Hate Speech Detection on Twitter: A Comparative Evaluation of Different Machine Learning Techniques

Aryan Rastogi, Arjit Kumar, Daarshik Dwivedi, Abhishek Pratap Singh, Suruchi Saberwal, and Mehboob Alam

Abstract The necessity for robust and fast detection techniques has become critical as social media hate speech has grown. This study investigates ways to identify hate speech comments present on Twitter using language processing methods. In this work, we suggest a cutting-edge method for effectively identify hate speech in tweets that combines linguistic elements and machine learning techniques. Using a sizable dataset of annotated tweets, we test our model, and we get good F1-score and accuracy. The findings of this study present the possibilities of using techniques for processing natural language to identify hateful speech on Twitter and can assist in direct the creation of efficient regulations and interventions to lessen the negative consequences of hate speech on social media sites.

Keywords Twitter · Machine learning · Tweets · Hateful speech

1 Introduction

From the rapid growth of the online ecosystem, it can be observe that a large amount of user data is produced every second in the form of images, videos, text posts. This data is commonly generated from social media platforms. The large amount of data generated also includes hate speech between different groups within and across countries to spread prejudices and disputes. Several social media platforms have measures to counter this problem. Twitter's rules state that users cannot use tweets to threaten or harass people as a result of their, gender, race, religion, or

A. Rastogi (✉) · A. Kumar · D. Dwivedi · A. P. Singh · S. Saberwal · M. Alam
Department of Computer Science, JSS Academy of Technical Education, Noida, Uttar Pradesh, India
e-mail: aryanrastogi97@gmail.com

S. Saberwal
e-mail: suruchi@jssaten.ac.in

M. Alam
e-mail: mahboobalam.fet@mriu.edu.in

K. C. Santosh et al. (eds.), *Advances in Artificial-Business Analytics and Quantum Machine Learning*, Lecture Notes in Electrical Engineering 1191,
https://doi.org/10.1007/978-981-97-2508-3_11

any other characteristic. Along with content that is blocked by gender, class, and handicap, YouTube screens information that incites hatred or hostility against specific individuals or groups.

Social media channels like Twitter and Reddit now represent the most popular mediums for the dissemination of hate speech due to their vast user base and ease of use. Twitter and Reddit have 280 million daily active users combined. The anonymity provided by social media can also encourage others to take part in hate speech, as they feel that they can express their views without fear of being made answerable for their deeds. As a result, hateful speech has increased on social media sites, posing a serious problem for online communities.

With such a wide reach of social media channels, the effects of hate spreading speech can be far-reaching, as it can lead to psychological harm, social isolation, and even physical violence. It can also create an atmosphere of fear and intimidation, which can prevent individuals from expressing their views and participating in public discussions. The impact of hate speech can be particularly devastating for vulnerable communities that may already be marginalized and discriminated against (Fig. 1).

To combat the rise of hateful speech, many researchers have conducted studies to understand the nature and extent of this problem. Machine learning-based automated hate speech identification systems have attracted increasing attention in recent years. These systems have the potential to identify hate speech and provide early warnings to online communities, thereby reducing the spread of hateful speech.

This study emphasises the critical need for efficient hate speech identification and mitigation on social networks. This study contributes to the creation of more potent methods for preventing hate speech online by putting forth a unique approach to hate speech identification that combines machine learning algorithms with natural language processing techniques. The findings of this study may influence the creation of automated technologies for identifying hate speech as well as the formulation of regulations and initiatives aimed at fostering more civil and welcoming online communities.

The research paper is organized logically and simply to effectively convey the goals and conclusions of the investigation. A literature review that gives the context

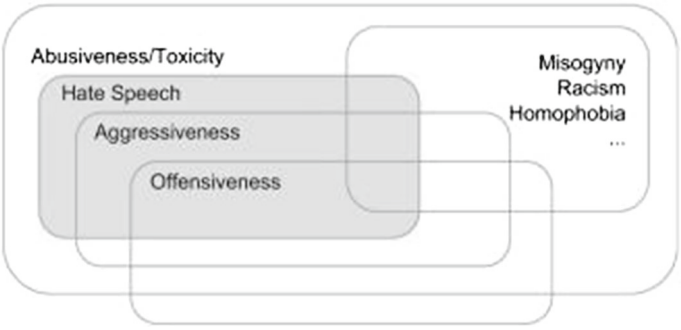

Fig. 1 Relationships between abusive language phenomena. *Source* Poletto et al. (2020)

