Advancing Hate Speech Detection

Updesh kumar

SRM University, Amaravati, AP, India
updesh_kumar@srmap.edu.in

Abstract
This research report is dedicated to the precise and thorough examination of hate speech within the context of textual content, excluding audio and video mediums. In an era where written communication prevails across various digital platforms, the identification and management of hate speech in text have become paramount. This study adopts state-of-the-art deep learning techniques, specifically bidirectional long short-term memory networks with additional custom layers, to develop a robust hate speech detection model.

The research process entails the collection and curation of a representative Twitter dataset, carefully annotated to distinguish between hate speech and non-hate speech instances. Subsequently, advanced natural language processing (NLP) techniques and bidirectional LSTM-SNP-based model is employed to extract meaningful features from the textual data, aiming to discern patterns indicative of hate speech.

The outcomes of this study offer a substantial contribution to the development of effective tools for combating hate speech in digital environments, specifically within written communication channels. By automating the identification and flagging of hateful content, this research endeavours to foster responsible digital citizenship and safeguard individuals from the harmful consequences of hate speech. The insights gained from this report hold significant promise for enhancing content moderation strategies and promoting a more inclusive and secure online sphere, particularly in the realm of text-based communication.

Keywords: Text-based communication, hate speech detection, bidirectional LSTM- SNP

1 Introduction

In contemporary society, effective communication is predominantly reliant on language, whether spoken or written. While everyday interactions commonly occur verbally, formal and authentic communication often necessitates written documentation. Societal order is governed by established rules and regulations, often enshrined in constitutions, that safeguard fundamental rights and empower individuals to address violations.

In the current landscape, the ease with which individuals can express their thoughts has led to an influx of uncensored content, potentially fueling conflict and societal unrest. Consequently, there is a critical need for a system that ensures the appropriateness

of speech and identifies individuals who may warrant intervention due to offensive or violence-provoking expressions. Analysing and categorising public speeches has become imperative for maintaining stability and fostering peace.

This research centres on the crucial task of hate speech detection, wherein the objective is to determine whether various forms of communication, including text and audio, harbour sentiments of hatred or incite violence against individuals or groups. Such prejudiced expressions often target 'protected qualities,' such as age, gender, sexual orientation, or race. The overarching goal is to contribute to societal harmony by developing robust mechanisms for identifying and mitigating hate speech.

1.1 Background

Over the years, research has suggested a number of methods for detecting hate speech; most recently, deep learning methods [1][2] have been used. It is commonly established that deep learning classification models function and generalise successfully when trained on a vast, varied, and superior data set [3][7]. Due to the lack of such a resource, earlier comparable efforts created datasets that were manually annotated and have tight size limitations.

This section outlines a number of well-known publicly labelled datasets that have been utilised by scholars to test and train hate speech detectors. To train the various hate classification algorithms featured in our trials, we take advantage of a large-scale corpus of created hate and non-hate sequences that extend these datasets.

This section also provides a general review of the methods currently used for detecting hate speech, along with a thorough explanation of the techniques we use to classify hate speech in our work.

We then employ this methodology to propose efficient and compact character-based hate detectors, and we analyse relevant research on character-based convolution neural networks.

1.2 Motivation for Advanced Machine Learning Techniques:

To get over the shortcomings of traditional methods, researchers are using state-of-the-art machine learning algorithms to identify hate speech. Massive volumes of social media data can be scanned by machine learning algorithms to find trends that might point to hate speech.

To ensure that the data is appropriately tokenized, encoded, and formatted to meet the requirements of LSTM, a good text representation takes into account the text's key features and semantic meaning. The model can learn contextual word and sentence representations as a result, making it easier to detect hate speech. The models can learn from this information more effectively and produce accurate predictions if the text data is meaningfully encoded, formatted, and represented. The incentives for using machine learning, such as the requirement for scalability, effectiveness, and objective detection, are examined in this subtopic. It highlights the potential of cutting-edge tools to gather

semantic understanding and contextual information, which are essential for correctly recognizing hate speech detection.

1.3 Bidirectional LSTM

The choice of Bi-LSTM stems from its ability to understand the connections within sequences by examining data in both forward and backward directions. In simpler terms, it looks at the words that come before and after a particular word in a sentence to better grasp its meaning.

The Bi-LSTM architecture involves two unidirectional LSTMs that process the sequence in both directions. Imagine having two separate LSTM networks: one reads the sequence of tokens as is, while the other reads it in reverse order. Each of these LSTM networks provides a probability vector as output, and the final result is a combination of these two sets of probabilities.

