

Profiling Hate Speech Spreaders on Twitter

Rakshita Jain¹, Devanshi Goel¹, Prashant Sahu¹, Abhinav Kumar² and Jyoti Prakash Singh¹

¹Department of Computer Science & Engineering, National Institute Of Technology Patna, India

²Department of Computer Science & Engineering, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar

Abstract

As today is the era of social media with nearly around 192 million daily active users on Twitter alone. With the increase in the number of people online, individuals inclined towards racism, misogyny, etc have led to the spread of hate speech online. It's high time that proper steps must be taken to curb this issue with one major step being to identify people who are spreading hate speech on Twitter. We have tried to perform the above task using natural language processing techniques for two different languages English and Spanish on the two datasets provided by PAN @CLEF 2021. Four machine learning classifiers (i) multinomial naive Bayes, (ii) K-Nearest Neighbors (KNN) classifier, (iii) logistic regression and (iv) linear SVM, along with three deep learning models (i) Long Short Term Memory (LSTM), Bidirectional Long Short term Memory (bi-LSTM) and Bidirectional Encoder Representations for Transformers (BERT) model were implemented for the identification of hate speech spreader. The experiments with all the mentioned models on the training dataset provided by PAN (by splitting it into training and testing datasets) revealed that the multinomial naive Bayes is the best model with an accuracy of 74% for the English dataset and 82% for the Spanish dataset. The multinomial naive Bayes model yielded an accuracy of 66% for the English dataset and 80% for the Spanish dataset with the unknown private dataset used by the organizers for the final evaluation of the models.

Keywords

Hate Speech, Online social media, Natural Language Processing, Machine Learning, deep-learning,

1. Introduction

Due to the excessive use of social media platforms by people belonging to different cultures and backgrounds, toxic online content has become a major issue in today's time. The emergence of social media platforms has given rise to an unparalleled level of hate speech in public conversations. The number of tweets containing hate speech and targeting one or another user is on the increase every day. Unfortunately, any user engaged on these platforms will have a risk of being targeted or harassed via abusing language, expressing hate towards race, colour, religion, descent, gender, nation, etc.

Hate Speech is no less than a felony that has been continuously and abruptly growing in recent years, and the rapidly growing availability of online platforms. The rise of social

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ rakshitaj.ug18.cs@nitp.ac.in (R. Jain); devanshig.ug18.cs@nitp.ac.in (D. Goel); prashants.ug18.cs@nitp.ac.in (P. Sahu); abhinavkumar@soa.ac.in (A. Kumar); jps@nitp.ac.in (J. P. Singh)


🌐 <https://www.linkedin.com/in/rakshita-jain-425106188/> (R. Jain);

<https://www.linkedin.com/in/devanshi-goel-81b252195/> (D. Goel);

<https://www.linkedin.com/in/prashant-sahu-7065831b1/> (P. Sahu)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

media has led users to publish and share any content, tell their views, show their liking or hatred towards people, community, race, non-living objects, etc. in an ever-growing fast way. The increased willingness of people to demonstrate their opinions publicly has contributed to the multiplication of hate speech also. The ease of getting access to these platforms and publishing content with minimal efforts have led to an increase in hate speech about every small thing that people criticize or do not like, influencing other people's minds and causing several negative consequences in society. On the internet and social network platforms, people are more likely to take on inappropriate or violent behaviour due to the anonymity provided by these environments. Since this type of prejudice can cause extreme harm to society, government, and social network platforms such as Twitter and Facebook can be benefited from hate speech detection and prevention tools.

Understanding whether a tweet is hate speech or not and hence finding out whether the user is a hate speech spreader or not, is a very tough task for the users, especially those who are not experts. Additionally, hate speech can also be present in the form of sarcasm [1] or indirect taunt, making it confusing for users to understand the intent behind the tweet.

Our work is based on an assumption that a user can be classified as a hate speech spreader if while analyzing a certain number of tweets of that user, we find that the majority of the tweets can be classified as hate speech content. For that, we have grouped all the tweets of the same id. Our ultimate target is profiling those users who spread hate speech based on the number of tweets containing any hateful content that they spread, for two languages - English and Spanish. This allows the social media platforms to identify hate speech spreaders on Twitter as an initial step towards preventing hate speech from spreading among social media users and preventing it from influencing the lives and work of target people. We focus on classifying users as hate speech spreaders or not hate speech spreaders (binary classification). Examples of each of these categories - taken from the user's tweet dataset (PAN21-Profiling-Hate-Speech-Spreaders-in-Twitter)[2] can be seen in Table 1.

Table 1

The table below lists some examples of hate and non-hate speech.

