INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Hate Speech Detection based on Word Embedding and Linguistic Features

Archika Jain[1]*, **Sandhya Sharma**[2]

1 Department of CSE, Suresh Gyan Vihar University, 302017, Jaipur, India
2 Department of ECE, Suresh Gyan Vihar University, 302017, Jaipur, India

OPEN ACCESS

Received: 22-08-2023
Accepted: 03-10-2023
Published: 12-11-2023

Citation: Jain A, Sharma S (2023) Hate Speech Detection based on Word Embedding and Linguistic Features. Indian Journal of Science and Technology 16(41): 3704-3713. https://doi.org/10.17485/IJST/v16i41.2128

* **Corresponding author**.

archikaagarwal@gmail.com

Funding: None

Competing Interests: None

Copyright: © 2023 Jain & Sharma. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To develop an improved hate speech detection method based on word embedding and linguistic features. **Methods:** Many machine-learning classifiers like Logistic Regression (LR), Gaussian Naive Bayes (GNB), Random Forest (RF), K-Nearest Neighbor (KNN) and Linear Support Vector Classifier (SVC) are trained on linguistic data for identifying hated speech. For this research two datasets has been used with the size of 24783 tweets and 6977 tweets for Tweet hate speech detection dataset and Hasoc19 dataset respectively. We have taken the size of training and testing dataset is 67/33 for both the dataset, in which size of training dataset is 67 and size of testing dataset is 33. **Findings:** On Tweet hate speech detection dataset we target the highest accuracy 0.90 and highest precision, recall and f1-score like 0.87, 0.85 and 0.90 respectively for label 0 and 0.98, 0.98 and 0.93 respectively for label 1 and 0.86, 0.85 and 0.74 for class 2 after applying random forest classifier. On Hasoc2019 dataset we achieve the highest accuracy 0.99 and highest precision, recall and f1-score values like 1.00, 0.99 and 1.00 for class 0 and 1.0, 0.99 and 0.99 for class 1 after applying Random Forest classifier with linguistic features TF-IDF word embedding technique. **Novelty:** Twenty linguistic features with term frequency-inverse document frequency (TF-IDF) word embedding technique make this research unique. Twenty linguistic characteristics have been chosen for detecting the despised information based on three groups of attributes which is complexity attributes, stylometric attributes and psycho-linguistic attributes have been chosen.

Keywords: Machine Learning Classifiers; Linguistic Features; Accuracy; TFIDF; Random Forest Classifier

1 Introduction

Social media platforms are used for a variety of daily activities, including news consumption, keeping up with popular topics, and sharing information. The widespread expansion of digital media accelerated the spread of hate speech. Due to the vast amount of heterogeneous material and varied writing styles, it is still difficult to discern hate speech from the comments on the internet. Techniques for artificial intelligence can be applied ethically or immorally. While taking into account the moral approach AI methods might be utilized to combat disinformation, deep-hated is exploited in an

immoral manner to propagate hate speech. To comprehend the facts and pattern of hate speech for prompt decision-making, artificial technologies like machine learning (ML) and sophisticated deep learning techniques must be used. There are several websites that examine the accuracy of information, including factcheck.org, FlackCheck.org, Ground News: News Comparison Platform, PolitiFact, Snopes, and The Washington Post's Fact Checker. Because hate speech is often produced with alluring headlines and graphics, people are readily drawn to it. Not only is it difficult to identify hate speech, but there is also an issue with public trust and a lack of critical thinking. People are less interested in the speech's veracity.

In recent years, people's reliance on social media has grown [1,2]. Platforms for the media as a source of communication and human interaction. Facebook, Instagram, and Twitter are among the social media platforms where people are increasingly sharing and expressing their thoughts, feelings, and opinions. However, on occasion, these messages contain offensive content and language that is intended at a specific individual or group. Social media is frequently used to share a variety of content. People frequently utilize social media to convey their ideas and opinions [1]. Despite the speed and openness of social media. It is also free, available, and easy to access because to its quick and spectacular growth. Nature is also rather delicate. It turns into a vehicle for individuals who do ill deeds to spread various forms of prejudice or hatred. Hate speech is merely words that may be extremely painful to an individual's or group's feelings [3] and may result in violence or insensitivity, both of which are irrational and inhumane behaviors. The use of hate speech, which is prohibited, has increased as online social media usage has increased. There is a connection between hate speech and hate crimes, and hate crimes are on the rise [4].

