International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Cyberbullying Detection on Social Media Platforms Using Natural Language Processing and Machine Learning

Chilla Gayatri
Reg. No. 24Q71F0006
gayatrichilla11@gmail.com
Department of Master of Computer Applications
Avanthi Institute of Engineering and Technology (Autonomous)
Vizianagaram, Andhra Pradesh, India
Under the guidance of Mr. P. Satyanarayana, M.Tech., (Ph.D.), Associate Professor
pukkallasatya84@gmail.com

***Abstract*—*Cyberbullying is a major problem on the internet that affects teenagers and adults and has led to serious consequences such as depression and suicide, making the regulation of content on social-media platforms a growing need. This study uses data from two different forms of cyberbullying—hate-speech tweets from Twitter and personal-attack comments from Wikipedia forums—to build a model for detecting cyberbullying in text data using Natural Language Processing (NLP) and machine learning. The problem is framed as a binary classification task, and three feature-extraction methods and four classifiers are studied to outline the best approach. A Support Vector Machine is used for Twitter hate speech and a Random Forest classifier for Wikipedia personal attacks. For the tweet data the model provides accuracies above 90%, and for the Wikipedia data it provides accuracies above 80%; the proposed Support Vector Machine approach reaches an accuracy of around 96% for detecting cyberbullying content, which is better than the existing systems. The system is implemented in Python with a Flask web framework. By going beyond simple pattern matching, the proposed approach provides more precise detection and can help protect users from social-media bullies and support the moderation of harmful content.*

Keywords*—*Cyberbullying Detection; Natural Language Processing; Machine Learning; Support Vector Machine; Random Forest; Text Classification; Hate Speech; Social Media.

I. INTRODUCTION

Cyberbullying is the use of technology to harass, threaten, embarrass, or target another person, and online conflict can escalate into real-life threats; in some cases victims have turned to self-harm. With the rapid growth of social-media usage, regulating harmful content has become a growing necessity, and automated detection of cyberbullying on social-media platforms has therefore become an important problem.

Machine learning is a system of computer algorithms that learn from examples through self-improvement without being explicitly programmed, combining data with statistical methods to predict outputs that can drive actionable insights. Unlike traditional programming, where every rule must be hand-coded and becomes unsustainable as the system grows, machine-learning models learn the relationship

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

between input and output data and adapt as new data arrives. Applied to text, NLP techniques transform language into features that classifiers can use to distinguish harmful content from benign content.

This project addresses cyberbullying detection as a binary classification problem over two major forms of cyberbullying: hate speech on Twitter and personal attacks on Wikipedia. Data from both sources is preprocessed and converted into features, and machine-learning classifiers are trained to label text as containing cyberbullying or not. Three feature-extraction methods and four classifiers are studied to determine the best approach, with a Support Vector Machine used for the Twitter hate-speech data and a Random Forest classifier used for the Wikipedia personal-attack data.

II. LITERATURE SURVEY

A great deal of research has explored solutions for detecting cyberbullying on social-networking sites. An early approach combined keyword matching, opinion mining, and social-network analysis and reported a precision of about 0.79 and recall of about 0.71 across datasets from four websites. Another study hypothesised that a troll operating a fake profile usually also maintains a real profile, and proposed a machine-learning approach to identify such profiles by selecting profiles, acquiring tweet information, selecting features, and attributing authorship; using around 1,900 tweets from 19 profiles it achieved roughly 68% accuracy for author identification and was validated in a school case study, though it has shortcomings when a troll has no real account or deliberately changes writing style.

Other work proposed a collaborative detection method with multiple connected detection nodes whose results are combined; a B-LSTM technique based on attention; and a KNN approach with new embeddings reaching a precision of around 93%. A Formspring dataset study reported about 78.5% recall using machine learning with oversampling to address class imbalance, while a BERT-based contextual-embedding model achieved an F1 score of about 0.94 on Formspring and 0.81 on Wikipedia data. Deep neural networks trained on Twitter, Wikipedia, and Formspring and tested on YouTube reached an F1 score of about 0.97 using a bidirectional LSTM, and related work studied swear words as features and how vocabulary varies across platforms. A mobile application, BullyBlocker, informed parents of cyberbullying against their children on Facebook by combining warning signs and vulnerability factors into a probability measure.

