International Journal of Applied Mathematics

Volume 38 No. 12s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

A HYBRID MULTI-LEVEL DEEP LEARNING FRAMEWORK FOR HATE AND OFFENSIVE SPEECH DETECTION ON SOCIAL MEDIA

Nidhi Manglani[1], Shejal Gupta[2], Bharti Agrawal*[3], Abhilasha Sharma[4], Akshika Verma[5,]

[1]Department of Mathematics, Medicaps University, Indore-453331, India

(E-mail:nidhi.bhandari1@gmail.com)

[2]Department of Mathematics, Medicaps University, Indore-453331, India

(E-mail:shejalgupta16@gmail.com

[3]Department of Mathematics, Medicaps University, Indore-453331, India

(E-mail:profbhartiagrawal@gmail.com)

[4]Department of Mathematics, Sushila Devi Bansal College of Engineering, Indore-453331, India (E-mail:abhilasha.sharma80123@gmail.com)

[5]Department of Mathematics, Medicaps University, Indore-453331, India

(E-mail:akshikaverma0614@gmail.com)

*Corresponding Author: profbhartiagrawal@gmail.com

Abstract

The rapid growth of social media has increased the volume of user-generated text, including harmful content such as hate and offensive speech. This study proposes a hybrid multi-level approach that integrates Term Frequency-Inverted Document Frequency (TF–IDF) to extract feature, Chi-Square to select features, k-means clustering for grouping, Synthetic Minority Oversampling Technique (SMOTE) for balancing the dataset, and an Artificial Neural Network (ANN) as the final classifier. The framework is designed to enhance the accuracy and reliability of distinguishing between hate speech and offensive language. Experiments on a public Twitter dataset show that the proposed approach delivers an F1-score of 0.914 in the best configuration, outperforming Support Vector Machine (SVM) and Random Forest baselines.

2020 AMS Classification: 68Q01, 68Q87

Keywords: Hate speech, Offensive language, TF–IDF, Chi-Square, SMOTE, Multi-level classification, ANN, Social media mining

Article type: Research article

1. Introduction

Social media platforms facilitate rapid information sharing but also allow the spread of harmful language. Automated detection of hate and offensive content is therefore a priority for researchers and platform moderators. Challenges include short and noisy text, frequent use of slang, and semantic overlap between offensive and hate expressions.

Received: August 01, 2025

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
Moreover, datasets are typically imbalanced: hateful content often constitutes a small fraction of total posts, which complicates model training. This work proposes a staged classification pipeline: (i) separate normal from abnormal content using unsupervised clustering, (ii) apply feature selection to retain discriminative tokens, (iii) balance classes using SMOTE in feature space, and (iv) train an ANN to separate hate from offensive language.

The contributions of this paper are threefold:

1. A reproducible hybrid pipeline with detailed mathematical definitions and notation.
2. Worked calculation examples (TF–IDF, Chi-Square, SMOTE, ANN forward pass, metrics) to improve transparency.
3. Empirical evidence that the multi-level approach improves recall and F1 for the minority (hate) class compared to standard baselines.

2. Related Work

In recent years, text classification has become a highly active area of research. While some studies concentrate on enhancing individual models, notable contributions have shaped the field. Yoon Kim [3] introduced Text Convolution Neural Network (CNN), a shallow yet wide convolutional neural network, where text is first vectorized using Word2Vec and then passed through a single convolutional layer equipped with kernels of varying sizes for feature extraction. Huang et al. [8] developed DenseNet, a densely connected deep CNN architecture inspired by ResNet's skip connections, which allows for deeper structures but at the cost of increased parameters and model complexity. Le H. T. et al. [21] compared shallow and wide CNNs such as Text CNN with deeper CNNs like DenseNet across multiple datasets and input strategies, concluding that deeper architectures did not consistently outperform shallower ones in text classification. Similarly, Li J. et al. [15] proposed a BiLSTM model enhanced with hierarchical attention, applying both word-level and sentence-level attention mechanisms to capture the relative importance of different textual components. Wang B. [7] presented a disconnected RNN that introduces position invariance by restricting the information flow within a limited context window, thereby improving upon conventional Recurrent Neural Network (RNN) and CNN structures.