The consequences of hate crimes are already enormous. Due to SM's widespread use and anonymity internet users take pleasure. In this big data era, manual processing is difficult and time-consuming for categories enormous amounts of text data and accuracy prediction is also tuff. To provide outcomes that are more precise and objective, machine learning classification techniques has been used [5–7]. Hate speech is a new word in the world of social media. As a result, there is no general definition of hate speech. Twitter defines hate speech (HS) as "anything that supports violence in case of directly or indirectly threatens other individuals on the basis of sexual, race, gender, ethnicity, nationality, age, orientation, handicap, religious affiliation, or serious sickness" [8–10]. People have become addicted to social media platforms in recent decades as a means of engaging to and connecting with others. Users are increasingly expressing and sharing their ideas, inner thoughts, and feelings via social networks [11] such as Twitter and Facebook. However, some messages may contain bias and harmful information directed at a specific people or group [12]. Fake news is defined as news items that are purposely and verifiably untrue and have the potential to mislead readers by providing purported, fictitious facts regarding social, health, economic, and political topics of interest [13]. Another viewpoint is to approach erroneous news directly as fake news, which includes fabrications, hoaxes, and satires [14]. The capacity of computers to understand the link between input and output without being explicitly programmed is known as machine learning. In contrast to conventional programming, which needs developing algorithms, machine learning requires finding the algorithm that learns patterns from a given dataset and constructs a predictive model, on the basis of which the computer learns the patterns between input and output [15]. For the purpose of identifying hate speech, many machine-learning classifiers [16] like Gaussian Naive Bayes, Linear Support vector classifier, Logistic Regression, Random Forest and K-nearest neighbor are trained on linguistic data.

The main contribution of this paper is:-

- We selected 20 linguistic characteristics using the Pearson correlation coefficient for the purpose of identifying hate speech
- We utilized TF-IDF statistical measures to obfuscate hate speech detection.
- Machine learning models were employed on two widely recognized datasets and surpassed the performance of state-of-the-art methods.

2 Methodologies

For hate speech detection [17] two datasets have been used. A dataset named as Twitter Hate Speech Detection Dataset is available on Kaggle repository and another dataset Hasoc2019 is available on GitHub. In this work, both datasets are used for identification of hate speech and these datasets are pre-processed using methods including stemming, null removal and stop word removal. The size of Twitter Hate Speech Detection Dataset is 24783 tweets. We have taken 3 labels like class 0, 1 and 2. Table 1 shows the dataset statistics for Tweeter hate speech detection dataset and Hasoc19 dataset. In Tweeter hate speech detection dataset the number of hated speech is 1430, offensive speech is 19190 and neither speech is 4163. The size of Hasoc2019 dataset is 6977 tweets in which the number of hated speech is 4440 as class 0 while non-hated speech is 2537 as class 1. We have taken the size of training and testing dataset is 67/33 for both the dataset, in which size of training dataset is 67 and size of testing dataset is 33. We have taken the more samples like 70/30, 80/20 and 90/10 of training and testing purpose. But we got better result on 67/33. We applied linguist features with TF-IDF technique so that we got highest accuracy for both the datasets.

Various new kinds of abusive language are emerging, including inflammatory language, negativity, sexism, and racism. The internet environment is becoming poisoned by misogyny, cyberbullying, etc. Several different research have been committed to creating automated solutions to the ability to spot this kind of stuff on social media (SM).

The proposed framework for identifying hated speech is presented in this section. The proposed methodology for predicting hated speech is shown in Figure 1. Preprocessing, feature set with TF-IDF, feature engineering, feature selection on the basis of 20 linguistic features, feature extraction methods, application of machine learning models, and detection of hated speech using machine learning classifier are some of the components that make up this system. For the purpose of identifying hate speech, many machine-learning classifiers like Logistic Regression (LR), Gaussian Naive Bayes (GNB), Random Forest (RF), K-nearest neighbor (KNN) and Linear Support vector classifier (SVC) are trained on linguistic data.

