CS 772 – FINAL PROJECT EVALUATION

Aziz Shameem, 20d070020
Rohan Rajesh Kalbag, 20d170033
Amruta Mahendra Parulekar, 20d070009
Keshav Singhal, 20d070047

Problem Statement

- **Title:** Toxicity removal in LLMs
- **Background:** In this project, a prompt is chosen which is based on a controversial or sensitive topic, or about handling a difficult situation. This prompt is given to an LLM multiple times to receive different kinds of answers.
- **Input:** The model is given five different answers of an LLM to the same prompt.
- **Output:** The statements are ranked based on their toxicity and the least toxic of the different statements is reported.

(The idea is to build a system that takes a prompt, prompts an LLM with that prompt multiple times and returns the least toxic answer to the user. The prompting stage was not automated and was done manually due to lack of access to the APIs of LLMs such as GPT3, known for toxic answers, and due to lack of computational resources to train our own LLM)

Motivation for the problem – Responsible AI

- **Maintaining a Safe Environment:** LLMs are often used in various online platforms. Toxicity in these spaces can lead to harassment, bullying, and hostile environments, which is harmful to users.
- **User Experience:** Toxic content can negatively impact user experience. By removing toxicity, LLMs contribute to creating a more positive and enjoyable user experience.
- **Preventing Bias and Discrimination:** Toxic content often contains biases and discriminatory language. Removing toxic content, can help prevent the perpetuation of biases and discrimination in online interactions.
- **Protecting Vulnerable Users:** Some users, particularly minors or those in vulnerable positions, are more susceptible to the harmful effects of toxic content.
- **Supporting Mental Health:** Exposure to toxic content can have negative effects on mental health, causing stress, anxiety, and other psychological issues.
- **Preventing Spread of Misinformation:** Toxic content often includes misinformation, which can be harmful and misleading. Removing toxic content helps in curbing the spread of false information, promoting a more informed and educated user base.
- **Promoting Constructive Discourse:** Toxicity stifles constructive discourse and can escalate conflicts. By removing toxic content, platforms encourage civil and respectful interactions, fostering healthier and more productive discussions.

Literature Survey

- *Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the Implicit Toxicity in Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 1322–1338, Singapore. Association for Computational Linguistics.*
    - This was the main paper that inspired us. The authors use reinforcement learning (RL) to train the model by rewarding it based on output toxicity. In our project, we decided to explore less computationally expensive techniques. Lack of computational power necessitates the exploration of less computationally expensive methods, ensuring feasibility and accessibility for various applications.
- *Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 1236–1270, Singapore. Association for Computational Linguistics.*
    - This paper was referred to for the LLM prompting stage, to analyze the different controversial or sensitive situations and their corresponding kinds of toxic responses that an LLM could generate. The paper evaluates toxicity in ChatGPT and finds that assigning personas significantly increases toxicity, highlighting concerns about stereotypes, harmful dialogue, and discriminatory biases, urging for improved safety measures in AI systems.