Tweet	Class
POC love talking about police brutality but no one talks about black on white crime	hateful
Hey Jamal (snickering uncontrollable) You want some (PFFF) LEMONADE!" What an IDIOT!	hateful
Romanian graftbuster's firing violated rights, European court says #URL#Russian ventilators sent to U.S.	non-hateful
#RT #USER#: "At least while Biden is bombing brown people, he's not being offensive on Twitter	non-hateful

The present work is investigating whether a user is a hate speech spreader or not using various conventional machine learning classifiers and deep learning models. In the case of conventional machine learning models, we used tf-idf and count vector features by varying the word n-gram range. In the case of deep learning models, word embedding, one-hot encoding vector, and BERT embedding vectors are used as input to the models. The performance of each of the models was finally compared to get the best performing model for the hate speech spreaders.

The rest of the sections are organized as follows: section 2 lists some of the state-of-the-art

works for hate speech detection. Section 3 discusses the proposed methodology in detail, section 4 list the finding of the proposed model and finally section 6 concludes the paper.

2. Related Work

Several works [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 3] have been proposed for hate speech detection in the last few years. Here, we are listing some of the state-of-the-art models for the identification of hate content from social media. SemEval-2019 task [15] was about the Detection of Hate speech against Immigrants and Women. In this task, the participant has been given tweets of two languages English and Spanish. The task included two subtasks. The first one was about identifying the hate speech and the later one was about identifying further features such as aggressiveness and the target group or individual. For the first task, the best result was with a macro F1-score of 0.65 for the English dataset, whereas it is 0.73 for the Spanish dataset. This was obtained by the SVM with RBF kernel using embeddings from Google's Universal Sentence Encoder as features. For the second task, the best result was with a macro F1-Score of 0.70 whereas for the English dataset it is 0.57. The main motive in [4] was to identify the user who is spreading the fake news, not to identify the message that it is fake news or not. From the evaluation of the approaches of the participants, it has been found that SVM with the combination of character n-grams and word n-grams is the best-suited approach for Spanish and logistic regression ensemble of five submodels: n-grams with Random Forest, n-grams with SVM, n-grams with Logistic Regression, n-grams with XGBoost and XGBoost with features based on textual descriptive statistics, is the best-suited approach for English. The best accuracy obtained for English was 75% and for Spanish was 82%. The authors [5] have investigated multiple approaches for the problem of hate speech, aggressive behaviour, and target group recognition. They have presented many models including Logistic regression, Convolutional Neural Network (CNN), Bidirectional Transformers (BERT) using word n-grams, character n-grams, word embedding, and psycholinguistic features (LIWC). Among these models, purely Data-Driven BERT model and to some extent hybrid psycho linguistically informed CNN outperformed all other models for all tasks in both languages English and Spanish. For English, the best F1-score of 0.60 for hate speech has been found by CNN using features word embedding and LIWC. For Spanish, the best F1-score of 0.72 for hate speech has been found by BERT using features cased word.

At EVALITA-2018 [6], several models were reported to detect hate speech in Italian Social Media. Linear SVM with word embedding as features had performed best for the given problem. For Twitter and Facebook, the best macro F1-score was 0.79 and 0.77. The paper [7] introduced a combined Convolution Neural Network (CNN) and Gated Recurrent Networks (GRU) to outperform many previously proposed methods. The problem addressed in [8] is about identifying hate speech and aggressive tweets on three publicly available datasets. The author used the TF-IDF vectors with different n-gram range as features. The author used the three model-Logistic Regression, Naive Bayes Classifier, and SVM. Among these models, Logistic Regression fed with TF-IDF vector with n-gram range (1,3) has given the best accuracy of 0.956%. The authors [9] have implemented deep learning models in sixteen different

datasets of nine different languages. They found that for small dataset Logistic Regression fed with LASER (Language-Agnostic SEntence Representation) embedding has performed best and for the larger dataset BERT based model has given better results. In this paper, the author has used features of LASER embedding and MUSE embedding and achieved an accuracy of 0.83%.

