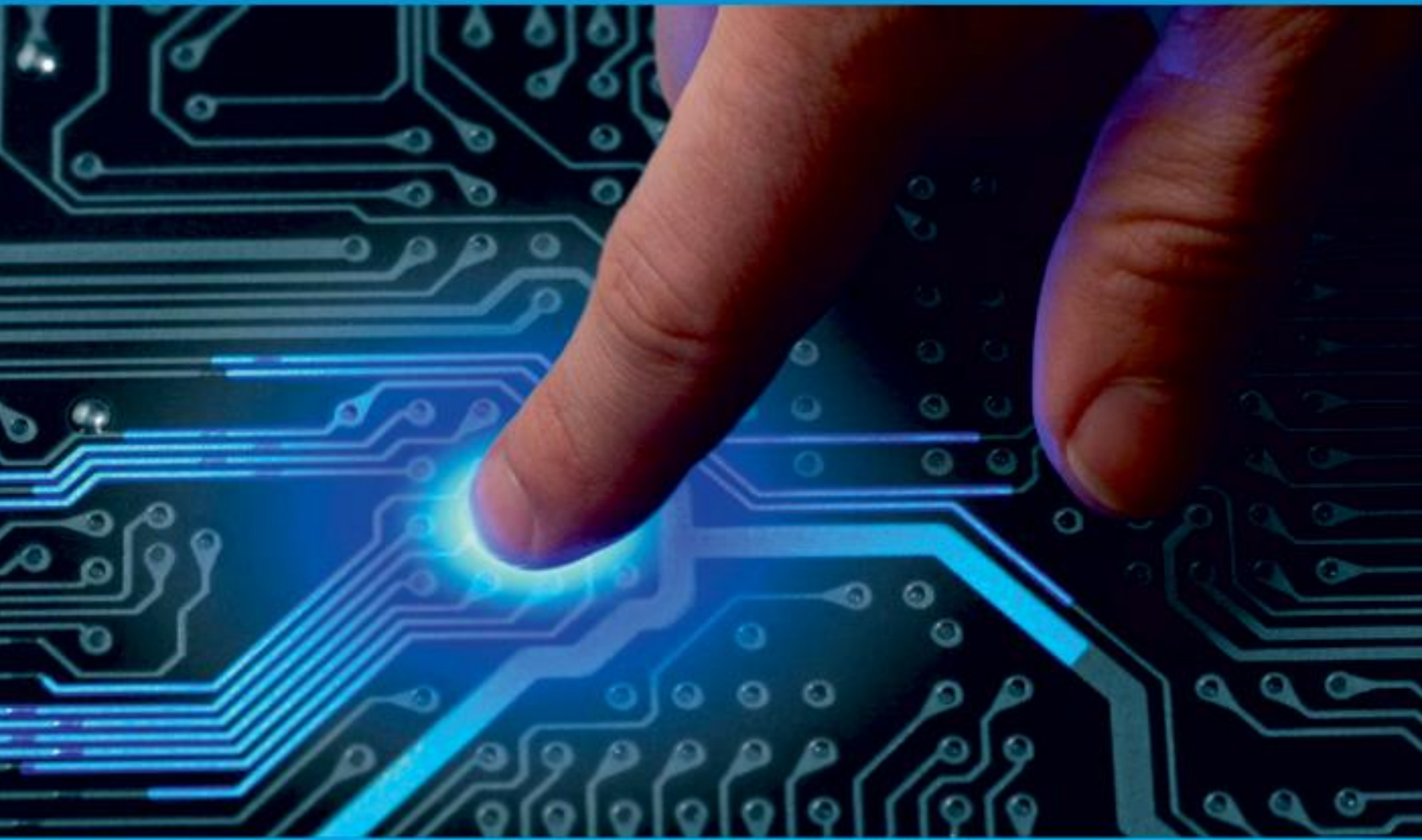
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 5, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105335 |

