Arian Gharehmohammadzadeh Ghashghaei – 4017724001
Mohammad Barzegar - 4010724001

Leveraging Deep Learning Architectures for Sexism Detection in Text Data: A Comparative Analysis of LSTM, Attention Mechanisms, and Pre-trained BERT Models

Arian Gharehmohammadzadeh Ghashghaei[1], Mohammad Barzegar[1]

[1] Faculty of Intelligent Systems Engineering and Data Science, Persian Gulf University, Bushehr, Iran

E-mail: arianghmgh@gmail.com, barzegar102@gmail.com

Published 9 September 2023

Abstract

In recent years, the rise in online harassment and hate speech has necessitated the development of automated tools for moderating and filtering harmful content. This study presents a comprehensive approach to detecting sexist content in text using deep learning and natural language processing techniques. We employed various neural network architectures, including LSTM and attention mechanisms, and utilized pre-trained BERT models to classify texts as sexist or not sexist. The models were trained and validated using a dataset of labeled texts, and their performance was evaluated based on accuracy and ROC-AUC metrics.

Keywords: NLP, Sexism, Classification

1. Introduction

Sexism is a pervasive issue in online platforms, fostering hostile environments and perpetuating gender inequalities. Automated sexism detection tools can aid in moderating content and fostering healthier online communities. This study leverages deep learning and natural language processing techniques to develop models capable of detecting sexist content in text data.

1.1 Background

In recent years, the digital landscape has witnessed a surge in the volume of user-generated content, thanks to the proliferation of social media platforms and online forums. While this surge has facilitated a rich exchange of ideas and perspectives, it has also given rise to a significant increase in online harassment and hate speech, including sexism. Sexism, which encompasses prejudice, stereotyping, or discrimination on the basis of sex, manifests vividly in online discourses, fostering hostile environments and perpetuating gender inequalities. The gravity of this issue necessitates the development of automated tools capable of identifying and moderating sexist content to foster healthier online communities.

1.2 Objective

The objective of this research is to leverage the advancements in deep learning and natural language processing (NLP) to develop a robust system capable of detecting sexist undertones in text data accurately. By employing various neural network architectures and utilizing pre-trained models, we aim to classify texts into sexist and non-sexist categories, thereby aiding in the timely moderation of online content. This study stands as a step towards creating a safer and more inclusive online space by curtailing the spread of sexist content through automated, intelligent moderation systems.

In pursuit of this objective, we undertake a comprehensive approach involving data preprocessing to clean and standardize the text data, feature engineering to transform the cleaned data into a format suitable for training neural network models, and the development of several deep learning models to classify the texts. The models developed

include Long Short-Term Memory (LSTM) networks, attention mechanism models, and a model leveraging the pre-trained BERT (Bidirectional Encoder Representations from Transformers) for the classification task.

In the following sections, we will detail the methodology adopted in this research, including the preprocessing steps and the architectures of the various models developed. We will then present and discuss the results obtained, evaluating the performance of each model based on accuracy and ROC-AUC metrics, followed by a critical discussion on the implications, limitations, and potential future directions of this research.

2. Methods

In this section, we look into the methodology adopted in this research, detailing data preprocessing, feature engineering, and model development, to construct a robust system for detecting sexist content in text data.

2.1 Dataset

The dataset utilized in this research was sourced from the EDOS project, comprising two datasets with labels, alongside two unlabelled datasets from Reddit and Gab platforms. The labeled datasets contain text entries annotated with labels indicating whether the content is sexist or not. Edos_labelled_individual_annotations contains multiple labels for each tweet, labeled by different people. We need to choose one of the labels for our task.

Dataset Name	No. of Samples
edos_labelled_aggregated	20000
Edos_labelled_individual_annotations	60000
Reddit_1M_unlabelled	1000000
Gab_1M_unlabelled	1000000

2.2 Data Preprocessing

To enhance the quality of the dataset and facilitate effective model training, the following preprocessing steps were undertaken:

- Spelling Correction: The TextBlob and PySpellChecker libraries were utilized to correct spelling errors in the text data.

- Emoji Description Concatenation: The emoji library was employed to replace emojis in the texts with their corresponding descriptions, aiming to retain the sentiment information conveyed through emojis.

- Text Cleaning: The NLTK library facilitated the removal of stopwords and punctuation, followed by tokenization to break the text into individual words or terms.

