Deep learning”.
- “Detection”.
- “Machine learning”.
- “Categorization”.
- “Classification”.
- “Text mining”.

Finally, we developed standard evaluation questions to serve as a guideline for the study and to confirm that it met the goal of this retrospective analysis.

- Q1: How thoroughly was the corpus identified and described (in terms of size and language)?
- Q2: Was the text classification technique precisely defined??
- Q3: Was the method's output easily outlined?

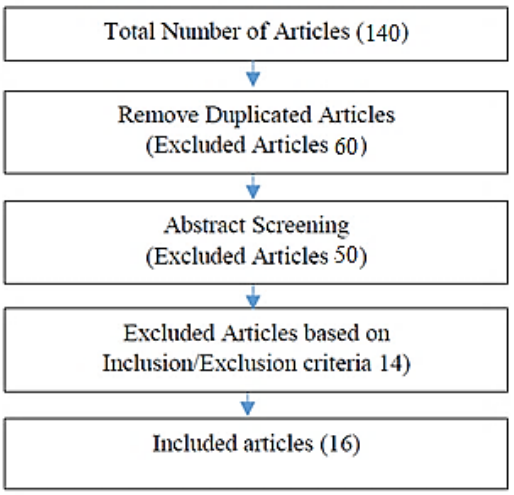


Figure 1. Data sources

III. CYBERBULLYING BACKGROUND

Cyberbullying, a well-known issue, has been linked to emotional disturbances as early as 2005 by researchers (Jaber *et al.*, 2022). These symptoms, including aggression, delinquency, despair, depression, self-destructive behaviour, personality issues, and self-harm, may pose risks associated with psychopathology, as noted by the authors. Cyberbullying, in particular, has significant impacts that can result in major and long-term consequences, especially for teenagers, who are often the victims. Studies by Urbaniak *et al.* (2022), and Alotaibi and Mukred (2022) have found that numerous victims of cyberbullying have attempted suicide due to the abusive, demeaning, and violent messages they receive.

According to Jhaver *et al.* (2022) and Willard (2007), cyberbullying can take on various forms, including:

1. Flooding: This type of cyberbullying involves the aggressor repeatedly making irrelevant comments to prevent the victim from participating in a dialogue.
2. Masquerade: The aggressor uses a fake profile with false information to bully a user in a forum, chat room, or social media platform, with the aim of degrading the victim's reputation.
3. Flaming: Involves two or more people personally attacking each other in a heated, short-lived disagreement, with all posts containing bullying language.
4. Trolling: This type of cyberbullying entails purposely posting opinions that align with other inflammatory remarks on an emotionally charged topic, with the goal of inciting conflict, even if the comments do not reflect the true beliefs of the poster.
5. Harassment: This type of cyberbullying closely resembles traditional bullying, as it involves sending frequent abusive texts to the victim over an extended period of time.
6. Cyberstalking and Cyber threats: This type of cyberbullying involves sending communications that contain threatening or excessively hostile remarks, as well as extortion.
7. Denigration: This type of bullying includes spreading vulgar, derogatory stories or making up false information about others and sharing it in a discussion forum, group chat, or website.
8. Outing: This type of cyberbullying involves the publication of sensitive, intimate, or humiliating material in a community chat or discussion board, with the aim of humiliating the victim. It is similar to denigration but usually requires a strong relationship between the abuser and the victim.
9. Exclusion: This type of cyberbullying involves deliberately ignoring or excluding certain individuals from discussions or interactions, particularly among young people or teenagers.

Due to the strong association between cyberbullying and various negative human behaviours, psychological research has extensively explored the interaction of cyberbullying with different factors in an effort to prevent cyberbullying. A study by Wang *et al.* (2021) found that showing compassion and building a strong relationship between caregivers and teenagers can have a positive impact on mitigating cyberbullying. Similarly, research reported by Zhang *et al.* (2022) demonstrated that increasing awareness about cyberbullying concerns plays a crucial role in preventing cyberbullying.

