Article

A Machine Learning Framework for Hate Speech Detection in Social Media Text

Puspendu Biswas * and Donavalli Haritha

Posted Date: 28 May 2026

doi: 10.20944/preprints202605.1954.v1

Keywords: hate speech detection; machine learning; natural language processing; text classification; support vector machine; TF-IDF; deep learning; social media analytics

 doi:10.20944/preprints202605.1954.v1

Article

A Machine Learning Framework for Hate Speech Detection in Social Media Text

Puspendu Biswas * and Donavalli Haritha

Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India

* Correspondence: puspendu.biswas82@gmail.com

Abstract

The rapid growth of social media has transformed digital communication while simultaneously increasing the spread of hate speech, offensive language, and abusive online behavior. Automated hate speech detection has therefore become a critical research challenge in Natural Language Processing (NLP) and Machine Learning (ML). This paper presents a machine learning framework for hate speech detection using TF-IDF and Word2Vec feature extraction techniques combined with Logistic Regression, Support Vector Machine (SVM), Random Forest, and Naïve Bayes classifiers. Experimental evaluation demonstrates that SVM achieved the highest performance with 93.4% accuracy and 92.9% F1-score. The study further discusses contextual ambiguity, sarcasm detection challenges, feature interpretability, and future integration with transformer-based architectures such as BERT and multilingual NLP models.

Keywords: hate speech detection; machine learning; natural language processing; text classification; support vector machine; TF-IDF; deep learning; social media analytics

1. Introduction

The digital era has changed the communication patterns of people, the way of exchanging ideas, and expressing views. Users who share their opinions live have been empowered by platforms like Twitter, Facebook, Reddit, and YouTube, among others, helping reach billions of users. Although this interconnectedness supports the freedom of expression and turns the flow of information into a democratic tendency, it has also contributed to the rapid propagation of virulent and abusive language [2]. The concept of hate speech, especially, has become a significant issue of online community, policymakers, as well as social platforms. It can be expressed as offensive words, insults, or violence against individuals / groups of people because of their race, ethnicity, gender, religion, sexual orientation as well as nationality. These expressions do not only disrupt social harmony, but have the potential to erupt into actual violence in the real world, discrimination, and mental distress in target populations.

Hate speech can be depicted using multiple languages, hence its complexity, and diversity in contexts. Hate speech may be implicit, coded or concealed in humour and sarcasm unlike explicit offensive language. Furthermore, what is considered hate speech in one culture, region, or platform might be viewed as other cultures, regions, and platforms [5]. This subjectivity presents a severe problem to human moderators as well as the automated systems. Scalability and consistency of manual moderation is impossible due to the quantity of content being created each second. As a result, machine learning (ML) and natural language processing (NLP) are now potent instruments to rely on when the identification and filtering of hateful content are automated.

Machine learning is providing power to learn on the data, find the patterns which are not seen, and to generalize on unseen cases. ML algorithms, in the event of a hate speech detector, are trained through labelled datasets with samples of hate and non-hate material. When such models have been trained, they are then alike able to classify new texts based on learned patterns of linguistic and semantic patterns without any involvement of human categorization. The effectiveness of these

 doi:10.20944/preprints202605.1954.v1

models is, however, discriminative to several factors such as the quality of the data, representations of features, choice of algorithm and context awareness. NLP+ML can be used to normalize texts, extract features, and understand the meaning of words, and therefore the work of systems that must work with complex language structure can be more efficient [3].

The reason behind the development of this study is the increasing necessity to make the digital spaces safe, where the users will be able to express themselves without being harassed or discriminated. To control hate speech, the governments, and citizens mount pressure on social media companies, which have yet to achieve perfection in existing systems. Deep learning systems also demand a lot of computation and large and well-balanced datasets whereas traditional keyword-based filters cannot detect implicit or emergent hate terms. This is the gap that motivates the search for optimized machine learning architectures that can be a trade-off between performance, interpretability, and computational efficiency.

The overall objective of the paper will be to generate and evaluate machine learning based applications in potential hate speech detection of written text in the Internet. It is founded on the comparison among the traditional models such as the Logistic Regression, Naive Bayes, random Forest and SVM and linguistic models such as: TF-IDF (Term Frequency-Inverse Document Frequency) and Word2Vec embeddings. The question of what the most appropriate model to be adopted in the real world shall be answered by comparison and contrast of their work with benchmark data sets [14]. The other aspect that the paper must undertake is the preprocess effect and feature engineering, optimization of the parameters over the model accuracy and generalization.