and rationale for the investigation is included after the introduction, which introduces the research challenge and motivation for the study. The data sources, preprocessing methods, and models utilized in the experiment are described in the methodology section. The methodology section of the article gives a thorough breakdown of the whole strategy adopted. The study's results are presented in the results section, which is followed by a discussion part that interprets the findings and sheds light on their relevance. The conclusion then emphasizes the research's contributions, summarises the main findings, and offers suggestions for future research. A reference list with a thorough breakdown of the sources used in the study is also included in the publication. Ultimately, the paper's structure is intended to aid the reader's comprehension and involvement with the study.

2 Related Work

Six research articles on the identification of hateful comments posted on social media are compiled in the literature review. Each study approaches the problem of identifying hateful speech using a different strategy or algorithm, employing a variety of datasets and linguistic factors Understanding the methodologies employed in each experiment and identifying the key conclusions and insights of each research publication are the goals of the analysis. In all, relevant work survey while taking 10 publications into account was conducted.

The influence of extra-linguistic factors on hate speech identification is examined in the first publication (Waseem and Hovy 2016). The study examined 16,914 messages for sexist, racist, or neutral content using Twitter data from 614 individuals. Tenfold cross-validation was used with the LR (Wenando et al. 2017) predictor to examine the impact of different features on prediction accuracy. The study discovered that character n-grams up to four characters long, along with demographic data other than gender, produced the greatest outcomes.

The second paper (Waseem 2016) explored the awareness of hate speech among annotators' effects on categorization models. The dataset was obtained by sampling tweets from the 130 k tweets extracted by Waseem and Hovy (2016). The features defined by the annotators were evaluated using fivefold cross-validation to assess the significance of characteristics in groups of novice and experienced annotators. The researchers found that their best-performing system was not able to substantially outperform previous classification efforts.

The significance of fine-grained labeling for hate speech identification is highlighted in the third study (Davidson et al. 2017). The dataset included 25 k tweets that were manually tagged by CrowdFlower staff from a sample of 85.4 million tweets sent by 33 k individuals. The resulting model, which had an F1 score of 0.90, a recall of 0.90, and an overall accuracy of 0.91, was an LR (Wenando et al. 2017) using L2 regularisation. The survey did discover that about 40% of hateful speech was misclassified.

The subject of the fourth study (Badjatiya et al. 2017) was whether or not to label a tweet as racist, sexist, or none. The researchers combined CNN- and LSTM-based architectures with text processing techniques such as TF-IDF (Shaikh and Doudpotta 2019), char n-grams, and BoWV. The Waseem and Hovy (2016) dataset contained 16,000 annotated tweets. The word TF-IDF approach outperformed the character n-gram approach according to the study, while the CNN model outperformed the LSTM and FastText models.

The fifth study (Gao and Huang 2017) investigated how to automatically detect hate speech while taking context into account. The dataset included 1528 comments from Fox News users, and both machine learning as well as deep learning techniques were taken into account. By include both word-level n-gram features as well as lexicon-derived characteristics, the researchers discovered that the LR (Wenando et al. 2017) method produced good results. Comparing the performance of utilising both ensemble models to using only one model, hate speech detection was substantially enhanced.

The sixth paper (Saha et al. 2018) focused on identifying misogynous posts in English on Twitter. The dataset consisted of 4000 labeled tweets for training and 1000 unlabelled tweets for testing. The researchers used LR (Wenando et al. 2017), XGBoost, and CatBoost for the classification task and observed that the LR (Wenando et al. 2017) system was the best performer among the model systems evaluated with an accuracy of 70.4%.

Using the "Hate and Abusive Speech on Twitter" dataset, researchers in Lee et al. (2018) tested conventional machine learning models versus neural network models to detect abusive language. The findings demonstrated that among models based on neural networks, RNNs with LTC modules had the highest accuracy ratings, while the LR (Wenando et al. 2017) was the most effective conventional model. The most trustworthy outcomes were provided by bidirectional GRU networks incorporating LTC.

