Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 02 (2026)

Comparative Evaluation of Feature Representation Techniques for Hate Speech Detection in Social Media Text

Priyanshu Jadon, Dr. Deepshikha Bhatia, Dr. Durgesh Kumar Mishra

Ph.D Scholar, CSE The IIS University, Jaipu,India

Sr. Assistant Professor, CSE The IIS University, Jaipu,India

Director, Symbiosis University of Applied Sciences, Indore, India

Abstract:- Social media platforms generate an enormous volume of short and highly dynamic textual content, much of which is posted without effective moderation. Although several computational approaches have been proposed for analyzing social media data, reliable hate speech detection remains difficult because model performance is strongly influenced by the quality of feature representation. This study examines the impact of multiple feature extraction techniques on a deep learning model for hate speech classification. Five feature representation methods—bi-gram features, part-of-speech (PoS) features, count vectorization, TF-IDF features, and word embeddings—were implemented and evaluated using a Sequential Convolutional Neural Network (SCNN). The models were assessed on a publicly available hate speech detection dataset through comparative experimental analysis. The findings indicate that count vectorizer and TF-IDF representations achieve superior training and validation accuracy, while changes in feature dimensionality significantly influence classification performance. The results also show that higher-dimensional feature spaces increase computational cost in terms of memory usage and execution time. Overall, the study highlights the importance of word-level textual information in the classification of short social media posts. . Based on these observations, future work will focus on designing a richer feature descriptor to further improve hate speech detection in short-text environments**.**

Keywords*: hate speech detection, social media text analysis, feature representation, deep learning, text classification, feature dimensionality*

1. Introduction

Social media text classification has received considerable attention because online platforms produce large volumes of short, informal, and noisy text. Unlike conventional documents, social media posts often contain slang, abbreviations, spelling variations, code-mixing, and context-dependent expressions, which make automatic classification more difficult. These characteristics become even more challenging when the objective is to identify hateful or offensive language with high reliability. For this reason, the selection of appropriate feature representation techniques plays a critical role in hate speech detection.

The present study is intended to examine how different text feature extraction methods influence the performance of deep learning-based classification of social media text. To achieve this objective, we first review recent contributions in social media text analytics, with particular attention to datasets, feature engineering strategies, and learning algorithms used for hate speech detection. We then conduct an experimental comparison of multiple feature representation techniques in order to analyse their effect on classification accuracy and computational efficiency. The outcome of this work helps establish the importance of feature design for reliable social media text classification and provides direction for future model improvement.

Online social platforms generate large-scale unstructured text that must often be organized automatically because manual analysis is impractical. Hartmann et al. [1] compared ten automated text classification methods across 41 social media datasets spanning different platforms, sample sizes, and languages. Their study showed that machine learning approaches generally outperform lexicon-based techniques, with Random Forest and Naive Bayes offering particularly competitive results under several settings.

Research on Roman Urdu sentiment analysis has remained relatively limited despite the widespread use of the language. Rana et al. [2] addressed this gap through an unsupervised framework that combines text normalization, sentiment lexicons, negation handling, and stemming. Their experiments on two public Roman Urdu datasets demonstrated that careful pre-processing and lexical modelling can significantly improve short-text sentiment classification.

Hate speech detection is especially difficult because the task requires appropriate datasets, clear annotation standards, and robust classifiers. Qureshi et al. [3] explored both baseline and self-discovered text mining features for multi-class hate speech classification and reported strong performance using dimensionality reduction with complex non-linear models. Their work emphasizes the value of balanced datasets and discriminative features for capturing subtle linguistic patterns.

A related challenge is the distinction between hateful content and merely offensive language. Khan et al. [4] proposed HateClassify, a CNN-based framework that reformulates the task beyond conventional multi-class categorization. Their results suggest that alternative labelling strategies can substantially improve hate speech detection accuracy.

Social media analytics has also been applied in disaster management, where the goal is to classify event-related messages for situational awareness. Yu et al. [5] evaluated a CNN model on geo-tagged Twitter datasets from multiple hurricanes and found that deep learning models generalize better than traditional classifiers such as SVM and Logistic Regression in cross-event classification settings.

