Detection Of Cyber Bullying on Social Media Using ML

G. Kokila[1], K. Abivarsha[2], P. Muthuselvam[3], A. Rahul sarath[4], M. R. Sneha[5]

[1]*Assistant Professor, Computer Science and Engineering, Tamilnadu College of Engineering, Coimbatore, India*

[2,3,4,5] *UG Students, Computer Science and Engineering, Tamilnadu College of Engineering, Coimbatore, India*

***Abstract*—On social media platforms, cyberbullying has become a widespread and alarming problem that affects people's mental health and general wellbeing all around the world. This paper suggests a cyberbullying detection system that makes use of the Support Vector Machine (SVM) algorithm in order to tackle this issue. The technology seeks to automatically detect and flag instances of cyberbullying in real-time social media content by utilizing machine learning. The first step in creating the detection system is gathering and classifying a large dataset of posts and comments that include instances of both cyberbullying and non-cyberbullying. The bag-of-words or TF-IDF approaches are used to extract important features from the text data after it has been pre-processed by eliminating unnecessary information, tokenizing it, and changing the text to lowercase. The SVM classifier, which looks for the best hyper plane to efficiently separate cyberbullying from non-cyberbullying content, is trained using these converted feature vectors as inputs. Metrics like accuracy, precision, recall, F1-score, and ROC-AUC are analysed to see how well the SVM model performs in detecting instances of cyberbullying using a different testing dataset. To improve the system's performance, model fine-tuning is done by experimenting with different SVM hyper parameters and cross-validation strategies**.

***Index Terms*—cyberbullying detection, distil Bert, machine learning, pre trained language models**

1. INTRODUCTION

Because social media is so widely used, cyberbullying is becoming a bigger problem that is impacting people's emotional and mental health. It entails harassing, threatening, or degrading people via digital communication means, and victims frequently suffer serious repercussions. Cyberbullying, in contrast to traditional types of bullying, may happen anonymously and spread swiftly, making it challenging to detect and manage. Manual monitoring becomes unreliable and ineffective as the number of online interactions rises. As a result, automated systems that are capable of efficiently identifying and categorizing dangerous content are now required. The use of machine learning techniques, especially text classification algorithms, allows for the analysis of vast volumes of data and the differentiation between abusive and non-abusive language. In addition to increasing detection accuracy, this strategy supports the development of a more secure and civil online community.

1.1 CYBERBULLYING DETECTION

Social media platforms' broad use in recent years has completely changed how people engage and communicate online. Although these platforms provide a wealth of chances for expression and connection, they have also given rise to a negative aspect known as cyberbullying. Cyberbullying is the practice of harassing, intimidating, or dehumanizing others via digital communication platforms like social media, text messaging, or online forums. Because cyberbullying can have serious and long-lasting impacts on its victims, including emotional discomfort, social isolation, and in some unfortunate cases, even suicide, its incidence has grown. As a result, preventing and identifying cyberbullying has become crucial to developing inclusive and secure online environments. Because social media sites create so much content, traditional manual approaches for identifying and stopping cyberbullying are frequently insufficient. Thankfully, new avenues for automated cyberbullying detection have been made possible by developments in machine learning (ML) and natural language processing (NLP). By utilizing these technologies, we can create intelligent systems that can instantly recognize potentially dangerous information, enabling prompt interventions and promoting a safer online environment.

