International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Hate Speech Detection using Deep Learning

Ibrahim Shaikh
Department of BECHLOR OF VOCATIONAL IN ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
Anjuman-I-Islam's AbduL Razzaq Kalsekar Polytechnic, New Panvel, Maharashtra, India

Abstract - With the exponential growth of social media platforms, the proliferation of hate speech has become a significant societal problem. Manual moderation of this vast amount of user-generated content is impractical, necessitating the development of automated detection systems. This research presents a deep learning-based approach for detecting hate speech in text data. We leverage a dataset of tweets labeled as hate speech, offensive language, or neither. Several deep learning models are explored, including Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM), to effectively classify the text. The methodology involves crucial preprocessing steps such as text cleaning, tokenization, and sequence padding to prepare the data for the models. Our experimental results demonstrate that the Bidirectional LSTM model achieves a high accuracy of 95.83% on the test set, outperforming other models. This study underscores the potential of deep learning techniques to create robust and scalable solutions for combating the spread of hate speech online.

Keywords: Hate Speech Detection, Deep Learning, Natural Language Processing (NLP), LSTM, Bidirectional LSTM, Social Media, Text Classification.

1.INTRODUCTION

The rise of social media has fundamentally changed how people communicate and share information. While these platforms offer numerous benefits, they have also become breeding grounds for hate speech and other forms of harmful content. Hate speech, which targets individuals or groups based on attributes like race, religion, ethnic origin, or gender, can incite violence and has a detrimental effect on society. The sheer volume and velocity of user-generated content make manual detection and moderation an unscalable and challenging task. Therefore, there is a critical need for automated systems that can accurately and efficiently identify hate speech.

This research focuses on developing a system for hate speech detection using deep learning techniques. We aim to build and evaluate different models, including LSTM and Bidirectional LSTM, which are well-suited for sequence data like text. The objective is to create a model that can effectively distinguish between hate speech, offensive language, and neutral text, thereby providing a tool to help mitigate the negative impact of hate speech on online platforms.

II. RELATED WORK

The detection of hate speech is a well-established area of research within Natural Language Processing (NLP). Early approaches often relied on traditional machine learning algorithms combined with feature engineering. These methods used features such as Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and n-grams with classifiers like Naive Bayes, Support Vector Machines (SVM), and Logistic Regression. While effective to an extent, these models heavily depend on the quality of the handcrafted features and may struggle to capture the complex semantic and contextual nuances of language.

With the advent of deep learning, there has been a paradigm shift in NLP tasks. Deep learning models can automatically learn relevant features from the data, eliminating the need for manual feature engineering. Recurrent Neural Networks (RNNs), and more specifically their advanced variants like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs), have proven highly effective for text classification. These models can process sequential data and maintain a memory of past information, which is crucial for understanding context. Furthermore, Bidirectional LSTMs (Bi-LSTMs) enhance this capability by processing text in both forward and backward directions, allowing for a more comprehensive understanding of the context surrounding each word.

III. METHODOLOGY

A. Dataset Description

The dataset used in this study is sourced from Kaggle and contains tweets that have been labeled into three categories:

* Hate Speech
* Offensive Language
* Neither

The dataset is loaded into a pandas DataFrame and consists of a column for the tweet text and a corresponding column for the class label.

B. Data Preprocessing

To prepare the text data for the deep learning models, a series of preprocessing steps were performed:

1.Text Cleaning: Removal of punctuation, special characters, and stop words using regular expressions and the NLTK library.

2.Tokenization: Breaking down the cleaned text into tokens.

3.Sequencing and Padding: Tokenized text was converted into sequences of integers and padded to a fixed length of 25 words.

C. Model Architecture

The models were implemented using the Keras library with a TensorFlow backend. The architecture includes:

*Embedding Layer: Converts integer-encoded sequences into dense vector representations.

*LSTM/Bidirectional LSTM Layer: Processes text sequences. Bi-LSTM processes in both directions.

*Dense Layers: Fully connected layers with ReLU activation.

*Output Layer: Softmax activation for three-class classification.

D. Training and Evaluation

The dataset was split into 80% training and 20% testing. The categorical crossentropy loss function and Adam optimizer were used. Training was done for 5 epochs with a batch size of 64. Accuracy and loss were tracked for both training and validation sets.

IV. EXPERIMENTAL SETUP

Experiments were conducted using Python 3 with the following libraries:

* TensorFlow and Keras
* Scikit-learn
* Pandas
* NLTK
* Matplotlib and Seaborn

V. RESULTS AND DISCUSSION

The Bidirectional LSTM model achieved the best performance with a test accuracy of 95.83%. Training history showed consistent improvement in accuracy and reduction in loss, indicating good model learning. A confusion matrix confirmed effective classification across all three classes.

VI. FUTURE WORK

Future enhancements include:

* Use of transformer-based models like BERT for improved accuracy.
* Incorporation of user profile and social network data.
* Development of real-time detection systems for direct integration with social media platforms.

VII. CONCLUSION

This paper presented a deep learning-based approach for hate speech detection. Through data preprocessing and implementation of LSTM and Bi-LSTM models, the study achieved high accuracy in classification. The Bi-LSTM model demonstrated a 95.83% test accuracy, confirming the effectiveness of deep learning in combating hate speech on social media.

REFERENCES

\[1] Z. Zhang and L. Luo, “Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter,” Semantic Web Journal, vol. 12, no. 4, pp. 543–563, 2021.

\[2] S. Davidson, D. Bhattacharya, and I. Weber, “Automated Hate Speech Detection and the Problem of Offensive Language,” in Proceedings of the 11th International Conference on Web and Social Media (ICWSM), Montreal, Canada, May 2017, pp. 512–515.