For Indonesian tweets, Putri et al. [6] compared several classification algorithms, including Naive Bayes, MLP, AdaBoost, Decision Tree, and SVM, on hate speech data related to politics, religion, ethnicity, and race. Their results indicated that Multinomial Naive Bayes achieved the best performance, while the study also highlighted the effect of data imbalance handling through SMOTE.

The problem becomes more complex when hate speech appears in code-mixed text. Santosh et al. [7] investigated hate speech detection in Hindi-English social media content using hierarchical and sub-word level LSTM architectures with attention mechanisms. Their work illustrates the importance of handling mixed-language structures in real-world online communication.

Capturing semantic context is another important direction in this area. Senarath et al. [8] combined vector space semantics, neural word embeddings, and domain knowledge to improve hate speech classification. They reported that hybrid semantic representations can increase F1-score across Twitter datasets, showing that contextual meaning contributes beyond surface-level token features.

Studies in languages other than English further demonstrate the diversity of the problem. Pronoza et al. [9] examined ethnicity-targeted hate speech in Russian social media using large-scale data and showed that fine-tuned RuBERT enriched with linguistic features produced the strongest results. Their findings confirm the usefulness of language-specific modelling for complex hate speech categories.

The HASOC shared task expanded multilingual hate speech research in Hindi, German, and English. Mandl et al. [10] reported that LSTM-based systems built on word embeddings were among the most frequently used and delivered strong macro-F1 performance, reinforcing the value of deep sequence models for offensive content detection.

Large labeled corpora are difficult to construct because hate speech often depends on context, poor writing quality, and subtle paralinguistic cues. Kovács et al. [11] proposed a deep neural architecture combining recurrent and convolutional components and demonstrated that performance can be improved through better resource utilization and innovative model design. Beyond textual features, some researchers have explored social network structure as an additional signal. Wich et al. [12] introduced a method for identifying hate networks on Twitter using a pre-trained model and contributed both a large anonymized offensive language dataset and a corresponding social graph for further analysis.

Modha et al. [13] framed offensive language detection as a multilingual, multi-level, and multi-class problem. Their results showed that deep neural and transfer learning approaches such as BERT, CNNs, and LSTMs generally outperform traditional SVM-based systems in more demanding classification settings.

Hate speech can also appear in multimodal form, where textual and visual cues jointly convey abuse. Kumar et al. [14] proposed a quaternion-based multimodal fusion architecture and demonstrated that efficient fusion can improve hate speech classification while reducing model complexity. Ethical moderation strategies have also been discussed in the literature. Ullmann et al. [15] proposed quarantining harmful content as a preventive mechanism that attempts to balance online safety with freedom of expression, highlighting that detection is not only a technical issue but also a governance challenge.

Multilingual hate speech detection remains an open challenge because features and architectures do not transfer equally well across languages. Corazza et al. [16] evaluated multiple feature sets and recurrent architectures across English, German, and Italian datasets, underscoring the importance of suitable feature selection in multilingual environments.

Automatic feature selection has also been studied as an alternative to manual engineering. Zhang et al. [17] combined convolutional and LSTM networks and reported strong results on several public datasets, while also showing that compact automatically selected features can outperform larger handcrafted spaces.

The Hateful Memes challenge introduced by Kiela et al. [18] demonstrated that unimodal text-only systems are often insufficient for complex hateful content. Their work established a difficult benchmark in which multimodal reasoning is essential for performance beyond simple baselines.

Class imbalance is another persistent problem in hate speech classification because hateful instances are typically sparse. Rizos et al. [19] explored text augmentation strategies for deep learning models and observed notable improvements in macro-F1 and recall, especially for minority classes.

For low-resource code-switched settings, Chopra et al. [20] proposed a three-stage pipeline combining author profiling, graph embeddings, and profanity modelling for Hindi-English hate speech detection. Their results showed that integrating social and linguistic signals can improve classification quality while also raising important questions about bias and reproducibility.

