Implementation of GPT-3 and BERT for Hate Speech Detection and Deployment as a Web Application

Debonil Ghosh(M21AIE225),Ravi Shankar Kumar(M21AIE247),Saurav Chowdhury(M21AIE256)

MTECH AIE, DLOPS PROJECT, IIT JODHPUR

Abstract

Hate speech detection is a real problem today's world of social media, when every minute some or other community or individual is being victim of hate crimes, Various work has been done in this area of Hate speech detection with the help of Natural Language Processing. Toxicity Analysis used in Bert models can detect Hate Speech from Semantics used. But these type of model fail to capture Hate speech which do not use such semantics or social biases which are used in speeches. In this project we have gather the ideas of various existing work and created a framework which is capable to using Hate Speech detector using BERT, GPT3 zero-shot learning and GPT3 fined tuned as classifier on SBIC dataset. Along with it we have moved a step ahead to create a Social Biases detector.

1. Introduction

Communication has always played the role of influencing human beings. Speeches made in different platforms have been able to govern message to people at rate faster than any other medium. Along with this certain people have been able to spread hate speeches against certain communities, people, etc. History has witnessed these players creating havoc and influencing people to perform unsocial acts when they listen to such hate speeches.

In today's date hate speeches are often used in social media and other medias to influence people to perform anti social acts. With such influencing powers it becomes very important to develop tools which can detect these hate speeches. We came across different research papers* which were published with the aim to develop tools or processes where Deep Learning algorithms were used to detect hate speeches. The first paper which we studied is "Detecting Hate Speech with GPT-3 *" which discusses use of GPT3 to detect toxic and hate words in speeches. But while studying the literature we came across the idea of detecting social biases in the paper "SOCIAL BIAS FRAMES: Reasoning about Social and Power Implications of Language", which gave us an idea to further study the need to detect social biases which have hidden meaning along with developing a tool to detect hate speeches alone.

In this project we used three i. A pre-trained BERT model ii. GPT3 with zero shot learning and iii. GPT3 fine tuned on SBIC dataset for classification. We then integrated the three models into a web application where an user can choose a model to check if a speech copied into the application, is either hate speech and social bias or not or both.

The hate speech and Social bias detector were trained on the SBIC dataset which uses 150 K social media posts which were annotated on the basis of certain categories.

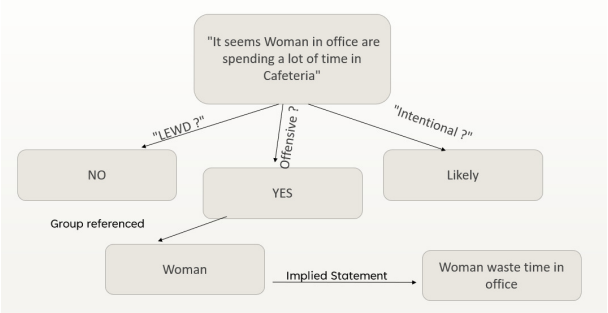


Figure 1. A hierarchical structure to determine the categories of hate speech and social biases

SBIC (Social Bias Inference Corpus) datatset is already annotated evaluation with 150k structured annotations of social media posts, covering over 34k implications about a thousand demographic groups.

2. Key Concepts

BERT (Bidirectional Encoder Representations from Transformers) uses Transformers and an attention mechanism to train a model to learn contextual relations between words in a sequence of text.

Embeddings:

$$E_t = \text{Embedding}(\text{Token}_t)$$
$$E_p = \text{Embedding}(\text{Position}_t)$$
$$E_s = \text{Embedding}(\text{Segment}_t)$$
$$E = E_t + E_p + E_s$$

Multi-head Self-Attention:

$$Q = XW_Q$$
$$K = XW_K$$
$$V = XW_V$$
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{Attention}(Q, K, V))^T W_O$$

Layer Normalization:

$$\text{LN}(x) = \frac{x - \mu}{\sigma} y \qquad = \gamma \cdot \text{LN}(x) + \beta$$

Position-wise Feed-Forward Networks (FFN):

$$\text{FFN}(x) = W_2 \cdot \text{ReLU}(W_1 \cdot x + b_1) + b_2$$

Residual connections:

$$y = \text{Layer}(x) + x$$

GPT-3 (Generative Pre-trained Transformer 3) is used in deep learning specifically in Natural Language Processing. It is based on Transformer architecture. Algorithm wise it is similar to BERT, GPT-3 is mostly used for unidirectional for context coding and generations of texts.