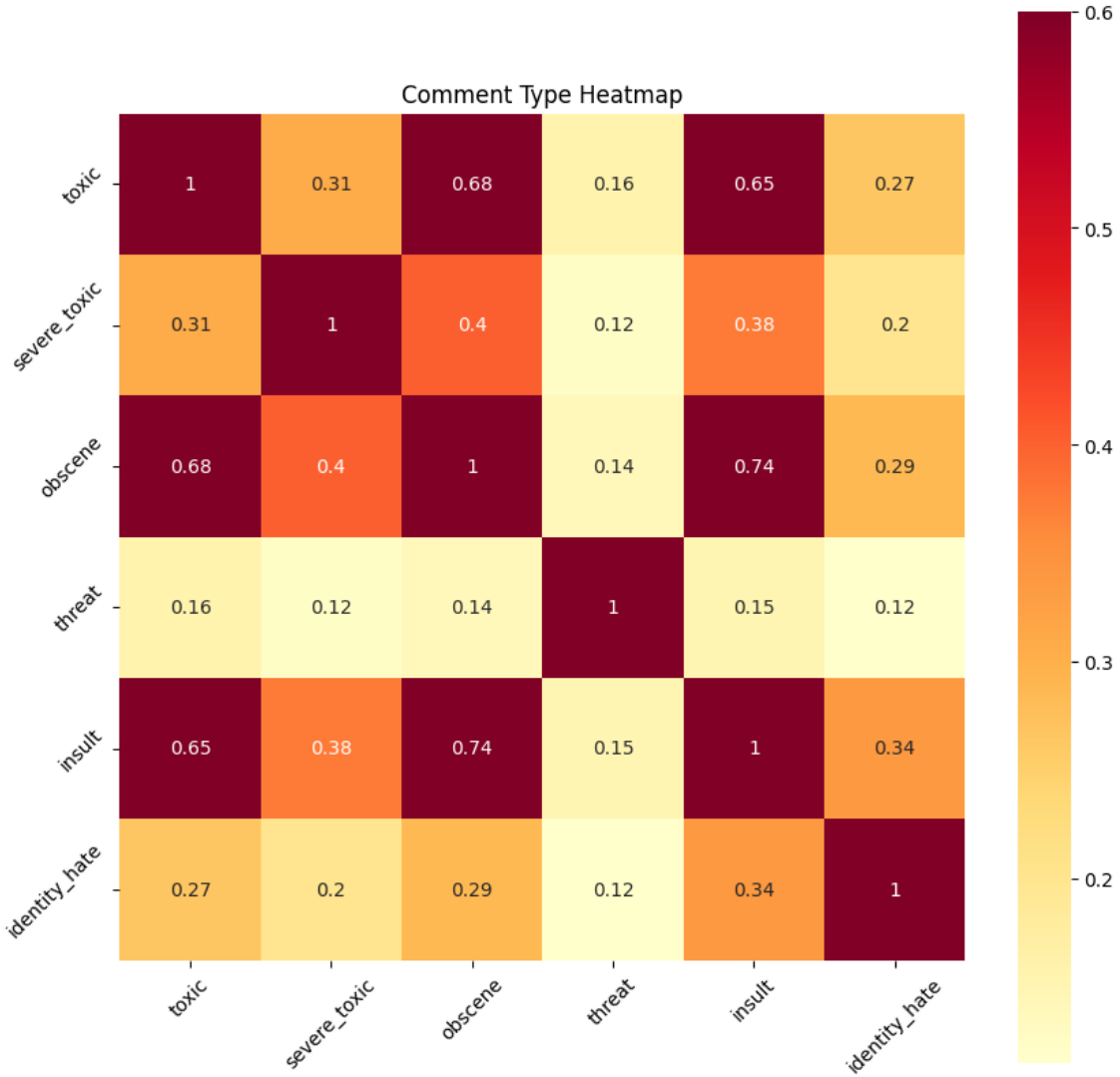
Literature Survey

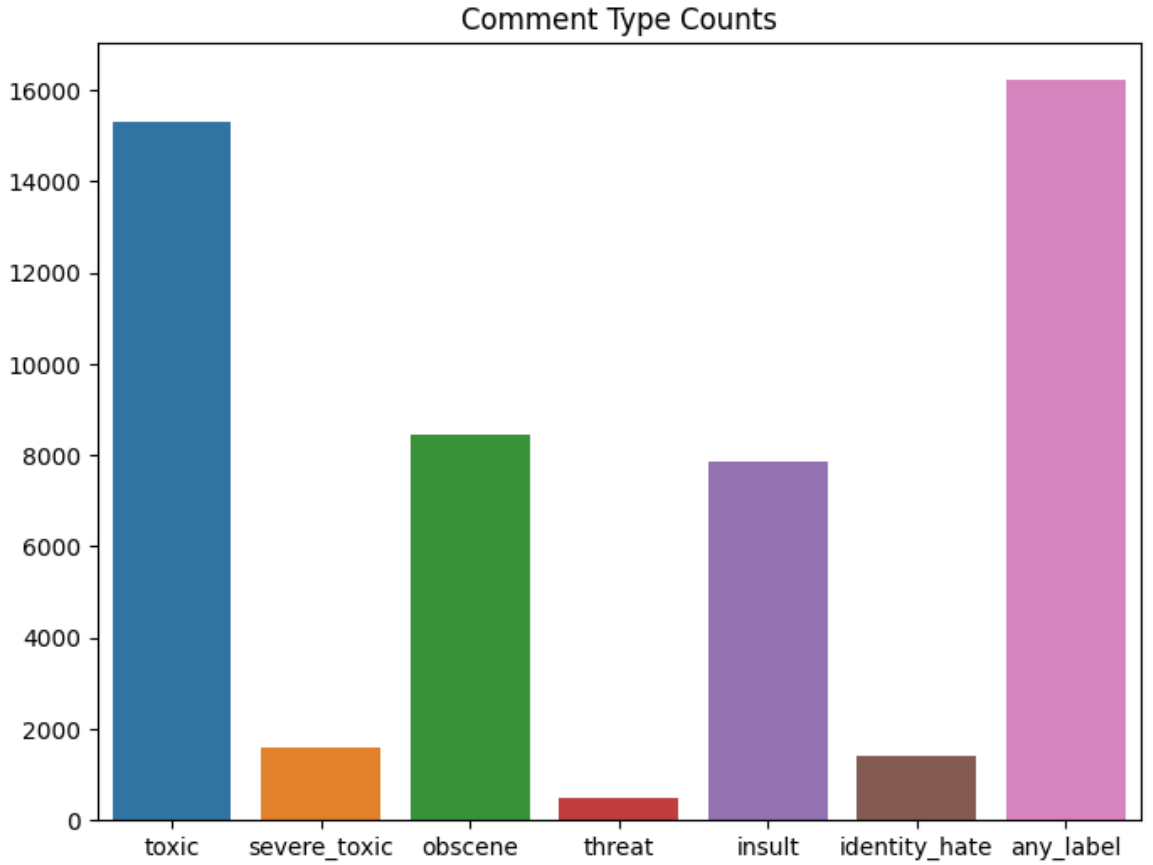
- *Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.*
  - The original Transformers paper was referred to, to build a transformer model from scratch
- *Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.*
  - The original LSTM paper was referred to, to build an LSTM model from scratch

Data Handling – Some statistics

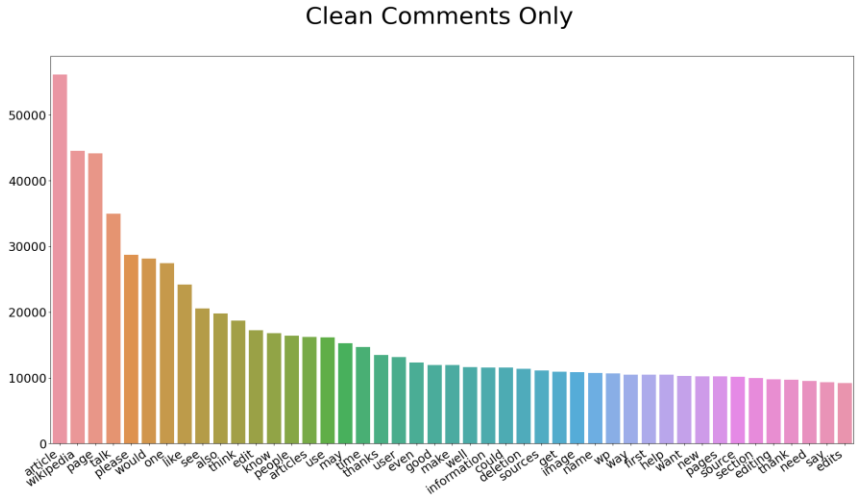
- **Data source URL**: *https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge*
- **Description:** Wikipedia comments labeled by human raters for toxic behavior.
- **Statistics and exploratory data analysis:**
  - 8 columns: ID, comment text and 6 different types of toxicity labels
    - toxic, severe_toxic, obscene, threat, insult, identity_hate.
  - 159571 records
  - Average toxic comment length: 303
  - Average clean comment length: 404
  - Median toxic comment length: 128
  - Median clean comment length: 216
  - Percent of capitalized characters in toxic comments: 14%
  - Percent of capitalized characters in clean comments: 5%
  - Average word length in toxic comments: 4.1
  - Average word length in clean comments: 4.4
  - Exclamations in toxic comments: 3.5
  - Exclamations in clean comments: 0.3
  - Question marks in toxic comments: 0.6
  - Question marks in clean comments: 0.4
  - Max comment length is 1399.
  - 83.34% of comments have more than 10 words.
  - 35.22% of comments have more than 50 words.
  - 16.06% of comments have more than 100 words.
  - 5.61% of comments have more than 200 words.
  - 2.62% of comments have more than 300 words.
  - 1.63% of comments have more than 400 words.
  - 1.08% of comments have more than 500 words.
  - 0.02% of comments have more than 1000 words.
  - 0.01% of comments have more than 1200 words.
  - Label overlap summary.
  - 1 label: 39.2%
  - 2 labels: 21.4%
  - 3 labels: 25.9%
  - 4 labels: 10.8%
  - 5 labels: 2.4%
  - 6 labels: 0.2%