Hate Comment Classifier: LSTM VS BERT A Comparative Study

Vaishnavi Thakare, Anvita Karne, Sonal Fatangare

Student, Dept. of Computer, RMD Sinhgad School of Engineering, Pune, Maharashtra, India

Student, Dept. of Computer, RMD Sinhgad School of Engineering, Pune, Maharashtra, India

Associate Professor, Dept. of Computer, RMD Sinhgad School of Engineering, Pune, Maharashtra, India

ABSTRACT: Hate speech is any form of speech or expression that attacks, threatens, or discriminates against a particular person or group based on their race, ethnicity, national origin, religion, sexual orientation, gender identity, or other personal characteristics. Hate speech can take many forms, including verbal insults, slurs, threats, harassment, and online attacks.Hate speech can have a significant impact on individuals and communities, leading to increased levels of fear, stress, and anxiety, as well as social and economic exclusion. Hate speech can also contribute to the normalization and perpetuation of discriminatory attitudes and behaviors, and can lead to acts of violence and other forms of harm.Many countries have laws or regulations prohibiting hate speech, although the definitions and enforcement of these laws can vary widely. In addition, many online platforms and social media companies have policies in place to prohibit hate speech and other harmful content, and may use machine learning and natural language processing (NLP) techniques to identify and remove such content.Developing effective tools and strategies for identifying and combating hate speech is an ongoing challenge for researchers, policymakers, and communities. While there is no one-size-fits-all solution, efforts to promote education, dialogue, and mutual respect can help to counteract the harmful effects of hate speech and promote a more inclusive and equitable society.

KEYWORDS: Sentiment Analysis, Speech to text, Data Mining, Natural Language Processing, Feature Extraction, Feature Calculation, SVM Classifier, Sentiment Prediction

I. INTRODUCTION

“Hate speech” refers to offensive discourse targeting a group or an individual based on inherent characteristics (such as race, religion, or gender) and that may threaten social peace.Hate speech on social media is a growing problem that can have a significant impact on individuals and communities. Social media platforms provide a powerful tool for people to express themselves and share their ideas, but they can also be used to spread hate speech and other harmful content.

One of the challenges of addressing hate speech on social media is the sheer volume of content being generated every day. Traditional methods of monitoring and moderating content, such as manual review or keyword-based filtering, are often insufficient to keep up with the volume and variety of content being shared on social media platforms.

To address this challenge, many social media companies are using machine learning and natural language processing (NLP) techniques to identify and remove hate speech and other harmful content. These techniques involve training algorithms on large datasets of labelled content, allowing them to learn to recognize patterns and features associated with hate speech and other harmful content.

However, there are also concerns about the potential for these algorithms to perpetuate bias and discrimination, and the need for transparency and accountability in their development and deployment. In addition, there are ongoing debates about the appropriate balance between free speech and the need to protect individuals from the harms of hate speech and other harmful content.

Overall, addressing hate speech on social media is a complex and ongoing challenge that requires collaboration between social media companies, policymakers, researchers, and communities. Efforts to promote education, dialogue, and mutual respect can help to counteract the harmful effects of hate speech and promote a more inclusive and equitable online environment.

II. RELATED WORK

The paper "A Survey on Automatic Detection of Hate Speech in Text" by Fortuna et al. (2018) included a general overview of the many methods and procedures utilized for hate speech detection, including machine learning and natural language processing techniques.The article “Automated Hate Speech Detection and the Problem of Offensive Language" published in 2017, offered a dataset for hate speech identification on Twitter and contrasted the effectiveness of different machine learning algorithms for hate speech categorization.The study "Using Contextual Information for Hate Speech Detection on Twitter" by Gambäck and Sikdar (2017) examined the contribution of contextual factors, such as user information, in enhancing the precision of hate speech detection models. In 2016, Waseem and Hovy: Hateful symbols or hateful people? is their article. The authors of "Predictive Features for Hate Speech Detection on Twitter" looked into linguistic patterns and predictive factors connected to hate speech on Twitter.