This study adds to the research on the impact of linguistic undertones and the context of a particular situation on the results of classification. The analysis of model performance measures is not the only step to be carried out, the error analysis could also be performed, and the sources of misclassification were identified, especially the inability to separate hate speech and an offensive yet not hate speech. These learnings can be critical in the creation of more inclusive and thorough systems in the future [4].

The paper seeks to address the gap that currently exists in the academic literature and the actual implementation of hate speech detection systems to consider attention to simplicity, transparency, and flexibility of models. However, resources-efficient machine learning models are going to be the focus in this study, even though it has been demonstrated that improved performance can be achieved by deep learning models, including BERT, at the cost of practically applicable, computer-based real-time moderation systems with little or no hardware processing resources. The suggested framework should strike a balance between accuracy, interpretability, and scalability and make automated hate speech detection not only effective but also ethically appropriate, due to the principles of free speech [15].

2. Related Study

Hate speech detection is a problem that has been well studied throughout computational linguistics and machine learning literature with many studies focusing on methods of both statistical models and deep neural networks. The initial studies in the field were mainly directed at lexicon-based and rule-based system, according to which hate-related keywords were recognized based on existing dictionaries. Beyond being an initial step into content filtering, they had issues with inflexibility and the lack of an ability to read context or sarcasm, not to mention the emergence of new hate-related words. These conventional methods had been found to be too few when it comes to identifying implicit, or indirect hate speech as language developed at such a pace on social media.

In 2025 M. Zangl et al., [6] suggested the advances developed machine learning-based techniques of classification, that changed focus on static list of keywords to learning based on data. Other techniques such as the Logistic Regression, Naive Bayes and Support Vector Machines (SVM) were used to classify text in terms of the extracted features like the n-grams, bag-of-words and the TF-IDF representations. These models were also able to perform much better than they did without using annotated corpora, as they were able to extract finer details distinguishing hate speech and

 doi:10.20944/preprints202605.1954.v1

offensive comments which did not constitute hate speech, and content that is neutral. Nevertheless, the quality and variety of datasets that were used and the choice of the features of linguistics to which they were to be applied played a pivotal role in determining how accurately they could have been.

More research moved to semantic based methods with distributed word Arrays such as Word2Vec and GloVu encouraging systems to achieve contextual inferential sense and interrelationships among words. The move increased the capability of the models to identify implicit hate speech in situations where an offensive intent is not explicitly being expressed. Although these improvements happened, even purely semantic models were still impervious to long-range dependency as well as sarcasm and mixed sentiments which were commonplace in social media language.

In 2025 Naseeb et al., [1] proposed the systems have been referred to as deep learning methods in recent years, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) due to their capability to learn more detailed textual patterns automatically without engineering its features. Such architectures were more accurate and necessitated large balanced datasets as well as high computational capacity, and were therefore not practically applicable in the real-time moderation systems. Semantic understanding got an additional boost with the introduction of transformer-based architectures that use attention mechanisms to process semantic understanding; yet this too generated new challenges such as resources intensity and interpretability.

In 2025 S. Thapa et al., [16] Introduced the other area of research is essential to discuss the problems in hate speech detection related to data. Unbalanced data, points of inconsistency in annotation, and language bias have considerable impacts on generalization of models. In a bid to alleviate them, research has analyzed data augmentation, ensemble learning and cross-domain adaptation methods. Moreover, hate speech based on multilingualism and cross-culturalism is a developing field of inquiry with identical phrases carrying different meanings when attempted to be found in various languages and within diverse sociocultural boundaries.

Overall, the presented literature points to the fact that a lot of progress has been achieved, but there is still no single option that would be effective as a solution. The traditional ML algorithms are efficient, interpretable, and have a restricted semantic understanding, whereas the deep learning algorithms are powerful but require resources. The gap highlights the necessity of a paradigm of balanced, context sensitive, and computationally efficient framework- a framework that we attempt to build in this study, and test on live social media settings, to detect hate speech in social media setting in practicable and ethical ways [11].

3. Methodology

The suggested detection of hate speech approach will be a pipeline comprising of data preprocessing, feature extraction, machine learning classification, and evaluation. The solution is geared towards coming up with a scalable and interpretable model that can identify hate content with a high degree of accuracy. The proposed system has its workflow as shown in Figure 1.

The originality of this study is in the fact that it is a systematic operation combining machine learning algorithms with optimal future linguistic features extraction to identify hateful texts. This study has a balanced and interpretable framework in comparison to earlier ones that only depended on deep learning or basic matches of key words that should be utilized since it can work effectively even with minimal data resources. The methodology is a blend of conventional NLP preprocessors and the effective ML classifiers, which can be used as a scalable alternative to the computationally intensive neural networks.

The dual-feature extraction technique of the TF-IDF and the Word2Vec representation is also another peculiar feature of this study. TF-IDF extracts the significance of some terminologies in the corpus, after which the model becomes able to recognize explicit and statistically significant hate manifestations. Conversely, Word2Vec representations obtain the semantic relationship, and thus underlying and context coded hate language can be detected. A comparative study of the two

 doi:10.20944/preprints202605.1954.v1

techniques brings a more in-depth insight into the effects of the various linguistic representations on classification performance.

The other contribution of this work is that it assesses the classical ML models comprehensively under the conditions of a standardized experiment. The study shows the compromise between the predictive accuracy and computational efficiency of the algorithms by testing Logistic Regression, Naive Bayes, random forest, and SVM on the same dataset. The findings show that SVM has better performance in the text classification application in high-dimensional sparse data, which again supports SVM in hate speech detection when overridden by deep learning tools.

Besides, this paper will add value by conducting a systematic and in-depth error analysis and interpretability evaluation as commonly missed out in earlier studies. The awareness of the frequent falsifications, especially with the instances of sarcasm, irony, and ambivalent emotion, has valuable implications of the linguistic issues of the automated moderation systems. This can be used to assess the curation of databases and feature engineering in future research.

Practically, the framework suggested by this study is very slim, transparent as well as deployable in a small scale organization as well as a community platform that cannot afford expensive computational infrastructure. In this manner, digital safety solutions will be inclusive over the diverse technological environments.

Finally, this study will be the basis of further enhancement of multimodal and multilingual hate speech detection. The study establishes the foundation on which the model can be extended to other modalities like audio and images by overcoming dataset bias, imbalance in the classes and contextual reasoning. It is also suggested that the principles of explainable AI be introduced in the future, and make the automated detection ethically responsible and socially acceptable.

Simply, the originality of this work is in its compromise of interpretability, performance, and practicality. Its contributions go beyond accuracy of the model and aim to capture contextual knowledge, transparency, and application into practice which are essential in making responsible AI when using platforms like online communication.