In Mozafari et al. (2020), researchers used a transfer learning methodology in conjunction with the supervised fine-tuning and unsupervised pre-trained BERT model methods to identify hate speech. They utilized (Waseem and Hovy 2016) and (Davidson et al. 2017) datasets, which included 25 k tweets from unbalanced groups. Utilizing a batch size of 32 for three training epochs as well as a dropout layer probability of 0.1 for all layers, the BERT model, text tokenizer, and WordPiece were used. A 2e−5 learning rate was used while using the Adam optimizer. F1 scores for BERT-based techniques were higher than baselines, while the addition of a CNN to the BERT model gave the greatest outcomes, with F1 scores for the Waseem and Davidson datasets of 88% and 92%, respectively.

In Mathew et al. (2020) researchers present a database that focuses on analyzing various aspects of hate speech. The dataset consisted of 20,000 posts collected from Twitter and Gab. Each post was physically annotated by multiple annotators, and the labels were chosen by a qualified majority. The resulting dataset provides a new benchmark for hate speech detection, with each data point labeled as offensive, normal, or hate, along with providing details about the target population and highlighted text passages that support the label. However, the study did not consider

demographic data that could have been useful for annotating the data, and only English language data was used.

In contrast, academics analyzed large-scale multilingual hate speech in 9 languages from 16 sources (Aluru et al. 2020). They trained models in multilingual environments using multilingual embeddings for words and sentences, such as LASER and MUSE. They evaluated different combinations of models such as BERT, mBERT, CNN + GRU, LASER + LR, MUSE + CNN-GRU, and Translation + BERT for the detection of hate speech. They found that BERT models performed better in high-resource scenarios than LASER + LR in low-resource situations. LASER + LR performed better than MUSE + CNN-GRU in most cases. They also discovered that there wasn't a one-size-fits-all solution, but Translation + BERT seemed to be an excellent compromise. In zero-shot analysis, mBERT outperformed LASER + LR significantly in three languages (Arabic, French, and German), while LASER + LR performed better than mBERT in Portuguese and Italian. In general, LASER + LR outperformed mBERT in low-resource situations for most languages. However, for high-resource situations, mBERT scored higher than LASER + LR for German, Arabic, and French. LASER + LR consistently delivered excellent results for Portuguese.

Overall, the literature review shows that researchers have employed a variety of techniques to help solve the issue of detecting social media hateful speech. While some studies focused on the impact of extra-linguistic features and fine-grained labels, others explored the use of context information and different text processing algorithms with machine learning and deep learning models. The studies' findings indicate that the problem of hate speech detection remains challenging, with misclassification rates ranging from 40 to 70.4%. The studies suggest that further research is necessary to help solve the issue of detecting hate speech and to improve the models' accuracy and performance.

3 Methodology

The suggested technique used to categorize tweets into two groups—"hate speech and non-hate speech"—is explained in this section. The whole research technique is shown in Fig. 2. Data collection, data preprocessing, feature engineering, data splitting, base model creation, best base model selection, hyperparameter tuning of selected models, and classification model assessment are the eight main procedures for the research methodology as indicated in this image. In the sections that follow, each process is covered in full.

A. *Data Collection*

For this study, dataset of hate speech tweets was collected that were made publicly available. The University of Aristotle Hate Speech Dataset, University of Copenhagen Hate Speech Dataset, HASOC 2019, HASOC 2020, and Georgia Tech Dataset were combined with the Davidson Dataset to create this dataset. The