2. Objectives

The primary objective of this study is to analyze and compare different approaches for hate speech detection in text classification, including lexicon-based methods, traditional machine learning models such as Naive Bayes, Support Vector Machines, Logistic Regression, and Random Forest, as well as deep learning architectures. The study aims to evaluate how effectively these methods capture contextual meaning, implicit aggression, sarcasm, and complex linguistic patterns such as code-mixed language, while also examining the impact of feature representation on model performance, particularly in handling challenges such as noisy user-generated text, class imbalance, multilingual expressions, and ambiguity between offensive and hateful content. Furthermore, the research seeks to identify the most effective techniques for improving classification accuracy and to provide a comparative understanding of different feature extraction approaches used in hate speech detection.

The literature on text classification and hate speech detection shows a clear progression from lexicon-based approaches to machine learning and, more recently, deep learning methods. Earlier studies largely depended on manually curated lexical resources and handcrafted statistical features. Although these methods were useful as

foundational approaches, they were often limited in their ability to capture contextual meaning, implicit aggression, sarcasm, and platform-specific language variation. With the growth of labelled datasets, traditional machine learning techniques such as Naive Bayes, Support Vector Machines, Logistic Regression, and Random Forest became popular for social media text analysis. These models improved predictive performance by learning from data rather than relying entirely on predefined lexicons; however, they still depended heavily on the quality of manually designed features and pre-processing pipelines. Deep learning later transformed the field by enabling automatic representation learning from text sequences. Architectures such as CNNs, RNNs, LSTMs, GRUs, and transformer-based models have shown strong performance because they can capture local patterns, long-range dependencies, and contextual semantics more effectively than conventional approaches. Despite these advances, several challenges remain unresolved. Hate speech detection continues to be affected by noisy user-generated text, class imbalance, multilingual and code-mixed expressions, limited contextual clues, and ambiguity between offensive and hateful content. These issues indicate that feature representation still plays a decisive role in overall system performance, which motivates the comparative investigation conducted in the present work. Some representative studies are summarized in Table I

Ref.	Research area	Work done	Algorithm used	Results
[1]	Large-scale social media analytics for sentiment/content organization	Benchmark study comparing automated text classification methods across 41 social media datasets.	SVM, Naive Bayes, Random Forest, LIWC and other automated methods	Random Forest performed strongly, Naive Bayes worked well on smaller samples, and lexicon-based methods generally lagged behind ML models.
[2]	Roman Urdu short-text sentiment analysis	Introduced an unsupervised framework for Roman Urdu sentiment classification.	Normalization, opinion lexicons, negation handling, stemming	The approach outperformed earlier models on two public datasets and showed strong domain adaptability.
[3]	Multi-class hate speech detection in social media	Built a balanced dataset and explored baseline as well as newly designed text features.	Text mining features with dimensionality reduction and non-linear classifiers	Feature-rich modeling with LSA-supported reduction produced competitive results, with CATBoost reported as the best-performing model.
[4]	Separating hate speech from offensive content	Reframed hate detection using the HateClassify framework.	Sequential CNN (SCNN)	The CNN-based method outperformed several alternatives and showed gains when the problem formulation moved beyond standard multi-class classification.
[5]	Cross-event social media text classification for disaster	Tested deep learning for topic	CNN vs. SVM and Logistic Regression	CNN achieved better generalization and

	response	classification across hurricane-related Twitter events.		higher accuracy than traditional baselines in both single-event and cross-event settings.
[6]	Hate speech detection in Indonesian tweets	Compared multiple classical classifiers on 4,002 tweets covering politics, religion, ethnicity, and race.	Naive Bayes, MLP, AdaBoost, Decision Tree, SVM, with SMOTE analysis	Multinomial Naive Bayes delivered the best accuracy and recall, while imbalance handling was also examined.
[7]	Hate speech in code-mixed social media text	Studied Hindi-English hate speech detection in mixed-language posts.	Hierarchical LSTM and sub-word level LSTM with attention	The study showed that specialized deep sequence models are useful for code-mixed hate speech detection.
[8]	Semantic feature learning for hate speech detection	Investigated whether richer semantic representations improve hate intent classification.	Vector space semantics, neural embeddings, domain knowledge features	Hybrid semantic features improved F1-score and better captured contextual meaning in Twitter data.
[9]	Ethnicity-targeted hate speech detection in Russian	Addressed fine-grained stance toward ethnic groups mentioned in the same text.	SVM, deep learning, RuBERT with linguistic features	Fine-tuned RuBERT combined with linguistic features achieved the strongest F1 performance for both binary and multi-class settings.
[10]	Multilingual HASOC hate speech evaluation	Created shared-task datasets in English, Hindi, and German.	Primarily LSTM with word embeddings	Reported strong macro-F1 scores and demonstrated the effectiveness of deep learning in multilingual offensive content detection.
[11]	Deep hate speech detection under limited resources	Combined recurrent and convolutional layers for abusive language classification.	Hybrid deep neural network	Performance on HASOC 2019 showed promise, and the study discussed strategies for improving results through better data and model design.
[12]	Hater network identification on social media	Used a pre-trained model to discover hateful user networks on Twitter.	Pre-trained offensive language model with social graph analysis	Released a large anonymized tweet dataset and a social graph for joint text-

