# A Comparative Study: Hate Speech Detection Models

**Tirthendu Prosad Chakravorty, Mobashra Abeer, Khan Abrar Shams,**
**Shaiane Prema Baroi, Sifat E Jahan, Annajiat Alim Rasel**
Department of Computer Science and Engineering, Brac University, Dhaka, Bangladesh
tirthendu.prosad.chakravorty@g.bracu.ac.bd, mobashra.abeer@g.bracu.ac.bd,
khan.abrar.shams@g.bracu.ac.bd, shaiane.prema.baroi@g.bracu.ac.bd,
sifat.jahan@bracu.ac.bd, annajiat@gmail.com

## Abstract

The ubiquity of toxic content on online platforms and its alarming increase due to the widespread targeting of individuals with hateful content has become a pressing issue of this modern era. Hence, it has become incredibly urgent to develop effective hate speech detection methods in order to combat this problem. As such, a comparative study has been conducted to evaluate various models in order to determine the effective approach for a given tweet-based dataset. The study involved preparing the dataset through various data pre-processing and data cleaning techniques before it was used in six different models: the Logistic Regression model, the fine-tuned Random Forest Classifier, the SVM model, a custom CNN model, a custom RNN model and the Transformer model (BERT). After evaluating the result of each model, the outcomes concluded that the pre-trained transformer model attained the best accuracy.

## 1 Introduction

Social media has surely made significant advancements in the ability to link individuals throughout time thanks to the present age of science and technology. However, this development has also caused severe difficulties in the lives of people, despite its immense benefits. Social media has led to numerous cases of harassment and abuse, which pose a threat to individuals' emotional well-being and have resulted in numerous fatalities due to cyberbullying. Recent studies show that social media has increased anxiety among young adults [25] which increases rate of social anxiety among victims. It has also been shown that youngsters who are bullied or subjected to cyberbullying are almost twice as likely to attempt suicide [13]. In order to stop at least one type of cyberbullying and harassment, it is necessary to identify toxic content on digital platforms. This is where Natural Language Processing (NLP) plays a crucial part as a branch of computer science that enables machines to understand human language and assist in language-related tasks. In this paper, different NLP techniques are explored along with various learning models to detect hateful speech and offensive language, with the aim of identifying the best model for the fastest and most efficient detection.

## 2 Related Work

Social media abuse is one of the complicated phenomenon that has many overlapping modes and objectives [8]. Due to their detrimental effects on our communities, scholars have recently become more interested in common forms of abusive languages, such as cyberbullying and offensive language. Numerous research has been conducted to discover these unwanted communications in social media amid other messages. The authors of this research [10] did a brief evaluation of eight methods and techniques for detecting hate speech. TF-IDF vectorizer, N-grams, sentiment analysis, Bag of the Word, part of speech, and rule-based approaches are some of the methodologies. The flaw of the study includes that it did not take into account methods like deep learning and the ensemble approach. The people who write about hate speech typically pick out the victims based on things like religion, politics, ethnicity, disability, gender, etc [12]. Every day, SM sites produce vast amounts of data that are growing geometrically [1]. It can be seen that a significant portion of the roughly 7.7 billion people on the globe [23] [21] are actively linked to one or more social media platforms [15] [16].

## 3 Dataset

The dataset for this paper has been obtained from a study conducted by multiple contributors, titled "Automated Hate Speech Detection and the Problem of Offensive Language", published in 2017 and named the Davidson Dataset [7].

| Class | Sample Size | Percentage |
|---|---|---|
| Hate | 1430 | 5.77% |
| Offensive | 19190 | 77.43% |
| Neither | 4163 | 16.80% |
| Total | 24783 | 100% |

Table 1: Dataset Statistics

A detailed summary of its major statistics can be found in Table I. The hate speech lexicon, which includes numerous terms and expressions labeled as hate speech by diverse internet users, is where the dataset is first compiled. A total of 85.4M tweets collected from 33,458 Twitter users were found when the researchers utilized the Twitter API to look for tweets that contained words from the dictionary. Employees of CrowdFlower (CF) manually code 24,783 randomly selected tweets from this dataset. Each tweet is required to be classified under one of three categories: Offensive, Hateful, or Neither by the employees.