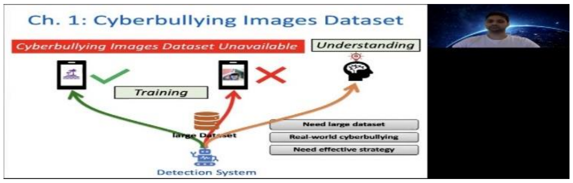

Figure 1. cyberbullying detection

1.2 DISTIL BERT

The creation of strong language models in the age of contemporary Natural Language Processing (NLP) has completely changed how computers comprehend and process human language. Distil BERT has become a leading candidate among these innovative models, providing exceptional performance and efficiency across a range of NLP workloads. The groundbreaking BERT (Bidirectional Encoder Representations from Transformers) model, first presented by Google in 2018, is distilled into Distil BERT. An important development in NLP was BERT's capacity to infer meaning and context from a word's left and right contexts. However, because of its enormous size, it was difficult to implement in contexts with limited resources and computationally costly. Hugging Face, a prominent NLP research organization, responded to these issues in 2019 by launching Distil BERT. By using a revolutionary technique known as "distillation," Distil BERT compresses the original BERT model while maintaining a large portion of its language processing capabilities. Distil BERT is consequently faster and smaller, which makes it more useful for real-world applications without sacrificing performance.

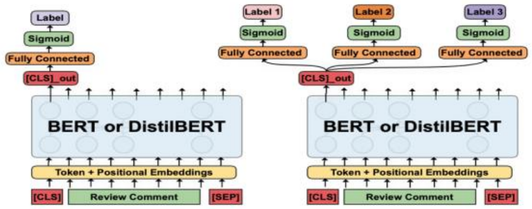

FIGURE 2. DISTIL BERT

1.3 MACHINE LEARNING

A state-of-the-art area of artificial intelligence called machine learning (ML) enables computers to learn from data and gradually enhance their performance without explicit programming. Machine learning algorithms allow machines to identify patterns, make decisions, and resolve complicated issues in a variety of fields by simulating human learning. Training the model using labeled data with input-output pairs is known as supervised learning. After learning to map inputs to matching outputs, the model is able to forecast outputs for previously unseen inputs. This method is frequently applied to problems including spam detection, picture recognition, and natural language processing. Conversely, unsupervised learning works with unlabelled data and looks for hidden structures, patterns, or groupings in the data without direct supervision. This is very helpful for applications like anomaly detection, dimensionality reduction, and clustering comparable data points. Behavior psychology serves as the basis for reinforcement learning, in which an agent interacts with its surroundings and learns to respond in a way that maximizes a cumulative reward signal. This paradigm is used in situations like resource allocation optimization, playing challenging games, and training self-driving cars.

1.4 PRE-TRAINED LANGUAGE MODELS

A new era of language generation and understanding has been brought about by pre-trained language models, which are now the foundation of state-of-the-art Natural Language Processing (NLP). These models have the amazing capacity to extract complex patterns and structures from enormous text corpora, thanks to the power of deep learning and enormous volumes of data. They get a general grasp of language through pre-training on a variety of large datasets, which may subsequently be refined for particular NLP applications. A paradigm change in NLP has occurred with the introduction of pre-trained language models, which do away with the necessity of creating task-specific models from the ground up. The creation of language-based applications has historically been difficult and time-consuming due to the meticulous feature engineering and domain-specific knowledge needed for NLP tasks. However, because pre-trained language models offer a strong basis of language understanding that can be used to a variety of activities, they provide a more effective and efficient method. In 2018, Google unveiled BERT (Bidirectional Encoder Representations from

Transformers), one of the most innovative pre-trained language models.

2. LITERATURE REVIEW

In this study, BARIS CAGIRKAN et al. have proposed Cyberbullying is a new type of bullying that has been moved from the physical to the virtual world and involves aggressive behavior by teenagers. It is a variation of classic bullying that has been transferred to electronic settings (social media, online gaming environments, blogs, etc.). In addition to determining the demographic and socioeconomic characteristics that contribute to bullying and cyberbullying, this study attempts to quantify the prevalence of cyberbullying among Turkish high school students residing in Eastern Turkey. 470 students between the ages of 15 and 19 make up the study population. To determine the scale's factor structure, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were used. It was shown that a one-factor structure best describes the Turkish version of the Cyberbullying Scale (CBS). Tukey HSD, one-way ANOVA, and the independent samples t test were used to compare the demographic and socioeconomic variables. According to the main conclusions, the following factors have a significant impact on students' CBS scores: gender, school type, number of siblings, mobile phone ownership and duration, private Internet access, family supervision, reason for Internet use, amount of time spent online, and type of messaging app [1].

