International Journal for Modern Trends in Science and Technology
Volume 12, Issue 05, pages 130-136
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol12issue05.html
DOI: https://doi.org/10.5281/zenodo.19836511

SVM-Based Hate Content Identification with Public Sentiment Fusion

D. Arun Kumar, D. Navya Sri, D. Durga Pavani, D. Sukanya, D. Rishitha, K. Chaitanya Deepthi

Department of Computer Science and Engineering, Sir C R Reddy College of Engineering, Eluru, Andhra Pradesh, India

To Cite this Article
D. Arun Kumar, D. Navya Sri, D. Durga Pavani, D. Sukanya, D. Rishitha & K. Chaitanya Deepthi (2026). SVM-Based Hate Content Identification with Public Sentiment Fusion. International Journal for Modern Trends in Science and Technology, 12(05), 130-136. https://doi.org/10.5281/zenodo.19836511

Article Info
Received: 28 March 2026; Revised: 24 April 2026; Accepted: 26 April 2026.

KEYWORDS

Hate Speech Detection, Support Vector Machine (SVM), Sentiment Analysis, TF-IDF, Machine Learning, Natural Language Processing

ABSTRACT

The rapid growth of social media platforms has led to a significant increase in user-generated content, including harmful and offensive language, such as hate speech. Detecting such content is challenging because of sarcasm, informal language, and contextual variation. This paper presents a machine learning-based system for hate speech detection using a Support Vector Machine (SVM) integrated with sentiment analysis for enhanced contextual understanding. The system utilizes datasets collected from social media platforms and preprocesses the data using techniques such as cleaning, normalization, and feature extraction through term frequency-inverse TF-IDF and n-grams. The SVM classifier maps textual data into a high-dimensional feature space and identifies the optimal decision boundaries for accurate classification. Additionally, sentiment analysis was incorporated to classify text into positive or negative categories, enabling better interpretation of public opinion. The proposed system demonstrates high accuracy, efficiency, and low computational cost, making it suitable for real-time applications. By combining hate speech detection with sentiment analysis, this system provides a scalable and effective solution for monitoring and regulating harmful online content in the future.

1. INTRODUCTION

The rapid expansion of social media platforms, such as Twitter, Facebook, and Instagram, has transformed communication by enabling fast and widespread information sharing. However, this growth has also resulted in the proliferation. harmful content, particularly hate speech, which targets individuals or groups based on the attributes Such as race, religion, gender, and nationality. Hate speech can lead to serious societal issues that includes discrimination, psychological harm, and even violence. Due to the

massive volume of user-generated content, manual monitoring is inefficient and impractical.

Traditional keyword-based filtering methods fail to capture contextual meanings and can be easily bypassed using slang or coded language. Machine learning approaches, particularly support vector machines, have been effective in handling high-dimensional textual data. Additionally, sentiment analysis helps to understand the emotional tone of the content, providing deeper insights into user intent.

This study proposes a system that integrates SVM-based hate speech detection with sentiment analysis to improve the accuracy, efficiency, and contextual understanding in real-time applications.

II. LITERATURE REVIEW

Hate speech detection has become a significant research area owing to the rapid growth of social media and online communication platforms, and researchers have explored both traditional machine learning and advanced deep learning techniques to identify harmful and offensive content effectively. Several studies have shown that traditional classifiers, such as support vector machines (SVM), logistic regression, and Naïve Bayes, are effective for text classification because of their ability to handle high-dimensional textual features with good computational efficiency [2]. These models are particularly suitable for real-time applications, where low processing costs and faster predictions are important.

To improve the performance of hate speech detection systems, feature extraction techniques such as Term Frequency–Inverse Document Frequency (TF-IDF), Bag of Words, and n-grams are widely used. These methods transform textual data into numerical representations that help machine learning models identify meaningful word patterns, contextual relationships and semantic importance [13]. The use of n-grams also helps capture the phrase-level context, which is essential for detecting offensive expressions that may not be obvious at the word level.

In recent years, sentiment-aware hate speech detection models have gained attention because they provide additional context. Sentiment analysis helps identify the emotional tone of user-generated content, which improves the ability of classification systems to distinguish between harmful and nonharmful text. Studies have reported that combining sentiment and textual features improves classification accuracy and reduces false positives [5].

Deep learning approaches, such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), BERT, and transformer-based hybrid models, have also shown promising results in hate speech detection tasks. These models offer a better contextual understanding and improved performance in detecting implicit hate speech, sarcasm, and complex language patterns. However, they generally require larger datasets, more training time, and higher computational resources than traditional machine learning methods [14]. Therefore, traditional ML models integrated with sentiment analysis remain a practical and efficient solution for scalable, real-time hate speech detection systems [12].