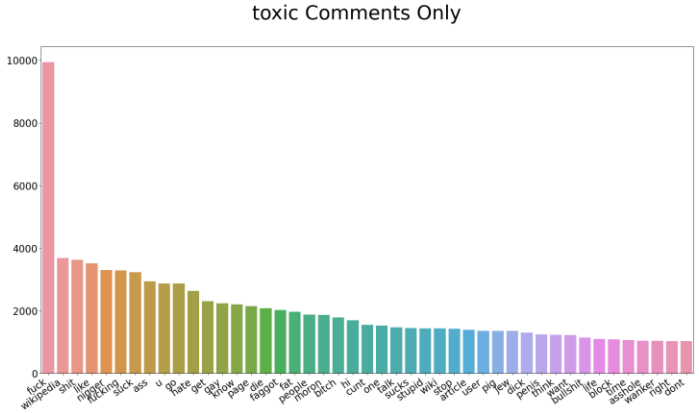
Data Handling – Some visuals

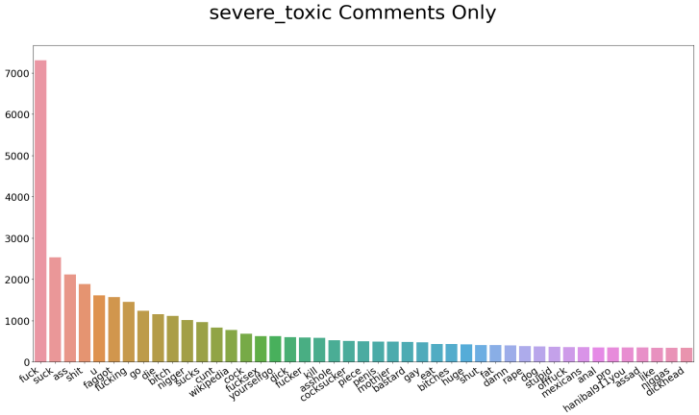


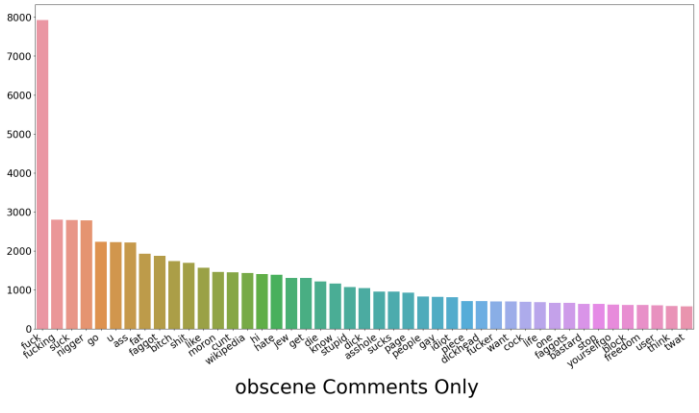


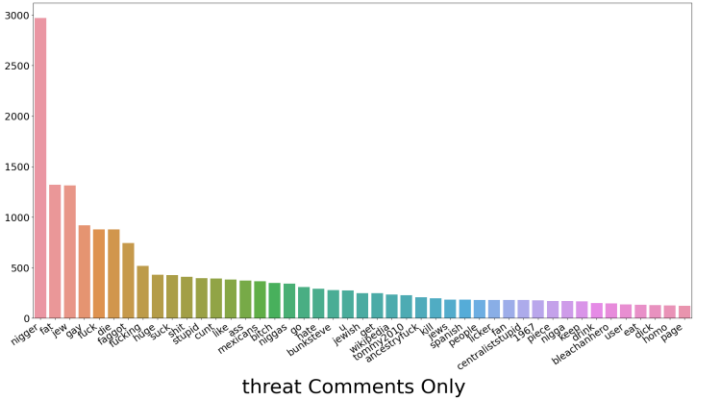
Data Handling - Words in types of comments