Saha al. et. [10] used (i) TF-IDF vectors, (ii) sentence embedding, and (iii) Bag Of Words Embedding with various machine learning models to report that Logistic Regression performed best. Two subtasks were performed in their work (i) to classify whether a text is hate speech or not and, (ii) to classify the texts in categories as a stereotype, sexual harassment, dominance, derailing, and discredit. The Logistic Regression model performed best for this task with an accuracy of 0.704. The authors [11] claim that the performance of various machine learning algorithms to detect hate speech is hampered by inefficient sequence transduction and the vanilla Recurrent Neural Networks (RNN). RNNs with attention also suffer from various problems such as lack of parallelization and long term dependency. Therefore, the authors proposed a transform-based model and used a public dataset containing 24,783 labelled tweets. The proposed DistillBERT transformer method was compared against other transformer baselines and recurrent neural networks for Hate Speech Detection in Twitter and results showed that DistillBERT transformers outperformed other models with an accuracy of 75%. The problem addressed in [16] is about recognizing hateful content in social media. Recurrent Neural Networks were ensembled and various user-related features were incorporated showing the users' tendency towards hate speech such as racism or sexism. Word frequency vectors along with these features and data were fed as input to the classifiers. The dataset used by them was a corpus of 16,000 tweets that is available publicly. The results were compared to existing state-of-art solutions. The model can successfully differentiate racism and sexism messages from the ones which do not fall in these categories. Finally, the highest F1-score of 0.9320 was achieved using the ensemble approach.

3. Methodology

This section discusses the proposed methodology in detail. The different classification and deep learning models used for profiling hate speech spreaders that learn the continuous representation of tweets and then pick features from them extracted using count vectorizer and tf-idf vectorizer. The performance of different models was compared for different n-gram ranges. In deep learning models like LSTM, we have used one-hot encoding for feature extraction. The detailed architecture and flow of different phases in which the computation is carried out are shown in Figure 1.

3.1. Data pre-processing

We have used the PAN21-Profiling-Hate-Speech-Spreaders-on-Twitter data provided by [17] maintained at [18] to validate our proposed model. The data contains user ids and their tweets. The data contains tweets of 200 users each for English and Spanish language. 200 tweets are provided for each user containing a combination of hate speech tweets and non-hate speech tweets. Therefore, a total of 40,000 tweets are used for the experimentation. Label information

