**DETECTION OF CYBER BULLYING ON SOCIAL MEDIA USING MACHINE LEARINING**

**Sri Ranga Lakshmi1, A.Priyavarshini2, K.PADMA3 , J.Lavanya**4  **, Bushra Tahseen5**

1BTech, CSE, KV Subba Reddy Engineering College, Kurnool, AP, India

2BTech, CSE, KV Subba Reddy Engineering College, Kurnool, AP, India

3BTech, CSE, KV Subba Reddy Engineering College, Kurnool, AP, India

4BTech, CSE, KV Subba Reddy Engineering College, Kurnool, AP, India

5BTech, CSE, KV Subba Reddy Engineering College, Kurnool, AP, India

**ABSTRACT (Font-Times New Roman, Bold, Font Size -12)**

Cyberbullying is a major problem encountered on internet that affects teenagers and also adults.It has lead to mishappenings like suicide and depression. Regulation of content on Social media platforms has become a growing need. The following study uses data from two different forms of cyberbullying, hate speech tweets from Twittter and comments based on personal attacks from Wikipedia forums to build a model based on detection of Cyberbullying in text data using Natural Language Processing and Machine learning. Three methods for Feature extraction and four classifiers are studied to outline the best approach. For Tweet data the model provides accuracies above 90% and for Wikipedia data it gives accuracies above 80%.

**Keywords:** **Cyberbullying, Social media, Natural Language Processing (NLP), Feature extraction, Machine learning**

**1.INTRODUCTION**

The primary problem addressed in this project is the automatic detection of cyberbullying content on social media platforms. With millions of messages posted daily on sites like Twitter and Wikipedia forums, it becomes nearly impossible to manually monitor each post. Many of these posts may contain hate speech, threats, or personal attacks that harm the targeted individuals. Traditional detection methods, such as keyword matching, fail to capture the complexity and variation in language use, sarcasm, or disguised abusive language. Furthermore, trolls may deliberately alter their writing style to bypass detection. Thus, there is a need for a smart, adaptive, and accurate system capable of identifying various forms of cyberbullying in textual data.

This issues developing a machine learning-based binary classification system that categorizes content as cyberbullying or non-cyberbullying. The problem is studied using two types of data: hate speech from Twitter and personal attacks from Wikipedia. Each dataset presents unique linguistic challenges, making it essential to explore various feature extraction techniques and machine learning models. The system must also be scalable and efficient, capable of being deployed in real-world applications for use by social media platforms and regulatory authorities to combat digital harassment.

1. **METHODOLOGY**

The main objective of this project is to develop a robust and accurate system for the detection of cyberbullying on social media platforms using Machine Learning and Natural Language Processing techniques. Specifically, the goals include:

**1. Preprocess Text Data:** Clean and normalize the input text by applying tokenization, stemming, and stop-word removal to ensure that the data is ready for analysis.

**2. Feature Extraction:** Use multiple methods such as Bag of Words, TF-IDF, and Word2Vec to convert textual information into numerical features suitable for machine learning models.

**3. Build and Compare Classifiers:** Train and evaluate multiple machine learning algorithms including Naive Bayes, Logistic Regression, Support Vector Machines (SVM), and Multi- Layer Perceptrons (MLP) to identify the most effective model.

**4. Accuracy Evaluation:** Measure the accuracy of the models on two types of data— Twitter hate speech and Wikipedia personal attacks—to identify which models work best under different conditions.

**5. Develop a Web-Based Interface:** Create a user-friendly interface where users can input text and receive predictions regarding whether the content is cyberbullying or not.

1. **MODELING AND ANALYSIS**
2. **Data Collection Module**

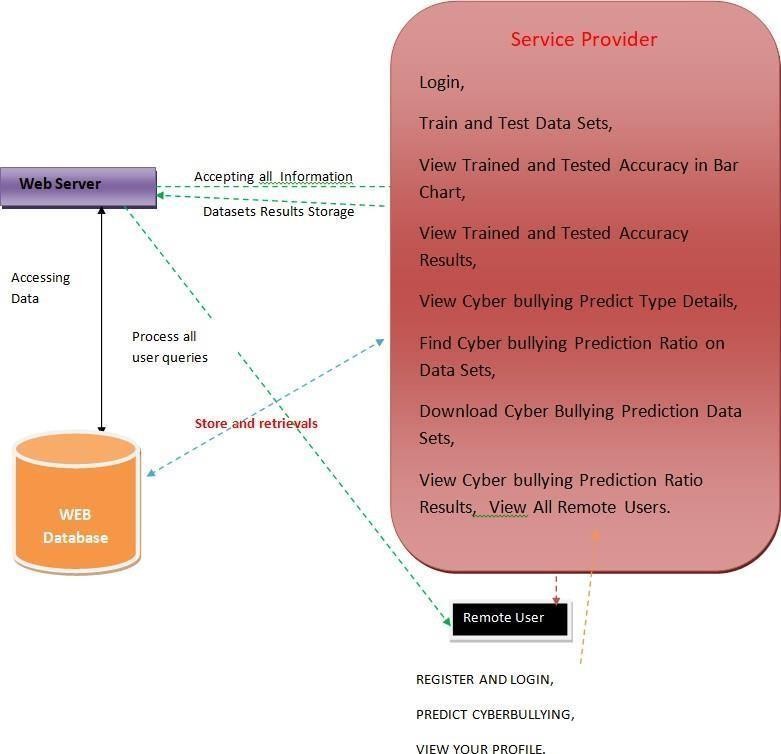
This module gathers textual data from social media platforms, forums, or datasets. The data may include user comments, posts, or chats. It supports various file formats such as CSV and TXT and integrates with APIs for real-time data acquisition.