In this study, Pham Thi Lan Ch et al. have suggested The purpose of this research is to find out how high school students in Hanoi, Vietnam, deal with cyberbullying and to investigate the relationship between the average amount of time they spend online each day and their vulnerability to cyberbullying. Using the respondent-driven sample technique, 215 students between the ages of 13 and 18 finished an online survey. The modified Patchin and Hinduja's scale was used to analyze the experience of cyberbullying. 45.1% of people reported having experienced at least one form of cyberbullying. Being called names or made fun of was the most prevalent kind of cyberbullying. There was a dose-response relationship between the average amount of time spent online each day and the likelihood of experiencing cyberbullying. 54% of participants who used the Internet for three hours a day reported having encountered cyberbullying, compared to 39% of those who used it for one to three hours and 30% of those who used Bullying is commonly described as an aggressive, deliberate act or behavior committed by a group or an individual against a victim who is unable to protect themselves on a regular basis over an extended period of time [2].

In this system, Amgad Muneer et al. have suggested The emergence of social media, especially Twitter, presents a number of problems because of a misinterpretation of the idea of free expression. One of these problems is cyberbullying, a serious worldwide problem that impacts both victims and communities. There have been numerous attempts in the literature to stop, prevent, or lessen cyberbullying; however, these efforts are realistic since they depend on the relationships between the victims. Thus, it is essential to detect cyberbullying without the victims' knowledge. We tried to investigate this problem in this work by gathering a global dataset of 37,373 distinct tweets from Twitter. Additionally, seven machine learning classifiers were employed: Random Forest (RF), AdaBoost (ADB), Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Light Gradient Boosting Machine (LGBM), and Stochastic Gradient Descent (SGD). In order to ascertain the classifiers' recognition rates when applied to the global dataset, each of these algorithms was assessed using accuracy, precision, recall, and F1 score as performance measures. The testing findings demonstrate LR's superiority, with a median accuracy of approximately 90.57%. SGD had the best precision (0.968), logistic regression had the best F1 score (0.928), and SVM had the best recall (1.00) among the classifiers [3].

In this system, Robin M. Kowalski et al. have proposed Bullying has existed in schools for a long time, but only recently has the negative effects of bullying come to light. Aggressive behaviors that are repeated over time and feature an imbalance of power between the perpetrator and their targets are generally referred to as bullying. Cyberbullying is a new form of bullying that has surfaced in recent years. Bullying via electronic means, including chat rooms, websites, online games, social networking sites, instant messaging, e-mail, and text messaging, is known as cyberbullying. Numerous children and young people have engaged in "traditional" types of bullying, according to research. According to Nansel and colleagues' findings from the first nationally representative study on bullying in the United States, 11% of sixth through tenth graders were "victims only," 13% were "bullies only," and 6% were "bully/victims"—that is, they had both been bullied and been bullied. More recently, a survey of 11 to 15-year-olds in 40 countries found that 26% of teenagers had experienced bullying on a regular basis as a victim, perpetrator, or both. Bullying also has a range of negative repercussions on physical health, according to research [4].

According to Yu-Chia Huang et al., the two most significant developmental psychologists are Lev Vygotsky and Jean Piaget. Despite their differences, their contributions to developmental psychology are equally noteworthy and distinctive. Notwithstanding these similarities, Piaget and Vygotsky's theories diverge in a significant and sometimes overlooked way, which influences how each author approaches the idea of cognitive growth. To put it briefly, which theory is more accurate? This essay will explore the theories of both psychologists, their similarities and differences, and the reasons they have both continued to be so prevalent in textbooks. The theories of Piaget and Vygotsky are frequently employed in contrast to one another, while never being in direct competition, because they both provide learning theories that differ significantly but nonetheless have an impact on our understanding of cognitive development. In psychology and neuroscience, cognitive development is the study of how people think, explore, and solve issues. Children are able to think about and comprehend the world around them thanks to the development of their knowledge, abilities, problem-solving skills, and dispositions. Jean Piaget and Lev Vygotsky's study has had a significant impact on psychology's methodologies and approaches to the cognitive-developmental problem [5].