Moreover, many countries have recognised cyberbullying as a criminal act and have implemented regulations to address and punish offenders. For instance, in the United Arab Emirates, individuals are encouraged to report incidents of bullying to competent authorities as a legislative measure. These efforts aim to prevent individuals, particularly children and teenagers, from engaging in cyberbullying and to support cyberbullying victims in coping with the detrimental consequences of cyberbullying.

However, despite the growing development of preventive measures, the availability of high-quality data limits the utilisation of advanced approaches in current research. Many studies rely on small and diverse datasets without a thorough evaluation of their applicability, as highlighted by Emmery *et al.* (2021).

The most effective way to detect cyberbullying in online messages is through the use of device-based automated advanced algorithms. These algorithms analyse information and generate reports when cyberbullying is detected, allowing for quick identification and intervention with minimal losses. Techniques, such as machine learning,

natural language processing (NLP), and deep learning (DL) are commonly used for cyberbullying identification. Desai *et al.* (2021) employed various NLP models, such as Bag of Words (BoW), Latent Dirichlet Allotment (LDA), and Latent Semantic Analysis (LSA), to identify bullying in social networks.

Furthermore, autodetection algorithms often utilise word encoding mechanisms, where a list of predefined objectionable phrases is extended and given different weights to identify bullying and latent features (Al-Marghilani, 2022).

IV. CYBERBULLYING DETECTION CATEGORIES

In this section, we present the most significant studies on cyberbullying detection, categorised based on the language used (Arabic or Latin) and the attributes and classifications employed in each case.

A. Arabic Language Recognition

The authors of the study conducted by Alakrot *et al.* (2018) investigated the impact of additional normalisations on the performance of a highly skilled Support Vector Machine (SVM) classifier for detecting offensive remarks. They used word-level and N-gram capabilities along with typical preprocessing algorithms. The dataset used in the study consisted of 15,050 comments from the YouTube site, which were collected from comments about notable Arabic celebrities and is widely available. The authors found that stemming and pre-processing phases improved the recognition of inappropriate content in ordinary Arabic text. Furthermore, the addition of N-gram capabilities further enhanced the efficiency of the classifier. However, it was observed that combining stemming and N-gram capabilities reduced precision and recall metrics. Therefore, the authors concluded that matching the preprocessing stemming and N-gram was the most effective technique, as it resulted in values between 1 and 5.

The study conducted by Elzayady *et al.* (2022) was the first to apply deep learning techniques for detecting abuse in the Arabic language. They used the current dataset with minor modifications, as demonstrated in a previous study by Ahmed *et al.* (2022). The dataset was tokenised into words to remove unnecessary characters before developing the model's layers, resulting in Word Embedding. The dataset was then partitioned using the Pareto concept, with 80 percent for the training phase and 20 percent for the test phase. For the training phase, the authors used a Feed Forward Neural Network (FFNN). The authors reported an accuracy of 92 percent, with a validation accuracy of 95 percent for the seven proposed layers in the neural network. However, they achieved an accuracy of 96 percent with three concealed layers, as shown in Figure 2.

Ahmed *et al.* proposed an approach for identifying cyberbullying in typed Arabic and English text. The first step involved applying text analysis techniques to analyse the tweet content, as well as language analysis techniques specific to English and Arabic languages. In the second stage, the researchers utilised a standard tweet package, specifically the TweetToSentiStrength Feature Vector filter. SentiStrength was used to evaluate the tweets, with positive sentiments indicated by an interval between two and five, negative sentiments represented by an interval between -2 and -5, and neutral sentiments represented by (1, -1). In this study, SentiStrength initially used English lexicon files, which were later replaced with Arabic files that included weighted vulgar terms created by Elzayady *et al.* for this purpose. Two datasets were used, one gathered from the Facebook platform weighing over 1 GB for validation, and the second obtained from Twitter to evaluate the algorithm. The dataset with the Arabic language included tweets from various dialects in Lebanon, Syria, the Gulf region, and Egypt, totalling 35,733 unique tweets after removing duplicates. The authors demonstrated that the precision of the two classifiers differed for the "yes" class, with SVM showing significantly higher accuracy. However, the overall system precision was over 93.56% for SVM and about 90% for NB, as shown in Figure 2.