1. **Data Preprocessing Module**

It cleans and prepares the raw text data for analysis. Processes include lowercasing, removing punctuation, stop words, and special characters, along with tokenization and stemming using tools like NLTK and RegexTokenizer.

1. **Feature Extraction Module**

Converts text into numerical vectors using Word2Vec embeddings. This helps in capturing the semantic meaning of the words for accurate machine learning model training.

****4.**Classification Module :**

**Fig 1 : Architecture of the System**

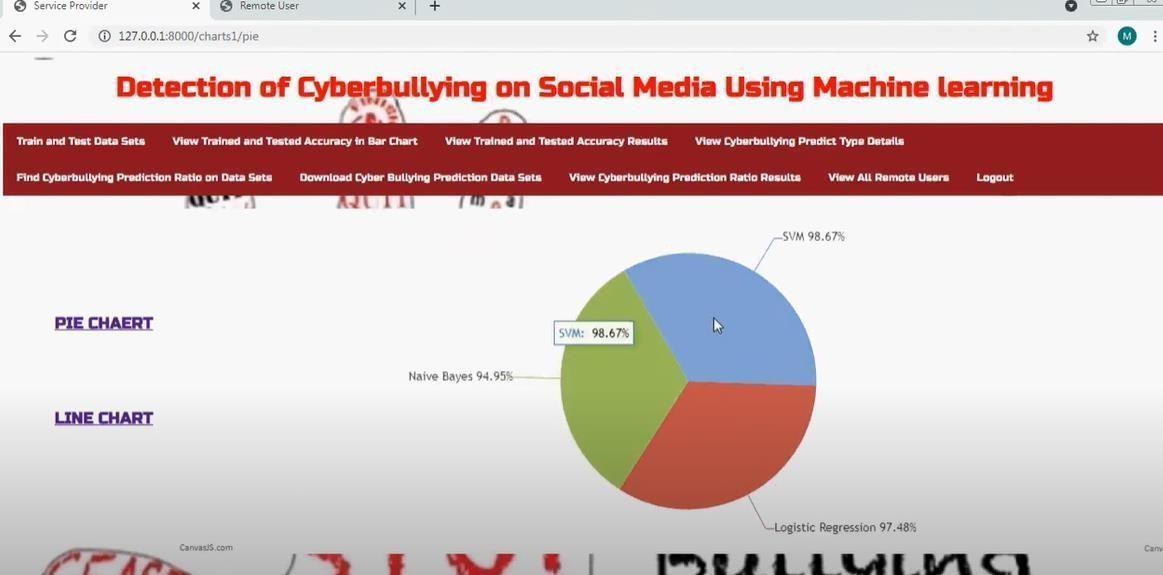
This is the core engine where algorithms such as Logistic Regression, Naïve Bayes, SVM, and MLP are trained and evaluated to detect cyberbullying. The best-performing model is used for prediction.

1. **Web Interface Module**

Built using Django, this provides a user-friendly interface where users can input text or upload files to check for cyberbullying content.

1. **RESULTS AND DISCUSSION**

**Fig 2 : Home Screen**



**Fig 3 : Results**

1. **CONCLUSION**

The proposed cyberbullying detection system offers a robust, scalable, and efficient solution to identify and flag abusive content on digital platforms. By leveraging powerful machine learning algorithms and natural language processing techniques, the system successfully classifies user-generated text into cyberbullying and non- cyberbullying categories. The use of Word2Vec enhances the contextual understanding of language, leading to improved classification accuracy. The system’s modular architecture, including data collection, preprocessing, classification, and web interface components, makes it easily adaptable for a wide range of platforms, including educational institutions, social media networks, and discussion forums.

1. **REFERENCES**
2. Zhong, H., Li, H., Squicciarini, A., Rajtmajer, S., Griffin, C., & Caragea, C. (2016). Content-driven detection of cyberbullying on the Instagram social network. Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI).
3. Cheng, L., Li, J., Silva, Y. N., Hall, D., & Liu, H. (2020). XBully: Cyberbullying Detection within a Multi-Modal Context. Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).
4. Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., & Oliveira, E. (2019). Automatic cyberbullying detection: A systematic review. Computers in Human Behavior, 93, 333–345.
5. Dadvar, M., de Jong, F., & Witteveen, L. (2015). Cyberbullying Detection: A Step Toward a Safer Internet Yard. In Advances in Artificial Intelligence – IEA/AIE.
6. Cambria, E., & White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research. IEEE Computational Intelligence Magazine, 9(2), 48–57.