3. EXISTING SYSTEM

Social networking and communication were made easier by information and communication technologies. However, there were negative effects of cyberbullying on the network. Online bullying posts can be reported, blocked, and removed manually, but these user-dependent methods are inefficient. Cyberbullying post text classification was hampered by the use of a bag of words text representation without metadata. This study used two methods—conventional machine learning and transfer learning—to create an automated system for detecting cyberbullying. This study used AMICA data, which included a structured annotation procedure and a substantial quantity of context related to cyberbullying. The traditional machine learning method used textual, sentiment and emotional, static and contextual word embedding, psycholinguistics, term lists, and toxicity characteristics. This study was the first to identify cyberbullying using toxicity features. Additionally, this study is the first to detect cyberbullying using Empath's lexicon and the most recent psycholinguistic features from the Linguistic Inquiry and Word (LIWC) 2022 tool. Gilbert, TN Bert, and Distil Bert contextual embeddings perform similarly, however Distil Bert embedding was chosen for its higher F-measure. When fed separately, the top three unique characteristics were toxicity features, Distil Bert embedding, and textual features that set a new benchmark. The Logistic Regression model outperforms Linear SVC with a faster training time and more effective handling of high-dimensionality features. The model's performance was increased to an F-measure of 64.8% after being fed a combination of textual, sentiment, DistilBert embedding, psycholinguistics, and toxicity features. The transfer learning strategy involved optimizing the version. It was discovered that the pre-trained language models, Distil Bert, Distil Roberta, and Electra-small, had faster training computations than their base form. The optimized DistilBert outperformed CML with the highest F-measure of 72.42%. Our study found that, when feature engineering and resampling were not used, Transfer Learning was the best option for improved performance and less work.

4. PROPOSED SYSTEM

The goal of the suggested method is to provide a precise and effective cyberbullying detection tool for social media sites. The system will use the Support Vector Machine (SVM) algorithm to automatically detect instances of cyberbullying in real-time social media content, utilizing machine learning. First, a diverse dataset comprising postings or comments that are either cyberbullying or non-cyberbullying will be gathered and labeled. Text cleaning, lowercasing, and tokenization are some of the preprocessing methods that will be used to convert the raw text data into a format that is appropriate for feature extraction. Meaningful features will next be extracted from the preprocessed text data using the bag-of-words or TF-IDF approaches. By identifying the best hyperplane in the feature space, the SVM classifier will be trained using these features to differentiate between content that is cyberbullying and content that is not. Using a variety of measures, the system's performance will be thoroughly assessed, and adjustments will be made to maximize its effectiveness. The SVM-based cyberbullying detection system will be implemented to function in real-time on social media platforms after it has been trained and assessed, offering prompt notifications and assistance to individuals who may be the victims of cyberbullying. The suggested solution seeks to adjust to changing cyberbullying trends by guaranteeing ongoing monitoring and updating, creating a more secure and civil online environment for all users.

