IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12681

Cyberbullying detection systems: a survey on methodologies and challenges

Anirudh Raj[1], Vandana R[2], Anitha H M[3]

Student, Information Science and Engineering, BMS College of Engineering, Bangalore, India[1]

Student, Information Science and Engineering, BMS College of Engineering, Bangalore, India[2]

Professor, Information Science and Engineering, BMS College of Engineering, Bangalore, India[3]

Abstract: Cyberbullying has emerged as a pervasive and harmful phenomenon in the age of technology, affecting individuals of all ages and across various online platforms. As the prevalence of cyberbullying continues to grow, the need for effective detection systems becomes crucial to protect and support victims. A comprehensive survey on methodologies and challenges related to cyberbullying detection systems is presented in the paper. The survey explores a diverse array of techniques and approaches employed in cyberbullying detection, including machine learning, natural language processing, social network analysis, and sentiment analysis. Various data sources and features utilised for detection are examined, such as text-based content, user behaviour patterns, and social interactions. Additionally, the paper discusses the challenges faced by cyberbullying detection systems, such as the evolving nature of cyberbullying tactics, the contextual complexity of online interactions, and the ethical considerations surrounding privacy and bias. The study expands on prospective topics for further research and development and points out the shortcomings of the approaches now in use.

Keywords: cyberbullying detection systems, social network analysis, sentiment analysis, natural language processing

I. INTRODUCTION

In today's digital era, the prevalence of cyberbullying has become a pressing concern, impacting individuals across various online platforms. The detrimental effects of cyberbullying necessitate the development of robust and effective detection systems to combat this growing problem. This survey paper aims to provide a comprehensive analysis of methodologies and challenges related to cyberbullying detection systems. This study's main goal is to examine the various approaches implemented in the domain of cyberbullying detection. By examining techniques such as NLP, machine learning, social network analysis, and sentiment analysis, we seek to gain insights into the diverse approaches used to identify instances of cyberbullying. NLP algorithms are utilised in classifying aggressive phrases from non-defamatory online messages and the general pattern of cyberbullying messages are utilised to train a model to predict the possibility of an aggressive or obscene comment. Other methodologies involve utilising sentiment analysis and social network analysis to identify underlying patterns with respect to frequency and timing of messages as well as user activity on the platform. Overall, the development of effective cyberbullying detection methods is an active area of research, and many different approaches are being explored. There is no one-size-fits-all solution, and different methods may be more or less effective depending on the specific context and type of online behaviour being targeted.

II. MOTIVATION

There's a glaring need for the cyberbullying detection systems to be implemented in all forms of digital platforms that enable their users to socialise and communicate with one another. Cyberbullying detection is crucial for the following reasons:

- **Rising prevalence of cyberbullying:** With the widespread use of digital platforms and social media, incidents of cyberbullying have increased significantly. The need for detection systems arises to address this growing problem.
- **Protection of victims:** Cyberbullying can have severe psychological, emotional, and even physical consequences for the victims. Detection systems can help identify and intervene in such instances, providing support and protection to those affected.
- **Prevention of harm:** Early detection of cyberbullying allows for timely intervention, preventing further harm to victims. Detection systems can help identify patterns and indicators of cyberbullying, enabling effective preventive measures.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023
DOI: 10.17148/IJARCCE.2023.12681

- **Safer online environment:** By implementing cyberbullying detection systems, online platforms can create a safer and more inclusive environment for users. This encourages positive online interactions and discourages the perpetuation of cyberbullying behaviours.
- **Educational institutions' responsibility:** Schools and educational institutions have a duty to protect their students from cyberbullying. Detection systems can assist in monitoring online activities, promoting a safe and conducive learning environment.
- **Legal and policy requirements:** Many jurisdictions have implemented laws and regulations addressing cyberbullying. Detection systems can aid in complying with these requirements and ensuring adherence to policies that safeguard against cyberbullying.
- **Parental involvement and awareness:** Cyberbullying detection systems can also help parents stay informed about their children's online activities and provide necessary guidance and support in cases of cyberbullying.
- **Research and analysis:** Systems for detecting cyberbullying produce useful data that may be utilised in research and analysis. This information can provide light on the frequency, trends, and effects of cyberbullying, which can help to improve preventive and intervention methods.