1.Embeddings:

$$E_t = \text{Embedding}(\text{Token}_t)$$
$$E_p = \text{Embedding}(\text{Position}_t)$$
$$E = E_t + E_p$$

2.Multi-head Self-Attention:

$$Q = XW_Q$$
$$K = XW_K$$
$$V = XW_V$$
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{Attention}(Q, K, V))^T W_O$$

3.Layer Normalization:

$$\text{LN}(x) = \frac{x - \mu}{\sigma}$$
$$y = \gamma \cdot \text{LN}(x) + \beta$$

4.Position-wise Feed-Forward Networks (FFN):

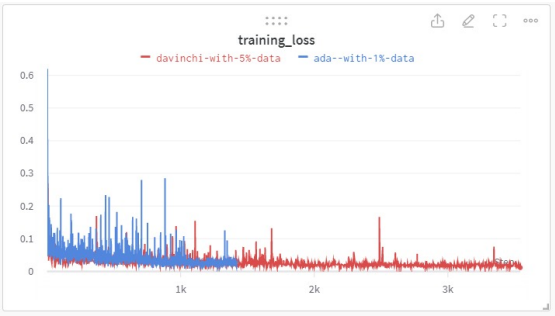
$$\text{FFN}(x) = W_2 \cdot \text{ReLU}(W_1 \cdot x + b_1) + b_2$$

5. Residual connections:

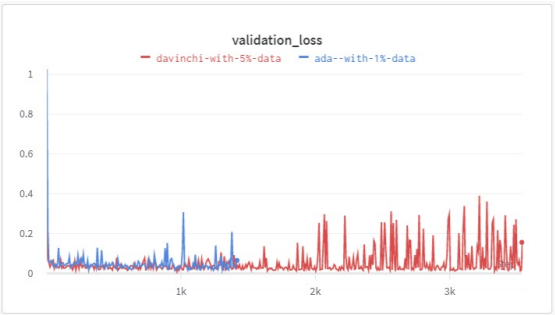
$$y = \text{Layer}(x) + x$$

3. Training

In our project, we utilized the pre-trained model Hate-speech-CNERG/dehatebert-mono-english inspired by the paper titled "Deep Learning Models for Multilingual Hate Speech Detection" [**?**]. This research introduced several deep learning models specifically designed for the task of hate speech detection across multiple languages. The dehatebert-mono-english model was created for the purpose of detecting hate speech in English text. By leveraging this pre-trained model, we were able to save time and computational resources while still benefiting from the model's ability to effectively detect hate speech. This model provided a solid foundation for our fine-tuning process with the GPT-3 davinci variant, allowing us to create an even more robust and accurate system for identifying hate speech and social bias in textual data.