The researchers in the study conducted by Mubarak *et al.* (2017) utilised a predefined set of vulgar terms as word stems, which were used to generate a larger list from a massive database of 186,000,000 tweets. This list was compiled to identify inappropriate and venomous language. The authors then created a new list of 3,000 words using the Log Odd Ratio (LOR) approach on both unigrams and bigrams. The authors evaluated the detection of

objectionable language using five methods, including Seeds Words (SW), SW+LOR (unigram), SW+LOR (bigram), LOR (unigram), and LOR (bigram). The highest accuracy was achieved by applying the LOR-generated list based on unigrams. The datasets used in the study, along with the list of vulgar terms and hashtags, were provided by the authors for research purposes and are depicted in Figure 2.

Figure 2 presents a comparison of the various methodologies in Arabic cyberbullying detection outlined previously (Numbers in the graph are written according to the French format).

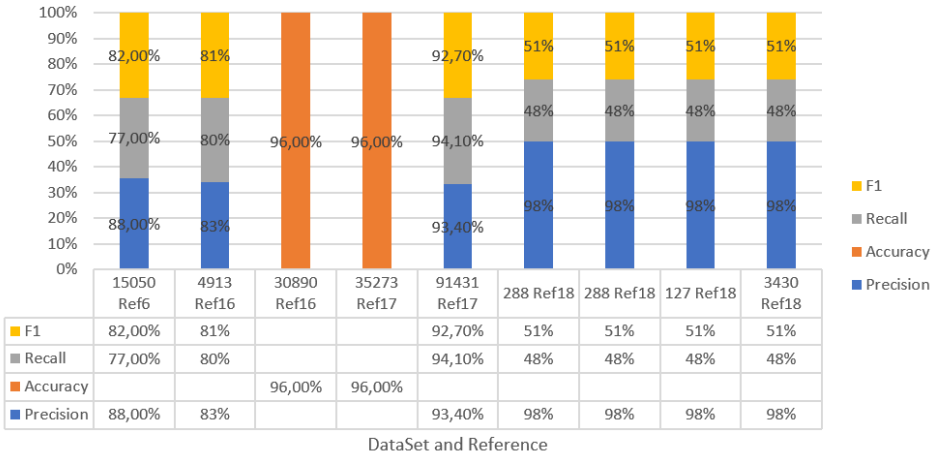


Figure 2. Performance analysis of cyberbullying detection in Arabic (social media)

B. Detection in Latin Language

The researchers in the study conducted by Sanchez and Kumar (2011) utilised sentiment analysis and the Naive Bayes classifier (NB) to identify instances of cyberbullying in a social network. They worked with a balanced sample of 5000 English tweets and gathered texts that contained specific phrases such as "Gay," "Homo," "Dike," and "Queer." These phrases were used as training data, where the presence of the term "queer" was used to categorise positive tweets, while the presence of "Gay," "Homo," or "Dike" was used to classify negative tweets. As a result, the NB classifier achieved an accuracy of 67.3 percent.

The researchers in the study conducted by Shetty *et al.* (2022) devised an unsupervised approach for identifying cyberbullying event logs on social media platforms. They utilised a growing hierarchical self-organising map as the basis of their model. The model incorporated machine learning techniques such as decision tree C4.5, support vector machine (SVM), and Naive Bayes (NB), as well as natural language processing (NLP) approaches that focused on semantic and syntactic aspects of textual phrases. The performance of the proposed model was evaluated on three distinct datasets and platforms, as depicted in Figure 3, which illustrates the range of performance for each classifier used in the study.