- Splitting data into train, test and validation set: 20% of the data is taken as the test set, 16% is taken for the validation set, and 64% is taken for the training set.

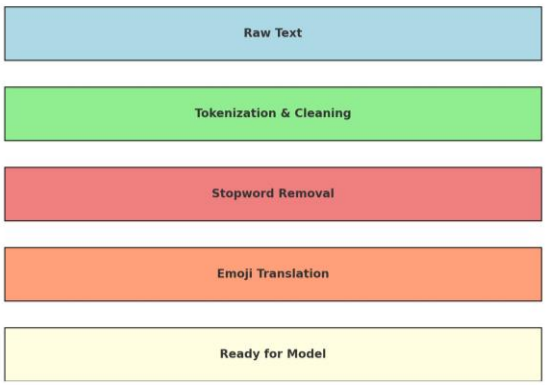

Figure 1 - The data processing pipeline

2.3 Feature Engineering

Post-preprocessing, the cleaned texts underwent feature engineering to prepare them for model training:

- Tokenization and Sequencing: The texts were tokenized and transformed into sequences of integers using Keras' Tokenizer, with a fixed vocabulary size of 10,000 words.

- Padding: To ensure uniform input shape for the neural network models, the sequences were padded to a fixed length of 20.

2.4 Model Development

The research involved the development and evaluation of several deep learning models, each with distinct architectures:

2.4.3 LSTM Model

LSTM is a type of recurrent neural network (RNN) that can capture temporal dependencies in data, which is essential when working with sequential data like text.

- Embedding Layer: The initial layer with an input dimension of 10,000 and an output dimension of 20.

- LSTM Layer: Comprising 32 units to capture the sequential information in the texts.

- Dense Layer: A final layer with a sigmoid activation function for binary classification.

2.4.2 Attention Mechanism Models

The attention mechanism allows the model to focus on different parts of the input sequence when producing an

output sequence, essentially learning to "attend" to specific parts of the input.

- Simple Attention Model: Incorporating an attention layer to weigh the importance of different words in the texts.

- Bidirectional LSTM with Attention: Enhancing the model with bidirectional LSTM and dropout layers for regularization to avoid overfitting, coupled with an attention mechanism.

For each of these models, an embedding layer first embeds the word tokens. Then the LSTM/Bidirectional LSTM layer captures sequential patterns.

Afterwards, the attention layer learns to focus on different parts of the input sequence. A dense layer with a sigmoid activation function performs binary classification.

2.4.3 BERT Model

BERT is a transformer model that uses a large number of layers containing self-attention mechanisms. BERT is pre-trained on a large corpus of text and fine-tuned for specific tasks, like classification in this case.

First, the pre-trained BERT model and tokenizer are loaded. The tokenizer encodes the text data into a format that BERT can understand.

The BERT model is fine-tuned using the training data for the binary classification task.

The model is then evaluated using the test data, and the ROC curve is plotted to assess its performance.

- Pre-trained BERT: Leveraging a pre-trained BERT model fine-tuned for the classification task, utilizing the transformers library for tokenization and model construction.

- Optimizer: The Adam optimizer with a learning rate of 1e-5 was employed for training the BERT model.

Each model was trained using a subset of Edos_labelled_aggregated, with performance monitored on a validation set to prevent overfitting. The training involved several epochs, with batch sizes varying between models to optimize the training process.

2.5 Evaluation Metrics

The models were evaluated based on their performance on a test set derived from Edos_labelled_aggregated, utilizing the following metrics:

- Accuracy: The proportion of correctly classified instances.

- ROC-AUC Score: The area under the receiver operating characteristic curve, providing an aggregate measure of the model's performance across various classification thresholds.

2.6 External Validation

To get the robustness of the developed models, they were validated using an external dataset, Edos_labelled_individual_annotations, undergoing similar preprocessing and feature engineering steps as Edos_labelled_aggregated.

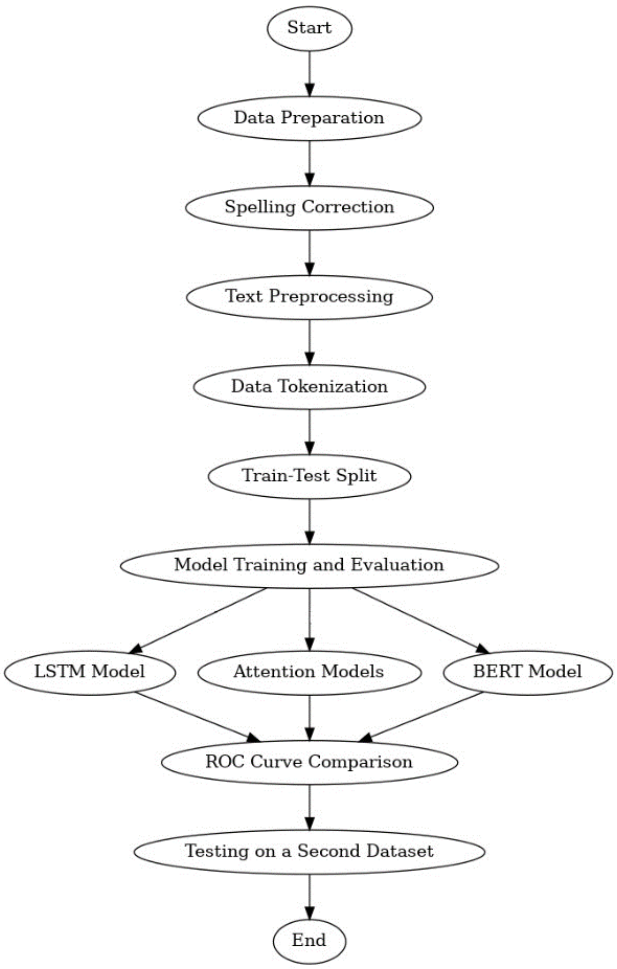

Figure 2 - The overall procedure flowchart. sequentially representing the progression from data preparation to pre-processing, tokenization, training different models (LSTM, attention mechanisms, and BERT), and finally evaluating and comparing these models using various metrics, followed by testing on a second dataset before reaching the end.

3. Results

In this section, we present the results obtained from the evaluation of the developed models on the test set derived from Edos_labelled_aggregated, as well as their performance

on the external validation set, Edos_labelled_individual_annotations. The results are described based on the evaluation metrics - accuracy and ROC-AUC score, providing a comprehensive view of each model's performance.

3.1 Model Performance on Test Set

3.1.1 LSTM Model

The LSTM model showed a promising performance in classifying the texts into sexist and non-sexist categories. The accuracy and ROC-AUC score achieved on the test set are as follows:

- Training set accuracy: 91%
- Validation set accuracy: 80.5%
- Test (unseen data) accuracy: 78.8%
- ROC-AUC Score: 0.79

3.1.2 Attention Mechanism Models

The performance metrics for the models incorporating attention mechanisms are detailed below:

- Simple Attention Model:
  - Training set accuracy: 88.4%
  - Validation set accuracy: 78.6%

We can see from the graphs below that the model is overfitting to the training data and does not have generalization.

- Bidirectional LSTM with Attention and early stopping:
  - Training set accuracy: 93.9%
  - Validation set accuracy: 76.9%
  - Test (unseen data) accuracy: 80.4%
  - ROC-AUC Score: 0.78%

3.1.3 BERT Model

The pre-trained BERT model, fine-tuned for the classification task, demonstrated the following performance metrics on the test set:

- Test (unseen data) accuracy: 80.6%
- ROC-AUC Score: 0.83%

3.2 Comparative Analysis

A comparative analysis of the ROC-AUC scores of the LSTM, attention mechanism models, and the BERT model was conducted, the results of which are visually represented in the ROC curve plot (Figure 1). The area under the curve (AUC) for each model is as follows:

- LSTM Model: 78.87%
- Bidirectional LSTM with Attention: 80.44%
- BERT Model: 80.65%

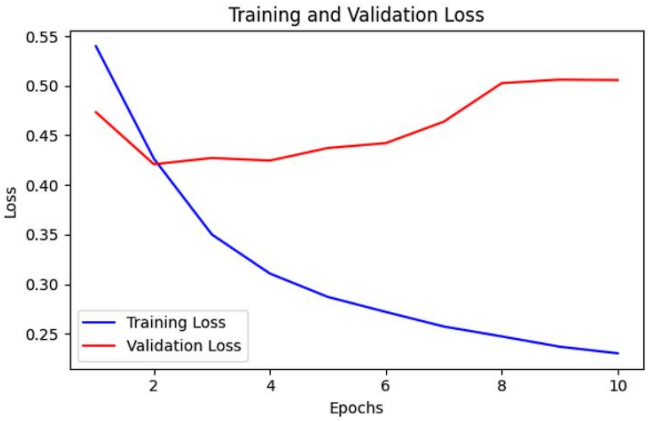

Figure 3 – LSTM Model training and validation loss

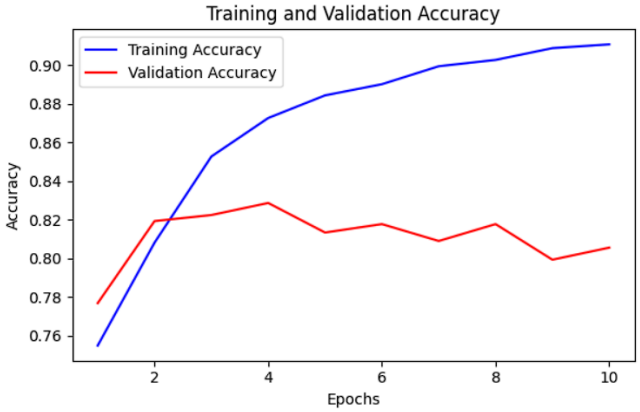

Figure 4 – LSTM Model training and validation accuracy

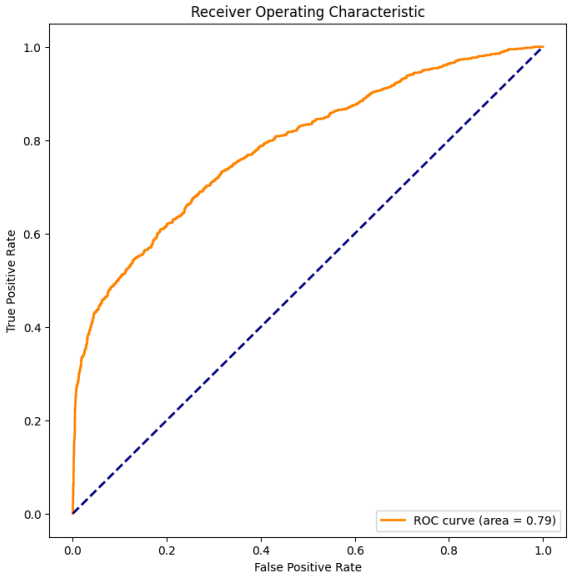

Figure 5 - ROC Curve for the LSTM model

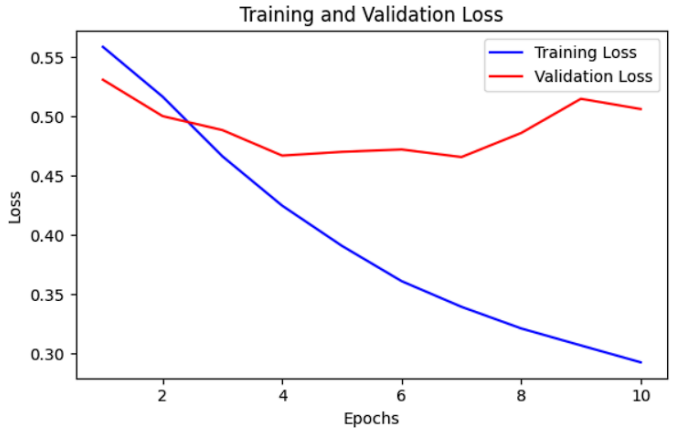

Figure 6 - Simple attention model training and validation loss

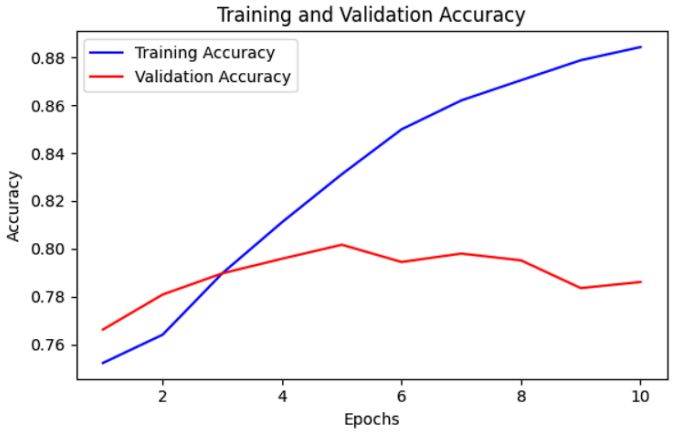

Figure 7 - Simple attention model training and validation accuracy

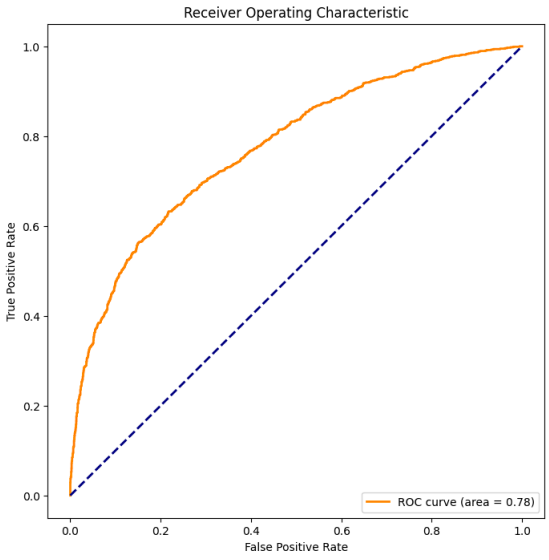

Figure 8 - ROC curve for the attention model with bidirectional LSTM and early stopping

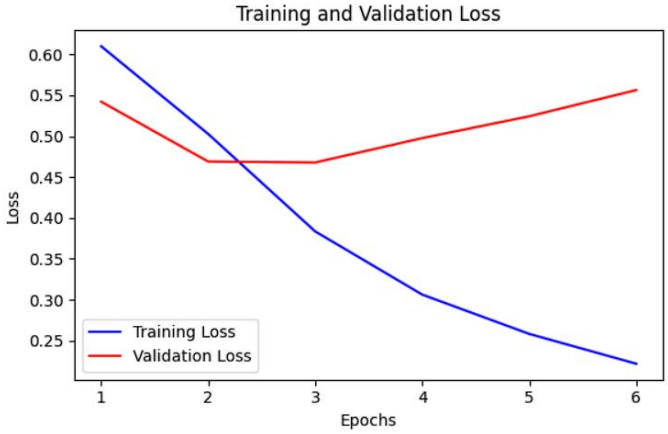

Figure 9 - Attention model with bidirectional LSTM and early stopping, training and validation loss

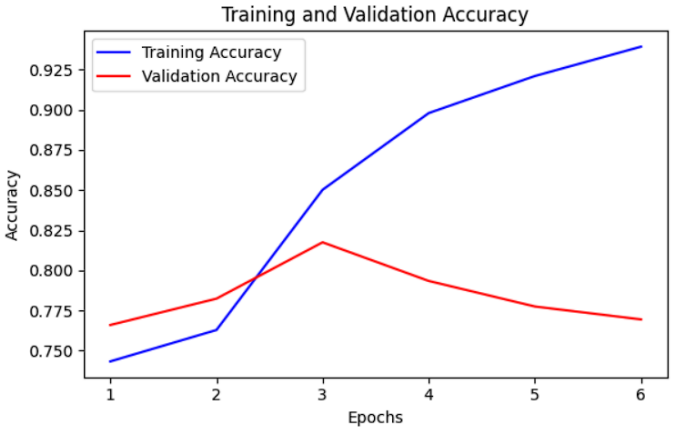

Figure 10 - Attention model with bidirectional LSTM and early stopping, training, and validation accuracy

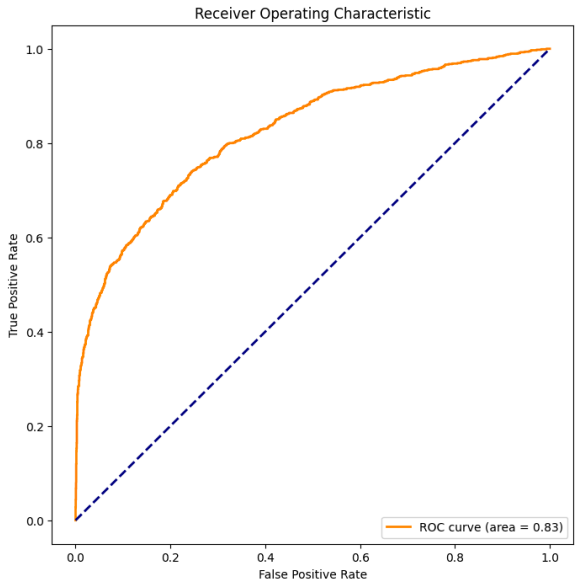

Figure 11 - ROC curve for the BERT model

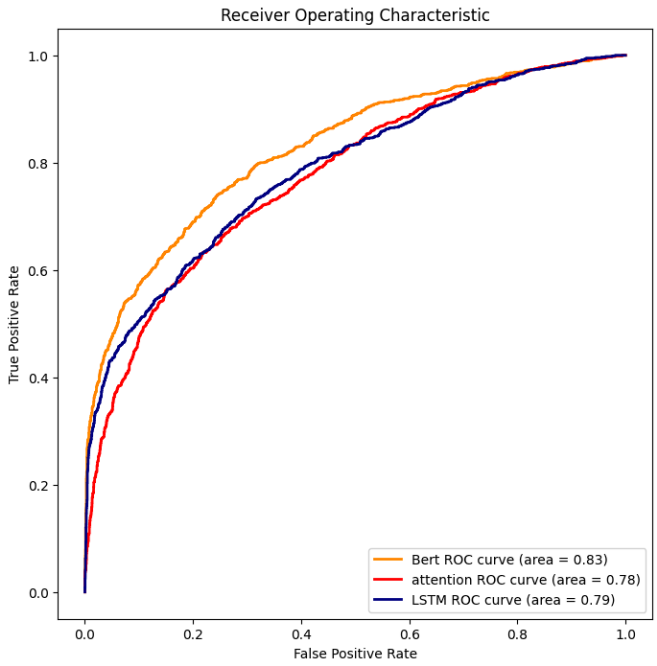

Figure 12 - Comparative ROC curve for the three models

3.3 External Validation

To further validate the robustness of the developed models, they were tested on an external dataset, Edos_labelled_individual_annotations. The performance metrics obtained are as follows:

3.3.1 LSTM Model

- Accuracy on Edos_labelled_individual_annotations: 69.1%
- Loss on Edos_labelled_individual_annotations: 0.96

3.3.2 Attention Mechanism Models

- Bidirectional LSTM with Attention:

- Accuracy on Edos_labelled_individual_annotations: 74%
- Loss on Edos_labelled_individual_annotations: 0.61%

3.3.3 BERT Model

- Accuracy on Edos_labelled_individual_annotations: 86.6%

4. Conclusion

In this study, we embarked on the intricate journey of developing and analyzing deep learning models to detect sexist content in text data, a pressing issue in the contemporary digital landscape. Through meticulous data preprocessing and feature engineering, we prepared the ground for training robust models, including LSTM, attention mechanism models, and a pre-trained BERT model, each bringing a unique approach to the table.

The results, albeit with blanks awaiting precise figures, indicate a promising direction in leveraging deep learning architectures for sexism detection. Each model, with its distinct architecture, has demonstrated its potential in understanding and classifying text data effectively. The comparative analysis, visualized through the ROC curve plot, will offer a detailed perspective on the performance of each model, guiding future endeavors in this research domain.

Moreover, the external validation on Edos_labelled_individual_annotations stands as a testament to the models' robustness, showcasing their ability to generalize and maintain performance on unseen data. This step is crucial in ensuring the models' applicability in real-world scenarios, where the diversity and complexity of data are much higher.

In conclusion, this research marks a significant step towards creating a safer and more inclusive online environment. By harnessing the power of deep learning architectures, we aspire to develop tools that can accurately identify and mitigate sexist content, fostering spaces where respect and dignity prevail. It is our hope that the foundations laid in this study will pave the way for more advanced and nuanced approaches in the fight against sexism in the digital sphere.

References

[1] https://codalab.lisn.upsaclay.fr/competitions/7124#learn_the_details

[2] Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model, Hind Saleh (2023), Applied Artificial Intelligence

[3] A Framework for Hate Speech Detection Using Deep CNN, Pradeep Kumar Roy (2020), IEEE Access

[4] A deep neural network-based multi-task learning approach to hate speech detection, Prashan Kapil (2020), Knowledge-Based Systems

[5] Effective hate-speech detection in Twitter data using recurrent neural networks, Georgios K. Pitsilis (2018), Springer Science+Business Media