TABLE I. SUMMARY OF REPRESENTATIVE PRIOR WORK

S.No	Approach	Technique	Reported Result / Note
1	Keyword matching + opinion mining	Social-network analysis	Precision ~0.79, recall ~0.71
2	Troll-profile identification	Machine learning + authorship	~68% accuracy for author ID
3	Collaborative detection	Multiple detection nodes	Combined multi-node results

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

S.No	Approach	Technique	Reported Result / Note
4	Attention-based model	B-LSTM	Sequence-based detection
5	Embedding-based classifier	KNN with new embeddings	Precision ~93%
6	Contextual embeddings	BERT	F1 ~0.94 Formspring, ~0.81 Wikipedia
7	Deep neural networks	Bidirectional LSTM	F1 ~0.97 (cross-dataset)

III. EXISTING SYSTEM AND PROPOSED SYSTEM

A. Existing System

Existing approaches to cyberbullying detection include troll-profile identification using machine learning, collaborative multi-node detection, attention-based B-LSTM techniques, and KNN with new embeddings (reported precision around 93%). Many of these techniques mainly look for patterns that already exist in the data and rely substantially on manual processes and human decision-making, which limits accuracy and scalability.

Disadvantages of the existing system:

- Lower accuracy in detecting cyberbullying content.
- Techniques mainly search for patterns that already exist in the data.
- Many existing techniques are manual and depend on human intervention.

B. Proposed System

The proposed system solves cyberbullying detection as a binary classification problem, detecting two major forms—hate speech on Twitter and personal attacks on Wikipedia—and classifying text as containing cyberbullying or not. It uses a Support Vector Machine for Twitter hate speech and a Random Forest classifier for personal attacks. The linear SVM plots a separating hyperplane in the feature space and is optimal for linearly separable data, using a hinge loss to optimise the margin; the Random Forest aggregates many decision trees and outputs the majority-voted class, providing a more accurate and robust decision through an ensemble of uncorrelated trees.

Advantages of the proposed system:

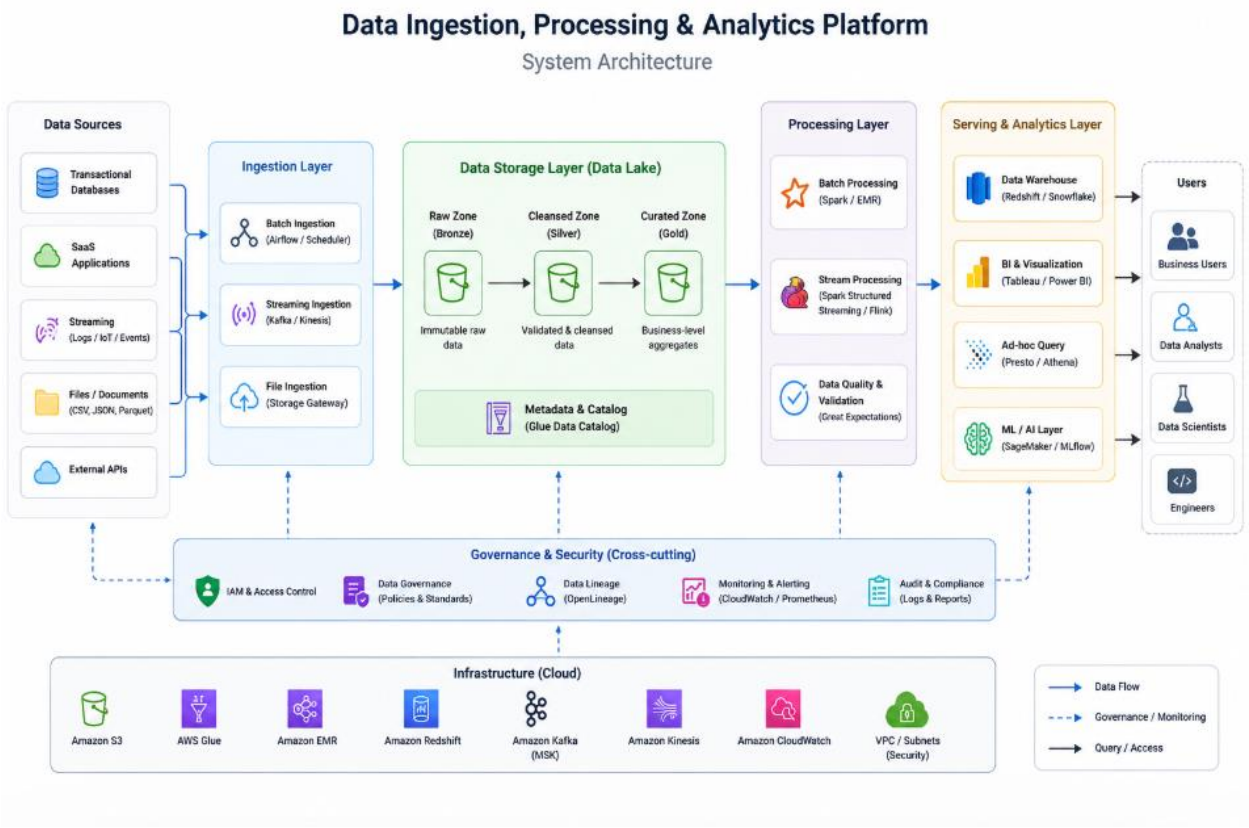
- Higher accuracy: the SVM reaches around 96% for detecting cyberbullying content, better than existing systems.
- Goes beyond simple pattern matching to predict outcomes from pre-existing data.
- Produces more precise results than the existing system.
- Helps protect users from social-media bullies and supports content moderation.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

IV. SYSTEM ANALYSIS AND DESIGN

A. System Architecture

The architecture uses a Support Vector classifier and a Random Forest classifier over the Twitter hate-speech and Wikipedia personal-attack datasets. Input text undergoes preprocessing and feature selection, after which the appropriate classifier predicts whether the content is offensive or non-offensive (for Twitter) or a personal attack or not (for Wikipedia). A performance-analysis stage produces evaluation metrics and graphs. The system thus consists of a data layer (the two datasets), a preprocessing and feature-extraction stage, a classification stage with the two models, and a performance-analysis and presentation stage.



B. Feature Extraction and Classifiers

Three feature-extraction methods and four classifiers are studied to outline the best approach. For hate speech, bag-of-words and TF-IDF representations are effective because such tweets often contain profanity that makes them easier to detect, whereas Word2Vec models that capture feature context proved effective on both datasets, giving similar results with comparatively fewer features when combined with multi-layer perceptrons. The final configuration uses a linear SVM for Twitter hate speech and a Random Forest classifier for Wikipedia personal attacks.

C. Requirements

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

The system is implemented in Python with a Flask web framework and runs on a standard Windows environment. Functionally, it must accept text input from the relevant dataset or user, preprocess and vectorise the text, classify it using the trained model, and present the prediction together with performance metrics and graphs. Non-functional considerations include reasonable accuracy on both datasets, usability through the web interface, and the ability to extend to additional data and forms of online abuse.

V. SYSTEM IMPLEMENTATION

A. Technology Stack

TABLE II. TECHNOLOGY STACK

Component	Technology / Tool
Programming Language	Python
Web Framework	Flask
Machine-Learning Models	Support Vector Machine, Random Forest
Feature Extraction	Bag-of-Words, TF-IDF, Word2Vec
Datasets	Twitter hate-speech tweets; Wikipedia personal-attack comments
Operating System	Windows 10

B. Processing Pipeline

Text from the Twitter hate-speech dataset and the Wikipedia personal-attack dataset is preprocessed and converted into features using bag-of-words, TF-IDF, or Word2Vec representations. The Support Vector Machine is trained on the Twitter data to classify content as offensive or non-offensive, and the Random Forest classifier is trained on the Wikipedia data to classify comments as personal attack or not. The linear SVM optimises a separating hyperplane using a hinge loss, suitable for linearly separable data, while the Random Forest aggregates the predictions of many decision trees by majority vote. After classification, a performance-analysis stage computes evaluation metrics and produces graphs that compare feature-extraction methods and classifiers.