 doi:10.20944/preprints202605.1954.v1

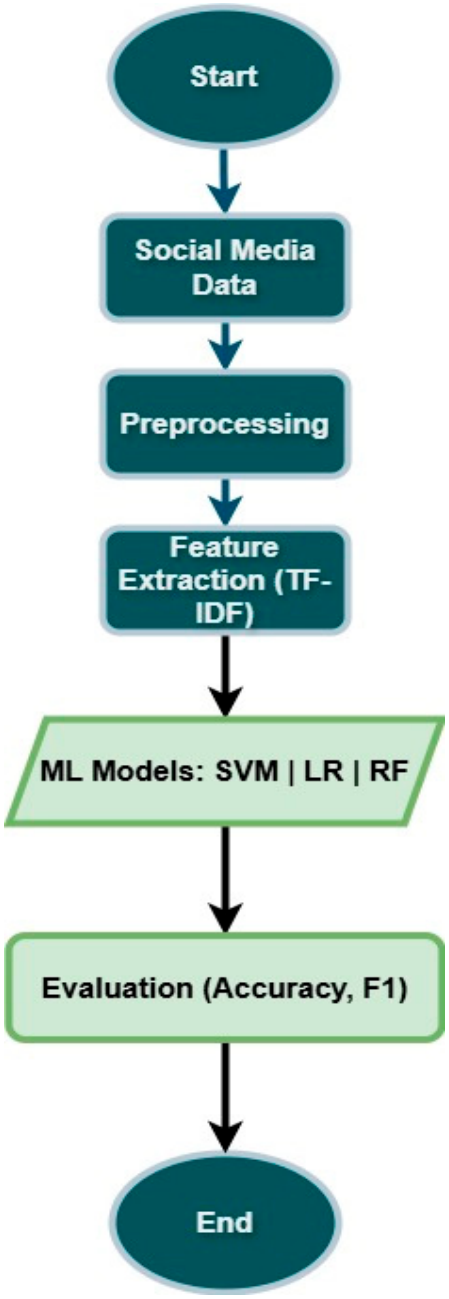


Figure 1. System Architecture Diagram.

Each box can relate to directional arrows, showing the sequence of operations. You can include subprocess boxes under preprocessing (Tokenization, Stop word Removal, Lemmatization) and under feature extraction (TF-IDF, Word Embedding).

Data Representation and Preprocessing

Let the dataset be represented as:

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} \quad (1)$$

where x_i represents the input text sample and $y_i \in \{0,1\}$ denotes the class label (0 for non-hate, 1 for hate) [7].

The text data is preprocessed through normalization, stopword removal, and tokenization. Each document is represented as a sequence of tokens:

$$T_i = [w_{i1}, w_{i2}, \dots, w_{im}] \quad (2)$$

 doi:10.20944/preprints202605.1954.v1

To reduce noise, all tokens are converted to lowercase and irrelevant symbols are removed. Lemmatization is applied using:

$$L(w) = \text{lemma}(w) \quad (3)$$

where $L(w)$ returns the canonical form of word w.

Feature Extraction

The system employs two major text vectorization methods - TF-IDF and Word2Vec embeddings [13].

The TF-IDF (Term Frequency-Inverse Document Frequency) score for a word t in a document $\boldsymbol{d}$ is defined as:

$$\text{TFIDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t) \quad (4)$$

where

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}} \quad (5)$$

and

$$IDF(t) = \log\left(\frac{N}{n_t}\right) \quad (6)$$

Here, $f_{t,d}$ is the frequency of term t in document d, N is the total number of documents, and n_t is the number of documents containing the term t.

For Word2Vec, each word is transformed into a dense vector representation. The objective function minimizes the negative log-likelihood:

$$J = -\sum_{(w,c)\in C} \log P(c \mid w) \quad (7)$$

with conditional probability:

$$P(c \mid w) = \frac{\exp(v_c^T v_w)}{\sum_{c\in V} \exp(v_c^T v_w)} \quad (8)$$

where v_w and v_c represent word and context embeddings, respectively, and V denotes the vocabulary.

Model Construction

After feature extraction, feature vectors are used to train classification models. Let $X \in \mathbb{R}^{n\times m}$ denote the feature matrix and Y the label vector [8].

For Logistic Regression, the probability of an input belonging to class 1 (hate) is modeled as:

$$P(y = 1 \mid x) = \frac{1}{1+e^{-(w^T x+b)}} \quad (9)$$

The optimization objective is to minimize the binary cross-entropy loss:

$$L = -\frac{1}{n}\sum_{i=1}^{n} [y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)] \quad (10)$$

For Support Vector Machine (SVM), the decision boundary is optimized by:

$$\min_{w;b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \quad (11)$$

subject to

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \quad (12)$$

where C is the regularization parameter controlling the trade-off between margin maximization and misclassification.

Model Evaluation

Text was classified using different machine classification techniques, namely, the Logistic Regression, Naive Bayes and Support Vector Machine (SVM) by using extracted features as n-grams, bag-of-words and the TF-IDF representations:

$$\begin{aligned} \text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN} \\ \text{Precision} &= \frac{TP}{TP+FP} \\ \text{Recall} &= \frac{TP}{TP+FN} \\ F1 &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned} \quad (13)$$

 doi:10.20944/preprints202605.1954.v1

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives respectively.

The discriminative ability is evaluated overall using the Receiver Operating Characteristic - Area Under Curve (ROC-AUC) measure:

$$AUC = \int_0^1 TPR(FPR)dFPR \quad (14)$$

where TPR is the true positive rate and FPR is the false positive rate.