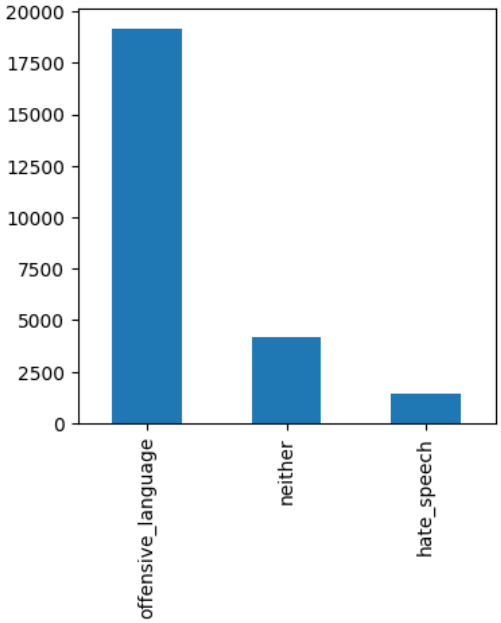


Figure 1: Class Distribution

# 4 Methodology

## 4.1 Data Preprocessing

Data Preprocessing is a crucial step of Natural Language Processing before applying any desired model to the dataset. This framework involves several essential steps to clean and transform the data for effective hate speech detection in the respective models. At first, stemming and removal of stop words was performed using the Snowball Stemmer and the NLTK library of Python. Moreover, further data cleaning methods were applied to remove unnecessary words starting with #, &, @ and html tags, URLs, punctuation, and empty new lines. Next, the TFIDF Vectorizer was used to convert the cleaned data into numerical values for feature extraction to be used in the later steps. Afterward, the dataset into test and train data was split, allocating 10% of the data for testing and 90% for training. Finally, a Vectorizer was used to transform the train and test data or features to be used in the models afterward.

## 4.2 Logistic Regression

Logistic Regression is a supervised learning model applied to solve multi-class classification problems. After preprocessing our data, we used the Logistic Regression model to make predictions on the testing data. Our model achieved an accuracy of 89.75% and produced a classification report, which showed the performance of our model for each class.

## 4.3 Random Forest Classifier

Another supervised learning model that can be used for both classification and regression issues is Random Forest. It combines the output obtained from multiple decision trees to arrive at a single output. In this study, we have used the Random Forest Classifier model by fine-tuning certain Hyperparameters through RandomizedSearchCV to obtain the best possible combination of Hyperparameters. We then used this optimized model to predict the testing data, which resulted in an accuracy of 89.47%. Moreover, a classification report also provided information about the performance of our model for each class, after applying it to our pre-processed data.

## 4.4 Support Vector Machine

Next, we utilized the Support Vector Machine (SVM) model, which also falls under the category of a supervised learning model that can be used to solve both classification and regression problems. We used the SVM model to predict the test data, resulting in an accuracy of 89.14%. We also produced a classification report that provides information about the model's performance for each class.

### 4.5 CNN

A comparison was made between a convolutional neural network (CNN) and its ability to classify text from Twitter comments. The CNN architecture includes an embedding layer, a 1D convolutional layer with a ReLU activation function, a Max Pooling layer, and a Dropout layer. The input is first passed into the embedding layer and then into the 1D convolutional layer. The output from the convolutional layer serves as the final output of this process. This output is subsequently fed into two dense layers, both utilizing ReLU and Sigmoid activation functions. Below is a simple sigmoid function:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

To train the model, an Adam optimizer with a learning rate of 0.0001 is employed, while early stopping is used as a regularization technique to avoid overfitting.

### 4.6 RNN

This paper also compares the text classification with Recurrent Neural Network (RNN) models leveraging Gated Recurrent Units (GRU) and pre-trained GloVe embeddings. This architecture relies on an embedding layer which was then initialized with pre-trained GloVe embeddings. Then it transforms the input word sequences into dense vectors while the Dropout layer prevents overfitting. The GRU layer (64-filter) processes the input sequence and generates a fixed-length output vector. The first Dense layer (64-filter) along with the ReLU activation function is used to capture higher-level features of the input, while the second Dense layer (3-filter) and Softmax activation function produce the output probabilities for the 3-class classifications. When the validation loss reaches a plateau, the learning rate is reduced by 0.1 using the ReduceLROnPlateau callback. Below is a softmax function:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad for\ i = 1, 2, \ldots, K \tag{2}$$

### 4.7 Transformer (BERT)

The Transformer technique has proved to be highly effective in Natural Language Processing (NLP), with more information available in reference [19]. Transformer models have been particularly useful for multilingual applications due to their versatility. These models can efficiently calculate vector-space representations that are useful in deep learning. BERT, for example, is composed of Encoders that are stacked on top of one another, with no decoders present. This means that the model only understands the context of a language, as Encoders learn the context of a language. Furthermore, each token is considered in the full context of all prior and subsequent tokens. As a result, transformer models are evident to be highly effective in NLP. The BERT model has pre-trained weights for every English word and is also capable of learning to map words to vectors on its own during training. In the code provided, the tokenizer used is a BERT tokenizer from the pre-trained model "bert-base-case". The optimizer used for training the model is AdamOptimizer, which is a popular optimization algorithm used in deep learning. Moreover, Categorical Cross Entropy was used as the loss function. For this research, the BERT model has been imported, but an intermediate layer of Relu and a dense layer of Softmax have also been added. These components are crucial in the training and evaluation of the BERT model, enabling it to make accurate predictions on different NLP-related works.

