2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS) | 979-8-3503-4233-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICTACS59847.2023.10390417

Deep Learning Approach for Hate and Non Hate Speech Detection in Online Social Media

Parshuram Sharma[1], Rakesh Kumar Tiwari[2]
[1]*Dept. of Computer Science Engineering,*
Technocrats Institute of Technology & Science, Bhopal, India
[2]*Dept. of Computer Science Engineering, Technocrats Institute of Technology & Science, Bhopal, India*

***Abstract*—People's ability to freely and anonymously share their thoughts and feelings online on social media platforms is contributing to a rising issue of hate speech. Hate speech has the potential to hurt both people and groups, contribute to the polarisation of society, and even provoke acts of physical violence. Therefore, identifying and removing hate speech from online social media is an important task for the purpose of maintaining an online environment that is healthy and respectful. A method based on deep learning is proposed in this study for differentiating between hate speech and other types of communication that may be found in online social media. The solution that has been developed makes use of a natural language processing (NLP) and long term short memory (LSTM) model that is trained on a huge dataset consisting of tweets that have been annotated. Tweets that include hate speech and tweets that do not contain hate speech were both included in the dataset and were labelled by human annotators. Python software with IDE of Spyder version 3.7 is used to carry out the simulation work. The accuracy level reached overall is 91.14%.**

Keywords: Hate, Twitter, Speech, Social Media, Machine Learning, LSTM, NLP.

I. Introduction

Online hate speech is a type of communication that assaults, threatens, or incites violence or hatred against people or groups on the basis of their race, ethnicity, nationality, religion, gender, sexual orientation, or other traits [1]. This kind of expression may take place anywhere online. Because users are able to quickly and easily share their thoughts and opinions with a wide audience via social media websites like Twitter, Facebook, and Instagram, these websites have become popular channels for the dissemination of hate speech [2].

Concerns have been made concerning the potentially detrimental effects that the spread of hate speech in online social media may have not only on individuals but also on society as a whole as a whole. Hate speech may exacerbate prejudice against marginalised groups, contribute to the polarisation of society, and even encourage acts of violence against such groups. Therefore, it is absolutely necessary to identify instances of hate speech in online social media and to take action against them [3].

Any kind of communication, on the other hand, that does not entail assaulting, threatening, or instigating violence or animosity against people or groups based on their personal traits is referred to as non-hate speech. It is possible for it to include a diverse assortment of material, such as views, facts, humorous anecdotes, and personal accounts [4].

It might be difficult to distinguish between speech that promotes hatred and speech that does not promote hatred via online social media due to the fact that hate speech can be communicated in covert or roundabout ways. It is also able to change and adapt to changing social norms and settings, making it difficult to identify using standard techniques of content control. Moreover, it may develop [5].

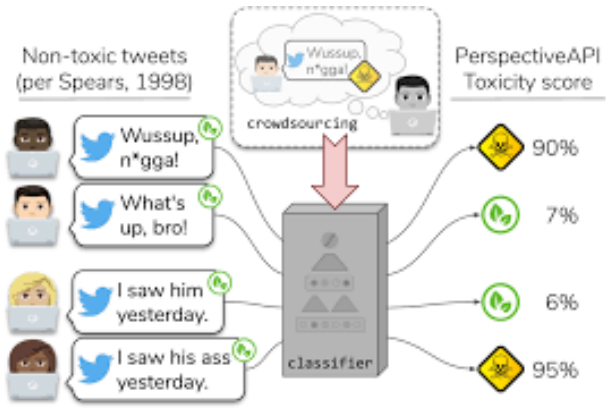

Fig. 1: Hate and Non Hate Speech

Figure 1 presents the picture of hate speech and non-hate speech found in the Twitter dataset together with the percentage rate for each category [6]. To find a solution to this problem, researchers have created models based on machine learning, particularly models based on deep learning, which can identify instances of hate speech in online social media. These models do an analysis of the text of postings made on social media platforms and categorise them as either hate speech or non-hate speech based on the results of the processing of natural language [7].

