

Analysis and Detection of Hate Speech in Social Media using Deep Neural Networks

Author: Mrs.A.Barveen M.E.,Ph.D,

Co-Authors: E.Ramanan¹, S.Shamsheer Ahamed², N.Suresh³

Department of Computer Science and Engineering,

Students of M.I.E.T. Engineering College, Trichy.







Abstract—In recent years, the use of hate speech increased rapidly in social media. Hate speech is a crime because it may cause violence in social media and also in real world too, so we are developing a deep learning model to detect the hate speech in social media. Hate speech detection is the process of automatically identifying and flagging speech that is deemed to be hateful or harmful towards a particular person or a group. It is a challenging task because it often involves highly contextualized language and can vary widely depending on the cultural and social norms of a particular community. We detect the hate speech in social media using deep neural network models such as Multi-Layer Perceptron (MLP) and Long Short – Term Memory (LSTM). The results shown in terms of Accuracy Metrics such as Precision and Recall scores.

Keywords—Hate Speech, Offensive Speech, Deep Neural Network, MLP, LSTM.

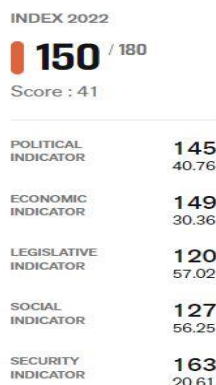
I. INTRODUCTION

Hate speech detection is the process of automatically identifying and flagging speech that is deemed to be hateful or harmful towards a particular individual or group. Hate speech can be divided into different categories as colour, disability, religion, sex orientation and soon on. This type of detection is often used in social media platforms, online communities, and other digital environments to help prevent harassment, discrimination, and other forms of harmful behaviour. Hate speech detection can be challenging because it often involves highly contextualized language and can vary widely depending on the cultural and social norms of a particular community. Additionally, there is often a fine line between freedom of speech and hate speech, which can make it difficult to develop automated detection systems that accurately identify hateful language without infringing on users' rights to express their opinions.

In 2022, india records the hate speech ranking has fallen to 150 out of 180 countries, according to Reporters Without Borders (RSF), where in last year's report, India was ranked 142 out of 180 countries. Hate speech is a crime and it has some ranges according to the content of speech. Hate speech not only in the form of text, also hate speech can in photo i.e, meme with hate meanings and harmful or hateful photos, aslo in audio format that some audio related speech posting social media. Hate speech is a big problem beacuae it may cause a critical problem in real world, that a person post a hate speech in social media that is related to a poiltical party which leads a problem between two poiltical party and also make a public disturbance. Though solve the problem regarding to hate speech, we developed a hate speech detection model using Deep Neural Network models.

Color	Title	Description	Examples
	6. Death	Rhetoric includes literal killing by group. Responses include the literal death/elimination of a group.	Killed, annihilate, destroy
	5. Violence	Rhetoric includes infliction of physical harm or metaphorical/aspirational physical harm or death. Responses include calls for literal violence or metaphorical/aspirational physical harm or death.	Punched, raped, starved, torturing, mugging
	4. Demonizing and Dehumanizing	Rhetoric includes subhuman and superhuman characteristics. There are no responses for #4.	Rat, monkey, Nazi, demon, cancer, monster
	3. Negative Character	Rhetoric includes nonviolent characterizations and insults. There are no responses for #3.	Stupid, thief, aggressor, fake, crazy
	2. Negative Actions	Rhetoric includes negative nonviolent actions associated with the group. Responses include nonviolent actions including metaphors.	Threatened, stole, outrageous act, poor treatment, alienate
	1. Disagreement	Rhetoric includes disagreeing at the idea/belief level. Responses include challenging claims, ideas, beliefs, or trying to change their view.	False, incorrect, wrong, challenge, persuade, change minds

Hate speech categorization



Hate speech Ranking in 2022 and 2021

II. LITERATURE SURVEY

They developed an automated system using the Deep Convolutional Neural Network (DCNN). The proposed DCNN model utilizes the tweet text with GloVe embedding vector to capture the tweets' semantics with the help of convolution operation and achieved the precision, recall and F1-score value as 0.97, 0.88, 0.92 respectively for the best case and outperformed the existing models. Initially, the machine learning based classifiers such as LR, RF, NB, SVM, DT, GB, and KNN were used to identify the HS related tweets on Twitter with the features extracted using *tf-idf* technique. [1]