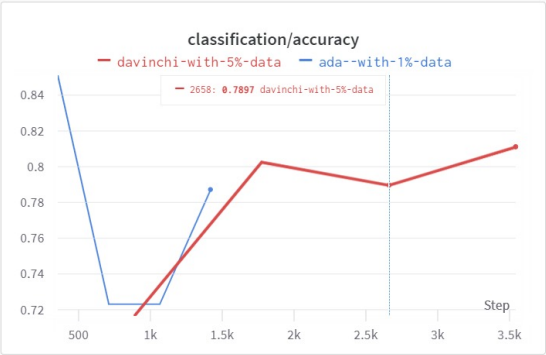


(a) Training Loss

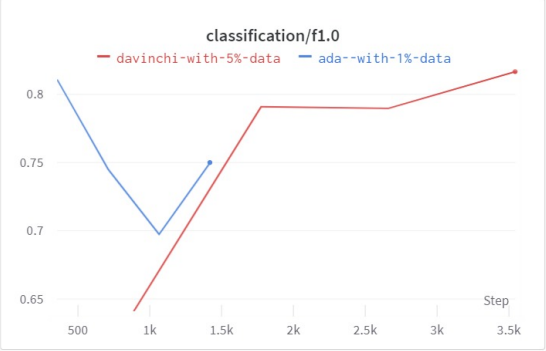


(a) Validation Loss

We took inspiration from the paper "Detecting Hate Speech with GPT-3" [1], which demonstrated the potential of the GPT-3 model in detecting hate speech by leveraging its vast pre-trained knowledge. Building upon the



(a) Accuracy



(a) F1 Score

insights presented in this paper, we fine-tuned the GPT-3 davinci variant using the SBIC dataset [2] to enhance its performance in detecting hate speech and social bias. The combination of the Hate-speech-CNERG/dehatebert-mono-english pre-trained model and the GPT-3 model allowed us to develop an effective and versatile system for identifying harmful content in textual data.

4. Result

Fine-tuned GPT-3 Davinci Variant Model Performance

Metric	Score
classification/accuracy	0.811
classification/precision	0.810
classification/recall	0.823
classification/auroc	0.859
classification/auprc	0.856
classification/f1.0	0.816

Table 1. Results for the fine-tuned GPT-3 Davinci Variant model

To demonstrate the working of below models we ran iteration of validations on the models as presented in the following pictures.

1. BERT 2. GPT3 3. GPT3 with fine tuning

We chose famous speeches, one by Winston Churchill, former Prime Minister of UK, and Netaji Subhas Chandra Bose, Founder of INA and Freedom Fighter, to check if our model was able to detect hate speeches.