## 5 Results and Discussion

For determining the outcomes, metrics like precision, f1 score, and recall were used to evaluate all the classes. In general, accuracy measures the overall performances of all the approaches. The following functions yield the above evaluation metrics:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1Score = \frac{2 \times TPR \times Precision}{TPR + Precision} \tag{5}$$

Out of all the ML learning techniques, SVC on average gave the best results and Random Forest yielded below-par outcomes. SVC attained average precision of 0.76, recall of 0.69, and f1 score of 0.70 over all classes. The scores are 0.79, 0.64, and 0.66 respectively for Random Forest. The scores obtained from generated confusion matrix were quite similar to each other. The best accuracy

| Model | Accuracy |
|---|---|
| Logistic Regression | 0.8975 |
| Random Forest Classifier | 0.8947 |
| SVM | 0.8914 |
| RNN | 0.89 |
| CNN | 0.87 |
| BERT | 0.98 |

Table 2: Accuracy Table

among all approaches, as can be seen in Table 2 was obtained by the BERT transformer.

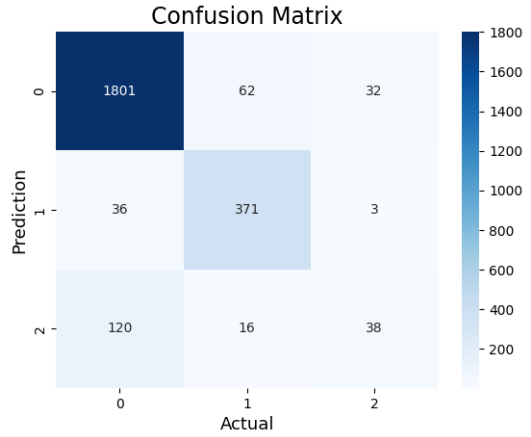


Figure 2: SVC Confusion Matrix

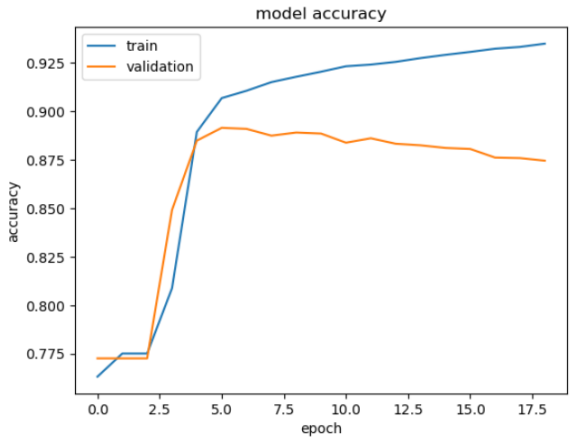


Figure 3: CNN Model Accuracy

## Conclusion

In conclusion, it is salient to acknowledge the growing issue of toxic content on digital platforms. Therefore, it is now more crucial than ever to build reliable hate speech detection techniques. This study investigated the efficiency of six various offensive language detection techniques on digital

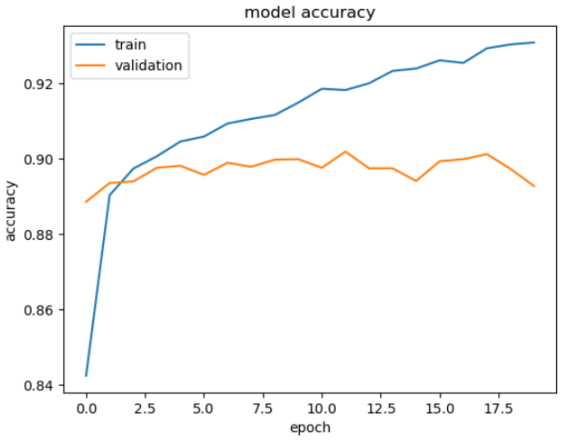


Figure 4: RNN Model Accuracy

media. After a careful analysis, the transformer model exhibited the greatest accuracy. Researchers and practitioners might use this study as a useful resource when creating more effective strategies to stop abhor speech on social networking sites. In the end, it is our duty to make the Internet a more secure and welcoming place for everyone.