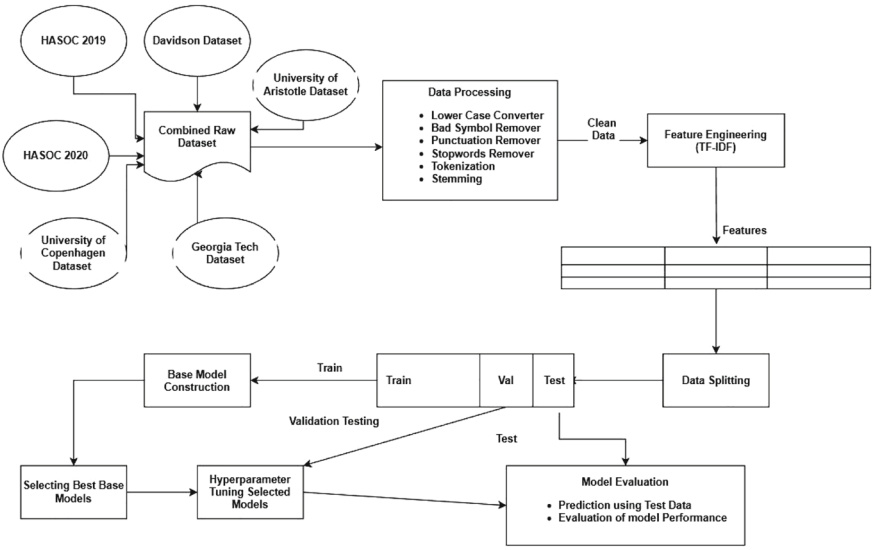

Fig. 2 System design

tweets in the merged dataset are categorised into two separate groups., namely hate speech and non-hate speech. There are 30,270 tweets in this dataset. Of these, 22.9% of tweets fall under the category of hate speech, while 77.1% fall under the category of non-hate speech (Fig. 3; Table 1).

B. *Text Processing*

Several research studies have explained that using text preprocessing makes better classification results (Dutta et al. 2022). So, in the dataset, different preprocessing techniques were applied to filter noisy and non-informative features from the tweets. In preprocessing, we changed the tweets into lowercase. Also,

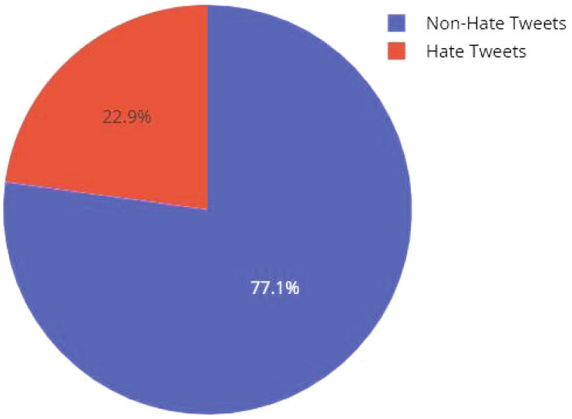

Fig. 3 Class distribution of combined dataset

Table 1 Datasets used

Name of dataset	Detail of instances used
Davisdon dataset (Banda et al. 2020)	Non-hate tweets: 23,353 hate tweets count: 1430
University of Aristotle hate speech dataset (Founta et al. 2018)	1614 hate speech tweets used
University of Copenhagen hate speech dataset (Waseem and Hovy 2016)	2684 hate speech tweets used
HASOC 2019 (Mandl et al. 2019)	58 hate speech tweets used
HASOC 2020 (Mandl et al. 2020)	131 hate speech tweets used
Georgia tech hate speech dataset (Banda et al. 2020)	1500 hate speech tweets used

we removed all the URLs, usernames, white spaces, hashtags, punctuations, and stop-words using pattern-matching techniques from the collected tweets. In addition, tokenization using preprocessed tweets was done. Each tweet is first tokenized into words or tokens, and then words are further transformed by the Porter stemmer into their root forms, such as offended to offend.

C. *Feature Engineering*

The categorization criteria from the raw text are incomprehensible to machine learning systems. For these algorithms to comprehend categorization rules, numerical characteristics are required. Hence, feature engineering is a crucial stage in text categorization. The primary features from the raw text are retrieved in this stage, and the characteristics are then represented numerically. In this inquiry, TF-IDF (Shaikh and Doudpotta 2019) has been utilized as a feature engineering technique.

D. *Exploratory Data Analysis*

Any research concerning hate speech must use exploratory data analysis (EDA). EDA includes looking at the data to find outliers, trends, and patterns. It enables researchers to comprehend the data's features more thoroughly and to spot any problems or biases that could exist. In the context of research on hate speech, EDA may entail determining the frequency and distribution of hate speech across various demographic groups, identifying the terminology or expressions that are most often used in hate speech, and assessing the environment in which hate speech is expressed. Examining the connections between hate speech and other factors, including the frequency of hate crimes or the usage of hate speech in political discourse, may also be part of EDA. Ultimately, EDA is a crucial instrument for understanding the type and extent of hate speech and for guiding the creation of successful solutions to this issue (Figs. 4 and 5).