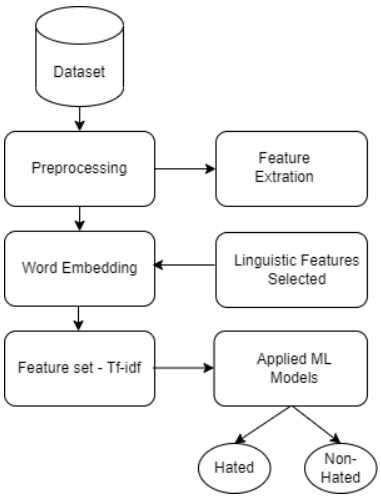

Fig 1. Proposed framework for Identification of Hate Speech

2.1 Dataset Collection

For hate speech detection[17] two datasets have been used. A dataset named as Twitter Hate Speech Detection Dataset is available on Kaggle repository and another dataset Hasoc2019 is available on GitHub. In this work, both datasets are used for identification of hate speech and these datasets are pre-processed using methods including stemming, null removal and stop word removal. The size of Twitter Hate Speech Detection Dataset is 24783 tweets. We have taken 3 labels like class 0, 1 and 2. Table 1 shows the dataset statistics for Tweeter hate speech detection dataset and Hasoc19 dataset. In Tweeter hate speech detection dataset the number of hated speech is 1430, offensive speech is 19190 and neither speech is 4163. The size of Hasoc2019 dataset is 6977 tweets in which the number of hated speech is 4440 as class 0 while non-hated speech is 2537 as class 1.

https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset
https://hasocfire.github.io/hasoc/2019/dataset.html

Table 1. Dataset Statistics

Dataset Name	Class	Meaning	No. of Speech
Tweet hate speech detection dataset	0	Hated speech	1430
	1	Offensive speech	19190
	2	Neither	4163
Hasoc19 dataset	0	Hated speech	4440
	1	Non-hated speech	2537

2.2 Feature of Linguistic

It is challenging to locate the key elements for detecting the hated speech on the web platform due to the varied nature of comments. Detection of the text's general structure is aided by linguistic factors.

2.2.1 Feature of Extraction

Hate speech detection discussed by significant linguistic features in this section.

- **Complexity:** To determine how complex a piece of news material is overall, complexity characteristics are used. It takes both word level and sentence level into account. Number of Character, Upper Case, Special Characters and Lower Case are some of the several metrics employed.
- **Psycholinguistic:** It has to do with how the brain works. These deciding elements aid in the evaluation or potency of particular emotions. Polarity and subjectivity are part of it. Speech sentiment, whether good or negative, is measured using polarity.
- **Stylo -metrics:** The semantic and grammatical components of the speech are taken into account by stylistic aspects. It aids in distinguishing false information from reliable information. It makes use of NLP approaches to understand syntax and speech style in order to extract grammatical information from material. The number of words, the number of syllables, the number of consonant, the number of articles, the noun count, the verb count, the adjective count, the determinant count, the rate of noun, the adverb count, the rate of verb, the number of negation, the rate of adverb and the number of interrogative text are some of the criteria used to detect the hate speech.

2.2.2 Feature of Selection

Linguistic features has been extracted after preprocessed the dataset. Twenty most significant linguistic features has been selected. On the basis of three categories of features twenty linguistic features has been selected for detecting the hated information. In complexity attributes 4 attributes have been selected named as number of lowercase, uppercase, special character and characters. In stylometric attributes total 14 attributes have been selected named as rate of verb, adverb count, words count, syllables count, consonant count, count of articles, number of noun, verb number, adjective number, determinant number, rate of adverb, count of negation, noun rate, number of interrogative text. Polarity and subjectivity these are 2 attributes which have been selected in Psycho-linguistic Attributes.

2.2.3 Word Embedding

In literature, various approaches have been used for feature extraction. A distinct feature extraction approach TF-IDF have been applied in this paper.

- **TF-IDF Technique:** The TF-IDF technique employed to determine a term's weight inside a document. It is typically utilized in information retrieval, text mining and feature extraction techniques.