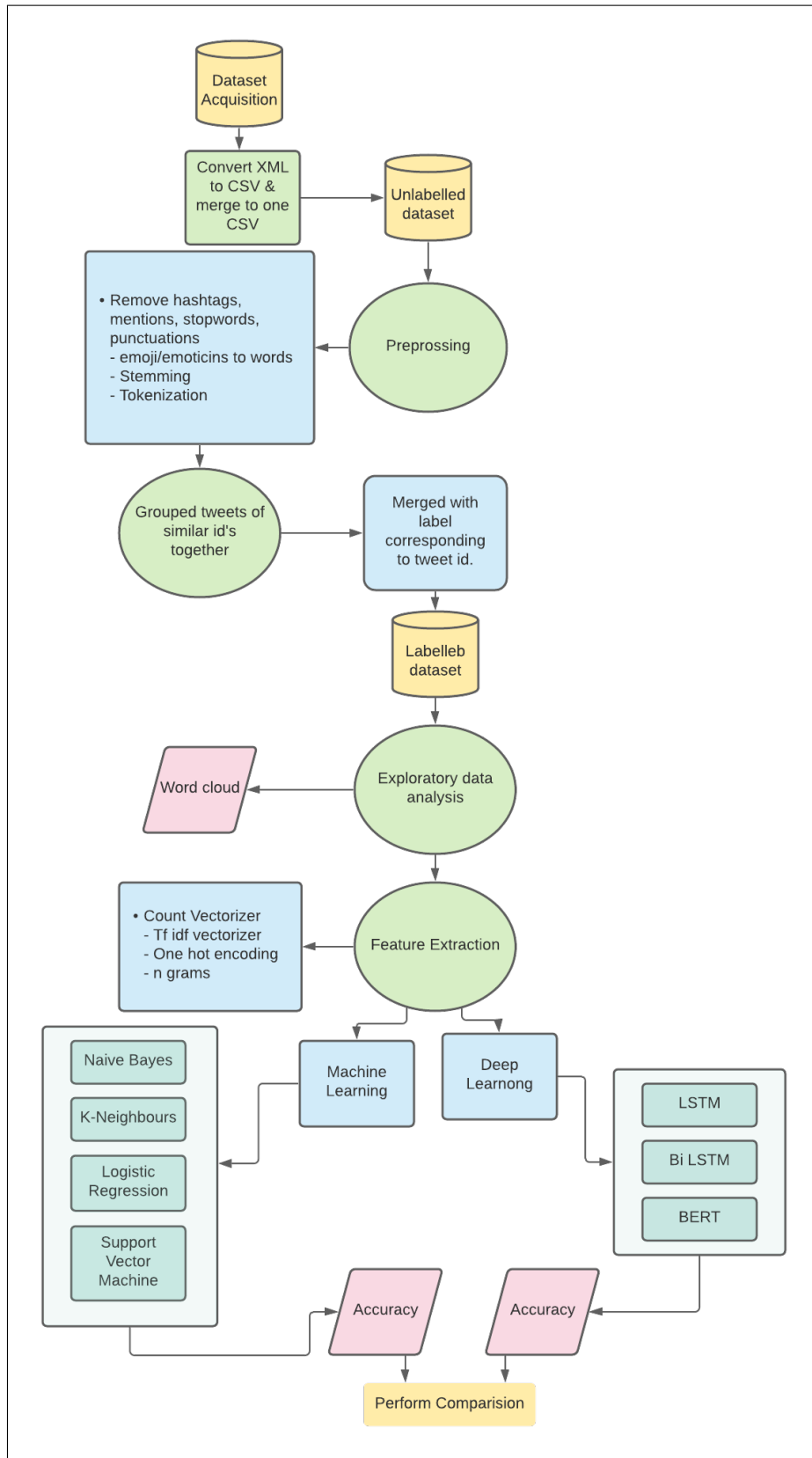


Figure 1: Overall flow diagram of the proposed model for hate speech spreaders



Figure 3: Word Cloud for the Spanish dataset

the LSTM network can learn which information is important to keep and which to throw away. By doing that it learns to use relevant information for doing predictions. In LSTM gates consist of a sigmoid activation function so for data to forget they multiply it by 0 and for data to keep they multiply it by 1. In LSTM the presence of forget gate, along with the additive property of the cell state gradients enables the network to update the parameter effectively so they emphasize only retaining the important information and discarding the rest. The architecture of the model used: Embedding Layer: Here we pass the vocabulary size as our first parameter, input feature size as the second parameter and the sentence length as the third parameter which in our case is 2500. This layer will give an output which we will pass through an LSTM layer, LSTM Layer: We have used 1 LSTM layer having 100 neurons, and Dense Layer: Since it is a classification problem, we will get an output from this dense layer. The hyper-parameters of the implemented LSTM model can be seen in Table 5.

Bi-LSTM: The bidirectional long-short-term-memory (Bi-LSTM) network is an extension of traditional LSTM that can improve model performance on sequence classification problems. The architecture of the model used: (i) Embedding Layer: Here we pass the vocabulary size as our first parameter, input feature size as the second parameter, and the sentence length as the third parameter which in our case is 2500. This layer will give an output which we will pass through a Bi-LSTM layer, (ii) Bidirectional-LSTM Layer: We have used one LSTM layer having 100 neurons, (iii) Dense Layer: Since it is a classification problem, we will get an output from this dense layer. The hyper-parameters of the implemented LSTM model can be seen in Table 6.

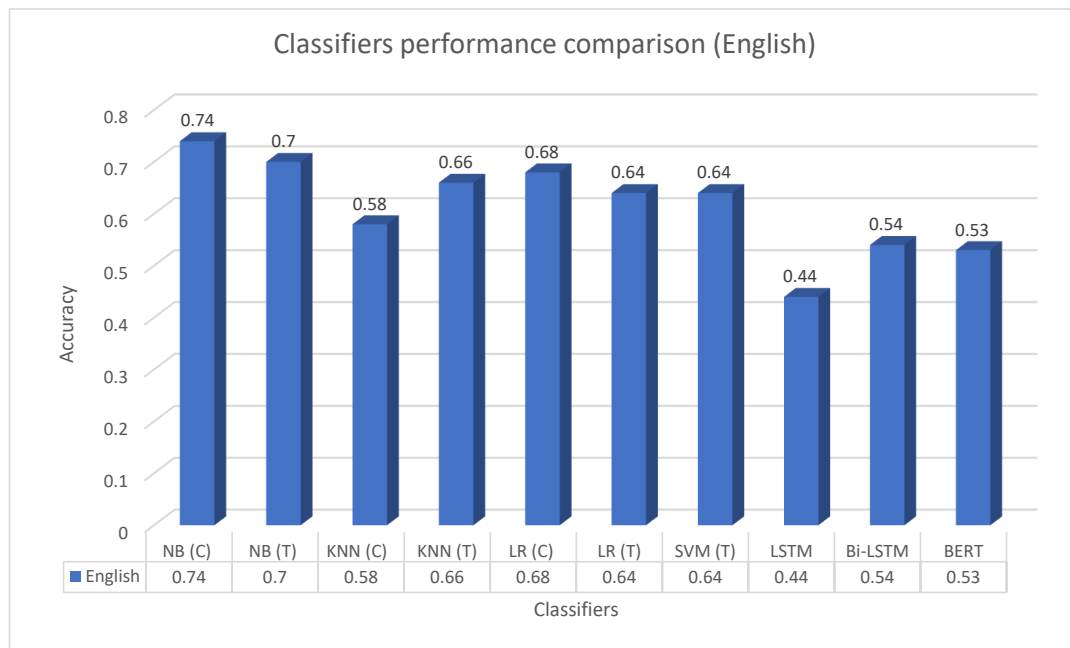


Figure 4: Classifiers comparison for the hate spreaders prediction for English Dataset