III. LITERATURE SURVEY

Maral Dadvar and Franciska de Jong [1] proposed that the incorporation of the users‘ information, their characteristics, and post harassing behaviour, alongside the content of their conversations, will enhance the accuracy and efficiency of cyberbullying detection. They investigated cyberbullying detection from two perspectives. First, which is the conventional way, the users‘ behaviour will be considered only in one environment, for instance, the user‘s comments on a video on YouTube.They envisioned an algorithm that would go through the comments‘ text and would classify them as either bullying or non bullying. At this phase of the experiment, they hypothesised that including the users‘ characteristics – either the bully or the victim - such as age and gender, will improve the detection accuracy.They looked into the gender-based method of detecting cyberbullying on MySpace, where gains in classification are seen. According to their investigation, author data can be used to increase the identification of inappropriate activity in social media profiles.

In the paper ‘Detecting A Twitter Cyber bullying Using Machine Learning‘ done by Rahul Ramesh Dalvi,Sudhanshu Baliram Chavan,Aparna Halbe [2], a machine learning model is proposed to detect and prevent bullying on Twitter.Two classifiers - SVM and Naive Bayes are utilised for training as well as testing the social media bullying content.Both Naive Bayes and SVM (Support Vector Machine) were able to detect the true positives with 71.25% and 52.70% accuracy,respectively.It shows that SVM overpowers Naive Bayes of similar work on the same data set.

In the paper published by Michele Di Capua,Emanuel Di Nardo, Alfredo Petrosino [3], a model of cyber-bullying aggression is proposed, based off of a hybrid set of features, starting with classical textual features but also dependant on the so-called ―social features‖.These features are related to social behaviour and their peculiarities are strictly related to the social platform ana- lyzed.They used data from formspring which got the accuracy of 73% and for the YouTube dataset, acquired the accuracy of 69%.

RULE BASED LEARNING AND BAG OF WORDS MODELLING:

Using data collected from Formspring.me, a social media site to ask one on one questions, a rule based learning approach as well as a bag of words model approach was proposed to detect aggression in the phrases by Reynolds, Kontostathis, and Edwards [4]. The presence of bad words and anonymity were used to generate a ruleset that helped in detecting bullying messages, as they considered that anonymity may promote or leverage the user‘s tendency to bully or harass. The percentage of false positives generated was a major drawback despite the model successfully detecting several instances of cyberbullying.

1. Rule-Based Machine Learning:

Rule-based machine learning is an approach that uses predefined rules or patterns to make predictions or classify data. These rules are typically created manually by domain experts or derived from prior knowledge. The rules consist of a set of conditions or logical statements that determine the outcome or decision. In the context of detecting cyberbullying, rule-based machine learning involves defining specific rules or patterns that indicate the presence of cyberbullying behaviour. These rules can be based on keywords, phrases, linguistic patterns, or other characteristics commonly associated with cyberbullying. For example, a rule might state that if a tweet contains explicit profanity or personal attacks, it is classified as cyberbullying. Rule-based machine learning has the advantage of being interpretable and easy

DOI: 10.17148/IJARCCE.2023.12681

to understand since the rules are explicitly defined. However, it may require manual effort to create and maintain these rules, and it might not effectively capture more complex patterns or nuances in cyberbullying behaviour.

2. Bag-of-Words Approach:

The bag-of-words approach is a popular technique used in natural language processing (NLP) for analysing text data. It represents a document (in this case, a tweet) as a "bag" or collection of words, disregarding grammar and word order but considering their frequency. The bag-of-words approach involves creating a dictionary or vocabulary of words that are relevant to cyberbullying. This dictionary includes both offensive or abusive words and contextual words that might indicate cyberbullying behaviour. Each tweet is then represented as a vector or a numerical representation indicating the frequency or presence of words from the dictionary.

ML algorithms, such as Naive Bayes or Support Vector Machines (SVM), can be trained using these bag-of-words representations to classify tweets as either cyberbullying or non-cyberbullying.The bag-of-words approach allows for automatic feature extraction and can capture the general sentiment or tone of a tweet. However, it does not consider the semantic meaning or context of the words and may overlook important contextual information that could impact the classification accuracy. Both the bag-of-words method and rule-based machine learning are frequently used strategies for detecting cyberbullying, and each has advantages and disadvantages of its own. To increase the precision and efficacy of cyberbullying detection systems, researchers may combine these strategies or look at additional machine learning techniques.

Chen [5] introduced an innovative approach for identifying offensive language used in social networks. By analysing various features encompassing users' writing styles, structures, and specific cyberbullying content, the technique aimed to identify potential bullies. The study primarily focused on a lexical syntactic feature that exhibited a remarkable ability to detect offensive content in messages sent by bullies, achieving a significant success rate. The results highlighted an impressive precision of 98.24% and recall of 94.34% in identifying offensive language constructs.

NAIVE BAYES AND SVM BASED APPROACH:

Naïve Bayes:

- Naive Bayes is a probabilistic classifier that assumes independence among features.
- It is computationally efficient and performs with appreciable efficiency even with scarce training data.
- Naive Bayes is particularly effective when the independence assumption holds true or when the features are independent conditionally given the class.
- It operates better on text classification tasks, including cyberbullying detection.
- Naive Bayes tends to handle noise and irrelevant features reasonably well.

Support Vector Machines (SVM):

- SVM is a supervised ML algorithm that finds an optimal hyperplane to separate data points of separate classes.
- By translating the input features to a higher-dimensional space, it is capable of handling complex feature spaces and performs well when the data is not thought to be linearly separable.
- Through the use of various kernel functions, SVM can handle both linear and non-linear classification issues.
- It generally performs well with a small number of samples and is effective in scenarios where the number of features is larger than the number of samples.
- SVMs have been used with great success for a variety of text categorization applications, including the identification of cyberbullying.

VandanaNandakumar et al. in her review on Twitter data using Bayes classifier algorithm and SVM model [7] compares the efficiencies of the two approaches and concludes that, for text data classification, Naive Bayes classifier performs superior than the SVM model.To evaluate the efficiency of the algorithms, the authors independently calculated probabilities for every feature set using the Twitter dataset. They plotted a graph to compare the performance of Naive Bayes and SVM algorithms. The comparison was based on the precision factor, considering how accurately the algorithms predicted the output variable, which in this case is the identifying of cyberbullying instances.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023
DOI: 10.17148/IJARCCE.2023.12681

DETECTION BY BERT MODEL:

A novel method for social media platform cyberbullying detection is put forth by J. Yadav et al. [8] using the BERT model which involves the following working steps:

1. Preprocessing: Textual information from social networking sites is preprocessed by removing unnecessary symbols, stopwords, and performing tokenization. This step prepares the data for input into the BERT model.

2. BERT Model Architecture: BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained deep learning model that utilises transformer architecture. It consists of encoder layers that capture contextual information from both left and right context of every word in a sentence.

3. Fine-tuning: The pre-trained BERT model is fine-tuned on a specific cyberbullying detection task. This involves training the model on a labelled dataset that includes instances of cyberbullying and non-cyberbullying content. The fine-tuning process adjusts the model's parameters to optimise its performance for the specific task.

4. Input Encoding: The text data is encoded into a numerical representation suitable for input into the BERT model. This typically involves tokenizing the text into subword units and converting them into corresponding embeddings.

5. Attention Mechanism: BERT utilises an attention mechanism to capture the relationships between words in a sentence. This mechanism allows the model to assign different weights to different words based on their relevance to the overall meaning of the sentence.

6. Classification: The fine-tuned BERT model is used for classification, where it predicts whether a given text instance represents cyberbullying or not. The model outputs a probability score indicating the likelihood of the text being associated with cyberbullying.

7. Thresholding: A threshold is set to determine the classification decision. If the predicted probability score exceeds the threshold, the text is classified as cyberbullying; otherwise, it is classified as non-cyberbullying.

Using data from the Formspring forum and Wikipedia, the suggested detection model was trained and assessed. The evaluation's findings revealed accuracy for the Formspring dataset of 98% and the Wikipedia dataset of 96%. It is important to note that these accuracy values are somewhat lower than those of earlier models employed in comparable investigations. Due to the Wikipedia dataset's bigger size and lack of oversampling requirements, the model fared better on it. To obtain equivalent findings, oversampling methods were used for the Formspring dataset.