2.2.4 Statisticians' Preference for Features

We combined the numerical values of the hate speech dataset with the established factors to provide them more training in order to get more accurate results. On unique important characteristics, we used the TF-IDF approach and it produced the best results. It offers both the significance of the words as well as their frequency in the corpus. Then, by removing the terms they are not as important for analyzing, we may make constructing models process simpler by lowering the size of the input space. Therefore, for future investigation of additional ML models, we favored the TF-IDF technique.

2.2.5 Machine Learning Models

In this part, we talk about the machine learning model that is used to identify hate speech.

- **Gaussian Naive Bayes (GNB):** The Bayesian theorem is the foundation of this procedure. Its foundation is the idea of feature independence. This technique may be used in a variety of fields, including text categorization, spam detection, and opinion mining.
- **Logistic Regression (LR):** With the use of a logistic regression equation, this method calculated the likelihood of a link between one or more independent variables and dependent variables. The loss function is convex at all times. It is necessary to choose characteristics properly.
- **Random Forest (RF):** The average outcomes after combining different decision trees (DTs) can be done by the supervised learning approach. The accuracy of the model might be improved by the large number of RF trees.

- **Linear Support Vector Classifier (SVC):** Both linear and non-linear applications are suitable for this supervised learning approach. The process determines the appropriate class for the current fact by determining the hyper plane that divides the new data point into classes.

$$y = (\overrightarrow{w}\,\overrightarrow{x}) = f(\sum j\, wj\, xj)$$

The coordinates or features in this case are called x, and the weights w, define the straight line slope.

- **K-Nearest Neighbor (KNN):** This approach works well for feature-based classification issues. The stated problem and statistics affect K's value. According to Varma et al. (2021), it employs a variety of distance metrics, including Euclidean, Manhattan, and Hamming metrics. Due to its complete reliance on every training set example, it has high memory needs and is time-consuming.

3 Results and Discussion

3.1 Algorithm for Identification of hate speech

The algorithm used to identify hate speech is covered in this section. We utilized a dataset for the suggested investigation. The dataset is preprocessed by deleting stop words and conducting stemming. Twenty linguistic characteristics from the dataset's title are retrieved after pre-processing. Based on the Pearson correlation coefficient, these twenty characteristics were chosen. The word embedding[18] for these qualities differs. These properties of each dataset are used to train a number of machine learning models.

Input: Dataset
Output: Hated or Non hated
1. Dataset pre-processing
2. Linguistic feature extraction
3. Select 20 feature based on Pearson correlation coefficient
4. Utilize the TF-IDF approach with datasets
5. Select best technique called as feature extraction
6. Apply various ML models like GNB, LR, RF, Linear SVC, K Neighbor

3.2 Experimental setup for Tweet hate speech detection dataset without Linguistic Features

In this we predict hated speech on the basis of machine learning models. Table 2 shows the ML classifiers that are used on Tweet Hate Speech Detection Dataset and find out precision, recall and f1-score for class 0, 1 and 2 and also find out accuracy.

Table 2. Result Evaluation for Tweet Hate Speech Detection Dataset

Tweet Hate Speech Detection Dataset										
ML Classifiers	**Class 0**			**Class 1**			**Class 2**			**Accuracy**
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	
GNB	0.07	0.74	0.13	0.89	0.22	0.35	0.44	0.56	0.49	0.30
LR	0.50	0.28	0.36	0.94	0.95	0.95	0.83	0.89	0.86	0.75
RF	0.53	0.25	0.34	0.91	0.96	0.92	0.83	0.75	0.79	0.86
Linear SVC	0.44	0.32	0.37	0.94	0.95	0.94	0.85	0.88	0.87	0.70
K Neighbor	0.42	0.26	0.32	0.87	0.95	0.91	0.75	0.51	0.61	0.74

3.3 Experimental setup for Tweet hate speech detection dataset with Linguistic Features

We examined the findings from our dataset in this section. For the purpose of predicting hated speech on linguistic data a number of machine-learning classifiers are trained. For enhance categorization other feature extraction methods are used like pre-processed dataset. Table 3 shows the ML classifiers that are used on Tweet Hate Speech Detection Dataset and find out precision, recall and f1-score for class 0, 1 and 2 and accuracy.