The researchers in the study conducted by Jacob *et al.* (2022) utilised a supervised technique to detect instances of cyberbullying. They employed various machine learning classifiers, TFIDF (Term Frequency-Inverse Document Frequency), and sentiment analysis techniques for feature extraction. The classifications were tested using different n-gram language models. The authors found that a neural network with three grams achieved a higher accuracy of 92.8 percent compared to an SVM with four grams, which had a lower accuracy of 90.3 percent. Additionally, in another study, the neural network outperformed other classifiers on the same dataset. The dataset used in the study was obtained from Kaggle (Formspring.me) and consisted of 1608 instances in English, classified into two categories: cyberbullying and non-cyberbullying, with 804 instances in each class. Figure 3 presents the average performance rates in the study conducted by Iqbal *et al.* (2022), a supervised technique was employed to identify bullying and harassment in Turkish language communications. Feature selection was done using information gain and chi-square approaches.

Kelly Reynolds *et al.* used the same labelled dataset obtained from Kaggle (Instagram and Twitter) as the authors. They evaluated the accuracy and running time of various machine learning classifiers, including SVM, Decision Tree (C4.5), Nave Bayes Multinomial, and K Nearest Neighbours (KNN). The authors compared the accuracy of classifiers under different conditions and found that the NBM classifier was the most efficient before implementing feature selection, while the IBK classifier was the most efficient after selecting 500 features. These findings are presented in Figure 3.

The authors of (Yuan & Lam, 2021) applied different text classification techniques to differentiate between hate speech, profanity expressions, and other types of texts. They utilised classical lexical features and a linear SVM classifier to identify the main concept. Three types of extracted features, namely surface n-grams, word skip-grams, and Brown clusters, were used with a linear SVM classifier. The character four-gram model yielded the highest accuracy, with approximately 78 percent.

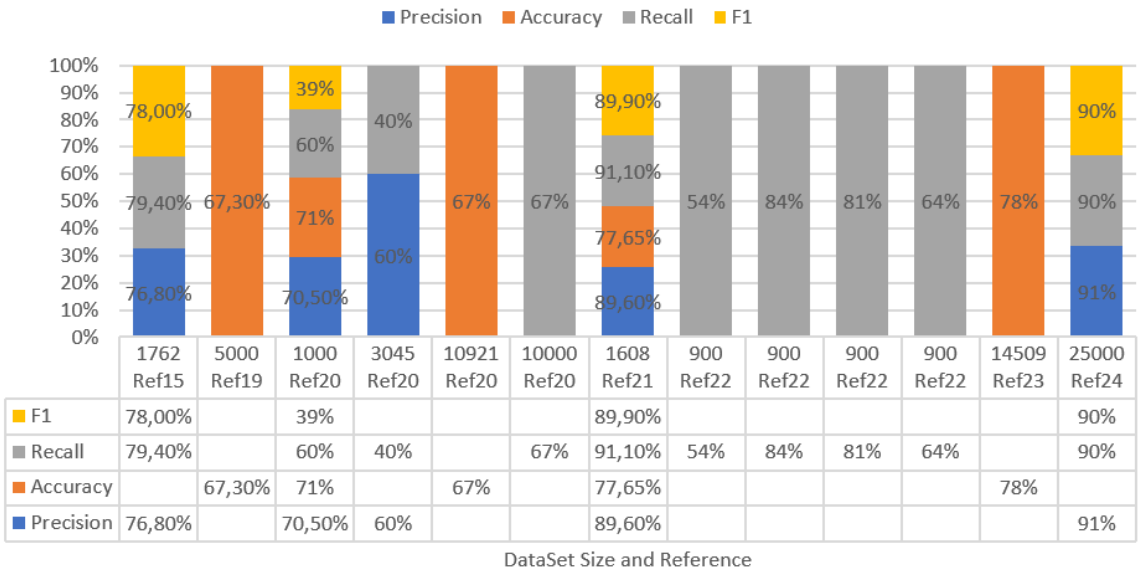
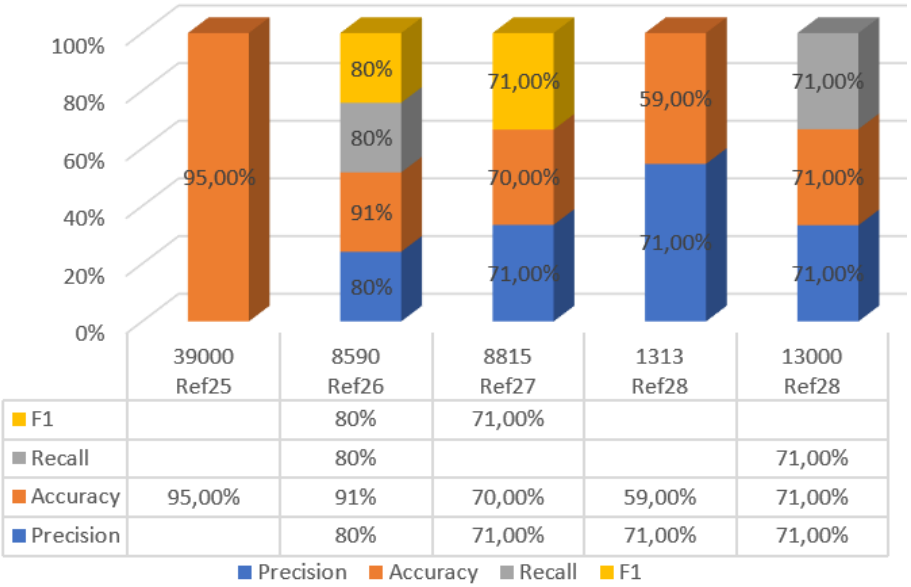