A machine learning model was created by Trana R.E. et al. [9] with the intention of addressing exceptional occurrences incorporating text retrieved from picture memes. A dataset made up of over 19,000 text samples taken from YouTube articles was gathered by the researchers. The study examines the performance of three machine learning algorithms when applied to the YouTube dataset: Naive Bayes, Support Vector Machine (SVM), and Convolutional Neural Network (CNN). The acquired results are contrasted with those from already published datasets, including the Form database. The authors carefully look at divisions of the YouTube database to find algorithms for identifying online bullying.

In four assessment categories—race, ethnic background, political thought, and generalism—Naive Bayes beat SVM and CNN. For a similar demographic group, SVM outperformed Naive Bayes and CNN, whereas all three algorithms showed comparable accuracy for the central body group. The study's conclusions offer priceless information for differentiating between hostile and non-hostile occurrences. Future research may focus on developing a two-step classification method to analyse text obtained from images and observe whether the YouTube data provides a more comprehensive context for identifying aggression-related clusters.

SENTIMENT ANALYSIS APPROACHES:

The Python module TextBlob offers a user-friendly API for analysing sentiment on text data. In order to identify cyberbullying behaviours from social network data, Nideeksha B K et al. [14] employed a pre-trained predictive model to categorise the sentiment of a given letter into either favourable, adverse, or neutral. Following is an explanation of the classification model's technique that was created around the sentiment analyzer.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12681

1. Text Preprocessing: The input text is pre-processed to remove any noise or irrelevant information. This often entails actions like lowercasing the text, eliminating punctuation and stopwords, and tokenizing (dividing the text into individual words or tokens).

2. Sentiment Polarity Calculation: TextBlob determines the sentiment polarity of the text using a lexicon-based strategy. Each word in the text is given a polarity value according to a predetermined sentiment lexicon or dictionary. The lexicon includes words and the sentiment ratings that correspond to them, which show whether a term is good, neutral, or negative.

3. Aggregation of Polarity Scores: The polarity scores of individual words in the text are aggregated to calculate an overall sentiment polarity score for the entire text. This can be carried out using a variety of methods such as taking the average of the scores or considering the highest or lowest score.

4. Sentiment Classification: TextBlob categorises the sentiment of the content into one of three categories—positive, neutral, or negative—based on the estimated polarity score. According to the use case or application, a different cutoff point may be used to categorise a feeling. It's crucial to keep in mind that since TextBlob's sentiment analysis relies on a lexical approach, it could not be as accurate at capturing the context or subtleties of the text as more sophisticated machine learning models. For activities that don't require fine-grained sentiment analysis, it offers a quick and simple solution to execute sentiment analysis.