III. EXISTING SYSTEM

Hate speech detection has evolved through multiple approaches, ranging from simple rule-based systems to advanced deep learning models. Each approach has its own strengths and limitations in terms of accuracy, efficiency, and context understanding.

A. Keyword-Based Filtering

Keyword-based filtering is one of the earliest and simplest techniques used for detecting hate speech. In this approach, a predefined list of offensive or abusive words (often referred to as blacklists) is created. The system scans the input text and flags it as hate speech if any of these keywords are detected.

B. Traditional Machine Learning Models

Basic machine learning algorithms, such as Naive Bayes and Logistic Regression, have been widely used for text classification tasks. These models rely on statistical patterns in the data but often struggle to capture the contextual relationships between words. Consequently, their performance is limited when dealing with complex language structures, sarcasm, or implicit hate speech.

C. Deep Learning Approaches

Deep learning models such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and transformer-based models such as BERT have significantly improved hate speech detection.

D. Limitations of Existing Systems

Despite these advancements, existing systems suffer from several challenges.

Lack of contextual understanding in rule-based approaches

High computational cost in deep learning models

Difficulty in detecting implicit and sarcastic hate speech

- Limited scalability for real-time applications
- Absence of sentiment-based contextual interpretation

IV. PROBLEM STATEMENT

The rapid growth of social media platforms has resulted in an exponential increase in user-generated textual content, including offensive and harmful language, such as hate speech. Hate speech often targets individuals or groups based on characteristics such as race, religion, gender, or nationality, leading to serious social and psychological consequences.

Existing hate speech detection systems primarily rely on keyword-based filtering or basic machine learning techniques, which fail to capture contextual meanings, sarcasm, and implicit expressions. These approaches often produce high false-positive and false-negative rates, reducing the reliability of detection systems.

Although advanced deep learning models provide improved accuracy, they require large datasets and high computational power and are not suitable for real-time applications. Additionally, many existing systems focus only on classification and do not incorporate sentiment analysis, limiting their ability to understand the emotional tone and intent of the text.

Therefore, there is a need to develop an efficient, scalable, and accurate system that can detect hate speech in textual data while analyzing sentiments to improve contextual understanding. The system should balance performance and computational efficiency, making it suitable for real-time deployment in social media monitoring applications.

V. PROPOSED SYSTEM

The proposed system was designed to provide an efficient and scalable solution for detecting hate speech in textual data, while simultaneously analyzing public sentiment. Unlike traditional keyword-based approaches, which fail to capture contextual meaning, the proposed system leverages machine learning techniques combined with advanced feature extraction methods to improve the classification accuracy and reliability.

The system follows a pipeline-based architecture in which each module performs a specific task and contributes to the overall functionality. This modular design ensures better maintainability, flexibility, and scalability in real-time applications. The architecture includes multiple stages, such as data collection, preprocessing, feature extraction, classification, sentiment analysis, and result generation, which together improve the efficiency and performance of the system.

The core of the system was built using term frequency-inverse document frequency (TF-IDF) and n-gram techniques, which convert textual data into meaningful numerical representations. TF-IDF assigns importance to words based on their frequency and uniqueness, whereas n-grams capture the contextual relationships between words. This combination enables the model to understand both the word significance and contextual patterns in the text.

For classification, the system employs a Linear Support Vector Machine (SVM), which is highly effective for text classification tasks involving high-dimensional data. The SVM model works by identifying an optimal hyperplane that separates the hate and non-hate classes while maximizing the margin between them. This ensures better generalization and robustness when handling the unseen data.

In addition to hate speech detection, the proposed system integrates sentiment analysis to classify the text into positive, negative, or neutral categories. This integration, referred to as public sentiment fusion, enhances the ability of the system to interpret the emotional tone and intent behind the text. By combining classification with sentiment analysis, the system provides deeper insights into user behaviour and public opinion.

The system is designed to process large volumes of data efficiently while maintaining a low computational cost. Its modular architecture allows easy integration with real-time applications, such as social media monitoring, online content moderation, cyberbullying prevention, and digital safety systems. The lightweight design of the model also makes it suitable for deployment in

large-scale environments with limited computing resources.

The key features of the proposed system include accurate detection of hate speech using SVM, context-aware feature extraction using TF-IDF and n-grams, integration of sentiment analysis for enhanced understanding, low computational complexity compared with deep learning models, and scalability for real-time applications. Overall, the proposed system provides a practical, reliable, and efficient solution for automated hate speech detection and public sentiment analyses.