Figure 3. Performance analysis of cyberbullying detection in Latin America (social media)

In (Pamungkas *et al.*, 2023), the authors collected tweets that contained hate speech keywords using a hate speech lexicon. These tweets were then classified into three types of speech, namely hate speech, offensive speech, and normal speech. To differentiate between these categories, the authors trained a multi-class classifier. They used various features, including bigrams, unigrams, and trigrams from each tweet, as well as characteristics such as the number of characters, words, and syllables. Binary and count indicators for hashtags, mentions, retweets, and URLs were also included as features. The authors performed 5-fold cross-validation to test multiple models and found that Logistic Regression and Linear SVM models outperformed other models. The final model was built using logistic regression with L2 regularisation, and it was trained by predicting the identifier of each trace across the entire dataset. The best performance of the final model is highlighted in Figure 3.

Al-Marghilani introduced Embedding Enhanced Bag-of-Words (EEBOW), a new learning method for cyberbullying recognition. EEBOW combines Bag of Words (BoW) features, latent semantic features, and bullying features. Word embedding, which captures the semantic details of words, is utilised to derive characteristics related to bullying. The authors claim that EEBOW performs significantly well compared to other models such as semantic-enhanced BoW (SEBOW), BoW, Latent Dirichlet Allocation (LDA), and Latent Semantic Analysis (LSA). Figure 3 provides a summary of the best-performing model (BoW), as well as a comparison of various approaches in cyberbullying detection discussed in the current sub-section.



	39000 Ref25	8590 Ref26	8815 Ref27	1313 Ref28	13000 Ref28
F1		80%	71,00%		
Recall		80%			71,00%
Accuracy	95,00%	91%	70,00%	59,00%	71,00%
Precision		80%	71,00%	71,00%	71,00%

Figure 4. Performance analysis of cyberbullying detection using deep learning algorithms (social media)

V. DEEP LEARNING FOR DETECTING

In (Al-Ajlan *et al.*, 2018a), the proposed approach utilises word embedding to construct a convolutional neural network (CNN) that incorporates semantic information, eliminating the need for a separate feature extraction process. The CNN-CB model consists of four layers, including embedding, segmentation, pooling, and dense layers. The authors applied this technique to a dataset of approximately 40,000 English tweets. The accuracy of the CNN-CB algorithm was then compared to that of an SVM classifier by the authors. According to the authors, the CNN-CB algorithm outperforms classical machine learning approaches, achieving an accuracy rate of 95 percent (as shown in Figure 4).