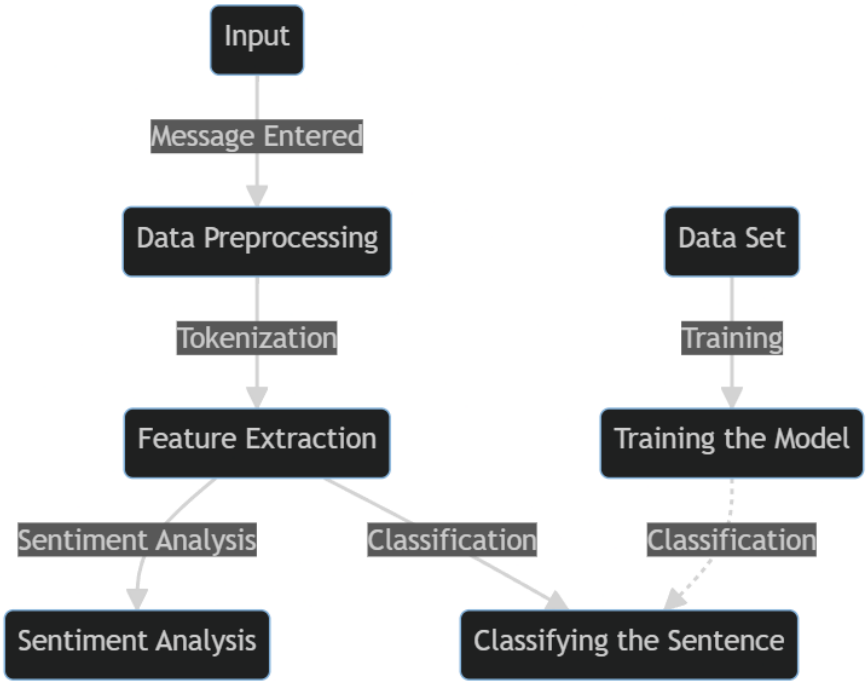

Fig 1.0 System architecture

After the required features are obtained according to their polarity scores, a classifier model is instantiated to start classifying messages based on negative polarity scores. The paper suggests that a chatbot could warn the bullies after the detection of a negative comment as a preliminary warning.

SESSION BASED CYBERBULLYING DETECTION:

P. Yi et al [15] has defined the Session-based Cyberbullying Detection framework that encapsulates the different steps and challenges of the problem. Based on this framework, we provide a comprehensive overview of session-based cyberbullying detection in social media, delving into existing efforts from a data and methodological perspective. The authors have also proposed evidence-based criteria for a set of best practices to create session based cyberbullying datasets.

DOI: 10.17148/IJARCCE.2023.12681

In addition, they have performed benchmark experiments comparing the performance of state-of-the-art session-based cyberbullying detection models along with large pre-trained language models across two different datasets. The framework proposed in this paper is Social Media Session-Based Cyberbullying Detection (SSCD) Framework.

Session-based cyberbullying detection refers to the task of identifying instances of cyberbullying within a session or conversation context, such as online chats, social media exchanges, or forum discussions. Unlike traditional cyberbullying detection that focuses on individual messages or posts, session-based detection takes into account the overall interaction patterns and dynamics within a conversation. The goal of session-based cyberbullying detection is to accurately classify a session as either cyberbullying or non-cyberbullying based on the content and behaviour exhibited during the conversation. This task is particularly challenging due to the complex and evolving nature of cyberbullying, where subtle patterns of aggression, harassment, or abusive behaviour can emerge over the course of a conversation.

Elucidating the process of session-based cyberbullying detection involves several steps:

- Data Collection: Relevant conversation data, such as chat logs, social media threads, or forum discussions, are collected and preprocessed. This may involve cleaning the data, removing noise, and anonymizing user information to ensure privacy.

- Feature Extraction: In order to distinguish between cyberbullying and non-cyberbullying sessions, a number of characteristics are collected from the chat data. These features can include linguistic features (e.g., sentiment, profanity, word usage), behavioural features (e.g., response time, message frequency), social network features (e.g., user connections, community structure), and contextual features (e.g., topic modelling, conversation flow).

- Model Training: Machine learning algorithms or deep learning (DL) architectures are trained using labelled data, where sessions are annotated as cyberbullying or non-cyberbullying. The models learn patterns and relationships between the extracted features and the corresponding labels to make predictions.

- Prediction and Evaluation: The trained models are used to forecast the cyberbullying status of new, unseen sessions. The predictions are evaluated using metrics such as accuracy, recall, precision, and F1-score to assess the effectiveness of the detection system.

- Iterative Refinement: The detection system may undergo iterative refinement to improve its performance. This can involve adjusting feature selection, experimenting with various ML algorithms, incorporating user feedback, or leveraging ensemble techniques to strengthen the overall accuracy and durability of the system.

- Session-based cyberbullying detection plays a crucial role in identifying and addressing instances of cyberbullying within online platforms. By detecting and flagging potentially harmful conversations, it enables timely interventions, promotes safer online environments, and helps protect individuals from the negative impacts of cyberbullying.

IV. CONCLUSION

Technology revolution advanced the quality of life, however, it gave predators a solid ground to conduct their harmful crimes. Internet crimes have become very dangerous since victims are targeted all the time and there are no chances for escape. Cyberbullying is among the most critical internet crimes and research proved its critical consequences on victims. From suicide to lowering victims‟ self-esteem, cyberbullying control has been the focus of many psychological and technical research. As previously indicated, a variety of DL algorithms have been presented that aid in the detection and mitigation of cyberattacks, reduce the likelihood of data breaches, and guarantee the security and privacy of sensitive data. The system also uses various techniques like signature-based detection, behaviour-based detection to identify and flag suspicious activity.The deployment of a robust cyber detection system can lead to significant benefits such as reduced downtime, improved incident response time, and decreased financial losses due to cyber-attacks. The system can also provide valuable insights into the security posture of the organisation, enabling security teams to take proactive measures to mitigate potential threats.