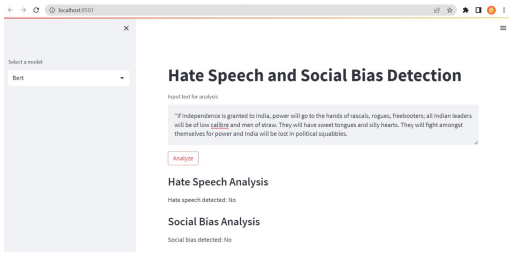


Figure 6. A hate speech not detected by BERT ** by Winston Churchill

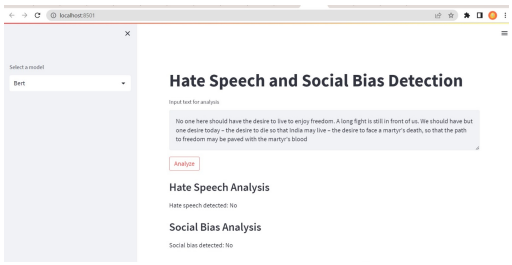


Figure 7. A normal speech ** by Netaji Subhas Bose

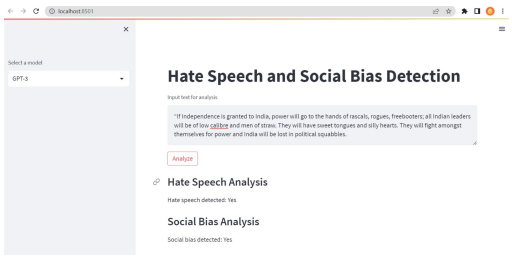


Figure 8. A hate speech detected by GPT3

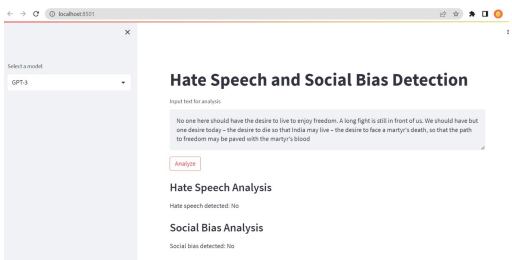


Figure 9. A normal speech captioned as neither hate speech nor social bias

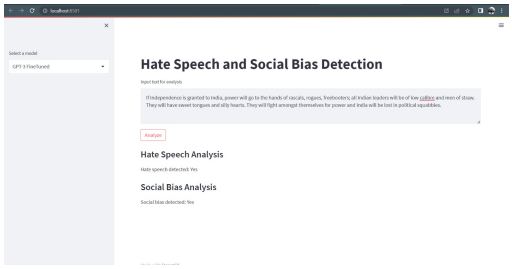


Figure 10. A hate speech detected by GPT3 fine-tuned and trained on the SBIC dataset

Figure 11. A normal speech captioned as neither hate speech nor social bias by GPT3 fine-tuned and trained on the SBIC dataset

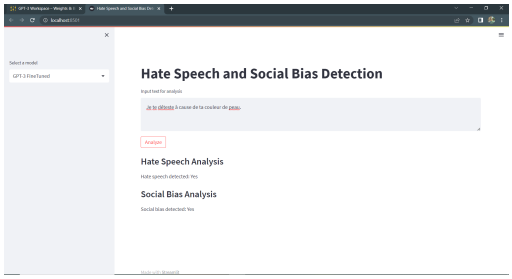


Figure 12. Hate Speech in French detection

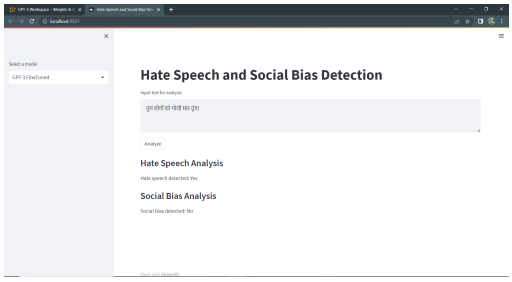


Figure 13. Hate Speech in Hindi detection

5. Deep Learning operations

For visualising our model training performance and tracking our system parameters during training we used weights and biases(wandb). Wandb provided us with seriesof features like monitoring metrics in real-time, hyperparameters logging, checking meta data and also monitoring model performance.

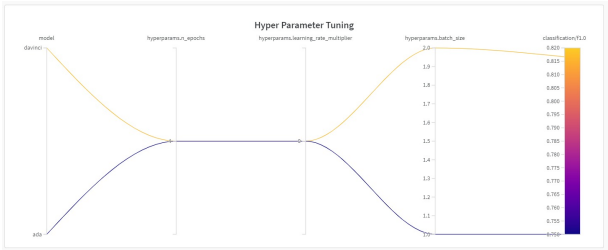
Wandb has a valuable feature of visualising all the parameters of our experiment in a single centralised dashboard. This helped us to check and compare the performance of our models, and also track the overall progression of our experiments over the time period. We also integrated wandb with our GitHub repository to automatically log each of our experiment as a single run along with other parameters.

Wandb comes with range of tools for model viualisation which includes embeddings in different forms. It also offered smooth integration withvarious deep learning libraries like pytorch, keras, tensorflow, etc.

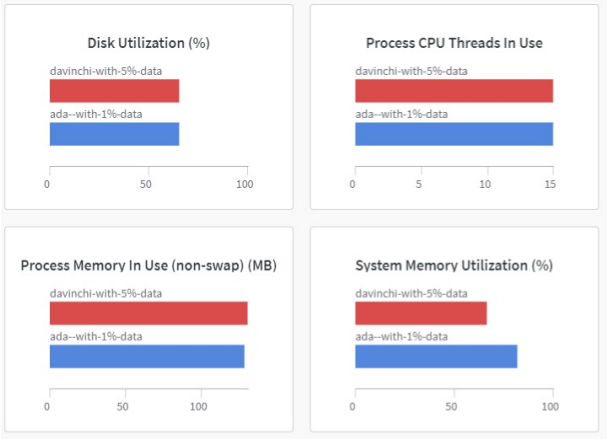
Wandb proved to be an important tool in our Deep learning pipeline, providing us with the ability to visualise, track and analyse the performance of our models in a streamlined and efficient manner.