In (Al-Ajlan *et al.*, 2018b), the author proposed a novel deep learning-based approach to enhance Twitter cyberbullying detection. The method aims to reduce the need for feature extraction and screening processes by preserving word meaning through word clustering and vector representation. The data is then fed into a convolutional neural network (CNN) for classification, along with meta-heuristic optimisation for parameter tuning, to select optimal or closely related values (as depicted in Figure 4).

In (Pericherla & Ilavarasan, 2021), Bu and Cho proposed an innovative approach that utilises two different deep learning models for cyberbullying detection. As reported in (Pericherla & Ilavarasan, 2021), the first model is a character-level convolutional neural network (CNN) designed to capture low-level grammar knowledge from character sequences. The second model is a word-level Long-Term Recurrent Convolutional Network (LRCN) aimed at extracting high-level semantic information from word sequences, thus complementing the CNN model. Bu and Cho used a dataset of over 9K comments from Kaggle for their experimentation. According to the authors, the hybrid approach achieved an awareness rate of 60% and an accuracy rate of 71%. Furthermore, the proposed agglomerated strategy significantly improved the efficiency and outperformed previous machine learning algorithms in the field of cyberbullying detection (as shown in Figure 4). In (Sadr *et al.*, 2022), Zhang *et al.* proposed a unique neural network model called Pronunciation-based Convolutional Neural Network (PCNN) for cyberbullying detection. The aim of this model is to overcome misinterpretation caused by improper phrases in cyberbullying texts. Phonology is utilised as an interface in a co-evolution neural network by the authors to correct spelling errors that do not affect pronunciation. The corrected text is then passed to the PCNN model for cyberbullying recognition. The authors conducted a comparative study to evaluate the efficiency of their model using unique data retrieved from Twitter (14K tweets) and Formspring.me (14K messages) datasets. They also addressed the issue of dataset balance using various methodologies and evaluated the results of balanced and

unbalanced datasets. Additionally, the authors compared the performance of PCNN with earlier studies and found that PCNN outperformed other approaches, particularly in the context of tweets compared to the Formspring.me dataset. Moreover, PCNN and CNN were considered as random models and outperformed CNN with pre-trained models, as shown in Figure 4.

VI. CONCLUSION

In our paper, we conducted a comprehensive analysis of the latest research in the field of cyberbullying analysis techniques, including characteristics, dataset size, dialect, and dataset repository. We also developed exemplary studies in order to provide a more comprehensive and aligned perspective on popular approaches. Our focus was on techniques that utilise algorithms and automation technologies. Through systematic queries and application of nomination criteria, we identified sixteen unique works for analysis. We observed that the CNN approach, specifically deep learning using CNN, consistently achieved the highest accuracy among the approaches studied, with five instances of its application in various settings. Our review also revealed that the majority of datasets used in these studies are derived from the Twitter platform, with no solitary samples as datasets, and the largest dataset size being 39K tweets, which may be insufficient for exploring deep learning techniques.

SVM was found to be the most commonly used classifier in both Arabic and Latin languages, outperforming other classifiers. N-gram, specifically bigram and trigram, was identified as the most commonly used feature. Additionally, the findings revealed that Twitter is the primary source of datasets used in these studies, and there are no harmonised datasets available. The standard method for determining accuracy, using precision, recall, and F1, is consistent with other languages. In a study by Ahmed *et al.* (2022), the authors achieved the highest accuracy by progressively using three techniques: SVM classifier, Language SentiStrength Lexicon feature, and NB classifier. It was also observed that deep learning algorithms have received less attention in Arabic datasets compared to English datasets, and there are fewer studies on identifying cyberbullying in Arabic literature compared to English literature. The maximum size of the dataset used in this study, reported by Ahmed *et al.* (2022), was about 36K tweets, which had limited information compared to the available English datasets.