Annotated datasets of social media posts labeled by human annotators may be used to train hate speech detection algorithms, which can then achieve a high degree of accuracy when recognizing hate speech [8]. These algorithms may be incorporated into social media platforms

to detect and remove hate speech posts automatically, as well as provide users with guidance on whether or not their own posts are appropriate[9].

One approach that may be used to detect hate speech is the use of machine learning algorithms. In particular, a model may be trained using supervised learning on a labeled dataset of hate speech and non-hate speech. To differentiate between the two forms of communication, this may be done. When applied to new content, the model may help assess whether or not it should be labeled as hate speech[10].

A big dataset consisting of annotated social media postings is necessary for training a model of this type. Items that have been manually annotated as hate speech and entries that have been mechanically annotated as not being hate speech must both be included in this collection. The process of labeling may be subjective and difficult to carry out due to the fact that hate speech may be conveyed in a number of various ways and be targeted at a range of different targets [11]. Because hate speech may be directed at many different groups, it is quite pervasive.

After compiling a dataset that is appropriate for machine learning, a number of different machine learning models, including support vector machines (SVMs), logistic regression, and neural networks, may be trained on the data. Specifically, neural networks, which are able to capture complicated patterns in text data, have showed findings that are promising in the identification of hate speech. This is because neural networks have the ability to recognize intricate patterns [12].

For the purpose of identifying instances of hate speech on social media platforms, deep learning models such as neural networks based on convolutional (CNNs) and based on recurrent (RNNs) have been used. In text data, CNNs are able to effectively capture local features, whereas RNNs are able to capture sequential information. It is also possible to employ hybrid models, which combine CNN and RNN, to take use of the benefits that come with using both kinds of models [13].

The evaluation of algorithms that identify hate speech might be difficult since there are no measures or standards that are generally accepted by everyone. Metrics like as accuracy, recall, and F1 score are used rather often, yet it is possible that they do not adequately reflect the richness and subtlety of hate speech identification [14].

The remainder of the article is broken up into four sections: section 2 discusses the context of the study, section 3 delves into the methodology, section 4 delves into the simulation and results, and section 5 discusses the findings of the research as well as the next steps for the investigation.

II. Literature Survey

H. S. Alatawi et al.,[1] presented a new framework based on the encoder representations in bidirectional transformers (BERT). The framework of BiLSTM obtained a score of 0.75 for F1, whereas the BERT model obtained a score of 0.80 for F1. Both models are validated by applying them to datasets obtained from Twitter (the balanced dataset) and Stormfront (the Stormfront dataset), a white supremacist forum.

J. Melton et al. [2] the resolution of problems has emerged as a key priority for major social media sources. The lack of properly labelled data, a developing language and the absence of baseline models for fringe sources like alt-right websites are three of the most significant issues associated with the automated identification and categorization of hateful material.

R. Tiwari et al.,[3] presented regression model based on the random forest and the decision tree for optimization of parameters from the selected dataset

Gab. L. Ketsbaia et al.,[4] Because of the ease with which others may engage in the harassment of victims and the opportunity for victims to remain anonymous, cyberbullying has emerged as a very problematic kind of online abuse. The research investigates several machine learning strategies, including word embeddings.

N. D. Srivastava et al.,[5] tackles the crucial problem of increasing hate speech and insulting remarks directed at people or groups on social media. This kind of activity has become commonplace on social media platforms, where individuals have easy access to expressing their animosity and communicating with a huge crowd—something that they may not even contemplate doing in world.

S. Gupta et al.,[6] The authors of the work believe that there is a lot of potential to enhance the accuracy of the present automated techniques of hate speech detection. Although a lot of manual effort and time is spent to finding a solution to the issue of identifying hate speech from online material, the authors of the study believe that there is a lot of potential to improve the accuracy of these approaches.

K. M. Hana et al.,[7] The findings of this study give a method for categorising hate speech that is posted on Twitter in other terminology. It also manages the messiness of the data that Twitter provides, such as multiple languages and material that does not conform to conventional formats. In addition to using SVM as a classifier, we evaluate its performance in comparison to those of various algorithms and techniques, including deep learning, CNN, and DistilBERT.

S. T. Luu et al.,[8] presented the ability to mitigate the negative impacts of online hate speech, protect vulnerable populations, and promote online civility, the study of hate speech prediction is an important but understudied field of research.