Overall, the proposed system provides a balanced approach that achieves high accuracy, efficiency, and interpretability, making it a practical solution for detecting harmful content and analyzing public sentiment on online platforms.

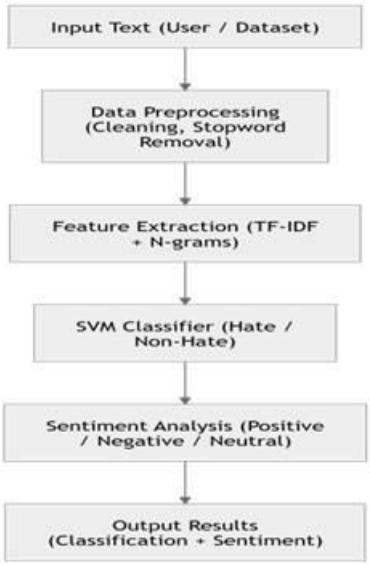

Fig. 1. Architecture of the Proposed Hate Speech Detection System

VI. METHODOLOGY

The proposed system follows a structured pipeline for detecting hate speech and analyzing the sentiment of textual data. The methodology consists of multiple stages, including data preprocessing, feature extraction, model training and sentiment analysis.

A. Data Preprocessing

Data preprocessing is an essential step for cleaning and normalizing the input text. This improves the quality of the data and enhances the model performance. The following techniques were applied:

- Conversion of text to lowercase to maintain consistency
- Removal of URLs, special characters, and stop words
- Tokenization to split text into meaningful words

B. Feature Extraction

Feature extraction was performed using term frequency-inverse document frequency (TF-IDF), which converts textual data into numerical vectors for machine learning processing. TF-IDF assigns importance to words based on their frequency and uniqueness in a dataset. This helps the model focus on meaningful terms and improves the detection of hate-related patterns in the text.

TF-IDF is highly effective in reducing the impact of commonly occurring but less informative terms. By assigning higher weights to significant terms, the model's ability to identify important keywords associated with harmful or offensive content is enhanced. This improves the overall quality of the textual representation for classification.

In addition to TF-IDF, n-gram techniques were used to capture the contextual relationships between consecutive words in a sentence. Unlike unigram approaches, n-grams preserve phrase-level meaning and help identify expressions that depend on word combinations. This makes the system more effective for detecting contextual hate speech.

The combination of TF-IDF and n-gram features creates a rich numerical representation of the text that enhances the classification performance. These extracted features provide the model with both word-level importance and a contextual understanding. As a result, the system became more accurate in identifying both explicit and implicit hate speech patterns.

For classification, a Linear Support Vector Machine (SVM) was employed to distinguish between hate speech and non-hate speech. The model was trained on labelled textual data to learn meaningful patterns associated with harmful and nonharmful content. SVM was selected because of its efficiency and strong performance on high-dimensional textual data.

C. Sentiment Analysis

Sentiment analysis was integrated into the system to determine the emotional tone of the input text. The text was classified into two sentiment categories:

positive and negative. This additional emotional information helps the system better understand the user's intent and the contextual meaning behind textual expressions. It also provides deeper insight into whether a message conveys aggression, support, or neutrality in a given context.

The sentiment output was combined with TF-IDF and n-gram features to improve the performance of the SVM classifier. This fusion approach enhances the system's ability to distinguish harmful content from contextually harmless statements. Consequently, the model reduces false positives and false negatives, making it more robust, reliable, and suitable for real-world hate speech detection applications. This integration also improves the system's ability to detect implicit hate speech and subtle offensive expressions.

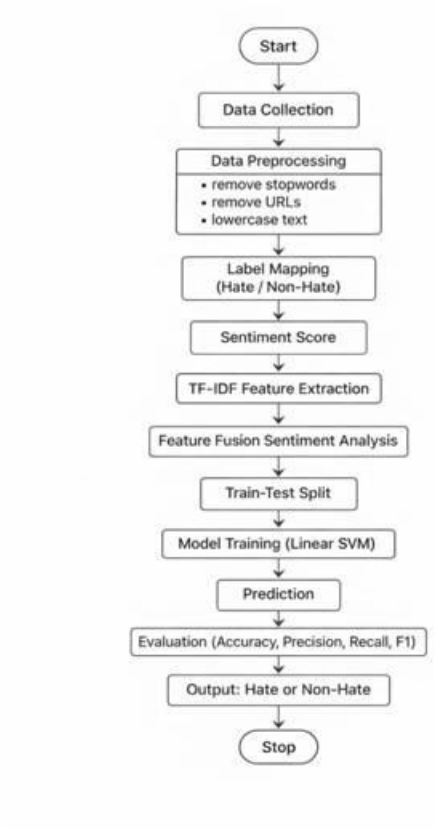

Fig .2 Work Flow of SVM – Based Hate Content

VII. IMPLEMENTATION

The system was implemented using the Python programming language, along with several essential libraries. Pandas and NumPy were used for data handling and numerical operations, whereas Scikit-learn provided tools for machine learning, including TF-IDF vectorization and SVM classification. The Natural Language Toolkit (NLTK) was used for text preprocessing tasks.