## Future Work

The study will subsequently concentrate on creating models and methods that can recognize hate speech in various linguistic and cultural contexts. The use of context, sarcasm, and subtle references in hate speech can make it difficult to identify. Using elements like user history, conversation flow, and overall content, future research will work to create algorithms that can more accurately interpret the context and intent underlying messages. This can entail making use of natural language processing (NLP) tools like sentiment analysis, emotion recognition, and conversational context comprehension. The research will also concentrate on creating reliable detection models that can withstand hostile attacks. Investigating techniques for creating and identifying adversarial examples, such as generative adversarial networks (GANs), may be one way to do this. Lastly, working on real-time detection and intervention will be one of the major focuses of this research work.

## References


[1] Mohammed Ali Al-Garadi, Mohammad Rashid Hussain, Nawsher Khan, Ghulam Murtaza, Henry Friday Nweke, Ihsan Ali, Ghulam Mujtaba, Haruna Chiroma, Hasan Ali Khattak, and Abdullah Gani. 2019. Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges. *IEEE Access*, 7:70701–70718.

[2] Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76. IEEE.

[3] Zahangir Alom, Tarek M Taha, Chris Yakopcic, Stefan Westberg, Shamima Nasrin, and Vijayan K Asari. 2017. Comprehensive survey on deep learning approaches.

[4] Leonidas Aristodemou and Frank Tietze. 2018. The state-of-the-art on intellectual property analytics (ipa): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (ip) data. *World Patent Information*, 55:37–51.

[5] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

[6] Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5:1–15.

[7] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[9] Zejin Ding. 2011. Diversified ensemble classifiers for highly imbalanced data learning and their application in bioinformatics.

[10] Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

[11] Aurélien Géron. 2017. Hands-on machine learning with scikit-learn and tensorflow: Concepts. *Tools, and Techniques to build intelligent systems*.

[12] Torstein Granskogen and Jon Atle Gulla. 2017. Fake news detection: Network data from social media used to predict fakes. In *CEUR Workshop Proc*, volume 2041, pages 59–66.

[13] Sameer Hinduja and Justin Patchin. 2018. Connecting adolescent suicide to the severity of bullying and cyberbullying. *Journal of School Violence*, 18:1–14.

[14] Mohamed Hosni, Ibtissam Abnane, Ali Idri, Juan M Carrillo de Gea, and José Luis Fernández Alemán. 2019. Reviewing ensemble classification methods in breast cancer. *Computer methods and programs in biomedicine*, 177:89–112.

[15] Gaoyang Liu, Chen Wang, Kai Peng, Haojun Huang, Yutong Li, and Wenqing Cheng. 2019. Socinf: Membership inference attacks on social media health data with machine learning. *IEEE Transactions on Computational Social Systems*, 6(5):907–921.

[16] Jitendra Singh Malik, Guansong Pang, and Anton van den Hengel. 2022. Deep learning for hate speech detection: a comparative study. *arXiv preprint arXiv:2202.09517*.

[17] Ricardo Martins, Marco Gomes, Jose Joao Almeida, Paulo Novais, and Pedro Henriques. 2018. Hate speech classification in social media using emotional analysis. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 61–66. IEEE.

[18] Piero Montebruno, Robert J Bennett, Harry Smith, and Carry Van Lieshout. 2020. Machine learning classification of entrepreneurs in british historical census data. *Information Processing & Management*, 57(3):102210.

[19] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2020. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, pages 928–940. Springer.

[20] Kristiawan Nugroho, Edy Noersasongko, Ahmad Zainul Fanani, Ruri Suko Basuki, et al. 2019. Improving random forest method to detect hatespeech and offensive word. In *2019 International Conference on Information and Communications Technology (ICOIACT)*, pages 514–518. IEEE.

[21] Shovon Paul, Jubair Islam Joy, Shaila Sarker, Sharif Ahmed, Amit Kumar Das, et al. 2019. Fake news detection in social media using blockchain. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, pages 1–5. IEEE.

[22] Andrés Suárez-García, Montserrat Díez-Mediavilla, Diego Granados-López, David González-Peña, and Cristina Alonso-Tristán. 2020. Benchmarking of meteorological indices for sky cloudiness classification. *Solar Energy*, 195:499–513.

[23] Lucia Tamburino, Giangiacomo Bravo, Yann Clough, and Kimberly A Nicholas. 2020. From population to production: 50 years of scientific literature on how to feed the world. *Global Food Security*, 24:100346.

[24] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28.

[25] Anna Vannucci, Kaitlin M Flannery, and Christine McCauley Ohannessian. 2017. Social media use and anxiety in emerging adults. *Journal of affective disorders*, 207:163–166.