Y. Zhou et al. [9] presented hypothesise that neural networks may be able to perform better in another job involving text categorization, namely the identification of hate speech. However, the deconstruction of ambiguity becomes more challenging when there are fewer hateful keywords present, despite the fact that the fuzzy control

system was able to compensate for this in the majority of instances.

K. Kumaresan et al.,[10] The current system will make use of human reasoning methods like ontologies and fuzzy logics in conjunction with sentiment analysis in order to identify instances of hate speech and deconstruct the ambiguity that is now present. The findings of the current method indicate that the system is capable of performing well when it comes to discriminating between material that is hateful and content that is offensive, and it is also capable of outperforming existing systems in critical variables.

B. R. Amrutha et al., [11] The presence of harmful information online is one of the most obvious issues plaguing today's social media platforms. This has continued uninterrupted as individuals from a wide variety of cultural backgrounds utilise the Internet, where they may remain anonymous while hiding their identities.

H. M. S. T. Sandaruwan et al., [12] presented a methods that use lexicons and machine learning to identify harmful and hateful remarks that are being published in Sinhala on social media platforms automatically. The lexicon-based method was started in our research with the lexicon-generating process, and the results showed that the lexicon had accuracy of 77.4% for detecting hate-speech, offensive-speech, and neutral-speech.

L. Jiang et al.,[13] The most important contribution that we made with this work was to examine various data-to-method ratios concurrently using a variety of different approaches. Because of this, high performance may be attained via the use of machine learning even when there is a limited amount of data. When conducting experiments, it is necessary to use more data in order to obtain the good results that can be obtained through the use of deep learning. When compared to the other methods that we used, BiRNN has the potential to produce the best results. Even if this approach is better than previous models, we still need to think about the kinds of data sets that could be used in the years to come.

C. Abderrouaf et al.,[14] The current strategy recommends using a method that is similar to classification in conjunction with a specialised data design process. The latter method explores the negativity that was present in the initial dataset in order to ensure a balanced training plan. This results in the creation of new paradigms for transfer learning, A comparison is made between two classification strategies, one utilising convolution neural networks and the other using LSTN architecture, both of which employ FastText embeddings as input features. These two classification strategies are compared with baseline models, which are composed of logistic regression and Naive Bayes classifiers.

M. Sajjad et al.,[15] This piece of work presents a method for classifying tweets into one of three categories: racism, sexism, or none of the above. In our classification technique, we combine the most cutting-edge syntactic and word n-gram features with the deep features that were recovered from a CNN that had been trained on semantic word embedding. We conduct in-depth tests on a conventional dataset that is comprised of 16,000 tweets that have been carefully annotated. Our current method achieves far higher levels of accuracy than any other state-of-the-art method, making it the most effective method available.

III. Proposed Methodology

The flow chart of the proposed method is followings-

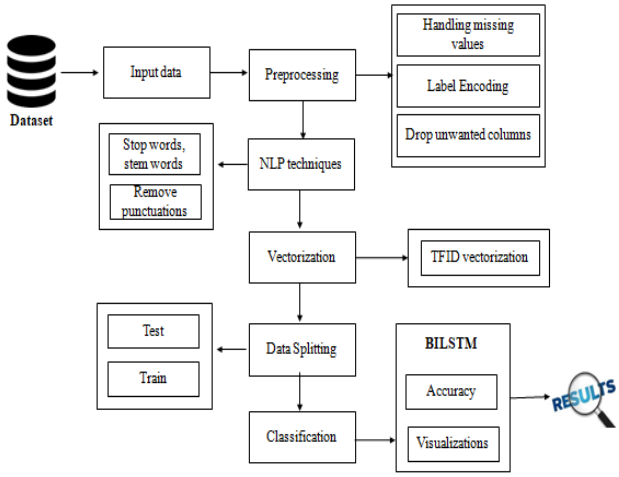

Fig. 2: Method Flow Chart

Step-

- To begin, go to the website of kaggle, which is a big research organisation that serves as a source of datasets, and download the hate speech dataset.
- We are now doing the preparation of the data, and here we are passing over the dataset that was lacking. Either get rid of the null value or replace it with a standard 1 or 0 value.
- At this point, use the natural language processing method to determine the language and eliminate any stop words, punctuation words, or other similar terms.
- At this point, the dataset is partitioned into the information used for training and testing purpose
- Utilise the categorization approach that is derived from deep learning and is known as long term short memory (LSTM).
- Now verify the performance characteristics and compute them in terms of the accuracy, F-measure, precision, recall and error_rate.
- Finally, the results of the simulation demonstrate that performance measures including accuracy, precision, recall, and confusion matrix value will be enhanced in the future.