In conclusion, a cyber-detection system is a crucial tool for organisations looking to strengthen their cybersecurity defences. The advantages of deploying a robust system are substantial, but organisations must be vigilant in making sure that the system is properly configured, maintained, and fed with high-quality data.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023
DOI: 10.17148/IJARCCE.2023.12681

REFERENCES

[1] M. Dadvar and F. de Jong. Cyberbullying detection: a step toward a safer internet yard. In Proceedings of the 21st International Conference on World Wide Web (WWW '12 Companion) ACM Digital Library. and Hypermedia and Web. Association for Computing Machinery. Special Interest Group on Hypertext, 2021.

[2] V. C. o. E. Rahul Ramesh Dalvi;Sudhanshu Baliram Chavan; Aparna Halbe, I. of Electrical, and E. Engineers. Detecting Twitter Cyberbullying Using Machine Learning. Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020) : 13-15 May, 2020.

[3] M. D. Capua, E. D. Nardo, and A. Petrosino. Unsupervised Cyberbullying Detection in Social Networks. 23rd International Conference on Pattern Recognition (ICPR), 2016.

[4] R. K, A. Kontostathis, and L. Edwards. Using machine learning to detect cyberbullying, 2011.

[5] Y. Chen, S. Zhu, Y. Zhou, and H. Xu. Detecting offensive language in social media to protect adolescent online safety, 2012.

[6] A. Singh and M. Kaur. Detection framework for content-based cyber- crime in online social networks using a metaheuristic approach. Arabian Journal for Science and Engineering, 45, 9 2020.

[7] Nandakumar, V. Kovoor, B. C, and S. M. U. Cyberbullying revelation in twitter data using naive bayes classifier algorithm. International Journal of Advanced Research in Computer Science, 9, 1 2018..

[8] J. Yadav, D. Kumar, and D. Chauhan. Cyberbullying detection using a pre- trained bert model. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pages 1096–1100, 2020.

[9] R. E. Trana, C. E. Gomez, and R. F. Adler. Fighting cyberbullying: An analysis of algorithms used to detect harassing text found on youtube. In T. Ahram, editor, Advances in Artificial Intelligence, Software and Systems Engineering, pages 9–15, Cham, 2021. Springer International Publishing

[10] N. Tsapatsoulis and V. Anastasopoulou. Cyberbullies in twitter: Afocused review. In 2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), pages 1–6. IEEE, 2019.

[11] G. A. Leo´n-Paredes, W. F. Palomeque-Leo´n, P. L. Gallegos-Segovia, P. E. VintimillaTapia, J. F. Bravo-Torres, L. I. Barbosa Santillan, and M. M. Paredes-Pinos. Presumptive detection of cyberbullying on twitter through natural language processing and machine learning in the Spanish language. In 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), pages 1–7, 2019.

[12] R. I. Rasel, N. Sultana, S. Akhter, and P. Meesad. Detection of cyber- aggressive comments on social media networks: A machine learning and text mining approach. In Proceedings of the 2nd International Conference on Natural Language Processing and Information Retrieval, pages 37–41, 2018.

[13] Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber Bullying Detection Using Social and Textual Analysis. In Proceedings of the 3rd International Workshop on Socially-Aware Multimedia (SAM '14). Association for Computing Machinery, New York, NY, USA, 3–6. https://doi.org/10.1145/2661126.2661133

[14] Cyberbullying Detection Using Machine Learning Nideeksha B K1, P Shreya2, Sudharani Reddy P3, Mohamadi Ghousiya Kousar4

[15] Yi, Peiling and Zubiaga, Arkaitz, Session-Based Cyberbullying Detection in Social Media: A Survey. Available at SSRN:https://ssrn.com/abstract=4208013 or http://dx.doi.org/10.2139/ssrn.4208013