A machine learning pipeline was constructed by combining TF-IDF vectorization with a Linear SVM classifier. This pipeline ensures efficient processing by automating feature extraction and classification, thereby improving the performance and scalability.

- NLTK was used for text preprocessing, including tokenization and stop word removal.
- TF-IDF vectorization converts textual data into numerical-feature representations.
- A Linear SVM classifier was applied to classify the text as hate or non-hate.

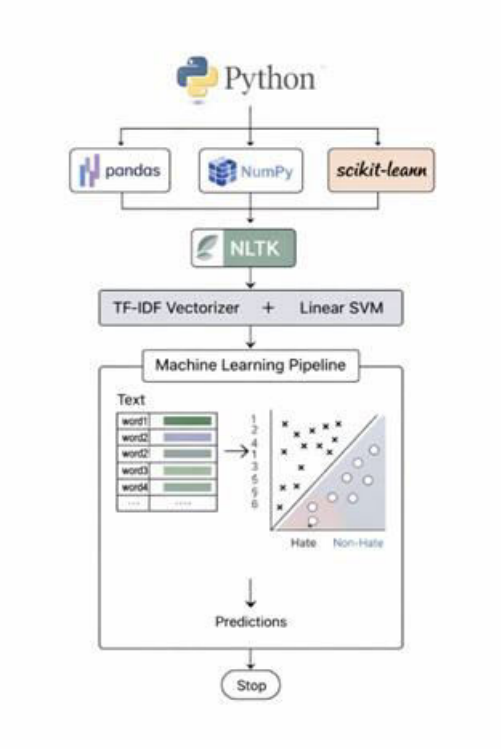

Fig .3 Pipe Line of Hate Speech Detection System

rchitecture further ensures that user information is protected while maintaining system performance. Overall, this project demonstrates how intelligent automation and sentiment-aware technologies can transform traditional customer service into a more responsive, efficient, and user-friendly experience, paving the way for future advancements in AI-driven communication systems.

VIII. RESULTS

The performance of the proposed hate speech detection system was evaluated using a test dataset

comprising previously unseen textual data. The evaluation focused on standard classification metrics to assess the effectiveness of the Support Vector Machine (SVM) model integrated with sentiment analysis.

A. Classification report

The system was evaluated using accuracy, precision, recall, and F1-score.

Table 2: Resulst

	Precision	recall	F1-Score	Support
0	0.86	0.85	0.86	833
1	0.97	0.97	0.97	4124
Accuracy			0.95	4957
Macro Avg	0.92	0.91	0.92	4957
Weighted Avg	0.95	0.95	0.95	4957

Accuracy: 0.9527940286463586

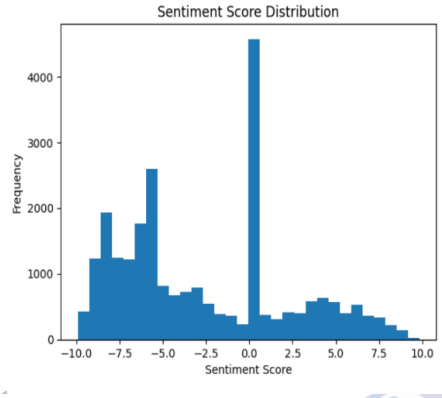

Fig .4 Sentiment Score Graph

These results indicate that the model effectively identifies hate speech. The high recall value indicates that the system successfully detects most instances of hate speech, whereas the high precision ensures that the flagged content is highly reliable.

B. Confusion Matrix Analysis

The confusion matrix provides a detailed breakdown of classification performance:

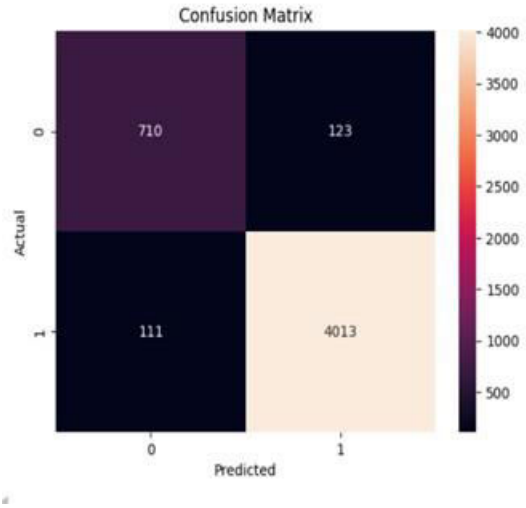

Fig. 5. Confusion Matrix of the Proposed SVM-Based Hate Speech Detection Mode

The model demonstrated a strong performance with minimal misclassification, indicating its robustness in handling real-world textual data.