BERT: BERT is different from the directional model which reads the input sequentially (left to right or right to left). The paper [20] showed that a bidirectionally trained model performs better than a single direction trained model. BERT can be easily fine-tuned for the classification problem, question answer problem, and named entity problem. We followed the following steps while training the BERT model: (i) First of all, we imported the BERT Tokenizer and Sequence Classifier, (ii) Convert each row of the data into an InputExample Object, (iii) We did tokenization of the InputExample objects and created the required input format from the tokens so that we can feed the data to the model.

4. Result

4.1. Evaluation Metrics

For evaluating the proposed models we used precision, recall, F1 Score, support, and accuracy. We have used classification reports and confusion matrix, these metrics are widely used for evaluating supervised machine learning models for classification when the dataset is multi-labelled.

Precision It is the ratio of accurately predicted users as hate speech spreaders to the total number of predicted users. It is computed as given in the equation below. The range of precision

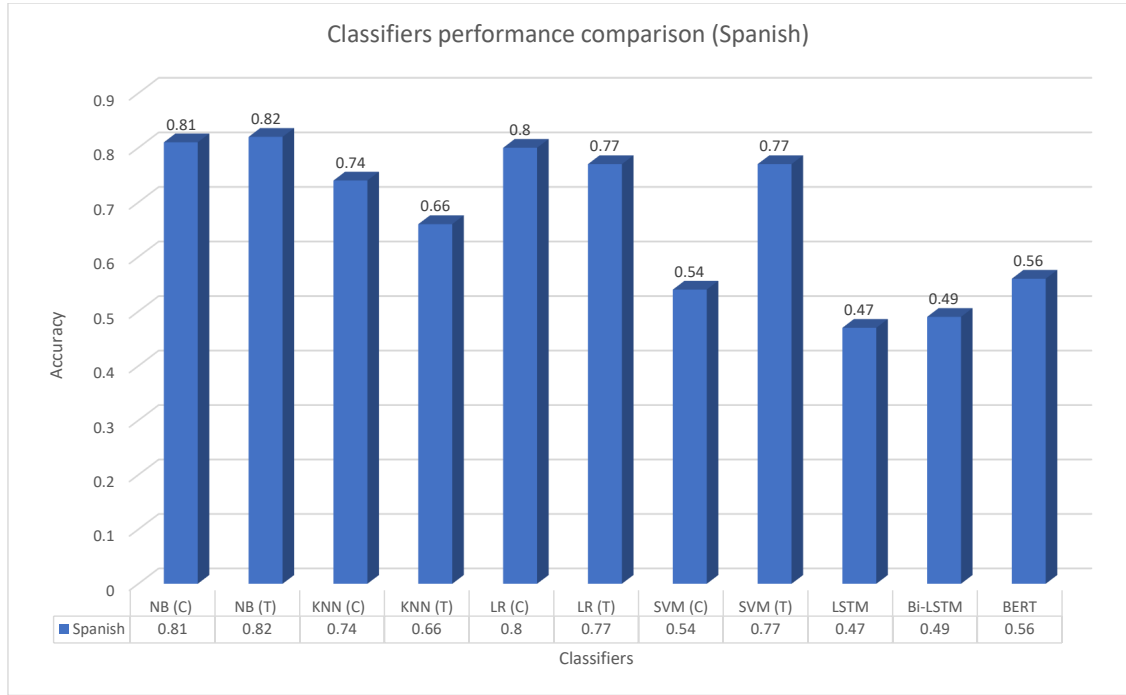


Figure 5: Classifiers comparison for the hate spreaders prediction for Spanish Dataset

varies between 0 and 1, where 1 is the best value and 0 is the worst value.

$$\text{Precision} = \frac{\text{Number of accurately predicted users}}{\text{Total number of predicted users}} \quad (1)$$

Recall It is the ratio of accurately predicted users as hate speech spreaders to the total number of real hate speech spreading users. It is computed as is given in the below equation. The range of recall varies between 0 and 1, where 1 is the best and 0 is the worst value.

$$\text{Recall} = \frac{\text{Number of accurately predicted users}}{\text{Total number of users}} \quad (2)$$

F1-Score The harmonic of Precision and Recall is called F1-Score. It can be represented by the following equation. The range of F1-score varies between 0 and 1, where 1 is the best and 0 is the worst value.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