The combination of LSTM and Natural Language Processing (NLP) approaches is an effective classification method for determining whether or not speech contains hate

speech. The text data may be preprocessed and transformed with the use of NLP methods, which makes it simpler for the LSTM model to recognise patterns in the text data and determine whether or not a post is hate speech or another kind of speech.

The NLP preprocessing methods include text normalisation, which is transforming text to a format that is standardised, eliminating stop words, which are frequent words that do not convey much significance, and stemming, which entails reducing words to their root form. Other approaches include removing stop words and removing common terms.

After the text data have been preprocessed, the LSTM model can be trained with the transformed text as the input after further processing. The model is taught to recognise patterns in the text data that are suggestive of hate speech, such as the use of language that is disparaging or the instigation to violence. For example, the model will learn to recognise patterns like these.

It is possible to extract additional features from the text data in order to better enhance the overall performance of the LSTM model. The usage of emojis, capitalization, punctuation, and several other style characteristics that may be suggestive of hate speech might be considered to be among these qualities.

It has been shown that the LSTM with NLP classification approach may obtain excellent accuracy and F1 scores on a variety of datasets for the purpose of detecting hate speech. This method may be included into social media platforms in order to automatically identify and delete postings containing hate speech. As a result, the damage caused by hate speech may be mitigated, and an online atmosphere that is more courteous and welcoming may be fostered.

There is a possibility that the efficiency of the LSTM with NLP classification approach would change based on the dataset as well as the environment in which the posts were published. As a result, the method has to be continuously improved upon and evaluated in order to guarantee that it will continue to be efficient in identifying instances of hate speech.

IV. Simulation Results

The suggested computation is carried out using software of python with IDE of Spyder version 3.7.

The dataset is shown in the Python environment in Figure 3, which shows the environment. The dataset contains rows and columns with a variety of different integers. Additionally, the name of the signal characteristic is stated.

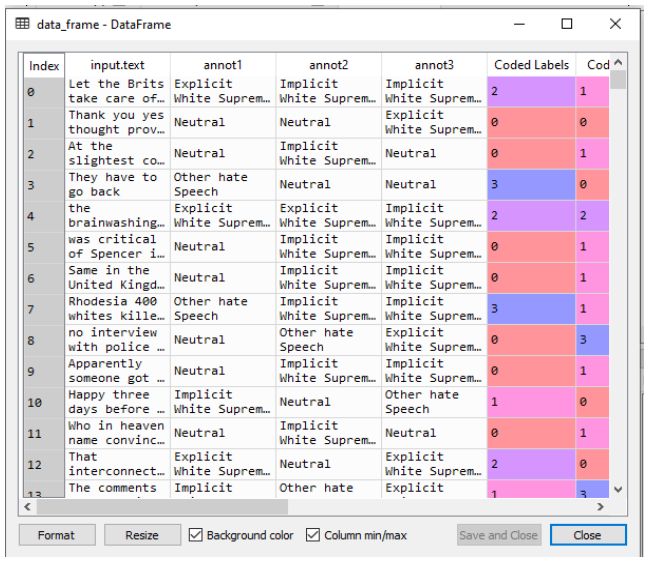

Fig. 3: Dataset Frame

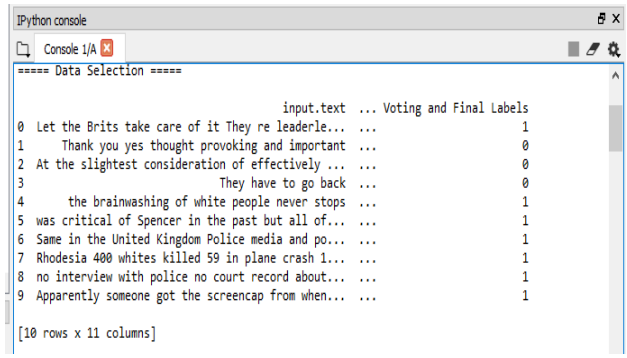

Fig. 4: Data Selection in Simulation

Figure 4 presents the simulation scenario in the spyder console, the data selection process is initializing.