C. Web Interface

The Flask web framework provides the interface through which input text is submitted and the classification result is displayed, along with the performance analysis. This makes the trained models accessible without requiring users to interact directly with the underlying code, supporting practical use for content screening.

VI. RESULTS AND DISCUSSION

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

The experiments compare three feature-extraction methods and four classifiers on the two datasets. For the Twitter hate-speech data, NLP techniques with basic machine-learning algorithms achieve accuracies of over 90%, because tweets containing hate speech often include profanity that makes them easier to detect; bag-of-words and TF-IDF models perform better than Word2Vec models on this data. For the Wikipedia personal-attack data, the model achieves accuracies above 80%; these comments are harder to detect because they generally do not use a common sentiment that can be learned, and the three feature-selection methods perform similarly. The proposed Support Vector Machine approach reaches an accuracy of around 96% for detecting cyberbullying content, which is better than the existing systems. Word2Vec models that use feature context proved effective on both datasets, giving comparable results with fewer features when combined with multi-layer perceptrons.

Representative screenshots from the prototype implementation:

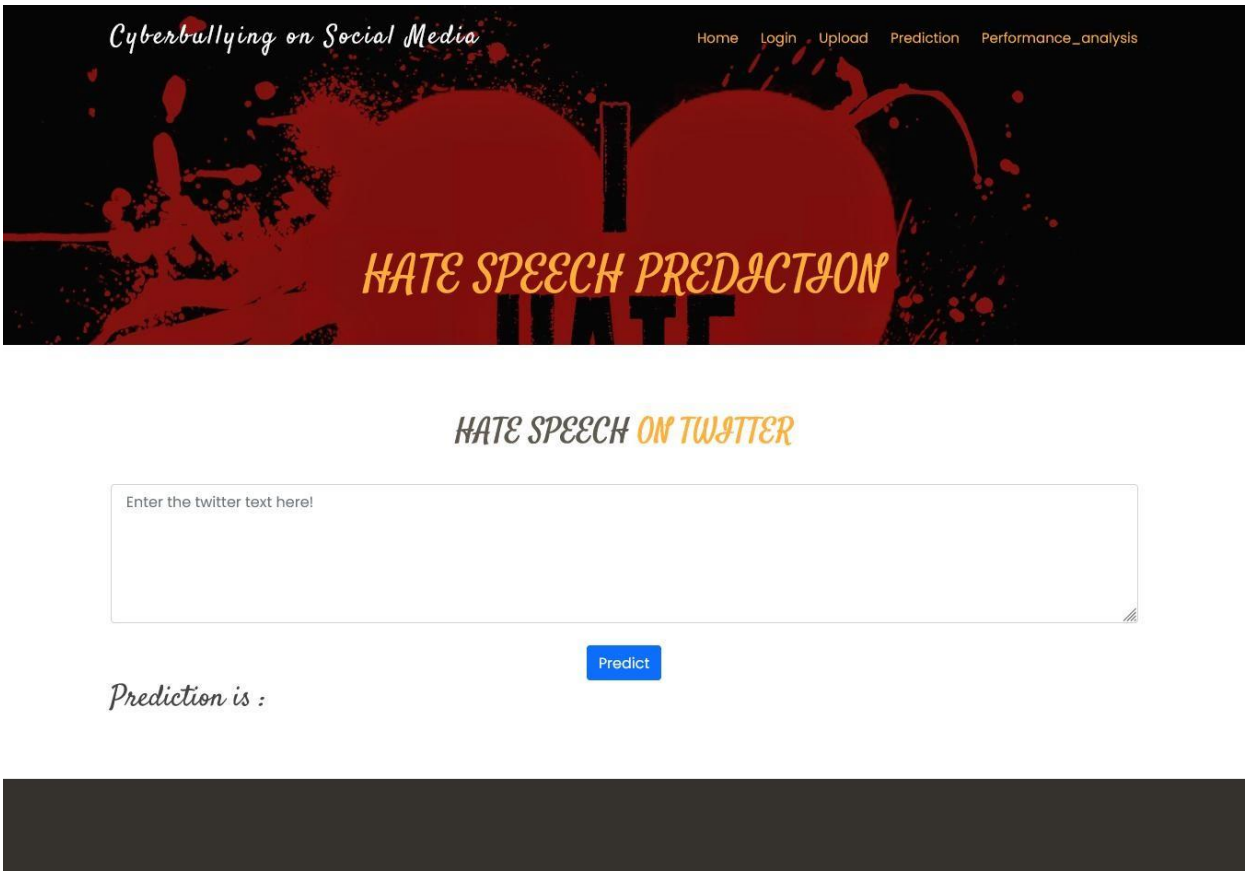


Fig. 1. Text input through the web interface.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