The performance of the proposed model is measured in terms of Precision (P), Recall (R), F_1 -score (F_1), and Accuracy (Acc.). The performance of different classifiers such as Naive Bayes (NB), K-Nearest Neighbour (KNN), Logistic Regression (LR), Support Vector Machine (SVM) with different count and TF-IDF n-gram features for English and Spanish datasets are listed in

Table 2

Results of different machine learning classifiers with count and TF-IDF features (English Dataset)

Models	N-Gram	Class	Count Vector				TF-IDF Vector			
			P	R	F_1	Acc.	P	R	F_1	Acc.
Naive Bayes	1-Gram	Non-hate spreader	0.79	0.72	0.75	0.74	0.72	0.72	0.72	0.70
		Hate spreader	0.70	0.77	0.73		0.67	0.67	0.67	
		Weighted Avg.	0.75	0.74	0.74		0.70	0.70	0.70	
	(1-3)-Gram	Non-hate spreader	0.76	0.53	0.62	0.65	0.72	0.58	0.65	0.65
		Hate spreader	0.59	0.80	0.68		0.59	0.73	0.66	
		Weighted Avg.	0.68	0.65	0.65		0.68	0.65	0.65	
KNN	1-Gram	Non-hate spreader	0.55	0.72	0.63	0.53	0.73	0.61	0.67	0.66
		Hate spreader	0.47	0.30	0.37		0.61	0.73	0.67	
		Weighted Avg.	0.52	0.53	0.51		0.68	0.67	0.67	
	(1-3)-Gram	Non-hate spreader	0.62	0.58	0.60	0.58	0.67	0.67	0.67	0.64
		Hate spreader	0.53	0.57	0.55		0.60	0.60	0.60	
		Weighted Avg.	0.58	0.58	0.58		0.64	0.64	0.64	
Logistic Regression	1-Gram	Non-hate spreader	0.67	0.61	0.64	0.62	0.75	0.50	0.6	0.64
		Hate spreader	0.58	0.63	0.6		0.57	0.80	0.67	
		Weighted Avg.	0.63	0.62	0.62		0.67	0.64	0.63	
	(1-3)-Gram	Non-hate spreader	0.76	0.61	0.68	0.68	0.88	0.39	0.54	0.64
		Hate spreader	0.62	0.77	0.69		0.56	0.93	0.70	
		Weighted Avg.	0.70	0.68	0.68		0.73	0.64	0.61	
SVM	1-Gram	Non-hate spreader	-	-	-	-	0.75	0.50	0.60	0.64
		Hate spreader	-	-	-		0.57	0.80	0.67	
		Weighted Avg.	-	-	-		0.67	0.64	0.63	
	(1-3)-Gram	Non-hate spreader	-	-	-	-	0.88	0.39	0.54	0.64
		Hate spreader	-	-	-		0.56	0.93	0.70	
		Weighted Avg.	-	-	-		0.73	0.64	0.61	

Tables 2 and 3, respectively. The results of the deep learning models such as LSTM, Bi-LSTM, and BERT for English and Spanish datasets are listed in Table 4. For the English dataset, out of all the different classification methods, naive Bayes performed best with an accuracy of 74% for n-gram range (1,1) and count vectorizer while 65% for n-gram range (1,3). For other criteria like for TF-IDF vectorizer with n-gram (1,1), it gave 70% accuracy and with n-gram (1,3) it gave 65% accuracy. Similarly, for the Spanish dataset, out of all the different classification methods, naive Bayes performed best with an accuracy of 79% for n-gram range (1,1) and count vectorizer while 81% for n-gram range (1,3). For other criteria like for TF-IDF vectorizer with n-gram (1,1), it gave 80% accuracy and with n-gram (1,3) it gave 82% accuracy.

The performance of classifiers for English and Spanish Datasets are plotted in Figures 4 and 5, respectively. In the figure, C represents classifier with count vector feature and T represents classifier with TF-IDF feature. From Figures 4 and 5, it can be seen that the Naive Bayes classifier with count vector features performed best for English Dataset and achieved an accuracy of 0.74 and Naive Bayes classifier with TF-IDF feature performed best for Spanish Dataset and achieved an accuracy of 0.82.