In the evaluation of seven research papers on cyberbullying detection in Latin, the SVM classifier was suggested as the optimal choice five times, followed by the NB classifier, which was tested three times. The majority of datasets used in these studies were obtained from the Twitter platform, and there is currently no standardised dataset available. The largest dataset size reported in a study by Pamungkas *et al.* (2022) was approximately 25,000 tweets. The most frequently researched characteristic was found to be Bigram and Trigram. The methods for assessing accuracy, such as F1, Recall, Accuracy, and Precision, were consistent with those used in other languages. Additionally, the highest accuracy was achieved by utilising an SVM classifier with bigram, trigram, and four-gram features.

Although preventive techniques for cyberbullying are widely accepted, a considerable portion of the literature focuses on enhancing the detection of cyberbullying by introducing new performance analysis or evaluation metrics. However, as the number of features increases, the selection and extraction phases become increasingly challenging.

However, it is worth noting that most of the research in this field is primarily focused on developing automatic solutions for the English language, despite the fact that each language has its own unique structure and laws. Additionally, there is a lack of standardised datasets or collections of offensive sentences that can be used for cyberbullying detection. Therefore, establishing an effective solution for identifying cyberbullying could greatly contribute to protecting individuals who are victims of bullying.

VII. REFERENCES

Ahmed, M, Rahman, M, Nur, S, Islam, AZM & Das, D 2022, 'Introduction of PMI-SO Integrated with Predictive and Lexicon Based Features to Detect Cyberbullying in Bangla Text Using Machine Learning', in Algorithms for Intelligent Systems: Proceedings of the 2nd International Conference on Artificial Intelligence Advances and Applications, India, 14 February 2022, researchers and practitioners in academia and industry, Jaipur.

Al-Ajlan, MA & Ykhlef, M 2018a, 'Deep learning algorithm for cyberbullying detection', International Journal of Advanced Computer Science Application, vol. 9, no. 9, pp. 199-205.

Al-Ajlan, MA & Ykhlef, M 2018b, 'Optimized twitter cyberbullying detection based on deep learning', in Computer Science: Proceeding of the 21st Saudi Computer Society National Computer Conference, Riyadh, 25 April 2018, Computer Science, Saudia Arabia.

Alakrot, A, Murray, L & Nikolov, NS 2018, 'Towards accurate detection of offensive language in online communication in Arabic', Procedia computer science, vol. 142, pp. 315–320.

Alotaibi, NB & Mukred, M 2022, 'Factors affecting the cyber violence behavior among Saudi youth and its relation with the suiciding: A descriptive study on university students in Riyadh city of KSA', Technology in Society, vol. 68, pp. 1-13.

Al-Marghilani, A 2022, 'Artificial Intelligence-Enabled Cyberbullying-Free Online Social Networks in Smart Cities', International Journal of Computational Intelligence Systems, vol. 15, no. 1, pp. 1-13.

Desai, A, Kalaskar, S, Kumbhar, O & Dhumal, R 2021, 'Cyber Bullying Detection on Social Media using Machine Learning', in Proceeding of the International Conference on Automation, Computing and Communication, 16 December 2020, computing.

Elzayady, H, Mohamed, MS, Badran, KM & Salama, GI 2022, 'Detecting Arabic textual threats in social media using artificial intelligence: An overview', Indonesian Journal of Electrical Engineering and Computer Science, vol. 25, no. 3, pp. 1712-1722.

Emmery, C, Verhoeven, B, De Pauw, G, Jacobs, G, Van Hee, C, Lefever, E & Daelemans, W 2021, 'Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity', Language Resources and Evaluation, vol. 55, no. 3, pp. 597-633.

Iqbal, I, Atay, T & Savitskaya, A 2022, 'Digital Literacy Gender Gap in E-Education Through Social Media During the COVID-19 Lockdown in Pakistan and Turkey', Research on Digital Citizenship and Management During Crises, vol. 2022, pp. 249-270.