Table 3 shows the model accuracy for tweet hate speech detection dataset by using TF-IDF feature extraction technique and got 90% highest accuracy when using Random Forest classifier.

Table 3. Result Evaluation for Tweet Hate Speech Detection Dataset

Tweet Hate Speech Detection Dataset										
ML Classifiers	**Class 0**			**Class 1**			**Class 2**			**Accuracy**
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	
GNB	0.37	0.04	0.07	0.82	0.84	0.83	0.35	0.42	0.38	0.73
LR	0.00	0.00	0.00	0.78	1.00	0.87	0.33	0.00	0.00	0.78
RF	0.87	0.85	0.90	0.98	0.98	0.93	0.86	0.85	0.74	**0.90**
Linear SVC	0.05	0.03	0.03	0.77	0.92	0.84	0.22	0.06	0.09	0.73
K Neighbor	0.12	0.04	0.06	0.78	0.94	0.85	0.17	0.04	0.06	0.74

3.4 Experimental setup for Hasoc2019 dataset without Linguistic Features

On Hasoc2019 dataset we apply the different ML classifiers for class 0 and 1. Table 4 shows the ML classifiers that are used on Hosac19 Dataset and find out accuracy, precision, recall and f1-score for class 0 and 1.

Table 4. Result Evaluation for Hasoc19 Dataset

Hasoc2019							
ML Classifiers	**Class 0**			**Class 1**			**Accuracy**
	P	**R**	**F1**	**P**	**R**	**F1**	
GNB	0.78	0.34	0.47	0.42	0.83	0.56	0.52
LR	0.71	0.81	0.76	0.58	0.44	0.50	0.67
RF	0.69	0.89	0.78	0.66	0.33	0.44	0.68
Linear SVC	0.70	0.76	0.73	0.53	0.45	0.49	0.65
K Neighbor	0.69	0.79	0.74	0.53	0.39	0.45	0.64

3.5 Experimental setup for Hasoc2019 dataset with Linguistic Features

We have taken another dataset Hasoc19. In this dataset we have taken 2 labels for measure like class 0 for hated speech and class 1 for non-hated speech. Table 5 shows the ML classifiers that are used on Hosac19 Dataset and find out accuracy, precision, recall and f1-score for class 0 and 1.

Table 5. Result Evaluation for Hasoc19 Dataset

Hasoc2019 Dataset							
ML Classifiers	**Class 0**			**Class 1**			**Accuracy**
	P	**R**	**F1**	**P**	**R**	**F1**	
GNB	1.00	0.99	1.00	0.99	1.00	0.99	0.98
LR	0.91	0.89	0.90	0.82	0.84	0.83	0.87
RF	1.00	0.99	1.00	1.00	0.99	0.99	**0.99**
Linear SVC	0.65	1.00	0.79	1.00	0.08	0.15	0.67
K Neighbor	0.81	0.84	0.83	0.70	0.66	0.68	0.77

Table 5 shows the model accuracy for hasoc19 dataset by using TF-IDF feature extraction technique and got 99% highest accuracy when using Random Forest classifier.

3.6 Evaluation Metrics

The model's performance is evaluated using a variety of measures are as follows:

- **Precision:** Precision is a description of random errors, a measure of statistical variability [(3)].

Precision = TP/ (TP+FP)

Where TP is called as true prediction
And FP is called as false Prediction

- **Recall:** Recall is the ratio of correctly predicted positive observations to the all observations in actual class [3].

Recall = TP/ (TP+FN)
Where TP is called as true positive
And FN is called as false negative

- **Accuracy:** It is the simplest straightforward performance metric, consisting of a ratio of accurately predicted observations to total observations [3].

Accuracy = (TP+TN)/ (TP+TN+FP+FN)
Here TP is called as true positive
TN is called as true negative
FP is called as false positive
And FN is called as false negative

- **F1-Score:** The F1 Score is calculated as the weighted average of Precision and Recall. As a result, this score considers both false positives and false negatives. It is not as intuitive as accuracy, but F1 is frequently more helpful than accuracy, especially if the class distribution is unequal [19].