However, the best ML model, i.e., SVM able to predict only 53% of HS tweets correctly on a 3:1 train-test dataset. The reason behind the low prediction of HS tweets may include the imbalanced dataset, hence the model biased towards the NHS tweets prediction as it is having most of instances. Deep learning-based CNN, LSTM, and their combinations C-LSTM models also have similar results with the fixed partitioned dataset. [7]

The experimental outcome on both the traditional machine learning based models and deep learning-based models confirmed that none of the models predicted the HS tweets with satisfactory accuracy on a fixed partitioned of train-test. Finally, 10-fold cross-validation was used with the proposed DCNN model and achieved the best prediction recall value of 0.88 for HS and 0.99 for NHS. The experimental results confirmed the k-fold cross-validation technique is a better choice with the imbalanced dataset.[9]

They presented a largescale empirical comparison of 14 shallow and deep models for hate-speech detection on three most commonly used datasets. In case of Detection accuracy, the combination of BERT, ELECTRA and AI-BERT and neural network-based classifiers perform consistently better than the other methods on the three benchmarks, especially in macro F1 score. [2]

In terms of computational cost, transformers-enabled classifiers are a lot more costly than the other models. AI-BERT is the most time-consuming model among the transformers, followed by the BERT model. Small BERT is the most efficient transformer model in our task. Deep classifiers like Bi-LSTM are also computationally costly. When considering both accuracy and efficiency, Electra-based MLP models seem to be the most practical method, which achieves among the best classification accuracy while being sufficiently computationally efficient.[8]

As transformers are pre-trained on significantly larger corpora than Glove, they learn embedding space with richer semantics, empowering significantly better detection performance. This can be observed by the performance difference of having the same MLP-based classifier trained upon respective TF-IDF, Glove, and transformers-based representations. However, it does not mean that the larger the pre-trained models are, the better performance we would obtain. For example, the ELECTRA transformer can normally perform better than, or on par with, the larger transformer BERT on a number of cases.[10]

Although the transformers-based hate speech detectors show promising performance, they are still weak in terms of macro F1 performance indicating the great difficulty in achieving high precision and recall rates of identifying the minority hatred tweets from the massive tweets. Another major challenge lies at domain generalization ability. It is difficult to collect a dataset that well covers all possible properties of hate speech in social media, which may lead to a domain difference between the source data that the detectors are trained on and the target data that the detectors are applied to. Thus, it is important for the detectors to have good domain generalization ability in practice. However, as

shown in their results, there is still a large gap between the same-domain performance and the cross-domain performance.[11]

They presented the principle of three types of text classification methods, ELMo, BERT and CNN, and applied them to hate speech detection, then improved the performance by fusion from two perspectives. The basic results of these three methods are fused through the combination rules by voting, meaning, maximizing and production. These performances are better than the performances of the original three methods, and the performance of the mean fusion method is the best.[3]

In that test, the CNN text classification performs best in the original three methods. Thus, three classifiers based on CNN are adopted in the next test. It can be concluded that whether by designing several different classifiers or using the same type of classifier with different parameters, the fusion of different classifier results can improve the classification accuracy and F1-score. And this approach improves the results with little additional cost. The results of the fusion of ELMo, BERT and CNN, and the fusion of three CNN classifiers with different parameters, showed that fusion processing is a viable way to improve the performance of hate speech detection.[13]

They proposed research that investigates the feasibility of automatically detecting white supremacist hate speech on Twitter using deep learning and natural language processing techniques. Two deep learning models are investigated in this research. The first approach utilizes a bidirectional Long Short-Term Memory (BiLSTM) model along with domain-specific word embeddings extracted from white supremacist corpus to capture the semantic of white supremacist slangs and coded words. The second approach utilizes one of the most recent language models, which is Bidirectional Encoder Representations from Transformers (BERT). The BiLSTM model achieved 0.75 F1-score and BERT reached a 0.80 F1-score. Both models are tested on a balanced dataset combined from Twitter and a Stormfront dataset compiled from white supremacist forum.[4]