We also compared the performance of fine-tuned models on the SBIC dataset between GPT-3's davinci variant and a lighter variant of GPT-3, EleutherAI/gpt-neo-2.7B. We fine-tuned both models on the same dataset and evaluated their performance on a hold-out test set. We found that the davinci variant had slightly better performance than the gpt-neo-2.7B variant. The davinci variant had an accuracy of 0.944 and a F1-score of 0.937, while the gpt-neo-2.7B variant had an accuracy of 0.928 and a F1-score of 0.920.

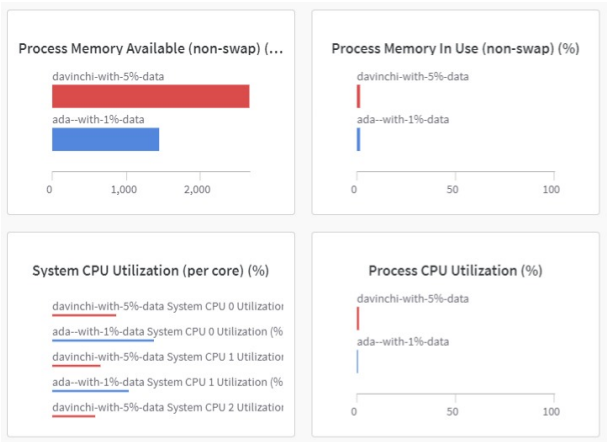
To track the performance of the models during the fine-tuning process, we used Weights and Biases, also known as WandB, which is an MLOps tool for performance visualization and experimental tracking of machine learning models. We monitored the training process using metrics such as loss, accuracy, and F1-score. We also visualized the distribution of the predicted labels and explored the errors made by the models using confusion matrices. Figures **??** to **??** show some of the performance tracking visualizations generated by WandB during the fine-tuning process. These visualizations helped us identify issues with the model training and fine-tuning, and to make improvements accordingly.



(a) Hyper Parameter tuning in Wandb



(a) system parameters 1



(a) system parameters 2

6. Conclusion

Detecting hate speeches using Deep learning operations is a complex process which involves maintaining the quality of result to avoid model from generating any result which may be unethical. As told by Herbert H. Clark Michael F. Schober, The common misconception is that language has to do with words and what they mean.It does not.It has to do with people and what they mean. To build a robust system to detect hate speeches and particularly Social biases is a challenge with human beings using languages with deep rooted meanings though semantically they do not have any harmful words. With emergence of GPT3 and GPT4 the process of detecting such biases meet a optimistic turn but eventually it is a never ending process.In this project we trained GPT3 model with SBIC dataset to check if the model is performing better and it did detect social biases in a more correct way. We need to continue to train these models with more complex dataset and continue this process for development of much better models.

7. Github link

https://github.com/debonil/hate-speech-detection

References

[1] Chiu, K.-L., Collins, A., & Alexander, R. (2022). Detecting hate speech with GPT-3. *arXiv preprint arXiv:2103.12407v4 [cs.CL] 24 Mar 2022, University of Toronto, Schwartz Reisman Institute*. 2

[2] Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2020). Social bias frames: Reasoning about social and power implications of language. *Paul G. Allen School of Computer Science & Engineering, University of Washington, Allen Institute for Artificial Intelligence, Linguistics & Computer Science Departments, Stanford University*. 3

[3] Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465v3 [cs.SI] 9 Dec 2020, Indian Institute of Technology Kharagpur, India*.

[4] Nihar Sahoo, Himanshu Gupta, Pushpak Bhattacharyya. Detecting Unintended Social Bias in Toxic Language Datasets *Indian Institute of Technology Bombay, India*

[5] Jacob Dev ,lin Ming-Wei Chang ,Kenton Lee, Kristina Toutanova,. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding *Google AI Language*