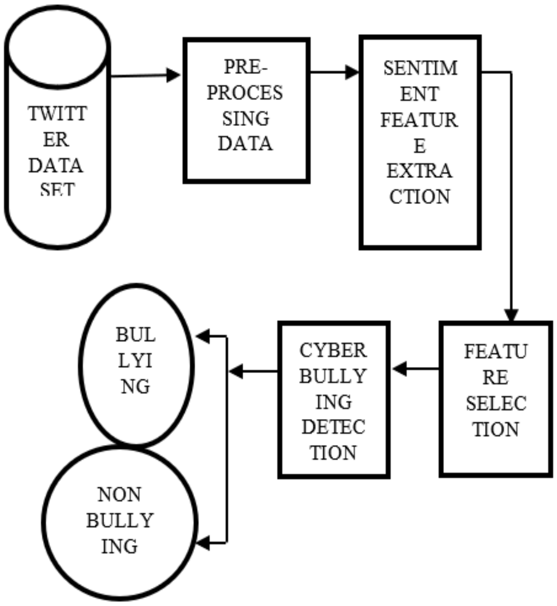

FIGURE 1. SYSTEM FLOW DIAGRAM

A. LOAD DATA

In order to train and test the cyberbullying detection system, this module loads the labeled dataset of social media posts or comments. It retrieves the text data and associated labels (cyberbullying or non-cyberbullying) by reading the dataset from a file or database.

B. DATA PRE-PROCESSING

The purpose of this module is to preprocess the unprocessed text input so that it may be used for SVM classification and feature extraction. Remove URLs, special characters, and other unnecessary information from the text data to make it cleaner. For case insensitivity, change the text to lowercase. The text should be tokenized into discrete words or tokens. Lemmatization or stemming can be used to reduce words to their most basic form (optional).

C. FEATURE SELECTION

From the pre-processed text input, this module extracts features and transforms them into numerical feature vectors that the SVM may use. To represent the text data as numerical vectors, use methods such as TF-IDF or bag-of-words. In order to train the SVM model, create feature matrices with the modified data.

D. TRAINING

This module is in charge of using the feature-selected and pre-processed data to train the SVM classifier. Create training and testing sets from the dataset. With the proper hyper parameters and kernel settings, train the SVM classifier using the training set.

E. TESTING

Using unseen data, this module evaluates the trained SVM classifier's performance. Assess the SVM classifier's ability to identify instances of cyberbullying using the testing set. Calculate the classifier's efficacy by calculating its accuracy, precision, recall, F1-score, and ROC-AUC.

F. EVALUATION AND PERFORMANCE

This module assesses the overall performance of the cyberbullying detection system by analyzing the testing module's results. Show the performance metrics and assessment metrics to give information about the accuracy and resilience of the classifier. Determine possible places where the system could be enhanced or adjusted.

6. RESULT ANALYSIS

To find out how well the suggested cyberbullying detection system identified harmful content, its results were examined using a variety of performance evaluation measures. The accuracy, precision, recall, F1-score, and ROC-AUC of the Support Vector Machine (SVM) classifier were assessed on a different set of data after it had been trained on a labeled dataset. Both cyberbullying and non-cyberbullying information were accurately classified by the model, which showed good performance with balanced precision and recall values. The ROC-AUC value demonstrated the model's capacity to differentiate between the two classes, while the F1-score showed a decent trade-off between false positives and false negatives. These findings imply that the system is capable of identifying abusive or insulting words in textual data. Through cross-validation and hyperparameter tuning, additional enhancements were made, maximizing the model's performance and guaranteeing its resilience across various data subsets.

7. CONCLUSION

In conclusion, the Support Vector Machine (SVM) algorithm-based suggested cyberbullying detection system provides a reliable and effective way to deal with the growing issue of cyberbullying on social media platforms. The system can automatically detect instances of cyberbullying in real-time social media content by utilizing machine learning techniques. This allows it to promptly inform users who may be experiencing cyberbullying situations and offer them support. Data loading, data pre-processing, feature selection, SVM training, testing, evaluation, and performance analysis are some of the key modules involved in system development. Furthermore, for proactive cyberbullying detection and action, the optional real-time monitoring module guarantees ongoing social media activity monitoring. The benefits of the system are found in its high accuracy in differentiating between content that is cyberbullying and content that is not, which allows people to act quickly to make the internet a safer place. Additionally, the system can adjust to changing cyberbullying trends and continue to be effective over time thanks to its scalability and continual improvement features.