The results of domain specific and agnostic word embedding with deep learning (BiLSTM) performs well for the problem of white supremacist hate speech. BERT model has also proved that it provides the state of the art for this problem. The experiment results show that BERT outperforms domain specific approach with 4 points, however, domain specific approach is able to detect intentionally misspellings and common slang from hate community while BERT model fails to detect as it is trained on Wikipedia and books.[12]

They developed a context-aware Roman Urdu Hate Speech detection model based on Bi-LSTM with an attention layer and used custom word2vec for word embeddings. First, they developed annotation guidelines for Roman Urdu Hate Speech. Second, we constructed a new Roman Urdu Hate Speech Dataset (RU-HSD-30K) that was annotated by a team of experts using the annotation rules. Different traditional as well as deep learning models, including LSTM and CNN

models, were used as baseline models. The performance of the models was assessed in terms of evaluation metrics like accuracy, precision, recall, and F1-score. The generalization of each model is also evaluated on a cross-domain dataset.[5]

Experimental results revealed that Bi-LSTM with attention outperformed the traditional machine learning models and other deep learning models with an accuracy score of 0.875 and an F-Score of 0.885. The results demonstrated that their suggested model (Bi-LSTM with Attention Layer) is more general than previous models when applied to unseen data. The results confirmed that lexical normalization of Roman Urdu words enhanced the performance of the suggested model.[14]

The current study proposes an approach to build a political hate speech lexicon and train artificial intelligence classifiers to detect hate speech. Their academic and practical contributions include the collection of a Chinese hate speech dataset, creating a Chinese hate speech lexicon, and developing both a deep learning-based and a lexicon-based approach to detect Chinese hate speech.[6]

They used the extended hate speech dataset to conduct a hate speech detection model based on the BERT deep learning model. Finally, they compared the performance of the BERT model and the lexicon-based approach. The results showed that the lexicon-based hate speech detection model yielded a precision of 55.4%, while the precision of the BERT model was 69.7%. However, the BERT model can obtain a better detection performance than the lexicon approach. Thus, the BERT deep learning model has the potential to detect hate speech.[16]

The current collected dataset focuses on comments to political news. Political hate speech is only one type of hate speech source. There exists a variety of hate speech types, such as hate speech focusing on race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. Future studies may use the approach developed by this study to collect hate speech datasets for different types of hate speech. However, in the current study, the lexicon contained only 153 terms; thus, the size of the hate speech lexicon is still limited. Future studies may attempt to extend the hate speech lexicon. In the current study, they did not consider the degree of hate for the hate speech terms. They only classified the 153 terms like hate speech terms. Future studies can determine the degree of hate and divide the hate speech terms into several intensity levels.[15]

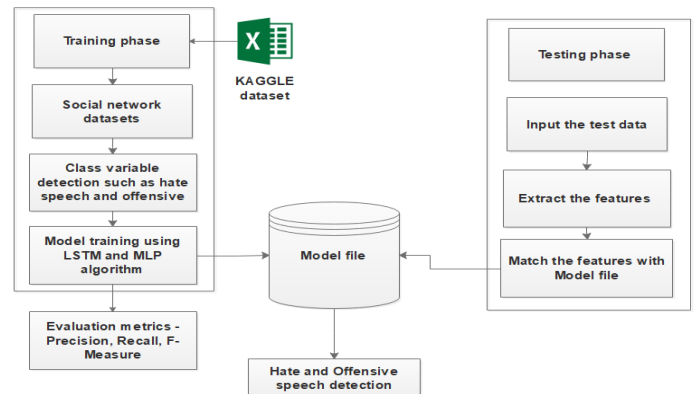
III. PROPOSED APPROACH

In proposed system we can implement multi-layer perceptron and Long Short-Term memory algorithm to classify the hate speech and provide comparative analysis in terms of accuracy parameter. Hate speech detection using a multi-layer perceptron (MLP) is a popular approach that uses a feedforward neural network with multiple layers to classify text as either hate speech or non-hate speech. The MLP approach involves feeding the input text into the neural network, which is then processed through a series of layers,

with each layer applying a set of weights and biases to the input data. The final layer outputs a probability score that indicates the likelihood that the input text is hate speech. Long Short-Term Memory (LSTM) is a classifier using recurrent neural network (RNN) approach. LSTM classifier classify the text as hate speech or non-hate speech. LSTM networks are designed to capture the temporal dependencies within sequential data, such as text, and are able to maintain long-term memory over many time steps, making them well-suited for natural language processing tasks.

MLP for hate speech detection is that it is a relatively simple and interpretable approach. It has a straightforward architecture and can be trained using a variety of algorithms, making them accessible to researchers and practitioners with varying levels of expertise. LSTM for hate speech detection is that they are able to capture the context of the text and take into account the sequence of words in the input text, allowing them to better identify patterns and features associated with hate speech. Additionally, LSTM networks can handle variable-length inputs, making them well-suited for natural language processing tasks where text can vary in length.

IV. ARCHITECTURE DIAGRAM



V. MODELS

- Datasets Acquisition
- Pre-processing
- Features Extraction
- Model Building
- Deployment the Model

A. Datasets Acquisition:

We upload the datasets related to hate speech that are collected from KAGGLE web sources which includes offensive, hate speech, and content details.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1																							
2																							
3																							
4																							
5																							
6																							
7																							
8																							
9																							
10																							
11																							
12																							
13																							
14																							
15																							
16																							
17																							
18																							
19																							
20																							
21																							
22																							
23																							
24																							
25																							
26																							
27																							
28																							
29																							
30																							

B. Pre-processing:

Data pre-processing is an important step in the process (data mining). If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. We eliminate the irrelevant values (like Symbols - @, #, \$, !... Numbers - 1, 2, 3, 4.... and some links in tweets) and also estimate the missing values (that some tweets contain nothing just a blank tweet) of data. Finally provide structured datasets.

C. Features Extraction:

Feature selection refers to the process of reducing the inputs for processing and analysis, or of finding the most meaningful inputs. Features are hating speech, offensive speech and normal speech in the dataset. Scoring to each feature is assigned by filter feature selection method which provides a statistical score. select the class variable and targeted variables from the datasets.

The features are ranked as 0,1,2 which is hate, offensive and normal speech and empty tweets are removed from dataset. Univariate - consider the feature independently, or with regard to the dependent variable. It can be used to construct the hate speech and offensive speech.

D. Model Building:

1) MLP Model:

Construct a Multilayer Perceptron (MLP) model for hate speech detection by performing the following steps:

- Define the input layer of the MLP with the appropriate input shape.
- Add one or more hidden layers with activation functions such as Rectified Linear Activation (ReLU), Logistic (Sigmoid).
- Add a final output layer with a softmax activation function for multi-class classification.
- Compile the model with an appropriate loss function, optimizer, and evaluation metric.
- Train the model on the preprocessed data and evaluate its performance on a validation set.

2) LSTM Model:

Construct a Long Short-Term Memory (LSTM) model for hate speech detection by performing the following steps:

- Define the input layer of the LSTM with the appropriate input shape.
- Add one or more LSTM layers with appropriate parameters such as number of units and dropout rate.
- Add a final output layer with a softmax activation function for multi-class classification.
- Compile the model with an appropriate loss function, optimizer, and evaluation metric.
- Train the model on the preprocessed data and evaluate its performance on a validation set.

E. Model Deployment:

Once the model trained, user can input the text to classify whether it is offensive, hate or normal text. Text can be classified with classifier which is provide high efficiency in model training. Finally, the provide label for detected text and the classification report for both MLP and LSTM algorithm.

VI. ACCURACY MATRICES

A. Precision:

- 1) Precision calculates the accuracy for minority(1).
- 2) Precision is the number of correct positive predictions made.precision will be 0 to 1.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (1)$$

B. Recall:

- 1) Recall provides an indictaion of missed postivies predictions(2). Recall will be 0 to 1.
- 2) Recall – Minimizing false negatives.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

VII. CONCLUSION AND FUTURE WORK

This research addresses the issue of hate speech detection and analysis on social media using a deep neural network such as MLP and LSTM. Comparison of two models have been done using the accuracy metrics in order to detect the speech with high performance model. The tweets can be detected accordingly as hate speech or offensive speech or normal text. The analysis of tweets shows the count of hate speech, offensive speech and normal speech in the dataset as graphical representation as pie chart.