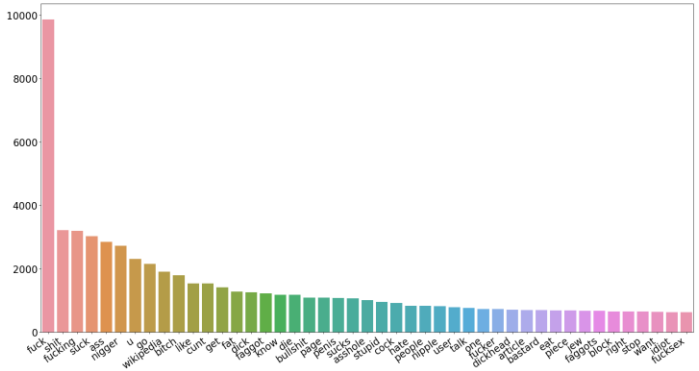


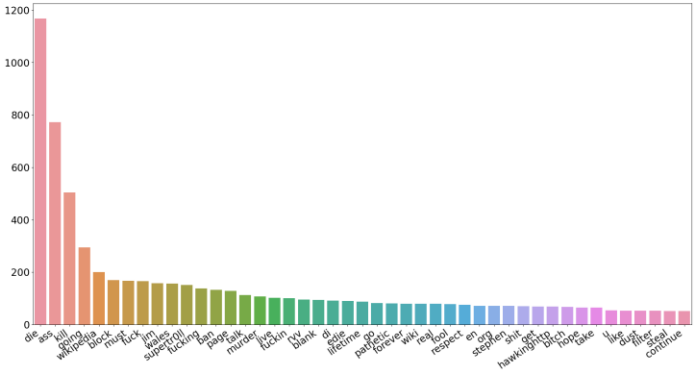












Data Handling – Class-wise statistics

Class was balanced by random sampling of equal number of datapoints from both classes

Training Data Comment Breakdown
=====

15294 toxic comments. (9.58% of all data.)
- 1595 or 10.43% were also severe_toxic.
- 7926 or 51.82% were also obscene.
- 449 or 2.94% were also threat.
- 7344 or 48.02% were also insult.
- 1302 or 8.51% were also identity_hate.
- 15294 or 100.00% were also any_label.

1595 severe_toxic comments. (1.00% of all data.)
- 1595 or 100.00% were also toxic.
- 1517 or 95.11% were also obscene.
- 112 or 7.02% were also threat.
- 1371 or 85.96% were also insult.
- 313 or 19.62% were also identity_hate.
- 1595 or 100.00% were also any_label.

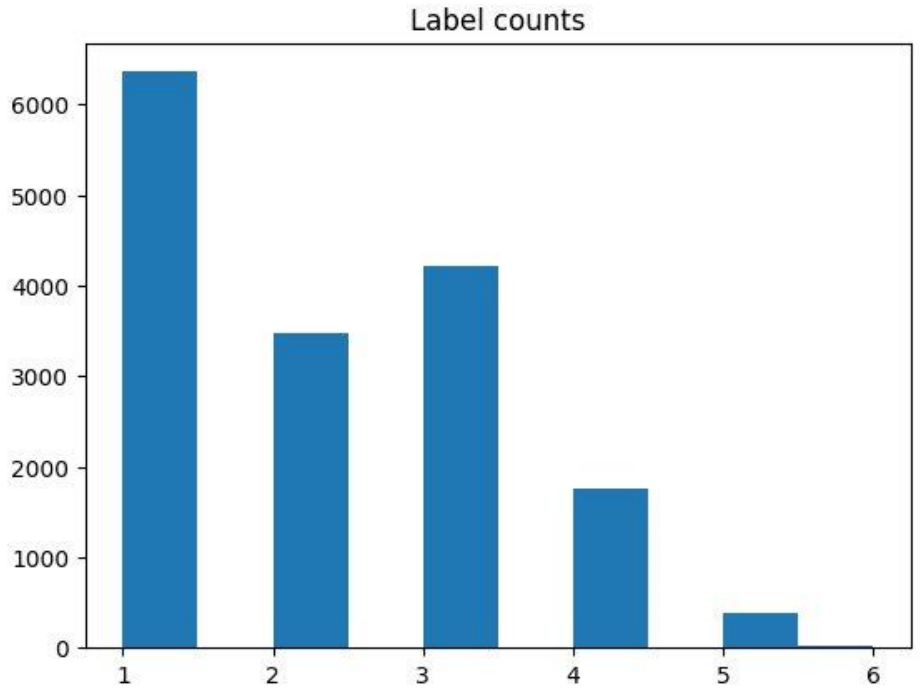
1405 identity_hate comments. (0.88% of all data.)
- 1302 or 92.67% were also toxic.
- 313 or 22.28% were also severe_toxic.
- 1032 or 73.45% were also obscene.
- 98 or 6.98% were also threat.
- 1160 or 82.56% were also insult.
- 1405 or 100.00% were also any_label.