1.4 Bidirectional LSTM-SNP model

The Bi-LSTM-SNP model is a bidirectional LSTM-SNP model that captures two-way semantic dependence more effectively by taking into account both forward and backward information. Bi-LSTM-SNP is able to concatenate the hidden states of the two LSTM-SNPs as the representation of each position while simultaneously utilising forward and backward channels. The LSTM-SNPs for the forward and backward directions are expressed as follows:

$$\vec{u}_t, \vec{h}_t = \text{LSTM-SNP}\left(\vec{u}_{t-1}, \vec{h}_{t-1}, W\right)$$
$$\overleftarrow{u}_t, \overleftarrow{h}_t = \text{LSTM-SNP}\left(\overleftarrow{u}_{t+1}, \overleftarrow{h}_{t+1}, W\right)$$

To extract contextual semantic information about aspect words, Bi-LSTM-SNP is employed.

Regarding aspect word, the front and back sentiment features can be obtained by the Bi-LSTM-SNP. The three gates of Bi-LSTM-SNP are controlled by the hidden states, allowing for the efficient extraction of contextual semantic information related to aspect words.

The forward and backward LSTM-SNPs in the following Bi-LSTM-SNP figure. The forward LSTM-SNP's hidden states are

$$\left(\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_i\right)$$

, and the hidden states for backward LSTM-SNP are

$$\left(\overleftarrow{h}_n, \overleftarrow{h}_{n-1}, \ldots, \overleftarrow{h}_i\right)$$

The ultimate representation will be a concatenation of the hidden states of the forward and backward LSTM-SNP.

1.5 Advancing Hate Speech Detection with Bi-LSTM Networks and Custom Neural Processing

In this research report, I introduce a novel approach to hate speech detection through the integration of a Bidirectional Long Short-Term Memory (Bi-LSTM) model, enhanced by a custom SimpleNeuralProcessor (SNP) layer. Placed strategically after the Bidirectional LSTM layer, the SNP layer introduces a tailored non-linear processing element, enriching the model's capacity to discern intricate patterns and relationships within the data. The Bi-LSTM architecture, trained on a consolidated dataset from Twitter Sentiment Analysis and Hate Speech and Offensive Language sources, employs preprocessing techniques like text cleaning and tokenization. The model's structure includes an embedding layer, spatial dropout, Bidirectional LSTM, the custom SNP layer, and a dense layer with a sigmoid activation for binary classification. Throughout training, binary cross-entropy loss and the RMSprop optimizer are utilized, while EarlyStopping and ModelCheckpoint callbacks optimize performance. Evaluation on a test set includes a printed confusion matrix for a comprehensive performance assessment. The research also showcases the preservation and utilization of the SNP layer's abstraction during the model's deployment, providing adaptability and contributing significantly to the model's effectiveness in hate speech detection.

2 Discussion

2.1 Experimental Setup

Language used: python

Performance evaluation tool: confusion matrix

Accuracy (all correct / all) = TP + TN / TP + TN + FP + FN

Dataset:

This research investigates the identification of hate speech using Twitter dataset (Hate speech and offensive language dataset [4] and Twitter sentiment analyses dataset [5]). The text in the dataset is categorized into three groups: neither, hate speech, or offensive language. It's important to note that the dataset includes text that could be seen as racist, sexist, homophobic, or generally offensive, considering the nature of the study.

In simpler terms, the aim is to predict labels for the test dataset based on a training set of tweets and labels. A label '1' indicates the tweet is racist/sexist, while '0' indicates it is not.

2.2 Methodology

For the purpose of detecting hate speech, the offered code carries out a number of operations relating to data preprocessing, model training, and evaluation. Below is the outline figure of the work:

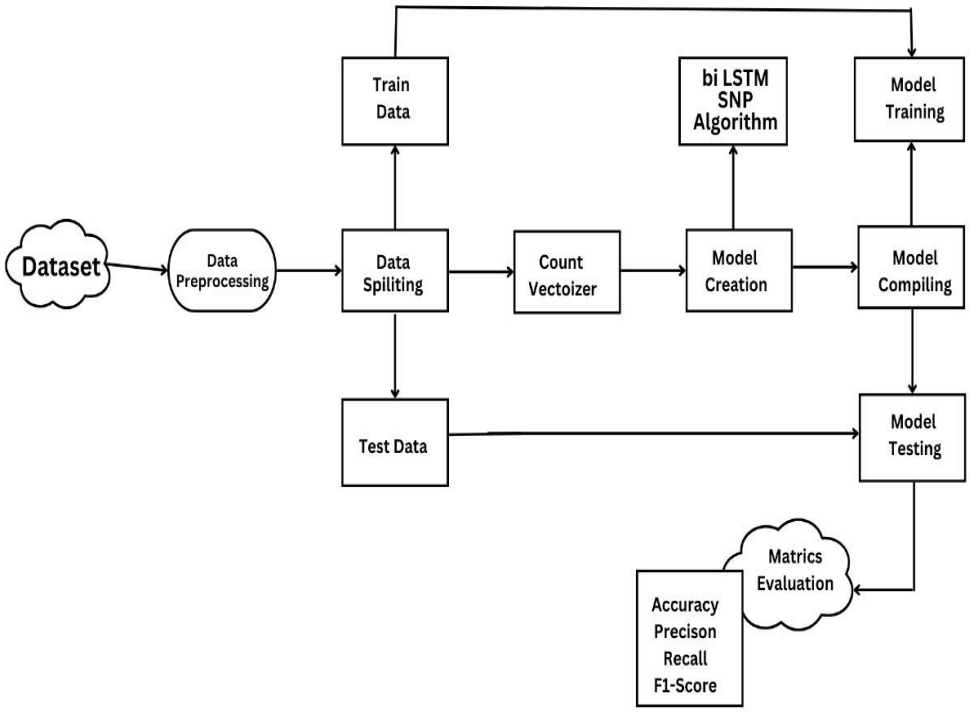

Fig.1. Procedure

- Import libraries
- Data collection
- Preprocessing
- Splitting the data
- Adding custom SNP layer
- Building model and prediction
- Evaluating result
- Testing the model

2.3 Experimental Result

```
Confusion Matrix:
[[7919  534]
 [ 408 5326]]
Precision: 0.9088737201365188
Recall: 0.9288454830833623
F1 Score: 0.9187510781438675
Accuracy: 0.9336011841827024
```

2.4 Result analyses

Comparing our analyses with machine learning algorithms using accuracy as a comparison measure.

Table 1. Analyses

Models	Accuracy
Proposed Model (Bi-LSTM SNP)	0.9336
Naive Bayes(unigram)	0.8521
SVM	0.4002
Decision Tree	0.8137
Word2vec model using LSTM	0.7815
GloVe Model using LSTM	0.829

The suggested model outperforms multiple benchmark models with a noteworthy accuracy of 93.36% in the research undertaken. With an accuracy of 85.21%, the Naïve Bayes (unigram) model performs admirably but lags behind the suggested model. Conversely, the SVM model exhibits a notable lag with an accuracy of 40.02%, indicating potential constraints in its predicting ability for the specified job. Comparing the Decision Tree model against the SVM, it shows its usefulness with an accuracy of 81.37%, which is rather good.

With LSTM, the Word2vec model achieves comparable accuracies of 78.15%, while the GloVe model achieves 82.9%. Although these models demonstrate the potential of using LSTM in conjunction with embedding approaches, the accuracy of the proposed model outperforms them.

3 Concluding Remark

Our study concludes that the proposed Bi-LSTM SNP model performs exceptionally well, outperforming benchmark models such as Naive Bayes (85.21%), Decision Tree (81.37%), Word2vec using LSTM (78.15%), and GloVe using LSTM (82.9%). The model's accuracy of 93.36% is demonstrated. The SVM model's limitations in detecting hate speech are highlighted by the notable performance gap (40.02%) that was observed.

Our study tackles problems in hate speech detection research, such as dataset size, reliability, and linguistic diversity issues, in addition to model evaluation. Significant obstacles arise from the lack of agreement on definitions of hate speech and the complexity of the model architecture, highlighting the necessity of ongoing dataset improvement and expert participation.

We support the use of fine-grained feature set selection for both traditional and deep learning models, and draw attention to the lack of established best practices for comparing cross-dataset methods. In summary, our study not only offers a sophisticated model for detecting hate speech, but also highlights the critical issues that require the focus of the scientific community in order to advance this area of study.

In conclusion, this research work presents a robust approach to identifying hate speech and offensive language in text data. The bi-LSTM SNP model, trained on a combined dataset of Twitter sentiment and hate speech data, provides a valuable tool for automated content moderation and addressing online toxicity. The results obtained through this analysis can be used to enhance online platforms' content filtering and promote a safer online environment.

Literature review

Hate speech detection has emerged as a crucial research domain, garnering interest from both political entities and individuals. In this era, natural language processing (NLP) has gained widespread acceptance among researchers, particularly in text classification. Beyond hate speech detection, text classification holds significance as artificial intelligence (AI) transitions towards generative AI. It is imperative to train generative AI models to effectively manage sentiments, ensuring control over content generation.

However, existing models grapple with limitations stemming from constrained datasets, including issues of consensus, bias, and reliability. As AI evolves, advancing methodologies become pivotal in addressing these challenges. Enhanced methods promise not only to refine hate speech detection but also to elevate content filtering on online platforms, fostering a safer digital environment. This literature review underscores the growing importance of NLP, the shift towards generative AI, and the necessity for robust methods to overcome current dataset limitations for the betterment of online content moderation.