				network analysis.
[13]	Multilingual and multi-class offensive content detection	Presented offensive language detection as a multilingual multi-level problem.	Traditional ML, deep neural models, transfer learning	Deep neural and transfer learning methods generally performed better than SVM in complex settings.
[14]	Multimodal hate speech classification	Proposed a fusion architecture for text-image hate speech detection.	Quaternion neural network with multimodal fusion	The model reduced parameter count substantially while maintaining effective performance.
[15]	Ethical control of online hate content	Discussed quarantining as a proactive strategy for handling harmful posts.	Detection framework with quarantining mechanism	Suggested a moderation strategy that balances user protection with censorship concerns.

3. Methods

The primary objective of the proposed work is to analyze how different text feature representation techniques influence classification performance on short social media text. Social media posts are inherently difficult to model because they often contain informal expressions, misspellings, abbreviations, context-sensitive meaning, and linguistic variation. At the same time, such data is widely used in applications such as opinion mining, marketing analysis, disaster management, abusive language detection, and public sentiment monitoring. Across these applications, feature representation remains a central requirement for reliable text classification.

In this work, we consider social media text analysis as a three-stage process: text pre-processing, feature extraction, and classification. Pre-processing is used to remove noise and improve the consistency of raw text. Feature extraction then transforms the cleaned text into a machine-readable representation that preserves meaningful linguistic information. Finally, a classifier is trained to predict the target class from the derived feature space. The focus of the present study is the second stage, namely the comparative evaluation of feature extraction techniques and their impact on deep learning-based hate speech detection.

Figure 1 presents the experimental framework adopted to measure the influence of different feature extraction techniques on a Sequential Convolutional Neural Network (SCNN). The study uses a publicly available Kaggle hate speech and offensive language dataset [21] containing 24,783 instances grouped into three classes: hate speech, offensive language, and neither. After basic text cleaning, the processed posts are transformed using five feature representation techniques, each of which is then supplied to the SCNN for training and validation.

1. Word Count Vectorizer: This method converts each document into a sparse vector based on token occurrence counts. When used without a manually restricted dictionary, the feature space is determined by the vocabulary observed in the corpus.

2. TF-IDF Vectorizer: TF-IDF produces a weighted representation in which terms are scored according to their importance within a document relative to the entire corpus. Compared with raw count vectors, this approach provides a more discriminative representation of informative words.

3. Word2Vec: Word embedding methods transform tokens into dense numerical vectors that capture semantic similarity. In this study, tokenization, indexing, and sequencing are used to prepare embedding-based inputs for deep learning.

4. POS Tag Features: Part-of-speech tagging converts each sentence into a distribution of grammatical categories. The resulting feature vector captures structural information by counting the occurrence of different POS tags.

5. Bi-Gram Features: A bi-gram represents an ordered sequence of two adjacent tokens. Bi-gram frequency patterns are useful for modeling short-range linguistic context and phrase-level usage in text classification.

The selected representation is then converted into a feature vector for the subsequent classification stage.