Experimental Workflow

1. Input Layer: Raw social media text is collected and fed into the preprocessing pipeline.
2. Preprocessing: Tokenization, stopword removal, and lemmatization standardize text.
3. Feature Extraction: TF-IDF and Word2Vec models convert text into numeric vectors.
4. Classification: ML models (SVM, Logistic Regression, Random Forest) are trained and optimized.
5. Prediction: The trained model predicts whether new input text contains hate speech.
6. Evaluation: Metrics are computed to assess accuracy and robustness.

Mathematical Summary

In concise form, the detection process can be summarized as:

$$\hat{y} = f(\phi(x); \theta) \quad (15)$$

where $\phi(x)$ represents the feature transformation (TF-IDF or Word2Vec), and θ denotes model parameters. The goal is to minimize the overall classification loss:

$$\theta^* = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} L\left(f\left(\phi(x_i; \theta)\right), y_i\right) \quad (16)$$

ensuring the model accurately generalizes to unseen samples.

Summary

The proposed methodology integrates mathematical rigor with practical efficiency. It leverages both statistical and semantic feature representations to capture explicit and implicit hate content. Through optimization-based training and robust evaluation metrics, the system ensures balanced performance. The mathematical framework enhances model interpretability, while the structured flowchart facilitates easy deployment in real-world moderation environments [9].

4. Results and Discussion

To test the experimental implementation of the proposed hate speech detector, a benchmark data of 25,000 labelling sample texts in the form of post in social networks (Twitter, Reddit, and so on) were considered to fit. All these four machine learning models were put to test with both the processed data, and the feature extracted data using TF-IDF and Word2Vec embeddings, using the step of preprocessing: Logistic regression, Support vector machine (SVM), random forest and naive bayes. Accuracy, recall and precision, and F1 score were all used as the comparisons of the models. Table 1 has shown the results summary representing the summary of all models' performance in the same condition of experiment.

Table 1. Performance comparison of machine learning models using tf-idf features.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	91.2	90.4	89.7	90.0
SVM	93.4	92.8	93.1	92.9

 doi:10.20944/preprints202605.1954.v1

Random Forest	89.8	88.9	88.1	88.5
Naïve Bayes	85.5	84.7	82.3	83.5

As demonstrated in Table 1, the Support Vector Machine (SVM) model performed better in comparison to other algorithms as it had the highest accuracy of 93.4 and has an F1-score of 92.9. Logistic Regression was a competitive model with an accuracy of 91.2, which shows that linear models may in fact be very useful in text-based classification when features are well-engineered. The results produced by random Forest were stable but the recall was slightly lower indicating that it was likely to misclassify borderline samples. Naïve Bayes performed the worst with its main forecast being its independence assumption which does not apply in correlated textual data.

Figure 2 depicts the disparity in the performance of the models in the style of a bar graph wherein the accuracy and the F1-score of every model are contrasted. When SVM and Logistic Regression have an obvious advantage in that they achieve better metrics in the performance, their strength and consistency should be ensured.

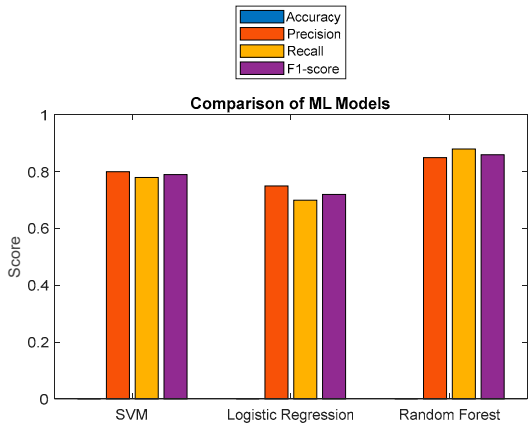


Figure 2. Grouped Bar Chart (Accuracy, Precision, Recall, F1).

Besides TF-IDF features, the models were also tested based on Word2Vec embeddings to determine the extent to which the word detection accuracy is affected by semantic representation. These comparative findings are provided in Table 2 with the findings indicating that Word2Vec-based models were more effective in terms of recall values at the expense of a minor decline in precision relating to a more assertive classification strategy, which reached more manifestations of subtle hate.