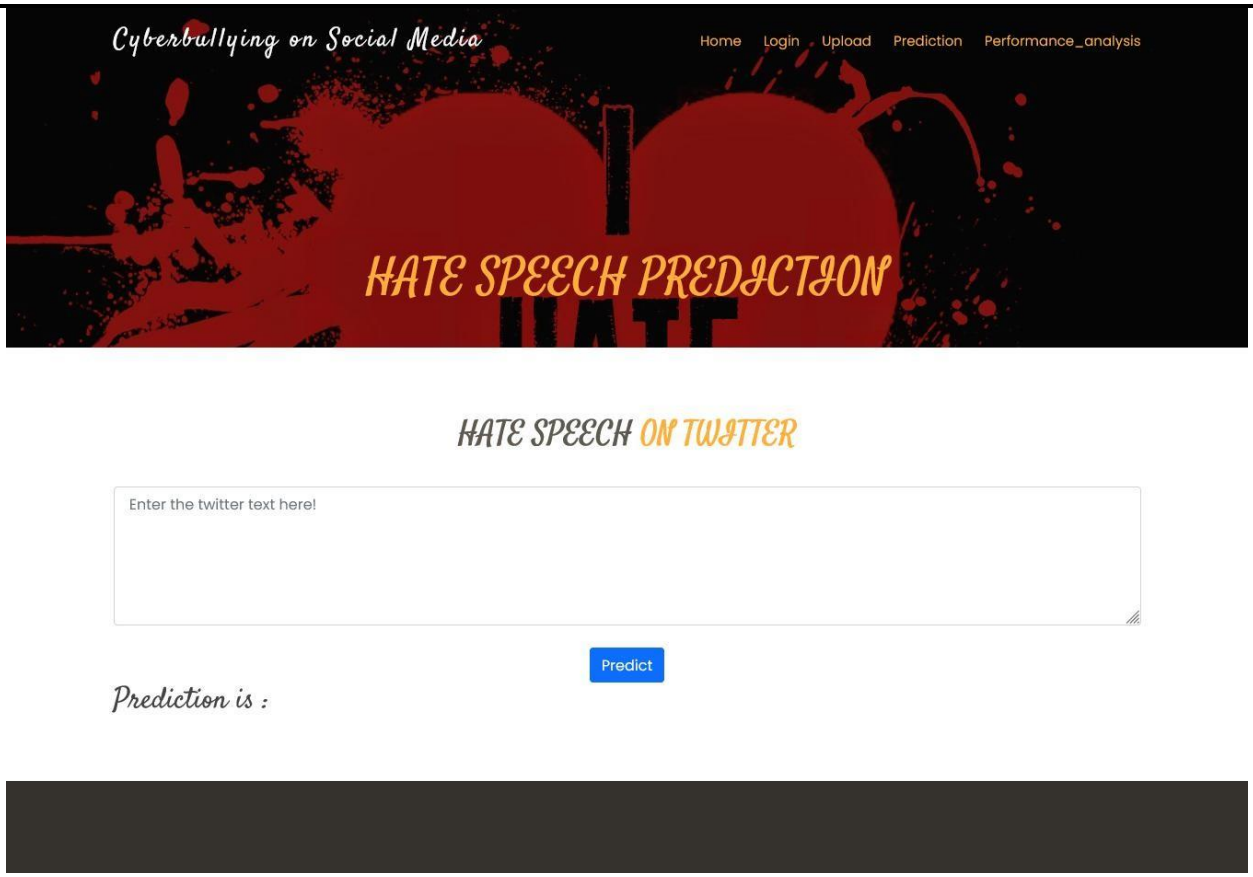


Fig. 2. Hate-speech classification result (Twitter).

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