This research detects the hate speech, only in text format that we will add some extra features to detect the hate speech in audio, memes and video. A feature that while a person posting a tweet in twitter displaying a alert message with "Hate Speech" or "Offensive Speech" then posting it, that the person posting a hate speech with his knowledge that it is a hate speech.

References

- [1] Pradeep Kumar Roy, Asis Kumar Tripathy, Tapan Kumar Das, Xiao-Zhi Gao, "A Framework For Hate Speech Detection Using Deep Convolutional Neural Network" 20-November-2022.
- [2] Jitendra Singh Malik, Guansong Pang, Anton Van Den Hengel "A Comparative Study On Deep Learning For Hate Speech Detection" 22-February-2022.
- [3] Yanling Zhou, Yanyan Yang, Han Liu, Xiufeng Liu, Nick Savage "Deep Learning Based Fusion Approach For Hate Speech Detection" 23-July-2020.
- [4] Hind S. Alatawi, Areej M. Alhothali, Kawthar M. Moria "Detecting White Supremacist Hate Speech Using Domain Specific Word Embedding With Deep Learning And Bert" 04-August-2021.
- [5] Muhammad Bilal, Atif Khan, Salman Jan, Shahrulniza Musa "Context-Aware Deep Learning Model For Detection Of Roman Urdu Hate Speech On Social Media Platform" 22-November-2022.
- [6] Chih-Chien Wang, Min-Yuh Day, Chun-Lian Wu "Political Hate Speech Detection And Lexicon Building: A Study In Taiwan" 29-April-2022.
- [7] A. Kamal and M. Abulaish, "An LSTM-based deep learning approach for detecting self-deprecating sarcasm in textual data," in Proc. 16th Int. Conf. Natural Lang. Process. (ICON), Hyderabad, India, 2019, pp. 201–210.
- [8] A. Kamaal and M. Abulaish, "CAT-BiGRU: Convolution and attention with bi-directional gated recurrent unit for self-deprecating sarcasm detection," *Cognit. Comput.*, vol. 14, pp. 91–109, Jan. 2021.
- [9] A. Kamal and M. Abulaish, "Self-deprecating humor detection: A machine learning approach," in Proc. 16th Int. Conf. Pacific Assoc. Comput. Linguistics (PACLING), Hanoi, Vietnam. Springer, 2019, pp. 483–494.
- [10] P. K. Jain, R. Pamula, and E. A. Yekun, "A multi-label ensemble predicting model to service recommendation from social media contents," *J. Super comput.*, vol. 66, no. 1, pp. 1–20, Sep. 2021.
- [11] P. K. Jain, E. A. Yekun, R. Pamula, and G. Srivastava, "Consumer recommendation prediction in online reviews using cuckoo optimized machine learning models," *Comput. Electr. Eng.*, vol. 95, no. 10, pp. 1–10, 2021.
- [12] M. Abulaish, N. Kumari, M. Fazil, and B. Singh, "A graph-theoretic embedding-based approach for rumor detection in Twitter," in Proc. WI, Thessaloniki, Greece, 2019, pp. 466–470.
- [13] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in Proc. 26th Int. Conf. World Wide Web, Perth, WA, Australia, Apr. 2017, pp. 1391–1399.
- [14] A. R. Gover, S. B. Harper, and L. Langton, "Anti-Asian hate crime during the COVID-19 pandemic: Exploring the reproduction of inequality," *Amer. J. Criminal Justice*, vol. 45, no. 7, pp. 647–667, 2020.
- [15] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on Twitter using a convolution-GRU based deep neural network," in Proc. Eur. Semantic Web Conf. Heraklion, Greece. Cham, Switzerland: Springer, 2018, pp. 745–760.
- [16] A.-M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, "A unified deep learning architecture for abuse detection," in Proc. 11th Int. Conf. Web Sci., Boston, MA, USA, 2019, pp. 105–114.