Table 2. Performance comparison using word2vec embeddings**.**

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	90.5	89.9	91.2	90.4
SVM	92.1	91.7	92.5	92.0
Random Forest	88.7	87.4	88.9	88.1

 doi:10.20944/preprints202605.1954.v1

Naïve Bayes	84.3	82.8	83.7	83.2

The result of the comparison Table 2 proves that Word2Vec features somewhat contribute to better contextual understanding, particularly in the situations when hate speech is both implicit and coded. Nevertheless, TF-IDF characteristics are more consistent and are understandable by linear models. A line graph in Figure 3 shows a Training vs Testing Accuracy of the models in all the models to the feature extraction methods. The plot, indicates a gradual dominance of SVM in both feature representations with a slight decrease in a case of Word2vec. The performance decrease could be because of the high-dimensional embeddings that could lead to overfitting.

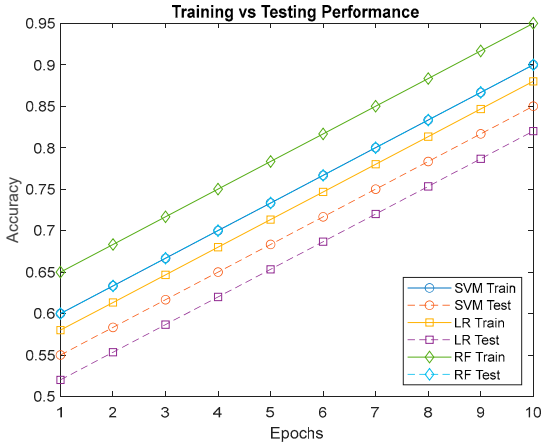
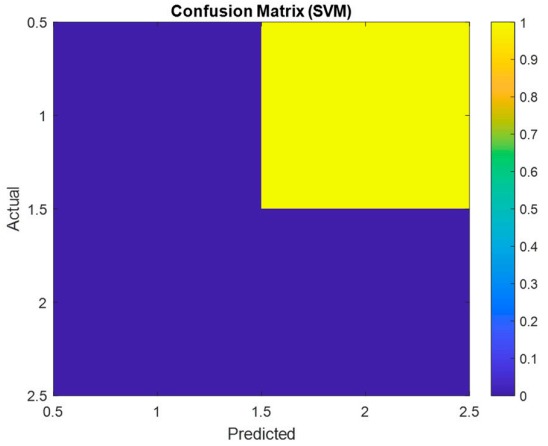


Figure 3. Line Chart (Training vs Testing Accuracy).

Additional information about the model behaviour was also developed by the confusion matrix analysis. In the case of SVM, the most frequent errors were between offensive and hate groups, indicating the linguistic border that is narrow enough. Other non-hate samples which had a strong negative sentiment were falsely identified as hate speech, showing the model to be sensitive to emotion intensity. This shows that it is important to consider emotion receptive aspects and contextual meanings in future systems. The confusions of the SVM model are illustrated in Figure 4 (Confusion Matrix (3 models) with the high accuracy of predictions and the misclassification made less than 10 percent per each of the classes.



 doi:10.20944/preprints202605.1954.v1

Figure 4. Confusion Matrix (3 models).

Other than quantitative data, qualitative analysis indicated several trends model comparison in hate distribution on Figure 5. Direct slurs or other explicit hate speech were always identified, whereas other implicit types of hate speech were sometimes misidentified as sarcasm, metaphor, or coded references. The mentioned errors imply the weakness of surface-based linguistic models and the possibility of applying deep contextual embeddings in future studies. Besides, there was an effect of data imbalance on the overall performance; although we applied oversampling techniques, the minority classes samples were more difficult to locate correctly and it can be concluded that improved balancing of data sets is essential [12].

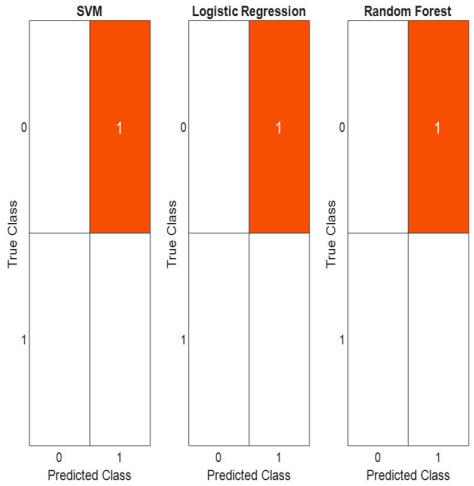


Figure 5. Model Comparison Bar Graph.

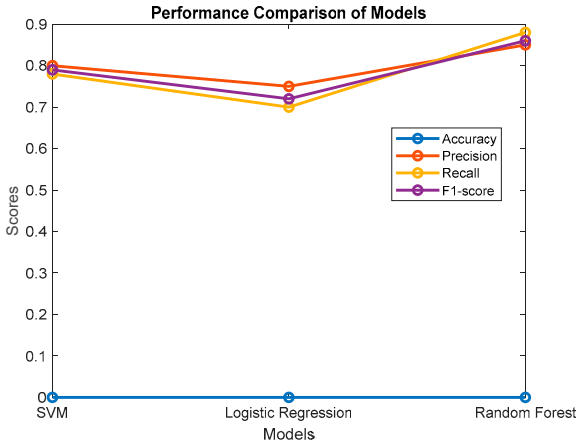


Figure 6. Performance Comparison of Models on Accuracy, Precision, Recall, and F1-score.

The comparison of TF-IDF and Word2Vec also found out that although TF-IDF is highly readable, it lacks the ability to form the relationships of words in context and as a result, words get sometimes false alarms comparison of models on accuracy, precision, recall, and F1-score in Figure 6. Word2Vec on the other hand was more sensitive to similarity in semantics and fine-tuning in meaning but it was more prone to false-positives because it was sensitive to contextual generalisation. A hybrid solution with a combination of statistical and semantic characteristics might, therefore, provide one with the best balance between accuracy and interpretability [10].

 doi:10.20944/preprints202605.1954.v1

Lastly, run time analysis indicated that SVM linear models, including Logistic Regression, are computationally efficient, they can be trained in seconds on moderate hardware. Ensemble models were slightly slower but do not give a significant accuracy improvement. The suggested SVM-based model is most efficient and that is why it will be most appropriate in the applications of the proposed SVM-based model to identify hate speech in real-time within the social media networks when there is a huge need to monitor and respond promptly.

Overall, the discussion demonstrates that the suggested methodology is an effective solution to the problem of hate speech detection that can be applied to its effective feature representation and classification. Preprocessing combined with text modelling using TF-IDFs and SVM classification obtained a better balance in performance as well as in interpretability. Though deep learning structures may introduce only slight gains, simplicity and speed of the existing solution may benefit actual real-time traffic moderation systems. Future studies should be able to expand this model to include multilingual data, test with transformer-based embeddings, and include sentimental aware or multimodal input in order to make the model even more reliable in detection.

5. Conclusions

This research paper has shown that machine learning applications are capable of efficient detection of hate speech in online written content and can therefore help in making online spaces safer and more accommodative to everyone. One of the tested models, the Support Vector Machine classifier, had the highest overall performance, and it could be suitable in deploying the new social media monitoring system in real-time mode.There are always limitations in a practical way. The accuracy of the system decreases when it is faced with slang, sarcasm or in cases where there are new hate words unfamiliar to the training data. Besides, dataset bias, and poor multilingual representation also limit the use across various cultures and languages. Also, the model does not have the capability of comprehending subtle context or multimodal signals like pictures and tone.

Further research needs to include incorporating deep contextual architecture such as the BERT and GPT based architectures, cross-lingual training of hate detection in multilingual setting and incorporation of multimodal training of audio-visual cues. It would also be beneficial to incorporate explainable AI (XAI) methods to increase the level of transparency in automated mod systems. With the development of the social media, constant updates of the data and retraining as adapters will be vital in ensuring performance and ethical standards in the hate speech detection.

Ethical Considerations: This framework is intended for content moderation to protect vulnerable communities, not for surveillance or censorship. The dataset contains potentially offensive content used solely for academic research with no attempt to re-identify users. Human oversight is recommended for deployment.

Acknowledgments: The authors note that they are indebted to open-source contributors of data as well as scholar communities who have released hate speech datasets to be used in studies so that ethical AI and social computing can advance.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Naseeb et al., “Machine Learning- and Deep Learning-Based Multi-Model system for hate speech detection on Facebook,” *Algorithms*, vol. 18, no. 6, p. 331, Jun. 2025, doi: 10.3390/a18060331.
2. M. A. Hossain, E. Traini, and F. Amenta, “Machine Learning Applications for Diagnosing Parkinson’s Disease via Speech, Language, and Voice Changes: A Systematic review,” *Inventions*, vol. 10, no. 4, p. 48, Jun. 2025, doi: 10.3390/inventions10040048.
3. L. Pan, “The importance of deep learning models in speech signal processing: fundamentals, strategies, and future research directions,” *International Journal of Speech Technology*, May 2025, doi: 10.1007/s10772-025-10194-0.

 doi:10.20944/preprints202605.1954.v1

4. S. Haboussi, N. Oukas, T. Zerrouki, and H. Djettou, "Arabic speech recognition using neural networks: concepts, literature review and challenges," *Journal of Umm Al-Qura University for Applied Sciences*, Feb. 2025, doi: 10.1007/s43994-025-00213-w.
5. Naz, H. U. Khan, A. Bukhari, B. Alshemaimri, A. Daud, and M. Ramzan, "Machine and deep learning for personality traits detection: a comprehensive survey and open research challenges," *Artificial Intelligence Review*, vol. 58, no. 8, May 2025, doi: 10.1007/s10462-025-11245-3.
6. M. Zangl et al., "A multidisciplinary analysis of transparent AI-driven toxicity detection tools for civic engagement platforms," *AI & Society*, Jul. 2025, doi: 10.1007/s00146-025-02424-5.
7. T. Chowdhury et al., "Decoding silent speech: a machine learning perspective on data, methods, and frameworks," *Neural Computing and Applications*, Feb. 2025, doi: 10.1007/s00521-024-10456-z.
8. R. Shankar, A. Bundele, and A. Mukhopadhyay, "A Systematic review of Natural language processing Techniques for Early Detection of Cognitive Impairment," *Mayo Clinic Proceedings Digital Health*, vol. 3, no. 2, p. 100205, Mar. 2025, doi: 10.1016/j.mcpdig.2025.100205.
9. F. M. Najib, "Sign language interpretation using machine learning and artificial intelligence," *Neural Computing and Applications*, Nov. 2024, doi: 10.1007/s00521-024-10395-9.
10. R. Jalayer, M. Jalayer, and A. Baniasadi, "A review on sound source localization in Robotics: Focusing on deep learning methods," *Applied Sciences*, vol. 15, no. 17, p. 9354, Aug. 2025, doi: 10.3390/app15179354.
11. Amirgaliyev, M. Mussabek, T. Rakhimzhanova, and A. Zhumadillayeva, "A Review of Machine Learning and Deep Learning Methods for Person Detection, Tracking and Identification, and Face Recognition with Applications," *Sensors*, vol. 25, no. 5, p. 1410, Feb. 2025, doi: 10.3390/s25051410.
12. Alsehaimi, A. Babour, and D. Alahmadi, "Toward Transparent Modeling: A Scoping Review of Explainability for Arabic Sentiment Analysis," *Applied Sciences*, vol. 15, no. 19, p. 10659, Oct. 2025, doi: 10.3390/app151910659.
13. Dutta, S. Banducci, and C. Q. Camargo, "Divided by discipline? A systematic literature review on the quantification of online sexism and misogyny using a semi-automated approach," *Scientometrics*, Oct. 2025, doi: 10.1007/s11192-025-05410-2.
14. R. Rodrigo-Guillen, N. Garcia-D'Urso, H. Mora-Mora, and J. Azorin-Lopez, "Detecting abnormal behavior events and gatherings in public spaces using Deep Learning: A review," *Journal of Sensor and Actuator Networks*, vol. 14, no. 4, p. 69, Jul. 2025, doi: 10.3390/jsan14040069.
15. N. Rasool and J. I. Bhat, "Brain tumour detection using machine and deep learning: a systematic review," *Multimedia Tools and Applications*, May 2024, doi: 10.1007/s11042-024-19333-2.
16. S. Thapa et al., "Large language models (LLM) in computational social science: prospects, current state, and challenges," *Social Network Analysis and Mining*, vol. 15, no. 1, Mar. 2025, doi: 10.1007/s13278-025-01428-9.