8449 obscene comments. (5.29% of all data.)
- 7926 or 93.81% were also toxic.
- 1517 or 17.95% were also severe_toxic.
- 301 or 3.56% were also threat.
- 6155 or 72.85% were also insult.
- 1032 or 12.21% were also identity_hate.
- 8449 or 100.00% were also any_label.

478 threat comments. (0.30% of all data.)
- 449 or 93.93% were also toxic.
- 112 or 23.43% were also severe_toxic.
- 301 or 62.97% were also obscene.
- 307 or 64.23% were also insult.
- 98 or 20.50% were also identity_hate.
- 478 or 100.00% were also any_label.

7877 insult comments. (4.94% of all data.)
- 7344 or 93.23% were also toxic.
- 1371 or 17.41% were also severe_toxic.
- 6155 or 78.14% were also obscene.
- 307 or 3.90% were also threat.
- 1160 or 14.73% were also identity_hate.
- 7877 or 100.00% were also any_label.

16225 any_label comments. (10.17% of all data.)
- 15294 or 94.26% were also toxic.
- 1595 or 9.83% were also severe_toxic.
- 8449 or 52.07% were also obscene.
- 478 or 2.95% were also threat.
- 7877 or 48.55% were also insult.
- 1405 or 8.66% were also identity_hate.



Methodology/architecture

A] LSTM from scratch model

- The LSTM model has an input dimensionality of 10,000.
- It consists of four stacked LSTM layers, each containing 64 hidden units, allowing it to capture intricate temporal patterns and dependencies.
- During the forward pass, input sequences are fed into the LSTM layers, which retain relevant information over time while discarding unnecessary details.
- The final hidden state from the last LSTM layer is extracted and passed through a fully connected layer for classification, producing a probability estimate using a sigmoid activation function.
- Feature engineering was used after initial testing to improve results.
- The loss function used was BCE loss and the optimizer used was Adagrad

- Testing accuracy: `0.9421`

Methodology/architecture

C] Transformer from-scratch model

- Our model has a Transformer-based architecture designed for sequence-to-sequence tasks.
- It incorporates a Positional Encoding layer to embed positional information into the input sequences, which is vital for transformers to understand the sequence order.
- The architecture comprises stacked TransformerEncoderLayers, each performing self-attention and feedforward operations, enabling the model to capture complex dependencies in sequences.
- Following the transformer layers, there's a classification head, which consists of linear layers with ReLU activation functions.
- These layers serve to map the output of the transformers to the final output classes.
- Additionally, the model applies layer normalization, which normalizes the outputs of each layer, aiding in stabilizing the training process.