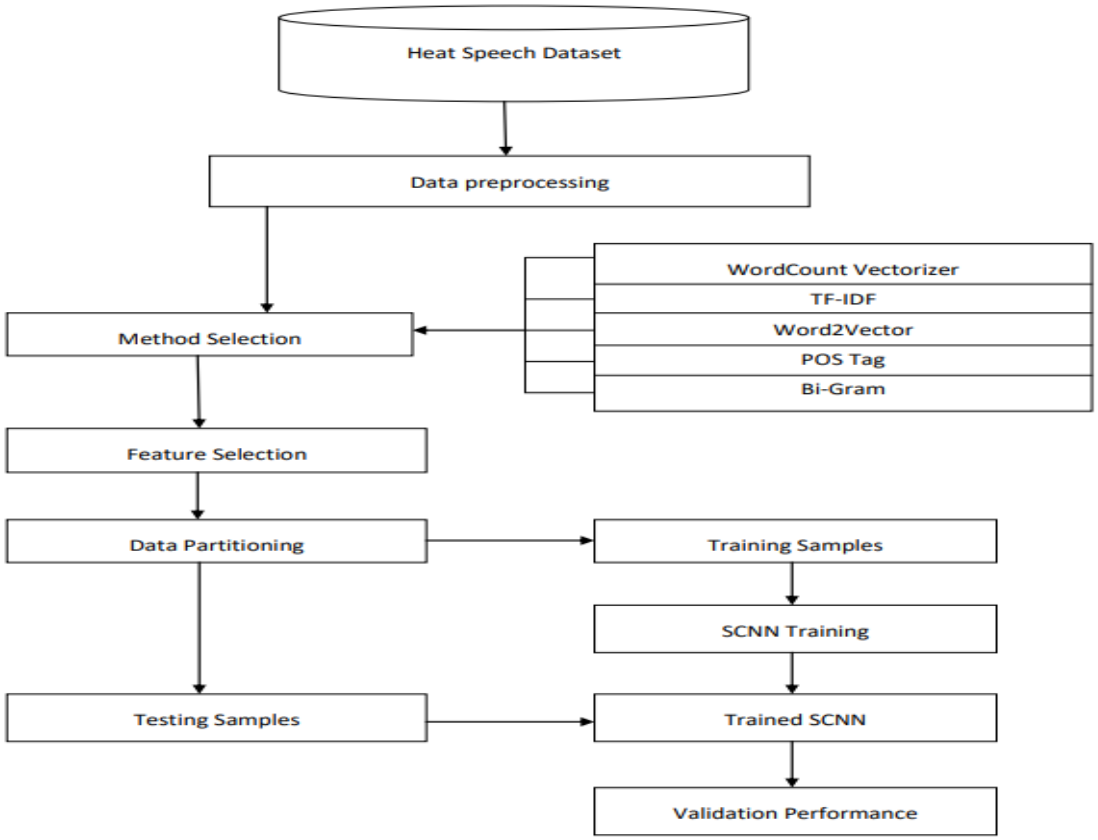

Figure 1 Proposed model for evaluating the effect of feature selection on SCNN

Here we have provided a provision to select a method for converting the clean data. After feature extraction, the complete dataset is divided into training and validation partitions for model development and performance assessment. The dataset is split in a 70:30 ratio for training and validation, respectively. A Sequential Convolutional Neural Network is then configured with an input layer matched to the feature dimension, followed by three hidden dense layers that use ReLU and sigmoid activation functions. The final output layer contains three neurons corresponding to the target classes. Each feature representation is used independently to train the SCNN, and the resulting validation performance is analyzed in the next section.

4. Results

This section presents the comparative performance of the implemented hate speech classification framework using five feature extraction techniques with the SCNN classifier. Figure 2 summarizes the model behavior in

terms of training accuracy, validation accuracy, training loss, and validation loss. Accuracy is reported as a percentage, while loss values are used to examine the convergence behavior of the learning process.

From both the training and validation perspectives, TF-IDF and Word Count Vectorizer consistently produce stronger results than the other evaluated representations. This observation suggests that explicit word-level information remains highly informative for hate speech and offensive language classification in short social media text.