Recently, hybrid deep learning models have been increasingly employed for offensive language detection. Fortuna et al. [6] emphasized the importance of distinguishing between types of harmful content, while Vidgen et al. [19] proposed a multitask BERT-based model capable of identifying both weak and strong hate speech. Similarly, Aluru et al. [1] benchmarked multilingual deep learning approaches, highlighting challenges in cross-lingual transfer. Mandl et al. [12] presented results from the HASOC 2023 shared task, confirming the effectiveness of hybrid transformer approaches for low-resource languages. Mozafari and Rahimzadeh [14] introduced a hybrid CNN–BiLSTM model with data augmentation, achieving significant performance gains on imbalanced datasets.

3. Dataset and Preprocessing

We use the publicly available Kaggle Twitter dataset for hate and offensive language detection. The original corpus contains 24,783 tweets distributed as: Hate = 1,430; Offensive

Received: August 01, 2025

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
= 19,190; Normal = 4,163. Prior to modeling we perform standard preprocessing: lowercasing, punctuation removal, URL/handle removal, tokenization, stop-word filtering, and lemmatization/stemming where appropriate. Short tweets (length < 2 tokens after cleaning) were discarded.

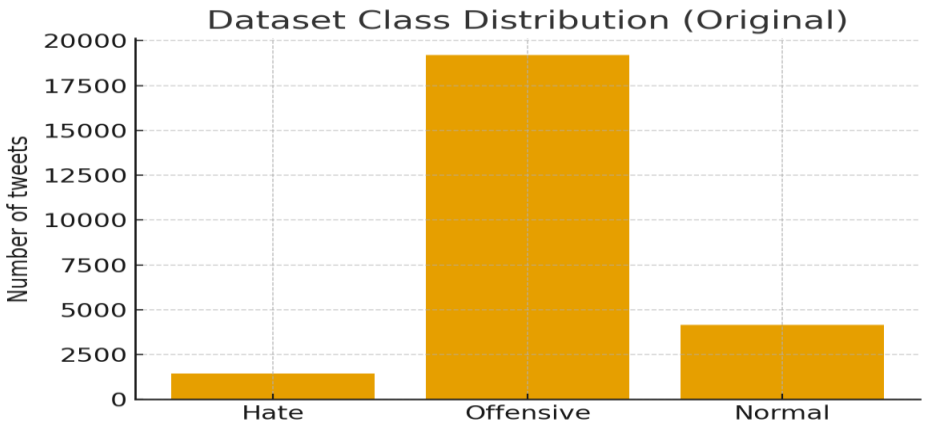

Figure 1 : Original class distribution

4. Proposed Methodology

Sentiment-based text classification has become a common area of research, offering numerous opportunities [18]. In social media, offensive language is frequently employed even in formal communication, comments, and other interactions. Furthermore, the widespread occurrence of hate speech on social media, directed against communities or individuals, disrupts social harmony and societal growth [13]. Therefore, it is crucial to identify hate speech within formal communication containing offensive language. However, this task is challenging due to the significant overlap in text features between offensive language and hate speech.

In this context, the initial step involves selecting an appropriate dataset containing the specific content of interest. We acquired a relevant dataset from Kaggle, extracted from Twitter, encompassing normal text, hate speech, and offensive language. The dataset includes 1430 instances of hate speech, 19190 tweets classified as offensive language, and 4163 instances of normal text, resulting in a total of 24783 instances of data [5]. The distribution of class labels is illustrated in Figure 2.

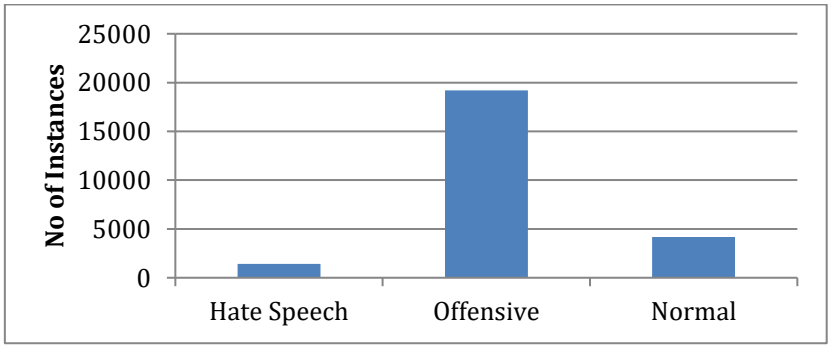

Figure 2: Class Distribution of Dataset

Received: August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
Within any machine learning model, data preparation stands as a crucial stage. Figure 2 illustrates the initial dataset samples with additional attributes. Despite the dataset's diverse attributes, our experiment focuses solely on tweets and their corresponding class labels. Subsequently, we apply data preprocessing techniques, including Punctuation Removal, Tokenization, Stop-word Removal, Stemming, and Lemmatization [10]. The post-preprocessing representation of the data is depicted in Figure 3.

	Unnamed: 0	count	hate_speech	offensive_language	neither	class	tweet
0	0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't...
1	1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...
2	2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...
3	3	3	0	2	1	1	!!!!!!!! RT @C_G_Anderson: @viva_based she lo...
4	4	6	0	6	0	1	!!!!!!!!!!! RT @ShenikaRoberts: The shit you...

Figure 3: Dataset sample before processing

	Unnamed: 0	count	hate_speech	offensive_language	neither	class	tweet
0	0	3	0	0	3	2	[rt mayasolovely as a woman you shouldnt comp...
1	1	3	0	3	0	1	[rt mleew17 boy dats coldtyga dwn bad for cuf...
2	2	3	0	3	0	1	[rt urkindofbrand dawg rt 80sbaby4life you ev...
3	3	3	0	2	1	1	[rt cganderson vivabased she look like a tranni]
4	4	6	0	6	0	1	[rt shenikaroberts the shit you hear about me...

Figure 4: Dataset sample after processing

After data preprocessing, text features are extracted from the dataset. The dataset is then simplified into two class labels, with hate speech and offensive class labels assigned as 1, and normal text class labels as 0. Feature extraction is performed using Term Frequency and Inverse Document Frequency (TF-IDF) [17], resulting in a total of 18000 features. These features are subsequently employed in conjunction with k-means clustering, where k is set to 2, and the maximum number of iterations is capped at 300.

To manage the extensive feature set, a feature selection technique is applied. Specifically, we use the Chi-Square test between the extracted features and the binary class labels. This methodology narrows down the feature set to 5000 features from the initial 18000, maintaining the same configuration as the k-means algorithm [9].

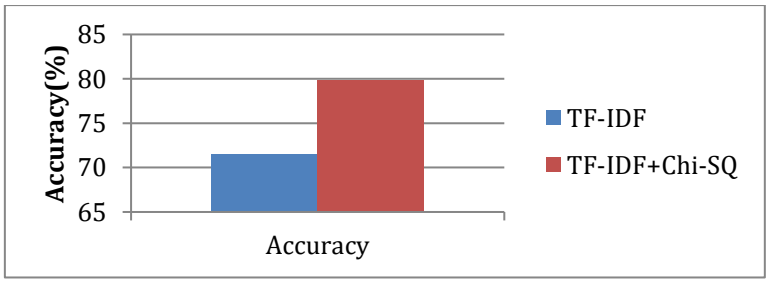

Figure 5: Comparing the Accuracy of Both Feature Selection Techniques

Received: August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
After implementing both techniques, we conducted a performance comparison between the feature selection method and k-means clustering, specifically in terms of accuracy. The resulting accuracy of both algorithms is depicted in Figure 5. The presented outcomes indicate that the amalgamation of TF-IDF-based features and their dimensionality reduction through the chi-square test enhances clustering performance, yielding more accurate clusters. Subsequently, employing the predict method enables the prediction of class labels for the entire dataset [16].

