**Detecting Offensive Language in Social Media Comments**

**- 2210080035(B. Madhu) , 2210080072(G. Madhava)**

**Abstract** With the rise of social media, detecting offensive language has become crucial to maintaining a safe online environment. This research explores different machine learning and deep learning techniques for classifying offensive comments. We compare traditional machine learning models like Naïve Bayes and SVM with deep learning models like LSTM and BERT, evaluating their effectiveness in detecting offensive content.

**1. Introduction** Social media platforms face challenges in moderating offensive language, which can include hate speech, cyberbullying, and toxic comments. This study aims to develop a model that can automatically classify comments as offensive or non-offensive using natural language processing (NLP) techniques.

**2. Literature Review** Previous research has explored rule-based filtering, supervised learning, and deep learning methods to detect offensive language. Traditional models like Naïve Bayes and SVM have shown effectiveness but struggle with contextual understanding, while deep learning models like LSTMs and Transformers (e.g., BERT) offer better accuracy by capturing contextual meaning.

**3. Methodology**

**3.1 Dataset Collection**

* Public datasets like Davidson's Hate Speech and Offensive Language dataset or Kaggle's Twitter Offensive Language dataset are used.
* Preprocessing includes lowercasing, stopword removal, stemming, and tokenization.

**3.2 Machine Learning Models**

* Traditional: Naïve Bayes, Support Vector Machine (SVM), Logistic Regression
* Deep Learning: LSTM, CNN, BERT
* Text vectorization techniques: TF-IDF, Word2Vec, GloVe

**3.3 Model Training & Evaluation**

* Split data (80% training, 20% testing)
* Evaluation metrics: Accuracy, Precision, Recall, F1-score
* Use confusion matrix for misclassification analysis

**4. Results & Discussion**

* Compare accuracy and performance of ML vs. DL models
* Analyze misclassified cases and discuss limitations
* Highlight deep learning’s advantages in contextual understanding

**5. Conclusion & Future Work**

* Deep learning models outperform traditional methods in detecting offensive language.
* Future work includes real-time deployment, multilingual support, and improved handling of sarcasm.