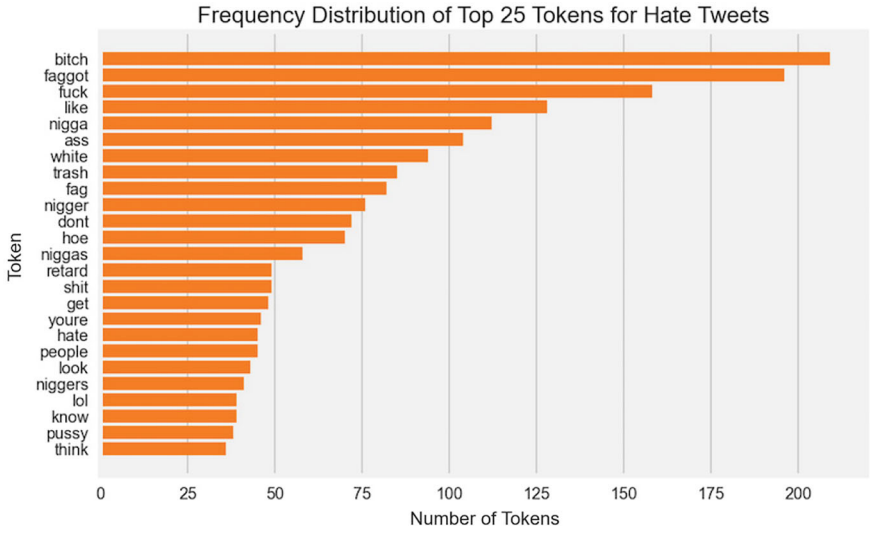

Fig. 4 Frequency distribution of top 25 tokens for hate tweets

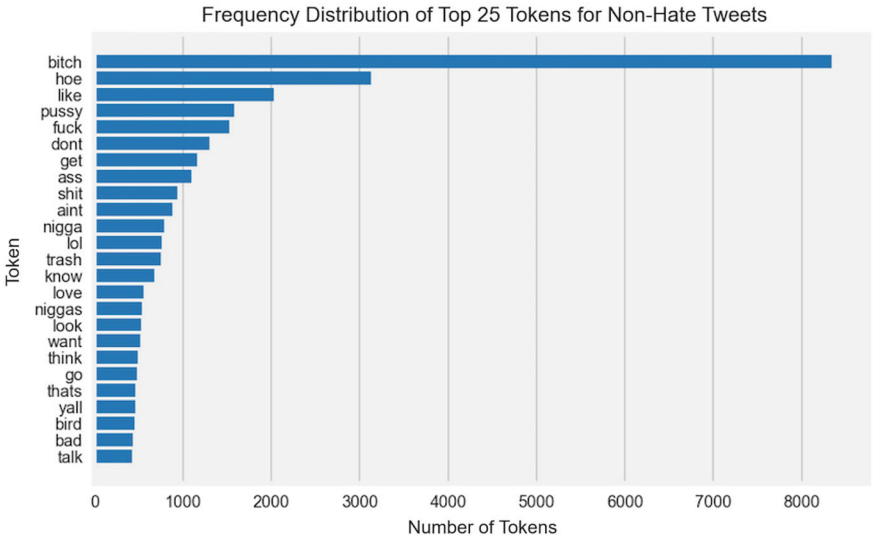

Fig. 5 Frequency distribution of top 25 tokens for non-hate tweets

There are common terms and expressions used in both forms of communication, according to an exploratory examination of the top 25 words in both the Hate and Non-Hate categories. This demonstrated that the vocabulary used to communicate hate speech and non-hate speech are not fundamentally dissimilar. It also

implied that some of the most popular terms in everyday speech may be appropriated and used to spread hate. This reinforced the necessity of supporting civil discourse and excellent communication in addition to monitoring hate speech in hopes of preventing the propagation of hate expression through social media.

E. *Data Splitting*

Table 2 displays the whole dataset's class-wise distribution as well as the data set following splitting (i.e. Training set, Validation set, and Test set). To divide the preprocessed data into training and validation sets, it utilized a 75:25 ratio, or 75% for training and 25% for validation. To create the test data, the validation dataset was once more divided in half using a 60:40 ratio. To learn the rule base, the classification model has to be trained using data for training. To evaluate the model's effectiveness, the validation data is used. Using test data on hypothetical scenarios, the categorization model is also assessed.

F. *Machine Learning Models*

It is thought that there isn't a single classifier that excels on all types of datasets when applying machine learning to a task. As a result, it is advised to test a variety of classifiers on a master feature vector to see which produces the best outcomes. So, the Multinomial Naive Bayes, LR (Wenando et al. 2017), SVM (Joachims 1998), RF (Xu et al. 2012), Adaboost (Ying et al. 2013), and GBDT are chosen as six distinct classifiers.