F1 Score = 2/ ([Recall]^ (-1) [+Precision]^ (-1))

- **Support:** The support is the number of samples of the true response that lie in that class.
- **Confusion Matrix:** On a test dataset, it is used to assess the classifier's performance in the form of a table. Given that the fact's results shows the classifier's values for false positives, true positives, true negatives, and false negatives.

3.7 Machine learning models' Confusion Matrix (CM) for Tweet hate speech detection dataset with Linguistic Features for Random Forest Classifier

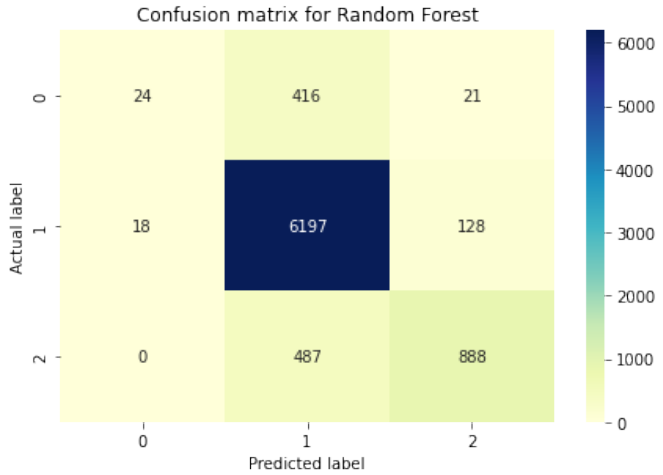

Fig 2. Confusion Matrix of RF

Figure 2 represent the confusion matrix values for Tweet hate speech detection dataset by random forest machine learning classifiers. It gives 6197 correct prediction.

3.8 Machine learning models' Confusion Matrix (CM) for Hasoc2019 dataset with Linguistic Features for Random Forest Classifier

Figure 3 represent the confusion matrix values on Hasoc2019 dataset by applying random forest machine learning classifiers. It 2294 true predictions and 0 wrong prediction.

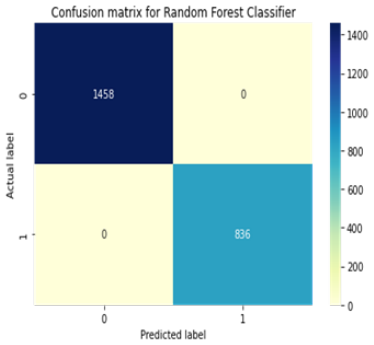

Fig 3. RF Confusion Matrix

3.9 Comparison with Existing Work

hows the comparison of our proposed framework with existing work. In the case of tweet hate speech detection dataset S. Kamble et-al[20] proposed CNN-1D model and find out precision, recall, f1-score and accuracy 0.83, 0.79, 0.81 and 0.83 respectively. P. K. Roy et-al[6] proposed DCCN model with k-fold and they got precision, recall and f1-score like 0.97, 0.88 and 0.92 respectively. We applied BERT and LSTM technique on both the dataset and found that accuracy is 60% for Tweet hate speech detection dataset and 65% for Hasoc19 dataset by using BERT technique. By using LSTM technique we got accuracy of 82% and 85% respectively for Tweet hate speech detection dataset and Haasoc19 dataset. For these dataset we have applied different machine learning models without linguistic features and with linguistic features. We have found that random forest with linguistic features machine learning classifier give best result out of without linguistic features, other technique like BERT and LSTM and the existing work.

In the case of Hasoc2019 dataset Wang et-al[21] proposed ordered neuron LSTM with k-fold ensemble method. They divide the dataset into two categories hated or non-hated and got the precision, recall and f11-score values for both the categories. G. Kovács et-al[22] proposed the RoBERT technique and found that micro f1 and weighted f1 values 0.79 and 0.84 respectively. Our proposed model random forest with linguistic feature gave the best result out of existing work and without linguistic features.