Future Work

There is a lot of scope in future to work on this. Hate speech detection is also possible through emoji and GIFs. Future project development may take a range of factors into account to improve the project. The model's performance may be further enhanced by additional research and optimization, including hyperparameter tuning. Using advanced deep learning architectures such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), the model is able to identify intricate patterns and connections in text data. The model might be improved by the addition of strategies like transformer models or attention processes.

Author Contribution

The entire scope of this research project was carried out by Updesh Kumar, the only author. They contributed their technical expertise to the development process by carefully implementing and coding algorithms for the detection of hate speech. In addition, Updesh Kumar skillfully authored the entire document, gathered pertinent resources, analysed data, and created findings. This extensive participation highlights the crucial part that Updesh Kumar plays in all phases of the study, from conception to the completed paper.

Conflicts of interest

The author declares no conflict of interest. The report is not influenced by financial or personal relationships.

References

1. Md. Saroar Jahan, Mourad Oussalaha, A systematic review of hate speech automatic detection using natural language processing, 2024. Neurocomputing. https://www.sciencedirect.com/science/article/pii/S0925231223003557
2. Fatima Alkomah (2022). A Literature Review of Textual Hate Speech Detection Methods and Datasets. https://www.mdpi.com/2078-2489/13/6/273
3. Sekolah.mu, 2021. Deep Learning Techniques for Text Classification. Towards Data science.
4. Andrii Samoshyn,2019. Hate speech and offensive language dataset https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset
5. Ali Toosi, 2015. Twitter Sentiment Analyses. https://www.kaggle.com/datasets/arkhoshghalb/twitter-sentiment-analysis-hatred-speech
6. Areei Al-Hassan, 2019.Detection of hate speech in social networks: A survey on multilingual corpus.6th International Conference on Computer Science and Information Technology.
7. Yanping Huang,Oian Liu,Hong Peng, Jun Wang ,Quin Yang(2023).Sentiment classification using bidirectional LSTM-SNP model and attention mechanism. Expeert System with applications. https://www.sciencedirect.com/science/article/pii/S0957417423002312?ref=pdf_download&fr=RR-2&rr=82c7fd970bc631e1#sec2
8. Mohamed Arbane, Rachid Benlarmi , Youcef Brik, Ayman Diyab Alahmr ,2023. Social media-based Covid-19 sentiment classification model using Bi-LSTM.Expert Systems with applications. https://www.sciencedirect.com/science/article/pii/S0957417422017353?ref=pdf_download&fr=RR-2&rr=82c7feee681731e1
9. Md Saraar Jahan, Mourad Ousallah (2021). A systematic review of Hate Speech automatic detection using Natural Language Processing. Neurocomputing https://www.sciencedirect.com/science/article/pii/S0925231223003557
10. Pitsilis, G.K., Ramampiaro, H., Langseth, H,2018. Effective hate-speech detection in Twitter data using recurrent neural networks. Appl. Intell. 4730–4742.https://link.springer.com/article/10.1007/s10489-018-1242-y,https://arxiv.org/pdf/1801.04433.pdf
11. Aziz, N.A.A.; Maarof, M.A.; Zainal, A. Hate Speech and Offensive Language Detection: A New Feature Set with Filter-Embedded Combining Feature Selection. In Proceedings of the 2021 3rd International Cyber Resilience Conference CRC 2021, online, 29–31 January 2021. https://ieeexplore.ieee.org/abstract/document/9392486?casa_token=r-ND6y6F6lkAAAAA:miTNG_Ic9Cqipflc7QIPvhkNiHU3zf8cikQnDx3DueAALIKVGghvJUfcFijjDIHf5Lu1TXWLYU
12. Malla, S., & PJA, A. (2021). COVID-19 outbreak: An ensemble pre-trained deep learning model for detecting informative tweets. Applied Soft Computing, 107, 107495. https://doi.org/10.1016/j.asoc.2021.107495
13. Ishani Priyadarshani, Sandipan Sahu Raghvendra Kumar (2023). A transfer Learning approach for detecting offensive and hate speech on social media platforms. Multimedia tools and applications. https://link.springer.com/article/10.1007/s11042-023-14481-3
14. Weijiang li, fang Qi, Ming Tang, Zhengtao Yu (2020). Bidirectional LSTM with self-attention mechanism and multichannel features for sentiment classification. https://www.sciencedirect.com/science/article/abs/pii/S0925231220300254
15. Alex Sherstinsky (2020). Fundamentals of Recurrent Neural Networks and Long short-term Memory (LSTM) network. Physica D:Nonlinear Phenomena. https://www.sciencedirect.com/science/article/abs/pii/S0167278919305974?via%3Dihub

16. Guxian Xu,Yueting Ming, Xiayou Qiu,Ziheng Yu, Xu Wu . Sentiment analyses of comment texts based on BiLSTM. https://ieeexplore.ieee.org/document/8684825