G. *Classifier Evaluation*

This phase involves utilizing the test set to predict the class of unlabeled text using the created classifier. The generated classifier's performance is evaluated using a variety of performance indicators. This is a quick discussion of a few standard text classification performance metrics.

(1) Precision: Another name for precision is the positive projected value. It is the percentage of predicted positives that really turn out to be positive. Equation (1) summarizes the precision calculation formula.

$$Precision = \frac{TP}{(TP + FP)} \tag{1}$$

(2) Recall: The percent measure of favorable outcomes that actually occur and are anticipated to occur. Observe "(2)".

Table 2 Data split

	Class label	Total instances	Training instances	Validation instances	Testing instances
0	Non-hate tweets	23,353	17,514	3503	2336
1	Hate tweets	7417	5562	1113	742

$$Recall = \frac{TP}{(TP + FN)} \tag{2}$$

(3) F1 Score: It represents the harmonic mean of recall and accuracy (as shown in Eq. 3). Precision and recall are given equal weight in the standard F1 Score. Equation (3) presents the formula to calculate F1 score.

$$F1\,Score = \frac{2 \times (precision \times recall)}{(precision + recall)} \tag{3}$$

(4) Accuracy: It is the number of cases that were appropriately categorized. The formula is given in Eq. (4).

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{4}$$

4 Experimental Observations

Aggregated datasets are trained on six classifiers and the model training and testing results and shown in Table 3. Further to this, the hyperparameter tuning of these classifiers has been performed using GridSearchCV to obtain optimal parameters and again the model training and testing results have been evaluated and presented in Table 4.

In terms of F1-score, accuracy, and recall, a comparison of the results in Tables 3 and 4 revealed that the hyperparameter tweaked models performed much better than the basic models. The greatest F1 score was 0.820634 for the LR (Wenando et al. 2017) approach using GridsearchCV, with recall and precision values of 0.778976 and 0.867000, respectively. Similar to how the SVM model with GridsearchCV worked well, with an F1 score of 0.828640, recall of 0.77987, and accuracy of

Table 3 Base model testing results

	Accuracy	F1 score	Recall	Precision
Multinominal Naive Bayes	0.863548	0.611111	0.44744	0.976331
Random forest (Xu et al. 2012)	0.882716	0.698413	0.563342	0.918681
Logistic regression (Wenando et al. 2017)	0.892788	0.743390	0.644205	0.878676
Support vector machine (Joachims 1998)	0.901884	0.779240	0.718329	0.851438
Adaboost classifier (Ying et al. 2013)	0.885965	0.738645	0.668464	0.825291
Gradient boosting classifier	0.878493	0.673647	0.520216	0.955446

Table 4 Hyperparameter tuned model results

	Accuracy	F1 score	Recall	Precision
Logistic regression w/ GridsearchCV (Wenando et al. 2017)	0.917894	0.820634	0.778976	0.867000
Random forest w/GridsearchCV (Xu et al. 2012)	0.895364	0.731219	0.590296	0.960526
Support vector machine w/ GridsearchCV (Joachims 1998)	0.922227	0.828640	0.77987	0.883910

0.883910, it likewise performed well. Our findings highlight the significance of hyperparameter optimization in hate speech identification as it may greatly enhance model performance.

The performance of the base models, on the other hand, varied. With an F1 score of 0.611111, the multinominal Naive Bayes model performed poorly in terms of detecting instances of hate speech. The RF (Xu et al. 2012) model fared better, earning an F1 score of 0.698413, but its recall was still just 0.563342, which was not particularly high. The gradient boosting classifier demonstrated the trade-off between accuracy and recall by having the highest precision (0.955446), but also the lowest recall (0.520216).

Overall, our findings highlight the utility of hyperparameter tweaking in enhancing model performance as well as the necessity of prioritizing F1 score, precision, and recall above accuracy in hate speech identification.

5 Conclusion

The investigation leads to the conclusion that detecting hateful speech is a difficult undertaking, especially when working with unbalanced datasets. A dataset comprising about 22 k non-hate tweets and 8 k hate tweets was produced for the study by combining hate and non-hate tweets.

Using default settings, the performance of six binary classification models was assessed. It was found that the SVM (Joachims 1998) model, which had an accuracy of 90.19%, was the most accurate, followed by Logistic Regression, which had an accuracy of 89.28%. It was pointed out, nonetheless, that accuracy is not the ideal statistic to assess a model's performance with an unbalanced dataset.