Table 6. Comparison of Our Dataset with Previous Work

Dataset	Model	Research study	Class Label	Precision	Recall	F1-Score	Accuracy
Tweet hate speech detection dataset	CNN-1D	S. Kamble et-al, 2018[20]	-	0.83	0.79	0.81	0.83
	DCCN with k-fold	P. K. Roy et-al 2020[6]		0.97	0.88	0.92	-
	RF without Linguistic Features		0	0.53	0.25	0.34	0.86
			1	0.91	0.96	0.92	
			2	0.83	0.75	0.79	
	Proposed	**RF with Linguistic Features**	**0**	**0.87**	**0.85**	**0.90**	**0.90**
			1	**0.98**	**0.98**	**0.93**	
			2	**0.86**	**0.85**	**0.74**	
Hasoc2019	Ordered Neuron LSTM with k-fold ensemble method	B. Wang et-al, (2019)[21]	0	0.66	0.71	0.69	-
			1	0.90	0.88	0.89	
	RF without Linguistic Features		0	0.69	0.89	0.78	0.68
			1	0.66	0.33	0.44	
	Proposed	**RF with Linguistic Features**	**0**	**1**	**0.99**	**1**	**0.99**
			1	**1.0**	**0.99**	**0.99**	

4 Conclusion

This research used a comprehensive technique to integrate linguistic characteristics with text in order to identify hate speech. The twenty most important linguistic features are picked in order to identify hate speech. Linguistic characteristics and TF-IDF word embedding techniques are coupled to achieve the best level of accuracy. In complexity attributes 4 attributes have been selected named as number of lowercase, uppercase, special character and characters. In stylometric attributes total 14 attributes have been selected named as rate of verb, adverb count, words count, syllables count, consonant count, count of articles, number of noun, verb number, adjective number, determinant number, rate of adverb, count of negation, noun rate, number of interrogative text. Polarity and subjectivity these are 2 attributes which have been selected in Psycho-linguistic Attributes. We achieve the highest accuracy 0.90 on Tweet hate speech detection dataset and 0.99 on Hasoc2019 dataset by applying Random Forest classifier. Also on Tweet hate speech detection dataset we achieve the highest precision, f1-score and recall values like 0.87, 0.85 and 0.90 respectively for class 0 and 0.98, 0.98 and 0.93 for class 1 and 0.86, 0.85 and 0.74 for class 2 respectively. On Hasoc2019 dataset we achieve the highest precision, recall and f1-score values like 1.00, 0.99 and 1.00 for class 0 and 1.0, 0.99 and 0.99 for class 1 after applying Random Forest classifier with linguistic features TF-IDF word embedding technique. For more accurate categorization in the future, we want to extract and research more language characteristics like user trustworthiness. In the future, we intend to look at different architectural layouts for hate speech detection in real-time and related methods. These data sets will include a wider variety of people and more samples overall.