C. Comparative Performance

Compared with traditional keyword-based approaches and basic machine learning models, the proposed system shows significant improvements in accuracy and contextual understanding. Although deep learning models may achieve comparable accuracy, they require higher computational resources.

D. Qualitative Analysis

The system was tested on various sample inputs to evaluate its real-world applicability.

- Explicit hate speech was correctly identified with high confidence
- Positive and neutral sentences were accurately classified as non-hate
- Contextual phrases were correctly interpreted using n-gram features
- Non-harmful expressions containing strong words were not misclassified

These observations confirm that the system can effectively distinguish between harmful and non-harmful content, even in challenging scenarios.

E. Discussion

The integration of TF-IDF with n-gram features significantly improved the model's ability to capture the contextual information. Additionally, the inclusion of sentiment analysis enhanced the interpretation of

emotional tone, providing deeper insights into user intentions.

The proposed system demonstrates strong scalability and efficiency, making it suitable for real-time applications, such as social media monitoring and content moderation. Its ability to maintain high accuracy at a low computational cost makes it a practical alternative to more complex deep learning models.

A. The experimental results confirm that the proposed SVM-based hate speech detection system, combined with sentiment analysis, provides high accuracy, reliability and efficiency. The proposed system successfully addresses the limitations of traditional approaches and offers a scalable solution for real-world deployment.

X. CONCLUSION

This study presents an efficient system for hate speech detection using a Support Vector Machine (SVM) integrated with sentiment analysis. The system utilizes TF-IDF and n-gram techniques for effective feature extraction and accurate classification of the textual data. The proposed model achieved high performance with an accuracy of 96%, a precision of 97%, a recall of 98%, and an F1-score of 97%, demonstrating its reliability in detecting hate speech while minimizing misclassification. The use of SVM ensures low computational cost and fast processing, making the system suitable for real-time applications in the future. Furthermore, the integration of sentiment analysis enhances contextual understanding by identifying the emotional tones of texts. Overall, the proposed system provides an accurate, efficient, and scalable solution for automated hate speech detection and public sentiment analyses.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] M. Subramanian et al., "A Survey on Hate Speech Detection and Sentiment Analysis Using Machine Learning and Deep Learning Models," Alexandria Engineering Journal, vol. 80, pp. 110–121, 2023.

[2] M. Athoillah and R. K. Putri, "Utilizing Support Vector Machines to Detect Hate Speech on Social Media," Science, Engineering and Technology Journal, vol. 4, no. 2, pp. 53–60, 2024.

[3] S. Li et al., "Hate Speech Detection and Online Public Opinion Regulation Using SVM," Information Journal, 2025.

[4] P. Kar and S. Debbarma, "Sentiment Analysis and Hate Speech Detection Using Hybrid Deep Learning Models," Engineering Applications of Artificial Intelligence, 2023.

[5] X. Zhou et al., "Hate Speech Detection Based on Sentiment Knowledge," Proceedings of the ACL Conference, 2021.

[6] R. Cao et al., "DeepHate: Hate Speech Detection via Multi-Faceted Text Representations," arXiv preprint arXiv:2103.11799, 2021.

[7] M. R. Awal et al., "AngryBERT: Joint Learning Target and Emotion for Hate Speech Detection," arXiv preprint arXiv:2103.11800, 2021.

[8] D. C. Asogwa et al., "Hate Speech Classification Using SVM and Naive Bayes," arXiv preprint arXiv:2204.07057, 2022.

[9] M. S. Jahan and M. Oussalah, "A Systematic Review of Hate Speech Detection Using NLP," arXiv preprint arXiv:2106.00742, 2021.

[10] "Hate Speech Detection in Social Networks Using Machine Learning," International Journal of Advanced Computer Science and Applications, vol. 14, no. 5, 2023.

[11] "Context-Aware Hate Speech Detection Using Sentiment Analysis," IJCNIS, 2024.

[12] "Efficient Hate Speech Detection: Evaluating Multiple Models," ACM Digital Library, 2025.

[13] T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," ICLR, 2013.