The F1 Score, Recall, and Precision results indicated that the SVM model performed superior to the other models. The model obtained F1 scores of 0.779, 0.718 for recall, and 0.851 for precision. Also, it was discovered that the SVM model's performance greatly increased following hyperparameter adjustment, with an F1 Score of 0.829, Recall of 0.780, and Precision of 0.884.

In conclusion, the study contends that SVM, particularly when working with unbalanced datasets, is a viable model for the detection of hateful speech on social

media. While assessing the model's performance on an unbalanced dataset, it is critical to place an emphasis on measures like F1 Score, Recall, and Precision rather than accuracy.

References

S.S. Aluru,, B. Mathew, P. Saha, A. Mukherjee, Deep learning models for multilingual hate speech detection (2020). ArXiv, abs/2004.06465

P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in *Proceedings of the 26th International Conference on World Wide Web Companion* (2017)

J.M. Banda, R. Tekumalla, B. Abrahao, M.A. Al-Garadi, D. Lewis, COVID-19 misinformation detection using sparse deep learning models. Appl. Intell., 1–15 (2020)

T. Davidson, D. Warmsley, M.W. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in *International Conference on Web and Social Media* (2017)

S. Dutta, S. Caur, S. Chakrabarti, T. Chakraborty, Semi-supervised stance detection of tweets via distant network supervision, in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (2022)

A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis, *Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior* (2018)

L. Gao, R. Huang, Detecting online hate speech using context aware models, in *Recent Advances in Natural Language Processing* (2017)

T. Joachims, Text categorization with support vector machines: learning with many relevant features, in *European Conference on Machine Learning* (Springer, 1998)

Y. Lee, S. Yoon, K. Jung, Comparative studies of detecting abusive language on Twitter, in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)* (Association for Computational Linguistics, Brussels, Belgium, 2018), pp. 101–106

T. Mandl, R. Jain, M. Illa, M. Zampieri, A. Bhardwaj, Overview of the HASOC track at FIRE 2019: hate speech and offensive content identification in Indo-European languages, in *Working Notes Proceedings of the Conference and Labs of the Evaluation Forum (CLEF)*, vol. 2380 (CEUR-WS.org, 2019)

T. Mandl, P. Goyal, R. Jain, M. Illa, M. Zampieri, A. Bhardwaj, Overview of the HASOC track at FIRE 2020: hate speech and offensive content identification in Indo-European languages, in *Working Notes Proceedings of the Conference and Labs of the Evaluation Forum (CLEF)*, vol. 2696 (CEUR-WS.org, 2020)

B. Mathew, P. Saha, S.M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, HateXplain: a benchmark dataset for explainable hate speech detection (2020). arXiv: Computation and Language

M. Mozafari, R. Farahbakhsh, N. Crespi, A BERT-based transfer learning approach for hate speech detection in online social media, in *Complex Networks and Their Applications VIII. COMPLEX NETWORKS 2019*, ed. by H. Cherifi, S. Gaito, J. Mendes, E. Moro, L. Rocha. Studies in Computational Intelligence, vol. 881 (Springer, Cham, 2020)

F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review. Lang. Resour. Eval. **30**, 1–47 (2020)

P. Saha, B. Mathew, P. Goyal, A. Mukherjee, Hateminers: detecting hate speech against women (2018). ArXiv, abs/1812.06700

S. Shaikh, S.M. Doudpotta, Aspects based opinion mining for teacher and course evaluation. Sukkur IBA J. Comput. Math. Sci. **3**(1), 34–43 (2019)

Z. Waseem, Are you a racist or am i seeing things? Annotator influence on hate speech detection on Twitter, in *Proceedings of the First Workshop on NLP and Computational Social Science* (Association for Computational Linguistics, Texas, Austin, 2016), pp. 138–142

Z. Waseem, D. Hovy, *Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter* (North American Chapter of the Association for Computational Linguistics, 2016)

F.A. Wenando, T.B. Adji, I. Ardiyanto, Text classification to detect student level of understanding in prior knowledge activation process. Adv. Sci. Lett. **23**(3), 2285–2287 (2017)

B. Xu et al., An improved random forest classifier for text categorization. JCP **7**(12), 2913–2920 (2012)

C. Ying et al., Advance and prospects of AdaBoost algorithm. Acta Automat. Sin. **39**(6), 745–758 (2013)