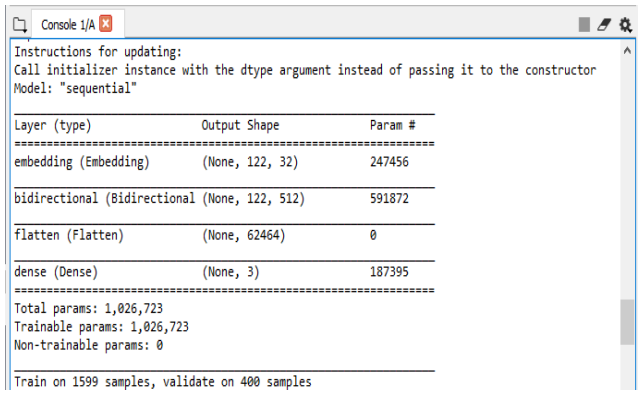

Fig. 5: Sample Selection

Figure 5 is showing the sample selection after the padding, total trainable parameters is 1026723.

Figure 6 is showing the iteration process after apply the proposed LSTM method. The time taken by each iteration is epoch.

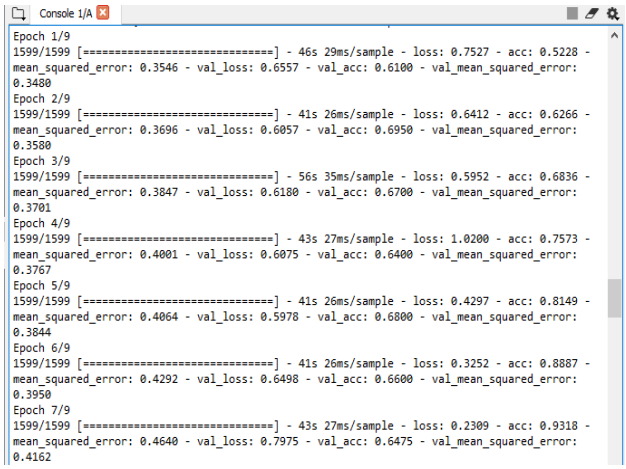

Fig. 6: Iteration

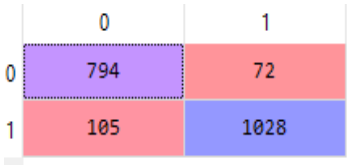

Fig. 7: Confusion Matrix

Figure 7 shows prediction or the confusion matrix of the proposed simulation process. It tells about the various predictions.

True_Positive(TP)=794
False_Positive(FP)=72
False_Negative(FN)=105
True_Nagative(TN)= 1028

Table 1: Result_Comparison

Sr No	Name of Parameters	Previous_work	Proposed_Work
1	Method Name	BERT-BiLSTM	NLP-LSTM
2	Hate Accuracy	82%	91.14%
3	Non Hate Accuracy	86%	91.68%
4	All Accuracy	84%	91.14%
5	F-measure	84%	89.97%

V. Conclusion

This research presents the use of a deep learning strategy to classify between hate_speech and other types of speech extracted from the Twitter dataset. We have used the Natural Language processing methods as well, the Classification Algorithms (also known as Deep Learning Algorithms). After that, other deep learning algorithms, such as LSTM, are suggested for use with NLP. The suggested method achieves an accuracy of 91.14%, which is an significant increase above the prior method's accuracy of 82%. The accuracy of the non-hate classification is now 91.68%, up from 86% before, and the overall accuracy optimisation is now 91.14%, up from 84% previously. As a consequence, the findings of the simulation demonstrate the accuracy of the previously described algorithm and provide an estimate of its performance metrics, which include accuracy for the whole task, hate accuracy, and non-hate accuracy.