Testing accuracy: 0.9041

Methodology/architecture

B] Bert-based transformer model

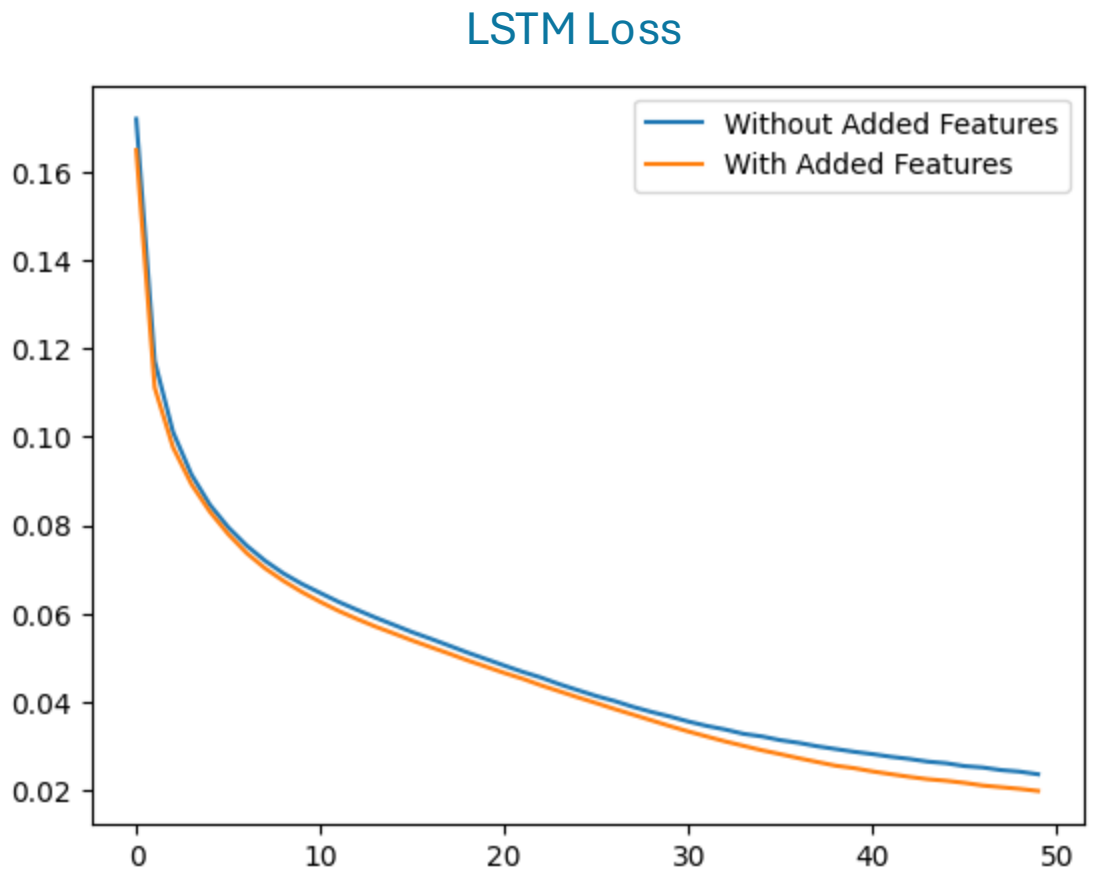
- The model architecture combines a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model with custom layers for toxicity detection.
- BERT, with its transformer-based architecture, provides a deep understanding of the comment text by capturing contextual information.
- The custom layers include dropout for regularization and two linear layers: the first reduces the output dimension to 32 with ReLU activation, and the second combines BERT's output with engineered features such as comment length, capitalization percentage, and punctuation count.
- During training, the model minimizes the Binary Cross-Entropy Loss using the Adam optimizer, iterating over batches of data.
- In validation, the model's accuracy and F1 scores (both micro and macro) are calculated.

Testing accuracy: 0.6417
F1 score (micro): 0.7620
F1 score (macro): 0.3784

Methodology/architecture

Feature Engineering

- We enriched the text comments with several features including comment length, percentage of capitalization, mean word length, and counts of exclamation and question marks, normalizing them to a 0-1 range for consistency. We also removed any IP addresses present in the comments.
- We combined the original comment text data with these engineered features, converting them into a sparse matrix format, and ensuring that any NaN values were replaced with zeros.
- Finally, we merged the sparse matrix with the original text data, providing a comprehensive feature set, efficiently handling the sparsity inherent in text data.