Table 3

Results of different machine learning classifiers with count and TF-IDF features (Spanish Dataset)

Models	N-Gram	Class	Count Vector				TF-IDF Vector			
			P	R	F_1	Acc.	P	R	F_1	Acc.
Naive Bayes	1-Gram	Non-hate spreader	0.78	0.78	0.78	0.79	0.79	0.81	0.80	0.80
		Hate spreader	0.79	0.79	0.79		0.82	0.79	0.81	
		Weighted Avg.	0.77	0.79	0.79		0.80	0.80	0.80	
	(1-3)-Gram	Non-hate spreader	0.86	0.75	0.80	0.81	0.86	0.75	0.80	0.82
		Hate spreader	0.79	0.88	0.83		0.79	0.88	0.83	
		Weighted Avg.	0.82	0.82	0.82		0.82	0.82	0.82	
KNN	1-Gram	Non-hate spreader	0.74	0.81	0.78	0.70	0.75	0.56	0.64	0.66
		Hate spreader	0.81	0.74	0.77		0.67	0.82	0.74	
		Weighted Avg.	0.78	0.77	0.70		0.71	0.70	0.69	
	(1-3)-Gram	Non-hate spreader	0.71	0.78	0.75	0.74	0.80	0.25	0.38	0.60
		Hate spreader	0.77	0.71	0.74		0.57	0.94	0.71	
		Weighted Avg.	0.75	0.74	0.74		0.68	0.61	0.55	
Logistic Regression	1-Gram	Non-hate spreader	0.77	0.84	0.81	0.80	0.79	0.72	0.75	0.77
		Hate spreader	0.84	0.76	0.80		0.76	0.82	0.79	
		Weighted Avg.	0.81	0.80	0.80		0.77	0.77	0.77	
	(1-3)-Gram	Non-hate spreader	0.76	0.81	0.79	0.79	0.84	0.66	0.74	0.77
		Hate spreader	0.81	0.76	0.79		0.73	0.88	0.80	
		Weighted Avg.	0.79	0.79	0.79		0.78	0.77	0.77	
SVM	1-Gram	Non-hate spreader	1.00	0.06	0.12	0.54	0.79	0.72	0.75	0.77
		Hate spreader	0.53	1.00	0.60		0.76	0.82	0.79	
		Weighted Avg.	0.76	0.55	0.41		0.77	0.77	0.70	
	(1-3)-Gram	Non-hate spreader	1.00	0.06	0.12	0.54	0.84		0.74	0.77
		Hate spreader	0.53	1.00	0.69		0.73	0.88	0.80	
		Weighted Avg.	0.76	0.55	0.41		0.78	0.77	0.77	

Table 4

Results for the different deep learning models for English and Spanish Datasets

Model	English Dataset	Spanish Dataset
	Acc.	Acc.
LSTM	0.44	0.47
Bi-LSTM	0.54	0.49
BERT	0.53	0.56

5. Discussion

The major finding of the current work is that multinomial naive Bayes with n-gram features is a better model for identifying hate speech spreaders. With the English dataset, a multinomial naive Bayes with one-gram tf-idf feature yielded the best accuracy value of 70% and 66% with known as well as unknown test dataset, respectively. While with the Spanish dataset, the tf-idf features vector of 1-3 gram reported the best result with an accuracy of 82% and 80% for known

Table 5

Description of Hyper parameters for implemented LSTM network

Hyper-parameters	Value
Activation	ReLU
Loss function	Binary Cross Entropy
Optimiser	Adam
Vocabulary Size	5000
Embedding Vector Feature	40
Sentence Length	2500
Epochs	15
Batch Size	64
Validation Split	0.26
Metrics	Accuracy

Table 6

Description of Hyper parameters for Bi-LSTM model

Hyper-parameters	Value
Activation	ReLU
Loss function	Sparse Categorical Cross Entropy
Optimiser	Adam
Epochs	15
Batch Size	64
Metrics	Accuracy

and unknown test datasets, respectively.

The deep learning models such as LSTM, Bi-LSTM and BERT models were not found to be performing well while predicting hate speech spreaders. One of the reasons may be inefficient features presented as input to the deep learning models were not capturing the semantics of the text. The other limitation of the current work is that only the textual contents of the tweets are used for the experiments. The other components of a tweet such as images, videos and URLs may augment the current input to yield better accuracy.

6. Conclusion

The prediction of whether a user is spreading hate speech or not from his combined tweets is a challenging task as tweets have various noise in terms of grammatical mistakes, spelling mistakes, and non-standard abbreviations. Along with that when the different tweets of a single user have merged together the sentiments of a particular tweet might counter the effect of others. We trained classification models using tf-idf and count vector as feature values. We have shown a comparative study of machine learning algorithms with respective feature sets. We have compared their accuracies for different n-gram ranges i.e (1,1) and (1,3) and also for tf-idf and count vectorizer. We have shown the accuracy estimated in each case in the result

section. We achieved our best result with an F1-score of 0.74 for the English dataset when we used Multinomial Naive Bayes with n-gram range (1,1) and count vectorizer and of 0.82 for the Spanish dataset again for Multinomial Naive Bayes with n-gram range (1,3) and for both tf-idf and count vectorizer.