References

[1] H. S. Alatawi, A. M. Alhothali and K. M. Moria, "Detecting White Supremacist Hate Speech Using Domain Specific Word Embedding With Deep Learning and BERT," in IEEE Access, vol. 9, pp. 106363-106374, 2021, doi: 10.1109/ACCESS.2021.3100435.

[2] J. Melton, A. Bagavathi and S. Krishnan, "DeL-haTE: A Deep Learning Tunable Ensemble for Hate Speech Detection," 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), 2020, pp. 1015-1022, doi: 10.1109/ICMLA51294.2020.00165.

[3] R. Tiwari, R. Sharma, and R. Dubey, "Microstrip Patch Antenna Parameter Optimization Prediction Model using Machine Learning Techniques ", IJRITCC, vol. 10, no. 9, pp. 53–59, Sep. 2022. doi. org/10.17762/ijritcc.v10i9.5691.

[4] L. Ketsbaia, B. Issac and X. Chen, "Detection of Hate Tweets using Machine Learning and Deep Learning," 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2020, pp. 751-758, doi: 10.1109/TrustCom50675.2020.00103.

[5] N. D. Srivastava, Sakshi and Y. Sharma, "Combating Online Hate: A Comparative Study on Identification of Hate Speech and Offensive Content in Social Media Text," 2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS), 2020, pp. 47-52, doi: 10.1109/RAICS51191.2020.9332469.

[6] S. Gupta, S. Lakra and M. Kaur, "Study on BERT Model for Hate Speech Detection," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1-8, doi: 10.1109/ICECA49313.2020.9297560.

[7] K. M. Hana, Adiwijaya, S. A. Faraby and A. Bramantoro, "Multi-label Classification of Indonesian Hate Speech on Twitter Using Support Vector Machines," 2020 International Conference on Data Science and Its Applications (ICoDSA), 2020, pp. 1-7, doi: 10.1109/ICoDSA50139.2020.9212992.

[8] S. T. Luu, H. P. Nguyen, K. Van Nguyen and N. Luu-Thuy Nguyen, "Comparison Between Traditional Machine Learning Models And Neural Network Models For Vietnamese Hate Speech Detection," 2020 RIVF International Conference on Computing and Communication Technologies (RIVF), 2020, pp. 1-6, doi: 10.1109/RIVF48685.2020.9140745.

[9] Y. Zhou, Y. Yang, H. Liu, X. Liu and N. Savage, "Deep Learning Based Fusion Approach for Hate Speech Detection," in IEEE Access, vol. 8, pp. 128923-128929, 2020, doi: 10.1109/ACCESS.2020.3009244.

[10] K. Kumaresan and K. Vidanage, "HateSense: Tackling Ambiguity in Hate Speech Detection," 2019 National Information Technology Conference (NITC), 2019, pp. 20-26, doi: 10.1109/NITC48475.2019.9114528.

[11] B. R. Amrutha and K. R. Bindu, "Detecting Hate Speech in Tweets Using Different Deep Neural Network Architectures," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 923-926, doi: 10.1109/ICCS45141.2019.9065763.

[12] H. M. S. T. Sandaruwan, S. A. S. Lorensuhewa and M. A. L. Kalyani, "Sinhala Hate Speech Detection in Social Media using Text Mining and Machine learning," 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer), 2019, pp. 1-8, doi: 10.1109/ICTer48817.2019.9023655.

[13] L. Jiang and Y. Suzuki, "Detecting hate speech from tweets for sentiment analysis," 2019 6th International Conference on Systems and Informatics (ICSAI), 2019, pp. 671-676, doi: 10.1109/ICSAI48974.2019.9010578.

[14] C. Abderrouaf and M. Oussalah, "On Online Hate Speech Detection. Effects of Negated Data Construction," 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 5595-5602, doi: 10.1109/BigData47090.2019.9006336.

[15] M. Sajjad, F. Zulifqar, M. U. G. Khan and M. Azeem, "Hate Speech Detection using Fusion Approach," 2019 International Conference on Applied and Engineering Mathematics (ICAEM), 2019, pp. 251-255, doi: 10.1109/ICAEM.2019.8853762.