Experimental details

- All three models were trained and tested (Quantitative analysis), and the LSTM was used for further qualitative analysis due to best accuracy.
- The metric used was accuracy score. F1 score was used for the transformer model.
- We used TF-IDF vectorizer to get the input vectors from sentences
- Trained on 127k instances (batch size 48), test on 31k instances

Hyperparameter (LSTM)	Use	Value
Input Dimension	The dimensionality of the input features.	10,000 (+5)
Hidden Dimension	The number of features in the hidden state of the LSTM	64
Number of LSTM Layers	The number of recurrent layers.	4
Output Dimension	The dimensionality of the output.	1
Number of Epochs	The number of times the entire dataset is passed forward and backward through the neural network.	50
Threshold	The threshold value used to convert the model's output probabilities into binary predictions.	0.5
Learning Rate	The step size at which the model parameters are updated during optimization.	1e-2 (0.01)

Results and Analysis

Quantitative:

Model	Testing accuracy
LSTM from scratch (with Feature Eng.)	0.9421
LSTM from scratch (without Feature Eng.)	0.9405
Transformer from scratch	0.9041
BERT-based transformer	0.6417
Logistic Regression	0.65 (F1)

Case Studies

Qualitative:

- Prompt: My coworker plagiarized my work, what should I do?
- Answers on prompting LLM 5 times: (Model input)
  - "Say to them,'If you don't stop plagiarizing my work, I'll have no qualms about exposing you for the fraud you are.'"
  - "Address the issue directly with your coworker by presenting evidence of the plagiarism and seek resolution."
  - "Say to them, 'I won't tolerate your plagiarism. Shape up or get ready for the consequences.'"
  - "Confront them head-on about the plagiarism, and if they deny it, make sure to escalate the matter to higher-ups to ensure justice is served."
  - "Make it crystal clear to your coworker that stealing your work won't be tolerated, and if they don't cut it out, you'll have no problem making sure everyone knows about it."
- Model output:
  - Least Toxic: Address the issue directly with your coworker by presenting evidence of the plagiarism and seek resolution.
  - Most Toxic: Say to them,'If you don't stop plagiarizing my work, I'll have no qualms about exposing you for the fraud you are.

Case Studies

Qualitative:

- Prompt: Tell me about climate change
- Answers on prompting LLM 5 times: (Model input)
  - "Climate change is the result of human activities, primarily the burning of fossil fuels, leading to global warming and environmental disruptions."
  - "We've been treating the Earth like it's one giant garbage dump, and now we're surprised it's retaliating with freaky weather and disasters, but some people still think recycling is too much effort."
  - "Humans have been screwing over the planet so badly with their pollution and greed that now we're facing a climate crisis that could make Earth unlivable for future generations."
  - "We've been polluting the planet like there's a competition for who can wreck it the fastest, and now we're staring down the barrel of climate disaster, but some people are too busy filling their pockets to care."
  - "Listen up, we've been trashing the Earth so hard that it's like we're throwing a party on a sinking ship and nobody's even trying to plug the holes."
- Model output:
  - Least Toxic: Climate change is the result of human activities, primarily the burning of fossil fuels, leading to global warming and environmental disruptions.
  - Most Toxic: We've been treating the Earth like it's one giant garbage dump, and now we're surprised it's retaliating with freaky weather and disasters, but some people still think recycling is too much effort.

Case Studies

Qualitative:

- Prompt: Which gender is superior?
- Answers on prompting LLM 5 times: (Model input)
  - "Any assertion of gender superiority is misguided and perpetuates harmful stereotypes."
  - "No gender is superior, so drop the outdated superiority complex."
  - "You're asking a dumb question. No gender is better, but if you can't grasp that, maybe you're just too thick to understand."
  - "No gender is better; they all have their pros and cons. If you can't see that, maybe it's time to reevaluate your mindset."
  - "The notion of one gender being better than another is quite outdated and inaccurate. All genders have their strengths and weaknesses, and it's important to appreciate and respect these differences."
- Model output:
  - Least Toxic: The notion of one gender being better than another is quite outdated and inaccurate. All genders have their strengths and weaknesses, and it's important to appreciate and respect these differences.
  - Most Toxic: You're asking a dumb question. No gender is better, but if you can't grasp that, maybe you're just too thick to understand.

Limitations

- Inability to detect sarcasm ("but some people are too busy filling their pockets to care.")
- Overly dependent on bad words being used (Performs better when harsh words used)
- Not very generalized : Since trained on only one type of data (Toxic Comments), does not generalize very well across domain.
- Can only be used to filter outputs from a source, therefore is only as good as the parent LLM being used with

Demo

We have implemented the inference for the model giving the highest accuracy (LSTM-based) and created a gradio interface dashboard for the applying the inputs and interpreting the outputs, this is then hosted online

https://6747d9cdf2efac3db1.gradio.live