The final accuracy, as calculated by PAN on the test dataset is 66% for the English dataset and 80% for the Spanish dataset making an average of 73%. These results were obtained using Naive Bayes Classifier with an n-gram range of (1,1) for the English dataset and (1,3) for the Spanish dataset.

This system can be utilized by different social media platforms to identify hate speech spreaders and remove such hate speech spreaders from their platform.

References

- [1] R. Pandey, A. Kumar, J. P. Singh, S. Tripathi, Hybrid attention-based long short-term memory network for sarcasm identification, *Applied Soft Computing* 106 (2021) 107348.
- [2] F. Rangel, G. L. D. L. P. Sarracén, B. Chulvi, E. Fersini, P. Rosso, Profiling Hate Speech Spreaders on Twitter Task at PAN 2021, in: *CLEF 2021 Labs and Workshops, Notebook Papers*, CEUR-WS.org, 2021.
- [3] F. Rangel, A. Giachanou, B. Ghanem, P. Rosso, Overview of the 8th author profiling task at PAN 2020: Profiling fake news spreaders on Twitter, in: *CLEF, 2020*.
- [4] J. Pizarro, Using n-grams to detect fake news spreaders on Twitter: Notebook for pan at clef 2020, in: *Cross-Language Evaluation Forum CLEF, 2020*, pp. 1–8.
- [5] S. Caetano da Silva, T. Castro Ferreira, R. M. Silva Ramos, I. Paraboni, Data driven and psycholinguistics motivated approaches to hate speech detection, *Computación y Sistemas* 24 (2020).
- [6] C. Bosco, D. Felice, F. Poletto, M. Sanguinetti, T. Maurizio, Overview of the EVALITA 2018 hate speech detection task, in: *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, CEUR, 2018, pp. 1–9.
- [7] Z. Zhang, D. Robinson, J. Tepper, Detecting hate speech on Twitter using a convolution-GRU based deep neural network, in: *European semantic web conference*, Springer, 2018, pp. 745–760.
- [8] A. Gaydhani, V. Doma, S. Kendre, L. Bhagwat, Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach, *arXiv preprint arXiv:1809.08651* (2018).
- [9] S. S. Aluru, B. Mathew, P. Saha, A. Mukherjee, Deep learning models for multilingual hate speech detection, *arXiv preprint arXiv:2004.06465* (2020).
- [10] P. Saha, B. Mathew, P. Goyal, A. Mukherjee, Hateminers: Detecting hate speech against women, *arXiv preprint arXiv:1812.06700* (2018).
- [11] R. Mutanga, N. Naicker, O. Olugbara, Hate speech detection in twitter using transformer methods, *International Journal of Advanced Computer Science and Applications* 11 (2020) 01.
- [12] A. Kumar, S. Saumya, J. P. Singh, NITP-AI-NLP@ HASOC-Dravidian-CodeMix-FIRE2020:

- A machine learning approach to identify offensive languages from dravidian code-mixed text., in: FIRE (Working Notes), 2020, pp. 384–390.
- [13] A. Kumar, S. Saumya, J. P. Singh, NITP-AI-NLP@ HASOC-FIRE2020: Fine tuned BERT for the hate speech and offensive content identification from social media., in: FIRE (Working Notes), 2020, pp. 266–273.
 - [14] S. Saumya, A. Kumar, J. P. Singh, Offensive language identification in Dravidian code mixed social media text, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 2021, pp. 36–45.
 - [15] V. Basile, C. Bosco, E. Fersini, N. Deborá, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, et al., Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 54–63.
 - [16] G. K. Pitsilis, H. Ramampiaro, H. Langseth, Effective hate-speech detection in Twitter data using recurrent neural networks, *Applied Intelligence* 48 (2018) 4730–4742.
 - [17] J. Bevendorff, B. Chulvi, G. L. D. L. P. Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, , E. Zangerle, Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter, and Style Change Detection, in: 12th International Conference of the CLEF Association (CLEF 2021), Springer, 2021.
 - [18] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA Integrated Research Architecture, in: N. Ferro, C. Peters (Eds.), *Information Retrieval Evaluation in a Changing World, The Information Retrieval Series*, Springer, Berlin Heidelberg New York, 2019. doi:10.1007/978-3-030-22948-1_5.
 - [19] J. P. Singh, A. Kumar, N. P. Rana, Y. K. Dwivedi, Attention-based lstm network for rumor veracity estimation of tweets, *Information Systems Frontiers* (2020) 1–16.
 - [20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).