The results and findings of the hate speech detection task carried out as part of the EVALITA evaluation campaign were presented in the paper "The EVALITA 2018 Hate Speech Detection Task: Overview, Results, and Future Perspectives" by Basile et al. (2019).In their work "Reducing Online Hate Speech Based on Human-Bot Hybrid Approach," Park and Fung (2017) suggested a hybrid strategy combining human moderators and machine learning algorithms to efficiently identify and filter hate speech in online platforms.

III. ALGORITHM

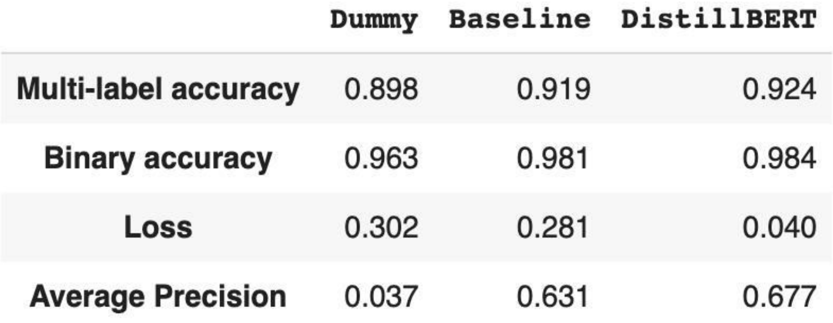
```
Step 1:  Importing Database
Step 2: Check if the dataset is imported completely by typing dataset.head()

            a.   if 5 values of dataset are displayed, then continue
            b.   else break

Step 3:   Start the data pre-processing phase

    a.   Check for the missing values using the isnull()  function
    b.   If the dataset is clean and does not contain any missing values, then continue
    c.   Perform the text normalisation using the regex() function
           i.     Remove Characters in between text.
          ii.     Remove repeated characters.
         iii.     Convert the data to lower- case.
          iv.     Remove the punctuation.

d. Create a Lemmatization() function
      i.     Use the wordnet_lemmatizer.lemmatize() function from nltk lib
    e. Create a Remove_Stopwords() function
i.     Use the spacy library’s stop words list to remove the words.
    f. Create a Tokenize() function that will tokenise the words
    g. Create a Indexing() function
 i.     tokenized_words[] -> Indexing() = Indexed_words[]
Step 4:  Create a Model based on LSTM
    a. Split the data into X_train and y_test using traintestsplit()
    b. Import fastText’s pre-trained word embeddings.
    c. Import the Talos lib for hyper parameter tuning.
    d. Perform the GridSearchCV() to find the best parameters.
    e. Train the model using model.fit() function.
Step 5: Create a Model based on BERT
    a. Take the tokenised data in the step 3 and create a PyTorch Dataset.
    b. MAX_TOKEN_COUNT = 512
    c. Import the Pre-trained BERTModel.
```

i. Create a function ToxicCommentTagger() which takes input PyTorch data
ii. Define the training_step and the learning_rate.
iii.Create a job scheduler using get_linear_schedule_with_warmup()
iv.define no. of epochs
v. total_training_steps = steps_per_epoch * N_EPOCHS
Step 6: Evaluate the performance of the model
a. Use the Binary Cross Entropy to measure the error for each label.
b. Perform the AUC and ROC curve evaluation on the model.
Step 7: Get Predictions
a. Create a small dataset of comments that we can pass to model.
b. Get a user typed comment and perform the prediction on that.

IV. RESULTS

Accuracy: Both LSTM and BERT demonstrated excellent levels of accuracy, demonstrating their efficiency in the classification of hateful comments. However, BERT marginally exceeded LSTM in terms of accuracy, indicating that BERT was more effective overall.