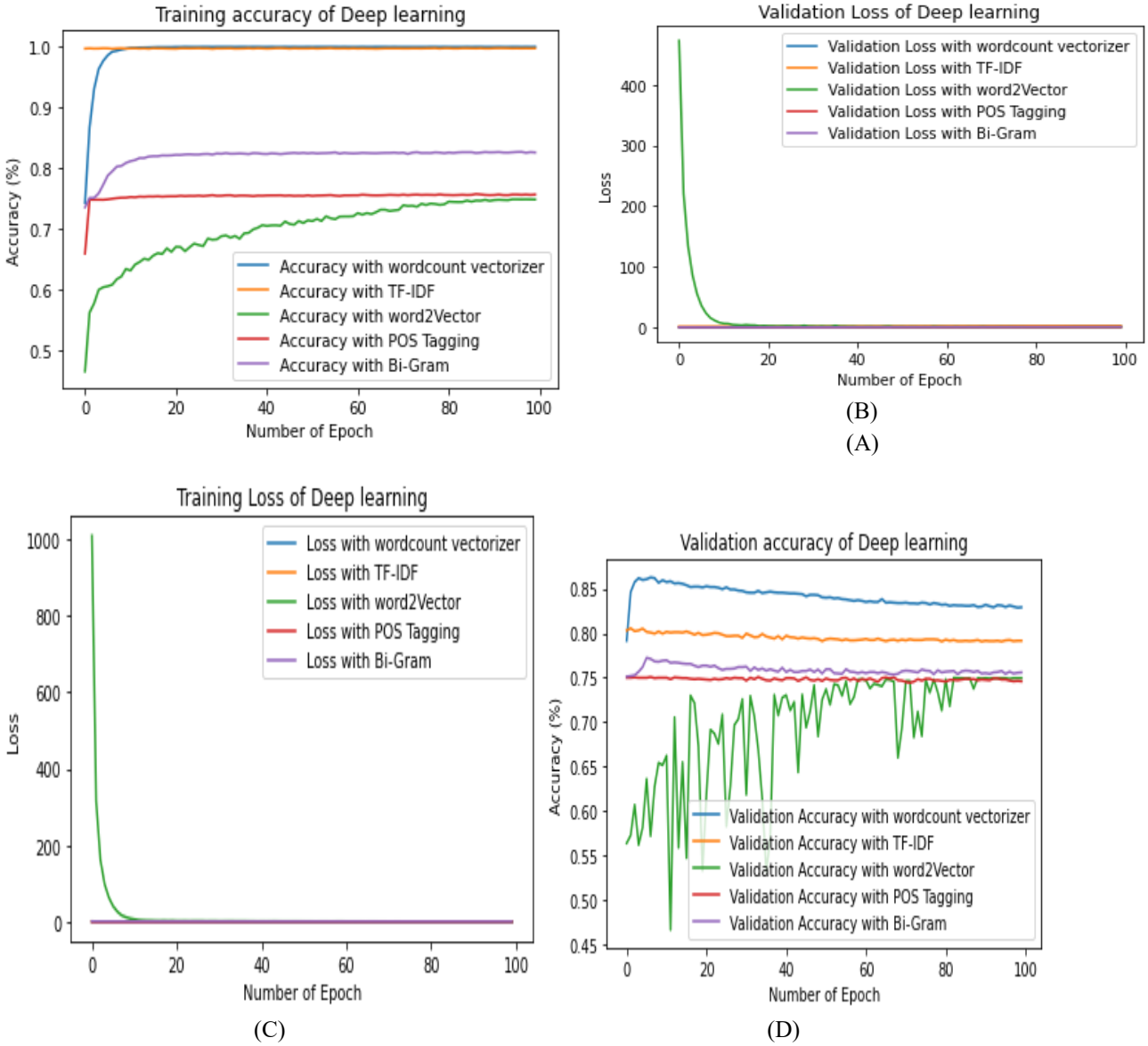

Figure 2 demonstrate the performance of the proposed model in terms of (A) Accuracy (B) validation accuracy (C) Training Loss and (D) validation loss

Computational efficiency is also an important aspect of feature representation. Figure 3 compares the training time required by the SCNN for different feature types, where the x-axis represents the selected representation method and the y-axis shows execution time in seconds.