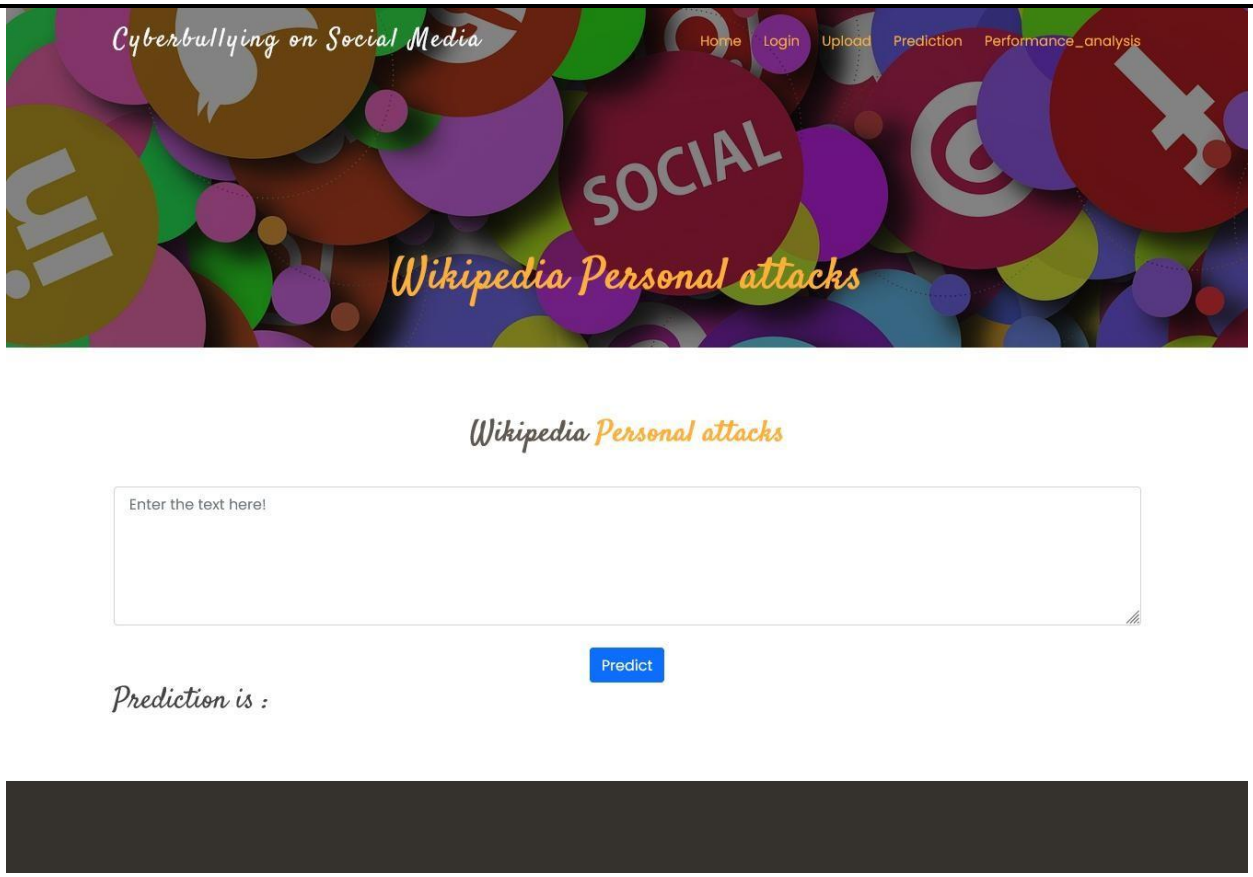


Fig. 3. Personal-attack classification result (Wikipedia).