References

1) Kumar A, Tyagi V, Das S. Deep Learning for Hate Speech Detection in social media. In: IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON). IEEE. 2021;p. 1–4. Available from: https://doi.org/10.1109/GUCON50781.2021.9573687.
2) Ahmed U, Lin JCW. Deep Explainable Hate Speech Active Learning on Social-Media Data. *IEEE Transactions on Computational Social Systems*. 2022;p. 1–11. Available from: https://doi.org/10.1109/TCSS.2022.3165136.
3) Luo J, Bouazizi M, Ohtsuki T. Data Augmentation for Sentiment Analysis Using Sentence Compression-Based SeqGAN With Data Screening. *IEEE Access*. 2021;9:99922–99931. Available from: https://doi.org/10.1109/ACCESS.2021.3094023.
4) Qureshi KA, Sabih M. Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text. *IEEE Access*. 2021;9:109465–109477. Available from: https://doi.org/10.1109/ACCESS.2021.3101977.
5) Mullah NS, Zainon WMNW. Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review. *IEEE Access*. 2021;9:88364–88376. Available from: https://doi.org/10.1109/ACCESS.2021.3089515.
6) Roy PK, Tripathy AK, Das TK, Gao XZ. A Framework for Hate Speech Detection Using Deep Convolutional Neural Network. *IEEE Access*. 2020;8:204951–204962. Available from: https://doi.org/10.1109/ACCESS.2020.3037073.
7) Zhou Y, Yang Y, Liu H, Liu X, Savage N. Deep Learning Based Fusion Approach for Hate Speech Detection. *IEEE Access*. 2020;8:128923–128929. Available from: https://doi.org/10.1109/ACCESS.2020.3009244.
8) Baydogan C, Alatas B. Metaheuristic Ant Lion and Moth Flame Optimization-Based Novel Approach for Automatic Detection of Hate Speech in Online Social Networks. 2021. Available from: https://doi.org/10.1109/ACCESS.2021.3102277.
9) Khan S, Kamal A, Fazil M, Alshara MA, Sejwal VK, Alotaibi RM, et al. HCovBi-Caps: Hate Speech Detection Using Convolutional and Bi-Directional Gated Recurrent Unit With Capsule Network. *IEEE Access*. 2022;10:7881–7894. Available from: https://doi.org/10.1109/ACCESS.2022.3143799.
10) Rodriguez A, Chen YL, Argueta C. FADOHS: Framework for Detection and Integration of Unstructured Data of Hate Speech on Facebook Using Sentiment and Emotion Analysis. *IEEE Access*. 2022;10:22400–22419. Available from: https://doi.org/10.1109/ACCESS.2022.3151098.
11) Rodriguez-Sanchez F, Carrillo-De-Albornoz J, Plaza L. Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data. *IEEE Access*. 2020;8:219563–219576. Available from: https://doi.org/10.1109/ACCESS.2020.3042604.
12) Plaza-Del-Arco FM, Molina-Gonzalez MD, Urena-Lopez LA, Martin-Valdivia MT. A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis. *IEEE Access*. 2021;9:112478–112489. Available from: https://doi.org/10.1109/ACCESS.2021.3103697.
13) Ilie VI, Truica CO, Apostol ES, Paschke A. Context-Aware Misinformation Detection: A Benchmark of Deep Learning Architectures Using Word Embeddings. *IEEE Access*. 2021;9:162122–162146. Available from: https://doi.org/10.1109/ACCESS.2021.3132502.
14) Lee E, Rustam F, Washington PB, Barakaz FE, Aljedaani W, Ashraf I. Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model. *IEEE Access*. 2022;10:9717–9728. Available from: https://doi.org/10.1109/ACCESS.2022.3144266.
15) Mehta H, Passi K. Social Media Hate Speech Detection Using Explainable Artificial Intelligence (XAI). *Algorithms*. 2022;15(8):291. Available from: https://doi.org/10.3390/a15080291.
16) Oriola O, Kotze E. Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets. *IEEE Access*. 2020;8:21496–21509. Available from: https://doi.org/10.1109/ACCESS.2020.2968173.
17) Naidu TA, Kumar S. Impact of Deep Learning Models On Hate Speech Detection. In: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE. 2021;p. 1–5. Available from: https://doi.org/10.1109/ICCCNT51525.2021.9579608.
18) Alatawi HS, Alhothali AM, Moria KM. Detecting White Supremacist Hate Speech Using Domain Specific Word Embedding With Deep Learning and BERT. *IEEE Access*. 2021;9:106363–106374. Available from: https://doi.org/10.1109/ACCESS.2021.3100435.
19) Soto CP, Nunes GMS, Gomes JGRC, Nedjah N. Application-specific word embeddings for hate and offensive language detection. *Multimedia Tools and Applications*. 2022;81(19):27111–27136. Available from: https://doi.org/10.1007/s11042-021-11880-2.
20) Kamble S, Joshi A. Hate speech detection from code-mixed hindi-english tweets using deep learning models. 2018. Available from: https://doi.org/10.48550/arXiv.1811.05145.
21) Wang B, Ding Y, Liu S, Zhou X. YNU_Wb at HASOC 2019: Ordered Neurons LSTM with Attention for Identifying Hate Speech and Offensive Language. *FIRE*. 2019;p. 191–198. Available from: https://ceur-ws.org/Vol-2517/T3-2.pdf.

22) Kovács G, Alonso P, Saini R. Challenges of Hate Speech Detection in Social Media. *SN Computer Science*. 2021;2(2):95. Available from: https://doi.org/10.1007/s42979-021-00457-3.