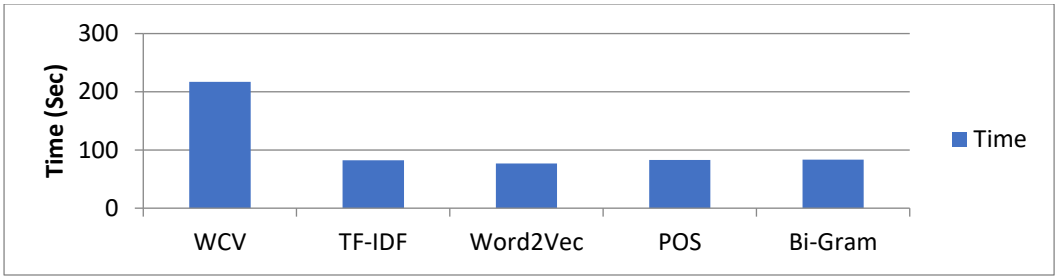

Figure 3 Training time of features

Although Word Count Vectorizer provides strong accuracy, it also requires comparatively higher training time. TF-IDF, by contrast, achieves similar or better predictive performance with lower computational cost. To further analyze scalability, the dimensionality of selected feature spaces was increased and the corresponding change in classification accuracy is illustrated in Figure 4.

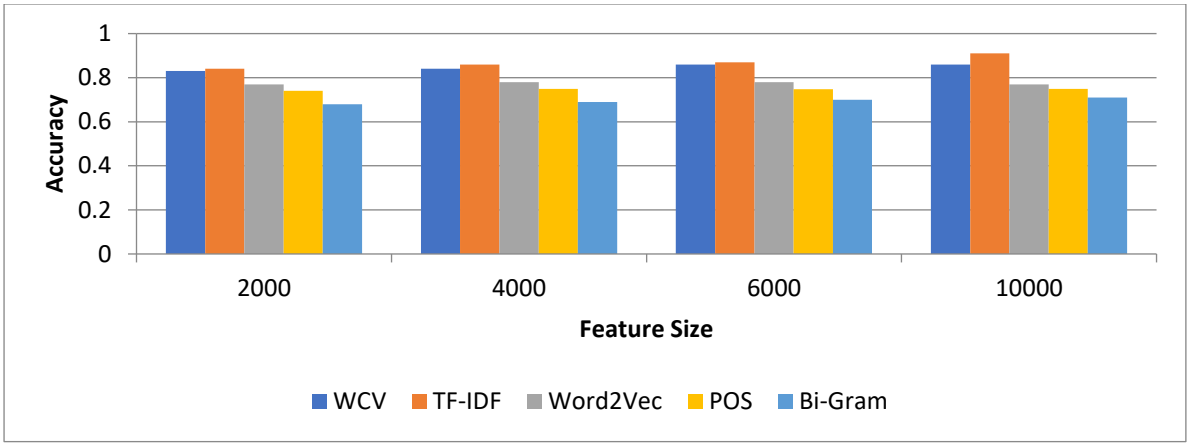

Figure 4 Accuracy with higher dimensional features

The experiments indicate that Word2Vec and POS-based features do not significantly benefit from increasing feature dimensionality under the present configuration, and therefore their performance remains relatively stable across settings. In contrast, count-based and TF-IDF-based representations can expand more effectively and show measurable accuracy improvements as the feature space grows. Among all methods, TF-IDF provides the most favorable balance between classification performance and computational efficiency. These findings suggest that accurate vocabulary modeling, together with careful preprocessing, is essential for improved social media text classification.

5. Discussion

This study investigated recent approaches to social media text analysis with a specific focus on hate speech detection and feature representation. The review and experiments together indicate that social media classification remains difficult because of informal writing style, spelling inconsistency, linguistic variation, limited context, and the noisy nature of user-generated content. The analysis also confirms that the selection of an appropriate feature representation has a substantial effect on classification quality.

To examine this issue empirically, five widely used text feature extraction techniques were evaluated using an SCNN classifier on a three-class hate speech dataset. The comparative results lead to the following key observations:

1. Feature representation has a strong impact on social media text classification performance.

2. Inclusion of informative lexical patterns can improve hate speech detection accuracy.

3. Context-aware and richer semantic features are still needed for further performance gains on short social media text.

Motivated by these findings, the following directions are identified for future work:

1. Incorporate additional contextual information from social media content to improve application-specific classification accuracy.