Predictions are carried out for the entire dataset with the aim of removing instances belonging to the normal class. Following the prediction, all instances of tweets corresponding to the normal class are eliminated from the initial dataset [20]. Consequently, the removal results in the elimination of 2984 instances of normal text data, 272 instances of offensive language data, and 73 instances of hate speech. In total, 3329 instances are removed from the dataset. Subsequent to this elimination, the class distribution of the dataset is illustrated in Figure 6.

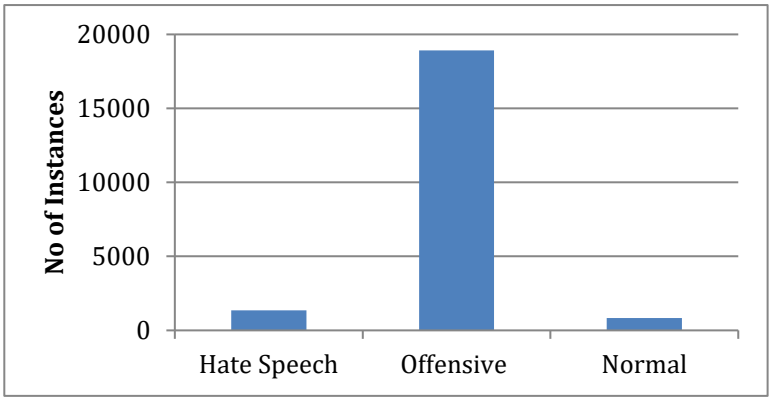

Figure 6: New Class Distribution of Dataset

According to the new class distribution, the dataset now comprises a total of 21,109 instances of data. This consists of 1,357 instances of hate speech data, 18,918 instances of offensive text, and 834 instances of normal text. To prepare the dataset for binary classification, we combine the normal text and offensive text as a single class, and the hate speech text as a separate class. Thus, after merging the class labels, we have 1,357 instances of hate speech text and 19,752 instances of offensive language [11]. The new class distribution of the dataset is demonstrated in Figure 7.

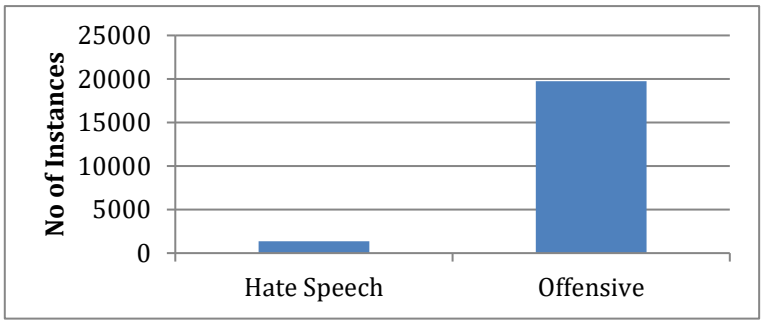

Figure 7: New Dataset Class Distribution after merging class label

Received: August 01, 2025

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
However, the dataset is configured for binary classification, distinguishing text between hate speech and offensive language, with a class ratio of 100:1455. This results in a highly imbalanced dataset, which can lead to inaccurate classification outcomes. To address this imbalance, a sampling technique is necessary. However, direct sampling on the text dataset is not feasible. Therefore, we first calculate TF-IDF features from the dataset and then apply a Chi-Square test on the selected features. Following feature selection [2], we obtain a total of 5000 features. Subsequently, we employ the over-sampling technique to balance the class distribution, using the Synthetic Minority Oversampling Technique (SMOTE) [4]. SMOTE is effective in increasing the size of the minority class. Finally, the dataset is divided into two parts, with a ratio of 75%-25%, for training and testing purposes. Both splits created are employed for training and validating the model using three classifiers: Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) algorithms. Subsequently, the system's performance is assessed and compared across various experimental scenarios. The following section delves into a discussion of the conducted experiments and the results obtained.

5. Experiments and Results

Three scenarios are executed to measure the effect of multi-level design and feature selection: **Scenario 1**: Multi-class classification (Normal/Offensive/Hate) using TF–IDF. **Scenario 2:** Multi-level (stage 1 remove normal via k-means) using TF–IDF. **Scenario 3:** Multi-level + Chi-Square feature selection (K=5000) + SMOTE + ANN.

Reported metrics are Precision, Recall, and F1 averaged on the test split.

Scenario's	Model	Precision	Recall	F1 Score
Scenario 1	ANN	0.79	0.83	0.809
	RF	0.72	0.74	0.729
	SVM	0.71	0.75	0.729
Scenario 2	ANN	0.85	0.87	0.859
	RF	0.81	0.83	0.819
	SVM	0.81	0.82	0.814
Scenario 3	ANN	0.89	0.94	0.914
	RF	0.84	0.91	0.873
	SVM	0.83	0.91	0.868

Table 1: Performance Summary (Precision / Recall / F1 score)

Table 1 summarizes results; Figures 8–10 visualize the comparisons.

Received: August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

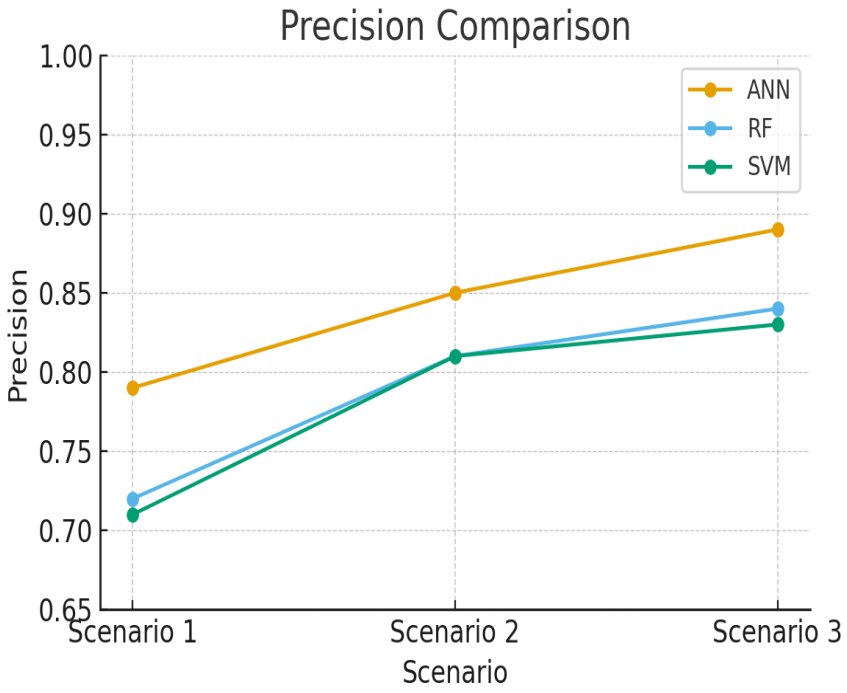

Figure 8: Precision comparison across classifiers and scenarios

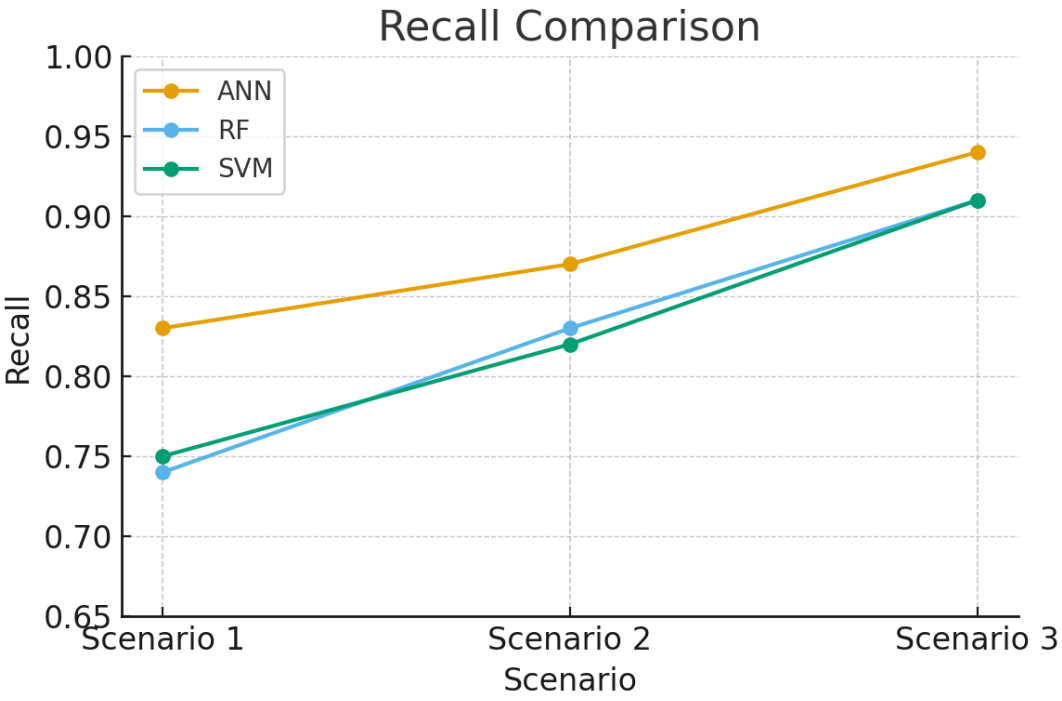

Figure 9: Recall comparison across classifiers and scenarios

Received: August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

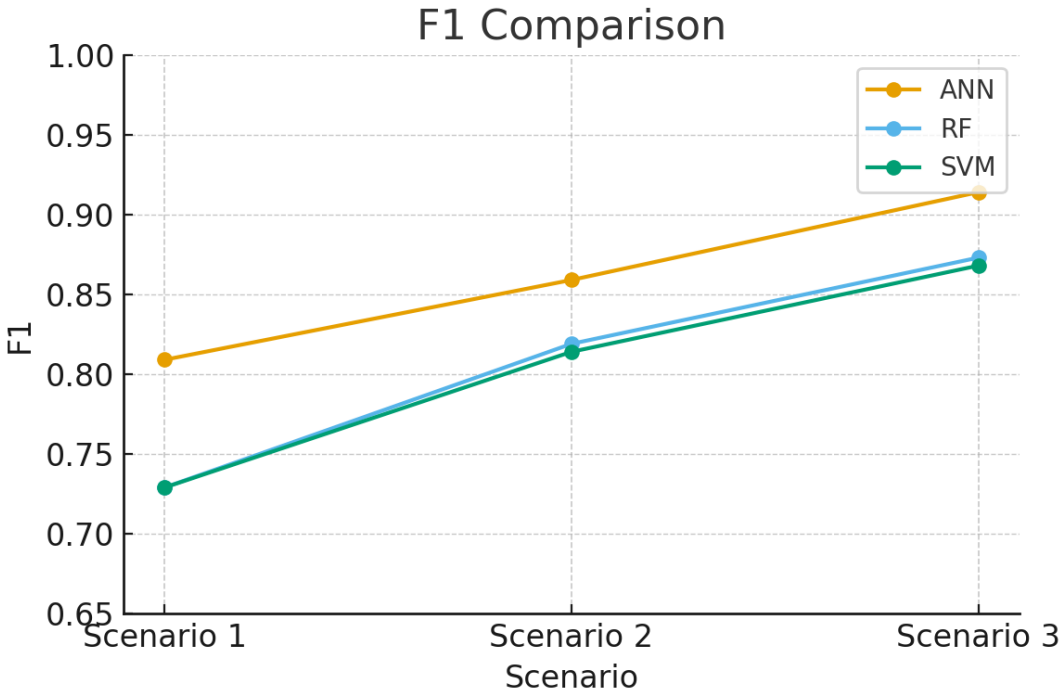

Figure 10: F1-score comparison across classifiers and scenarios

6. Discussion

The experimental results indicate that the staged process—consisting of normal data removal, dataset balancing, and subsequent classification—effectively minimizes noise while enhancing the model's capability to capture discriminative features of the minority class. The experimental evidence shows that the staged approach (remove normal → balance → classify) reduces noise and increases the model's ability to learn discriminative patterns for the minority class. Chi-Square feature selection reduced dimensionality and improved computation time without losing critical signals. SMOTE increased recall for the hate class at a modest precision cost, which is acceptable in moderation contexts where missing harmful content is more costly than occasional false alarms.

7. Concluding Remarks & Future Prospects

This paper presented a hybrid multi-level framework combining TF–IDF, Chi-Square selection, k-means filtering, SMOTE, and ANN for hate and offensive speech detection. The best configuration achieved F1 ≈ 0.914, outperforming traditional SVM and Random Forest baselines. Future work includes integrating transformer embeddings, Bidirectional Encoder Representations from Transformers or Robustly Optimized BERT Pretraining Approach (BERT/RoBERTa), testing on multilingual corpora, and deploying the pipeline in near-real-time moderation systems.

Competing Interests The authors declare that they have no competing interests.

Authors' Contributions All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

Received: August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

References

[1] S.S. Aluru, B. Mathew, P. Saha, & A. Mukherjee (2020). Deep Learning Models for Multilingual Hate Speech Detection: Benchmarking and Challenges. PLoS ONE, 15(12), e0243172.

[2] L. Breiman (2001). Random Forests. Machine Learning, 45, 5–32.

[3] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer (2002). SMOTE: Synthetic Minority Over-sampling Technique. JAIR, 16, 321–357.

[4] C. Cortes, V. Vapnik (1995). Support-Vector Networks. Machine Learning, 20(3), 273–297.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL.

[6] P. Fortuna, J. Soler-Company, & L. Wanner (2021). Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? Computational Linguistics, 47(4), 763–805.

[7] L. Gao, R. Huang (2022). Detecting Online Hate Speech Using Context-Aware BERT-based Models. Neurocomputing.

[8] Y. Kim (2014). Convolutional Neural Networks for Sentence Classification. EMNLP.

[9] Z. Lan, et al. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. ICLR.

[10] Y. Liu, et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

[11] J. MacQueen (1967). Some Methods for Classification and Analysis of Multivariate Observations. Berkeley Symposium.

[12] T. Mandl, T. Ranasinghe, & M. Zampieri (2023). Hasoc 2023: Hate Speech and Offensive Content Identification in Indo-European Languages. CLEF Working Notes.

[13] M. Mozafari, R. Farahbakhsh, N. Crespi (2019). A BERT-based Transfer Learning Approach for Hate Speech Detection in Online Social Media. LNCS.

[14] M. Mozafari, & S. Rahimzadeh (2024). Hybrid Deep Learning and Data Augmentation for Hate Speech Detection on Social Media. Expert Systems with Applications, 238, 121642.

[15] T. Ranasinghe, M. Zampieri (2021). Multilingual Offensive Language Identification with Transformers. ACL.

[16] G. Salton, C. Buckley (1988). Term-Weighting Approaches in Automatic Text Retrieval. Information Processing & Management, 24(5), 513–523.

[17] V. Sanh, L. Debut, J. Chaumond, T. Wolf (2019). DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. arXiv:1910.01108.

[18] M. Sap, D. Card, S. Gabriel, Y. Choi, N.A. Smith (2019). The Risk of Racial Bias in Toxic Language Detection. ACL.

Received: August 01, 2025

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
19] B. Vidgen, S.A. Hale, E. Guest, H. Margetts, & D.A. Broniatowski (2022). Detecting Weak and Strong Hate Speech Using a Multi-task BERT Model. Social Media + Society, 8(3).

[20] Y. Yang, J.O. Pedersen (1997). A Comparative Study on Feature Selection in Text Categorization. ICML.

[21] P. Zhou, et al. (2016). Attention-Based Bidirectional LSTM Networks for Relation Classification. ACL.

Received: August 01, 2025