8. FUTURE WORK

Future research into more sophisticated machine learning approaches may concentrate on improving the efficacy and performance of the cyberbullying detection system. Integrating deep learning models, like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), which are capable of identifying intricate patterns and semantic links in text data, is one possible approach. Furthermore, adding contextual data and sentiment analysis could enhance the system's comprehension of the purpose of social media posts. Increasing the system's adaptability to other languages and cultural quirks is another essential component for future research, since this will allow it to detect cyberbullying in various communities and geographical areas. To further improve the accuracy and user experience of the system, cooperative initiatives with social media platforms could be undertaken to introduce automatic user feedback mechanisms and real-time reporting. To ensure that the system stays a strong and pertinent weapon in the ongoing battle against cyberbullying, it will be essential to gather and analyze data continuously in order to keep it abreast of new trends in cyberbullying.

REFERNECES

[1] "Cyberbullying among Turkish high school students," by B. Cagirkan and G. Bilek 10.1111/sjop.12720 Scandin. J. Psychol., vol. 62, no. 4, pp. 608–616, Aug. 2021

[2] P. T. L. Chi, V. T. H. Lan, N. H. Ngan, and N. T. Linh, "Online habits, cyberbullying experiences, and coping mechanisms among Hanoi high school students," Art. no. 205510292093574 in Health Psychology Open, vol. 7, no. 1, January 2020, doi: 10.1177/2055102920935747

[3] "CyberDect. A novel approach for cyberbullying detection on Twitter," by A. López-Martínez, J. A. García-Díaz, R. Valencia-García, and A. Ruiz-Martínez, in Proc. Int. Conf. Technol. Innov., Guayaquil, Ecuador: Springer, 2019, pp. 109–121, doi: 10.1007/978-3-030-34989-9_9.

[4] R. M. Kowalski and S. P. Limber, "Academic, psychological, and physical correlates of cyberbullying and traditional bullying," Journal of Adolescent Health, vol. 53, no. 1, pp. S13–S20, July 2020, doi: 10.1016/j.jadohealth.2012.09.018

[5] In Proc. Adv. Social Sci., Educ. Humanities Res., 2021, pp. 28–32, Y.-C. Huang, "Comparison and

contrast of Piaget and Vygotsky's theories," doi: 10.2991/assehr.k.210519.007

[6] Anwar, A., Kee, D. M. H., and Ahmed, "Interpersonal deviance and workplace cyberbullying: Understanding the mediating effect of silence and emotional exhaustion," May 2020, pp. 290–296 in Cyberpsychol., Behav., Social Netw., vol. 23, no. 5, doi: 10.1089/cyber.2019.0407.

[7] "Cyberbullying on social media under the influence of COVID-19," B. M. H. Kee, M. A. L. Al-Anesi, and S. A. L. Al-Anesi DOI: 10.1002/joe.22175 Global Bus. Organizational Excellence, vol. 41, no. 6, pp. 11–22, September 2022

[8] "Cyberbullying and children and young people's mental health: A systematic map of systematic reviews," by I. Kwan, K. Dickson, M. Richardson, W. MacDowall, H. Burchett, C. Stansfield, G. Brunton, K. Sutcliffe, and J. Thomas February 2020, vol. 23, no. 2, pp. 72–82, Cyberpsychol., Behav., Social Netw., doi: 10.1089/cyber.2019.0370.

[9] "Associations between social media and cyberbullying: A review of the literature," by R. Garett, L. R. Lord, and S. D. Young December 2016, mHealth, vol. 2, p. 46, doi: 10.21037/mhealth.2016.12.01

[10] "Detecting cyberbullying in social commentary using supervised machine learning," in Proc. Future Inf. Commun. Conf. Cham, Switzerland: Springer, 2020, pp. 621–630, by M. O. Raza, M. Memon, S. Bhatti, and R. Bux