2. Develop more robust preprocessing pipelines to address spelling variation, informal language, and other linguistic irregularities found in online text.

3. Evaluate multiple deep learning architectures and hybrid feature representations for further improvement in hate speech detection.

Refrences

[1] J. Hartmann, J. Huppertz, C. Schamp, M. Heitmann (2019), Comparing automated text classification methods, International Journal of Research in Marketing, 20–38.

[2] T. A. Rana, K. Shahzadi, T. Rana, A. Arshad, M. Tubishat (Nov 2021), An Unsupervised Approach for Sentiment Analysis on Social Media Short Text Classification in Roman Urdu Sentiment analysis on short text classification in Roman Urdu, ACM Trans. Asian Low-Resour. Lang. Inf. Process., 21(2), Article 28.

[3] K. A. Qureshi, M. Sabih (2021), Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text, IEEE access, VOLUME 9.

[4] M. U. S. Khan, A. Abbas, A. Rehman, R. Nawaz (2020), HateClassify: A Service Framework for Hate Speech Identification on Social Media, IEEE Internet Computing Published by the IEEE Computer Society © 2020 IEEE, 1089-7801.

[5] M. Yu, Q. Huang, H. Qin, C. Scheele, C. Yang (2019), Deep learning for real-time social media text classification for situation awareness – using Hurricanes Sandy, Harvey, and Irma as case studies, International Journal of Digital Earth, 12(11), 1230–1247

[6] T. T. A. Putri, S. Sriadhi, R. D. Sari, R. Rahmadani, H. D. Hutahaean (2020), A comparison of classification algorithms for hate speech detection, IOP Conf. Series: Materials Science and Engineering (2020)

[7] T. Y. S. S. Santosh, K. V. S. Aravind (2019), Hate Speech Detection in Hindi-English Code-Mixed Social Media Text, 3–5.

[8] Y. Senarath, H. Purohit (2020), Evaluating Semantic Feature Representations to Efficiently Detect Hate Intent on Social Media, IEEE 14th International Conference on Semantic Computing (ICSC)

[9] E. Pronoza, P. Panicheva, O. Koltsova, P. Rosso (2021), Detecting ethnicity-targeted hate speech in Russian social media texts, Information Processing and Management.

[10] T. Mandl, P. Majumder, S. Modha, D. Patel, M. Dave, C. Mandlia, A. Patel (2019), Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages,12–15.

[11] G. Kovács, P. Alonso, R. Saini (2021),Challenges of Hate Speech Detection in Social Media.

[12] M. Wich, M. Breitinger, W. Strathern, M. Naimarevic, G. Groh, J. Pfeffer, Are Your Friends Also Haters? Identification of Hater Networks on Social Media

[13] S. Modha, T. Mandl, P. Majumder, D. Patel, Tracking Hate in Social Media: Evaluation, Challenges and Approaches, SN Computer Science (2020) 1:105.

[14] D. Kumar, N. Kumar, S. Mishra (2020), QUARC: Quaternion Multi-Modal Fusion Architecture For Hate Speech Classification", 2020

[15] S. Ullmann, M. Tomalin (2020), Quarantining online hate speech: technical and ethical perspectives,

[16] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, S. Villata (2020), A Multilingual Evaluation for Online Hate Speech Detection, ACM Transactions on Internet Technology, 20(2)..

[17] Z. Zhang, D. Robinson, J. Tepper (2018), Hate Speech Detection Using a Convolution-LSTM Based Deep Neural Network.

[18] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine (2020), The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes.

[19] G. Rizos, K. Hemker, B. Schuller (2019), Augment to Prevent: Short-Text Data Augmentation in Deep Learning for Hate-Speech Classification

[20] S. Chopra, R. Sawhney, P. Mathur, R. R. Shah, Hindi-English Hate Speech Detection: Author Profiling, Debiasing, and Practical Perspectives.

[21] Dataset link: https://www.kaggle.com/mrmorj/hate-speech-and-offensive-language-dataset, Last accessed [5 Feb 2022]