Jaber, LS, Rinaldi, CM, Saunders, CD & Scott, J 2022, 'The Intent Behind Bullying: An Application and Expansion of the Theory of Planned Behaviour', Contemporary School Psychology, vol. 27, pp. 1-15.

Jacob, RP, Manoj, K, Mohan, D, Issac, S & Sudarsan, D 2022, 'Cyberbullying Detection and Prevention Using Artificial Intelligence', Soft Computing for Security Applications, vol. 181, pp. 905-914.

Jhaver, S, Chen, Q Z, Knauss, D & Zhang, A 2022, 'Designing Word Filter Tools for Creator-led Comment Moderation', in Human Factors in Computing System: Proceeding of the CHI Conference on Human Factors in Computing Systems, New York, 29 April 2022, Association for Computing Machinery, United States.

Marx, JD 2016, 'Healthy communities: What have we learned and where do we go from here?', Social Sciences, vol. 5, no. 3, pp. 44.

Mubarak, H, Darwish, K & Magdy, W 2017, 'Abusive language detection on Arabic social media', in Proceedings of the first workshop on abusive language online, August 2017, Vancouver, Association for Computational Linguistics, Canada.

Pamungkas, EW, Basile, V & Patti, V 2023, 'Investigating the role of swear words in abusive language detection tasks', Language Resources and Evaluation, vol. 57, pp. 155-188.

Perera, A & Fernando, P 2021, 'Accurate cyberbullying detection and prevention on social media', Procedia Computer Science, vol. 181, pp. 605-611.

Pericherla, S & Ilavarasan, E 2021, 'Cyberbullying detection on multi-modal data using pre-trained deep learning architectures', Revista Ingeniería Solidaria, vol. 17, no. 2, pp. 1-20.

Refaee, EA 2021, 'Data-oriented Approach for Detecting offensive Language in Arabic Tweets', in Proceeding of the International Conference on Software Engineering & Computer Systems and 4th International Conference on

Computational Science and Information Management, 26 August 2021, Pekan, IEEE, Malysia.

Sadr, H & Nazari Soleimandarabi, M 2022, 'ACNN-TL: attention-based convolutional neural network coupling with transfer learning and contextualized word representation for enhancing the performance of sentiment classification', The Journal of Supercomputing, vol. 78, pp. 1-27.

Sanchez, H & Kumar, S 2011, Twitter bullying detection. Ser NSDI, Twitter bullying detection. Ser'. NSDI, vol. 12, no. 15.

Shetty, S, Dsouza, J, Rodrigues, A & Shetty, M 2022, 'Cyberbullying Detection in Native Languages', Soft Computing for Security Applications', vol. 1397, pp. 749-763.

Urbaniak, R, Ptaszyński, M, Tempska, P, Leliwa, G, Brochocki, M & Wroczyński, M 2022, 'Personal attacks decrease user activity in social networking platforms', Computers in Human Behavior, vol. 126, pp. 1-20.

Wang, X, Qiao, Y, Li, W & Dong, W 2021, 'How is Online Disinhibition Related to Adolescents' Cyberbullying Perpetration? Empathy and Gender as Moderators', The Journal of Early Adolescence, vol. 42, no. 5, pp. 704-732.

Willard, NE 2007, 'Parent guide to cyberbullying and cyberthreats. From Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress', Oregon: Center for Safe and Responsible Internet Use.

Yuan, Y & Lam, W 2022, 'Sentiment Analysis of Fashion Related Posts in Social Media', in Electronic Resource: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 15 February 2022, Arizona, ACM, USA.

Zhang, XC, Chu, XW, Fan, CY, Andrasik, F, Shi, HF & Hu, XE 2022, 'Sensation seeking and cyberbullying among Chinese adolescents: Examining the mediating roles of boredom experience and antisocial media exposure', Computers in Human Behavior, vol. 130, no. 107185.