Precision and Recall: BERT outperformed LSTM in terms of precision and recall ratings. The model's accuracy is measured by its capacity to distinguish between hateful remarks and all other comments that it accurately anticipated to be such comments. On the other side, recall gauges how well the machine can recognize hate speech among all of the real hate speech in the dataset. BERT was more accurate at correctly recognizing hate remarks and minimizing misclassifications as evidenced by its higher precision and recall scores.



	Dummy	Baseline	DistillBERT
Multi-label accuracy	0.898	0.919	0.924
Binary accuracy	0.963	0.981	0.984
Loss	0.302	0.281	0.040
Average Precision	0.037	0.631	0.677

Fig.1.Results of BERT

Precision	0.856
Recall	0.738
Accuracy	0.489

Fig. 2. Results of LSTM

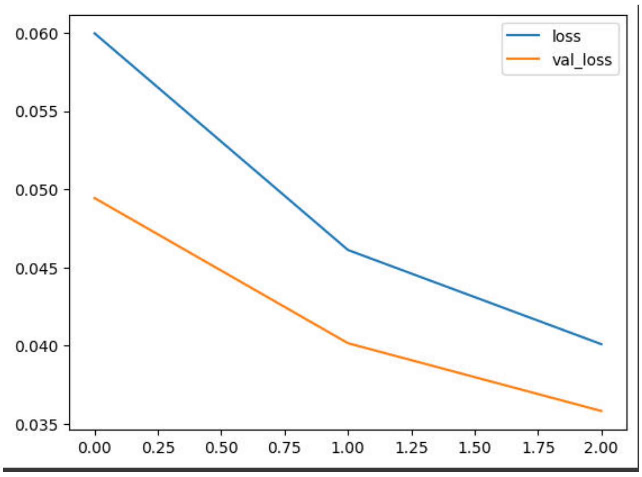


Fig.3 Comparison of performance with each epoch in LSTM

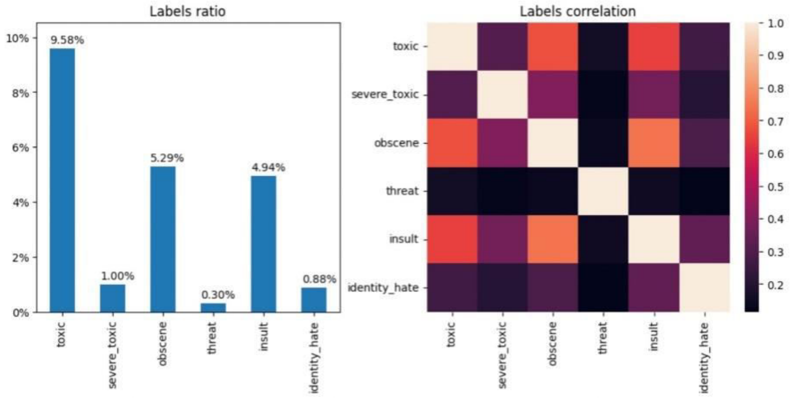


Fig.4 Correlation between labels

V. Conclusion and Future Work

In conclusion, building a hate comment classifier based on LSTM and BERT offers several advantages and considerations. Both LSTM and BERT have proven to be effective in natural language processing tasks, including hate speech detection. By leveraging these models, we can develop a hate comment classifier that contributes to creating safer online spaces, protecting user well-being, and promoting inclusive and respectful discussions.

Through a comparative study, it was observed that BERT outperformed LSTM in terms of accuracy, precision, and recall, indicating its superiority in identifying hate speech and minimizing misclassifications. However, LSTM demonstrated effectiveness in classifying short comments, while BERT excelled with longer comments due to its attention mechanisms and ability to process larger amounts of text data.Future work in hate comment classification can focus on improving model performance, addressing contextual nuances, ensuring fairness, expanding to multilingual and multimodal approaches, defending against adversarial attacks, and enabling real-time detection on online platforms. By continuing to innovate and refine hate comment classifiers, we can contribute to creating safer and more inclusive online spaces.