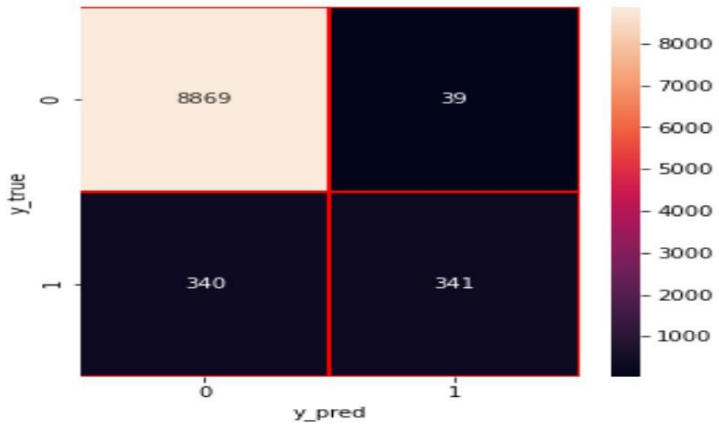
International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

performance analysis

Precision and recall

	Precision	Recall
Non offensive(0)	0.96	1.00
Offensive(1)	0.90	0.50

Confusion Matrix



International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Fig. 4. Performance analysis and comparison graphs.

VII. CONCLUSION AND FUTURE WORK

Cyberbullying across the internet is dangerous and can lead to severe consequences such as depression and suicide, so controlling its spread on social-media platforms is vital. This work proposed an architecture for cyberbullying detection covering two types of data: hate speech on Twitter and personal attacks on Wikipedia. For hate speech, NLP techniques proved effective with accuracies of over 90% using basic machine-learning algorithms, because such tweets contain profanity that makes them easily detectable, and bag-of-words and TF-IDF models gave better results than Word2Vec. Personal attacks were harder to detect through the same model because the comments generally lacked a common learnable sentiment, and the three feature-selection methods performed similarly. Word2Vec models that use feature context proved effective on both datasets, giving similar results with comparatively fewer features when combined with multi-layer perceptrons.

With the availability of more data and better-classified user information for other forms of online attack, cyberbullying detection can be deployed on social-media websites to identify and restrict users who engage in such activity. Future work can expand the datasets, incorporate more advanced contextual language models, extend detection to additional forms of online abuse and platforms, and integrate the model into real-time moderation pipelines so that harmful content can be flagged as it appears.

REFERENCES

[1] I. H. Ting et al., "Cyberbullying Detection Using Keyword Matching, Opinion Mining and Social Network Analysis," (datasets from four websites).

[2] P. Galán-García et al., "Supervised Machine Learning for the Detection of Troll Profiles in Social Networking Sites," (troll-profile identification study).

[3] A. Mangaonkar et al., "Collaborative Detection of Cyberbullying Behaviour in Twitter Data," (multi-node collaborative detection).

[4] P. Zhou et al., "Attention-Based Bidirectional LSTM for Text Classification."

[5] Banerjee et al., "KNN with New Embeddings for Cyberbullying Detection," (reported precision ~93%).

[6] K. Reynolds, A. Kontostathis, and L. Edwards, "Using Machine Learning to Detect Cyberbullying," (Formspring dataset).

[7] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying Detection Using BERT Contextual Embeddings."

[8] M. Dadvar and K. Eckert, "Cyberbullying Detection in Social Networks Using Deep Learning-Based Models."

[9] S. Agrawal and A. Awekar, "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms."

[10] Y. N. Silva, C. Rich, and D. Hall, "BullyBlocker: Towards an Interdisciplinary Approach to Identify Cyberbullying."