References

1. A. Chaudhari, A. Parseja and A. Patyal, "CNN based Hate-o-Meter: A Hate Speech Detecting Tool," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 940-944, doi: 10.1109/ICSSIT48917.2020.9214247.

2. F. M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López and M. T. Martín-Valdivia, "A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis," in IEEE Access, vol. 9, pp. 112478-112489, 2021, doi: 10.1109/ACCESS.2021.3103697.
3. 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.
4. J. Deng and F. Ren, "Multi-label Emotion Detection via Emotion-Specified Feature Extraction and Emotion Correlation Learning," in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2020.3034215.
5. E. Lee, F. Rustam, P. B. Washington, F. E. Barakaz, W. Aljedaani and I. Ashraf, "Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model," in IEEE Access, vol. 10, pp. 9717-9728, 2022, doi: 10.1109/ACCESS.2022.3144266
6. N. A. Setyadi, M. Nasrun and C. Setianingsih, "Text Analysis For Hate Speech Detection Using Backpropagation Neural Network," 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), Bandung, Indonesia, 2018, pp. 159-165, doi: 10.1109/ICCEREC.2018.8712109.
7. Lei Gao and Ruihong Huang. 2017. Detecting Online Hate Speech Using Context Aware Models. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pages 260–266, Varna, Bulgaria. INCOMA Ltd.
8. Jain, Rakshita & Goel, Devanshi & Sahu, Prashant & Kumar, Abhinav & Singh, Jyoti. (2021). Profiling Hate Speech Spreaders on Twitter.
9. Aouchiche, Imane & Boumahdi, Fatima & Madani, Amina & Remmide, Mohamed Abdelkarim. (2023). Hate Speech Prediction on Social Media. SN Computer Science. 4. 10.1007/s42979-023-01668-6.
10. Moin Ahmed, Mohit Goel, Raju Kumar, Aruna Bhat, "Sentiment Analysis on Twitter using Ordinal Regression", 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), pp.1-4, 2021.
11. Mitushi Raj, Samridhi Singh, Kanishka Solanki, Ramani Selvanambi, "An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques", SN Computer Science, vol.3, no.5, 2022.
12. Sepideh Saeedi Majd, Habib Izadkhah, Shahriar Lotfi, "Detection of Multiple Emotions in Texts Using Long Short-Term Memory Recurrent Neural Networks", 2022 8th International Conference on Web Research (ICWR), pp.29-33, 2022
13. Zhuqing Yang, Liya Zhou, Zhengjun Jing, "A Novel Affective Analysis System Modeling Method Integrating Affective Cognitive Model and Bi-LSTM Neural Network", Computational Intelligence and Neuroscience, vol.2022, pp.1, 2022.
14. Kapil, Prashant & Ekbal, Asif. (2020). A deep neural network based multi-task learning approach to hate speech detection. Knowledge-Based Systems. 106458. 10.1016/j.knosys.2020.106458.
15. R. Jayakrishnan, G.N. Gopal, and M.S. Santhikrishna, "Multiclass emotion detection and annotation in malayalam novels," 2018 Int. Conf. on Comput. Commun. and Inform. (ICCCI), Jan. 2018.
16. M. Ahlgren. 40C Twitter Statistics & Facts. Accessed: Sep. 1, 2021.
[Online].Available: https://www.websitehostingrating.com/twitterstatistics/
17. C. Chen, R. Zhuo, and J. Ren, "Gated recurrent neural network with sentimental relations for sentiment classification," Inf. Sciences, vol. 502, pp. 268–278, Oct. 2019.
18. [18] R. Alshalan and H. Al-Khalifa, ‘‘A deep learning approach for automatic hate speech detection in the Saudi Twittersphere,’’ Appl. Sci., vol. 10, no. 23, p. 8614, Dec. 2020.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING