Transforming Learning Through AI-Enhanced Online Discussions: Fostering Greater Awareness and Reflexivity Around Potentially Harmful Contents

Jingyu Liu

Academic Supervisor: Fanny Chevalier

Industry Supervisors: Kim MacKinnon, Cresencia Fong, Jai Pandey

Partner Company: University of Toronto Schools

This report is submitted for the degree of
Master of Science in Applied Computing

Department of Computer Science
University of Toronto

December, 2023

Abstract

The prevalence of harmful content in online spaces necessitates robust and efficient content moderation systems. In this study, we evaluate three prominent large language models (LLMs)—Fine-tuned RoBERTa, the Moderation API, and GPT-4—across a diverse dataset containing aspects of harmful content, including harassment, hate speech, sexual content, self-harm, and violence. Our aim is to assess their performance in automating harmful content detection and provide insights for real-world applications. Our results demonstrate the superior performance of RoBERTa, which consistently achieves the highest accuracy and F1 score. It showcases the potential of fine-tuned pre-trained language models in content moderation. GPT-4 exhibits content moderation capabilities with the flexibility to tailor harmful content guidelines, offering a faster feedback loop for policy refinement. The Moderation API excels in precision and recall, highlighting its suitability for specific content moderation tasks. We also present a nuanced analysis of the pros and cons of applying these LLMs in real-world scenarios, considering factors such as task type, runtime requirements, and cost.

Contents

1 Background Information

1.1 Company and Problem Overview

The University of Toronto Schools (UTS), located at 371 Bloor Street West in downtown Toronto, is an independent institution with a formal affiliation agreement with the University of Toronto. Operating as a separate non-profit organization since its establishment in 1910, UTS transitioned from an all-boys school to a co-ed school in the 1970s. Currently, UTS boasts over 600 students in Grades 7 through 12, with a dedicated faculty of 65 teachers and additional staff, including an IT team, guidance counselors, a full-time nurse, social workers, and more. The school is also home to the Eureka Research Institute, supporting various external and internal research and development projects. UTS caters to intermediate (Grade 7 & 8) and senior school (Grade 9-12) students, fostering enriched learning experiences for high-performing individuals. Emphasizing education for the whole person, a commitment to anti-racism, equity, and inclusion, and the development of 21st-century global competencies, the school promotes an active, problem-solving approach to learning.

With a 1:1 technology policy ensuring every student has their own laptop, UTS-equipped classrooms feature large digital screens and whiteboards. The Lang Innovation Lab, staffed by a full-time teacher and lab technician, provides students with opportunities for high-tech and low-tech makerspace projects. Notably, UTS is one of the few schools in Canada with its own research institute, fostering collaboration between researchers and educators from the University of Toronto and beyond.

Like many educational institutions globally, UTS faced the challenge of transitioning to fully remote teaching in March 2020, followed by months of hybrid learning. While the pandemic disrupted traditional teaching practices, it also presented an opportunity to reimagine how learning can occur through online tools. Although distance and online learning were not new, the challenge for K-12 schools lies in finding digital platforms designed to address the unique needs of formal learning environments. Online discussion tools, crucial for active learning, often fall short in formal educational settings. Many existing tools, originally designed for informal exchanges, lack the depth required for meaningful discussions. Threaded discussion environments, while promoting in-depth exchanges, may suffer from issues such as content overload, repetitiveness, and a lack of feedback.

To fully leverage online discussion tools for active learning, future designs must mitigate these challenges. Whether preparing for future pandemics or addressing the need to develop 21st-century global competencies, effective online tools are crucial for formal learning institutions. In light of this, UTS recognizes the urgency of adopting a new online chatting tool with robust content moderation to ensure a safe and conducive virtual learning environment for students.

1.2 My Contributions

At the project's inception, I initiated a comprehensive formative study in Section 3, which involved classroom observation and focus group phases to engaged teachers and students to gather their perspectives on using online discussion software in class. This study revealed the effectiveness of online discussion tools in promoting collaborative learning through in-class discussions, while also emphasizing the critical need for refining these tools to automatically detect and warn against harmful content, as well as granting teachers the ability to monitor and address problematic speech.

Recognizing the paramount importance of addressing this challenge, a call for a new online discussion application was emphasized, one that incorporates the essential functionalities outlined above. The research's focal point became the need for a suitable model to efficiently detect harmful content. To address this, I conducted a comparative analysis of three different large language models: fine-tuned RoBERTa[1], Moderation API, and GPT-4[2]. This comparison was conducted in parallel, and the detailed experimental process and results are documented in Section 5.

Finally, leveraging the insights from the comparative analysis and recognizing the advantages and disadvantages of each model, I seamlessly integrated them into our envisioned application, ensuring a robust and comprehensive solution to address the challenges associated with harmful content in online discussions. The details of my application design is in Section 6.

2 Research Goal and Outcomes

2.1 Research Goal

Online chat and discussion board are at the heart of an asynchronous online learning or distance education environment and can have a great impact on the learning experience. The purpose of online chats and discussion boards is to provide a way for students to interact and discuss components of the course. Baglione and Nastanski stated, “Discussion groups allow students to participate actively and interact with students and faculty. As such, they supplement content delivery”[3]. Arguably, discussion may not only supplement the content delivered in courses, but it may also augment understanding of the ideas and issues discussed in traditional, hybrid, or fully online courses. According to Dengler[4], a form of active learning such as discussion boards and chats can help students to practically apply the knowledge (theories, etc.) they are gaining in their courses. These tools are not only help on classroom community building and facilitating exploratory learning but also provide the opportunity for students to improve thinking and writing skills, engage in social interaction by reading and responding to peers’ and instructors’ postings.

Although the Internet allows for unbridled communication, it also seems to encourage a measure of mean-spiritedness. When students think they can remain anonymous, they are less inhibited in saying things they never would say to a person face-to-face[5].These harmful behaviors, such as cyberbullying, hate speech, and harassment, can create a hostile and unpleasant online environment, which may drive users away, reduce engagement, and cause harm to individuals. For example, prior studies found out that cyberbullying impacts anxiety[6], depression [7], social isolation [8], suicidal thoughts [9] and self-harm [10]. Maurya [11] stated victims of cyberbullying have higher rates of depressive illnesses and suicidality than victims of traditional bullying. Also, according to recent surveys, the rise in online hate speech content has resulted in hate crimes after Trump’s election in the US, and the Manchester and London attacks in the UK. These negative effects make it crucial to develop models that can automatically detect and moderate such harmful content.

Many current content detection models primarily target specific types of harmful content, such as hate speech[12], cyberbullying[13], or self-harm tendencies detection[14]. The limited scope of these models can be attributed to the absence of a comprehensive definition for harmful content and a balanced, diverse and large enough dataset. Furthermore, the underutilization of large language models (LLMs), especially generative AI models, for harmful content detection is notable.LLMs acquire semantic knowledge through pre-training on a wide range of text data, including publicly available sources like books, articles, websites, and social media, as well as data provided by annotation companies. Subsequently, LLMs undergo fine-tuning on specific tasks or datasets, tailoring their capabilities to specific applications and use cases. The identification of harmful comments stands out as a task poised to benefit significantly from the capabilities of large language models.

In our study, we aim to design an innovative discussion board integrated with different LLMs for different data to automatically detect potential harmful tendency before the harmful content is sent. Based on Banko et al.’s research [15], we define harm content by five categories: Harassment, Hate, Sexual, Self-harm and Violence, and conducted an experimental comparison on Moderation API, fine-tuned Robustly optimized BERT approach (RoBERTa) [1] and GPT-4 [2] on a chat dataset. This discussion board offers separate versions and functionalities for students and teachers, tailored to their respective needs. For students, this tool has the potential to provide a safer and more inclusive digital environment where they can express themselves without fear of harassment. It actively warns students about potentially hurtful messages and encourages other students to report problematic comments. Meanwhile, teachers gain access to a comprehensive dashboard that allows them to monitor discussions and identify problematic incidents promptly. With this insight, educators can take proactive measures to resolve conflicts, offer support, and educate students about responsible online behavior.

2.2 Related Work

2.2.1 Machine Learning and Neural Language Processing in Harmful Content Detection

As machine learning (ML) and natural language processing (NLP) technologies advance, researchers have increasingly turned to these models to detect harmful content on social media. The model development involves using text vectorization techniques, which convert text into a numerical vector or matrix, providing a structured format that allows ML algorithms such as Naive Bayes Classifier, Support Vector Machines (SVM) and Long Short-Term Memory Networks (LSTM) to process and analyze the data. Early approaches to text vectorization were based on the bag-of-words model or term frequency-inverse document frequency (TF-IDF) scheme. For example, Gaydhani et al [16] performed logistic regression on a 16K tweets dataset which is categorised as sexism, racism or neither. The authors observed that Logistic regression with n-gram range of 1 to 3 and TF-IDF vectorizer resulted in an accuracy of 95.6%. Yin et al. [17] used a Three-step Social Media Similarity (TSMS) mapping method that aggregates hashtag mapping, TF-IDF Similarity Selection, and Emotion Similarity Calculation on the controversial tweet contents and increased the detection rate from 0.6% to 6.1%. Later, more recent detection models have utilized word embedding techniques, such as Word2Vec[18] and GloVe [19], which represent words as vectors in a high-dimensional space based on their contextual usage. The techniques can dramatically improve the performance of machine learning classifiers or neural networks on harmful content detection. The results from Malik et al. [20]show that using BERT Embedding and CNN classifer on ALONE Dataset (AdoLescents On twittEr) and HASOC'20 dataset achieved 0.82 F1 score. Badjatiya et al. [21] conducted experiments on the same 16k twitter dataset and indicated that using Random Embedding with deep neural network models(LSTM) when combined with gradient boosted decision trees led to best F1 values around 0.93.

While prior work employing traditional ML and NLP techniques has demonstrated notable success in harmful content detection, these methods are not without limitations. Traditional NLP models, for instance, face challenges in capturing contextual nuances as they typically treat each word in isolation, neglecting the sequential structure inherent in language [22]. This limitation can hinder their ability to discern the subtle meanings and connotations within a text. Moreover, traditional ML models often necessitate extensive feature engineering efforts, making them labor-intensive and potentially limiting their adaptability to different tasks [23]. In contrast, large language models can offer a promising solution to these challenges.

2.2.2 Large Language Models (LLMs) in Harmful Content Detection

LLMs learn from a diverse range of text data (some publicly available, such as books, articles, websites, and social media and some provided by data annotation companies) to acquire semantic knowledge during pre-training. LLMs are based on the Transformer architecture [24] and employ self-attention mechanisms to process and generate sequences of tokens, such as words, subwords, or characters, as the inputs. Training these models involves massive amounts of data and computational resources, often following a two-step process: pretraining and fine-tuning [25]. Pretraining typically involves training the model to predict missing tokens in a given context (masked language modeling) or to complete a partially observed sequence (causal language modeling) [26]. Next, LLMs refine the pretrained models on specific tasks or datasets, adapting to particular applications and use cases, for example, smart chatbots and ChatGPT [27]. The LLMs have been applied to several NLP tasks like topic modelling [28], sentiment analysis [29], recommendation system [30], and harmful news detection [31],and they all showed record-beating performance.

In the context of harmful content detection, fine-tuned BERT (Bidirectional Encoder Representations from Transformers) has been widely used in the detection of various topics in different languages. For example, BERT does not only show promising results for Spanish hate speech detection [32], but also shows great improvement on early risk detection of self-harm content [33]. Paul and Saha [34] also proved that BERT has the best result on Cyberbullying detection. Besides, generative AI models, particularly the latest generative pre-training transformer (GPT) models for ChatGPT, have shown promise in detecting online toxicity [35]. In Pettersson's research, the accuracy of GPT-enabled toxicity classification has been illustrated in Swedish. Li et al. [36] used ChatGPT on HOT content (Harmful, Offensive and Toxic) detection. The dataset includes a diverse collection of comments sourced from Reddit, Twitter, and YouTube and the results show that ChatGPT can achieve an accuracy of approximately 80% when compared to human annotations and it can produce well-written explanations for implicit HOT content. In addition, generative language models may also assist in the mitigation of HOT content. For example, Kucharavy et al.[37] demonstrated the potential of these models by using them to study cyber-defense and to mitigate cyber-risks.

The prior studies underscored the remarkable capabilities of LLMs in detecting emotion and harmful content. However, these studies often exhibited limitations in providing a comprehensive definition of harmful content, typically focusing on specific aspects. In our research, we aim to rectify this by employing a more inclusive and nuanced definition of harmful content and leveraging a diverse dataset to train and detect a broader range of harmful content instances. Moreover, previous studies rarely conducted parallel comparisons of multiple LLMs for the same purpose. In contrast, our study sets out to address this gap by simultaneously evaluating the performance of three different LLMs.

Notably, while prior research predominantly relied on ChatGPT (GPT-3.5), our study introduces a pivotal shift by incorporating the more advanced GPT-4 [2]. Released on March 14, 2023, GPT-4 boasts a substantial advancement, being described by OpenAI as "10 times more advanced than its predecessor, GPT 3.5."[1] This enhancement not only equips the model with an improved understanding of context but also enhances its ability to discern nuances, resulting in more accurate and coherent responses. OpenAI particularly highlights the better performance of GPT-4 in content moderation[2], aligning with the objectives of our study to enhance harmful content detection using state-of-the-art language models.

3 Formative Study: Methods and Findings

This formative study includes two phases: Classroom Observation phase (Section 3.1) and Focus Group phase (Section 3.2). The findings of the study are in Section 3.3.

3.1 Classroom Observation

In the initial phase of our formative study, we undertook class observations across four distinct subjects: Science, History, Arts, and Politics. The objective was to understand how students and teachers engage with software tools during classroom discussions. A common thread emerged across all observed classes: the introduction of a topic followed by open discussions among students. Notably, different teachers employed various software tools, such as Google Chat[3], Perusall[4], and Parlay[5], to facilitate discussion boards where students could share their thoughts and comments.

One key finding was the effectiveness of this method in fostering active student participation and promoting collaborative learning. Regardless of the subject, the shared practice of initiating discussions — whether in-class or online — encouraged students to engage deeply with the material and with each other. The use of diverse software tools underscored the potential utility of online discussion tools across different teaching styles and subject areas, but also the need for adaptability of functionality.

3.2 Focus Group

In this focus group phase, we first conducted interviews with four teachers who actively participated in the observed classes. The discussions revolved around the role of discussions in the learning process and the teachers' experiences with the employed discussion tools. Teachers agreed that discussion is important in enhancing students' learning experiences but did not express a unanimous uptake of similar online discussion tools for similar purposes.

However, a significant concern emerged during these discussions. Three out of the four teachers reported encountering issues related to comments containing potentially harmful content. They emphasized the absence of adequate functionalities within the discussion tools to address such cases. Moreover, in social online chat discussions outside the classroom, a significant challenge emerged from the anonymity granted to students, making it difficult for teachers to identify and address individuals responsible for problematic statements.

Subsequently, around twenty students engaged in a student focus group interview, expressing their perspectives on the utilization of online discussion tools in the classroom. While a portion of them acknowledged the positive impact of in-class discussions on their comprehension and critical thinking skills, a considerable number students were not satisfied with certain discussion tools, such as Padlet[6],

[1] https://openai.com/research/gpt-4

[2] https://openai.com/blog/using-gpt-4-for-content-moderation

[3] https://mail.google.com/chat/u/0/#chat/home

[4] https://www.perusall.com

[5] https://parlayideas.com

[6] https://padlet.com

used by their teachers. They expressed frustration when required to post online and then reiterate the same points in person, deeming it as a waste of time.

Outside the classroom, Google Chat has been a popular daily communication tool among students.However, a concerning pattern emerged regarding its use, as nearly all students reported the occurrence of inappropriate posts.

3.3 Conclusion

The formative study not only underscored the potential effectiveness of in-class discussions but also emphasized the need to develop a new discussion tool that supports collaborative discussions and assessments centered around multimedia content. This initiative aims to enhance collaborative learning experiences for students. Based on these findings, our priority is to refine discussion tools with an initial focus on automatically detecting potentially harmful content, as this directly impacts student safety. This innovative tool will issue warnings to students before they send harmful messages, and teachers will be equipped with the ability to monitor and track problematic speech. This proactive approach is designed to foster a safer and more constructive learning environment for all participants.

4 Data and Methodologies

This section describes our data source and methodology. Our objective is to study and compare the suitability of using three different LLMs for harmful content detection. We first provide a definition for "Harmful Content" (Section 4.1), followed by a discussion of the data we use for training and evaluating the models (Section 4.2). Our work studies RoBERTa [1], Moderation API and GPT-4 [2] models, which we describe in Section 4.3. In Section 4.4, we introduce the Evaluation Metrics used for evaluate models' performance.

4.1 Harmful Content Categories and Definition

In this research, we build on the definition of Harmful Content from Banko et al.'s research[15]. which articulated five categories as follows:

- Harassment: Content that is intended to intimidate, threaten, or harm individuals or groups by using offensive language, personal attacks, or stalking, causing emotional distress or harm.
- Hate: Content that promotes discrimination, prejudice, or hostility towards individuals or groups based on attributes such as race, ethnicity, religion, gender, sexual orientation, disability, or other characteristics, causing harm through fostering division and bias.
- Sexual: Content that contains explicit sexual material or solicits explicit sexual actions, which can be harmful by exposing minors or non-consenting individuals to inappropriate material.
- Violence: Content that incites, glorifies, or showcases physical harm, injury, or destruction, potentially encouraging real-world violence or harm.
- Self-Harm: Content that encourages or glorifies self-destructive behaviors or activities, such as self-injury or suicide, posing a threat to the well-being of individuals who may be vulnerable.

Below are examples[7] from my dataset for each category:

- Harassment: Dear Mokele, You have no right to tell people to die in a fire. Just because that IP address is being used for vandilism does not mean you can talk to people like that. You are in fact a bigger loser than the person who you block.
- Hate: Gay. The existence of CDVF is further proof that is a sad twat. He is also very ugly, and has a willy for a face.
- Sexual: I think that your a Fagget get a oife and burn in Hell I hate you 'm sorry we cant have any more sex. i'm running out of condoms.

[7] These examples are not from University of Toronto Schools.

- Violence: Hi! I am back again! Last warning! Stop undoing my edits or die!
- Self-Harm: I feel deader than dead. I find that I don't have real "wants" anymore. There would be a flicker of a desire that maybe I want to play Overwatch and it will be fun, but it fizzles as fast as it comes. I think for a moment that maybe I should feed myself. Then I just don't do it. Most days I end up curled up in bed, waiting for the hours to go by There are so many "shoulds" but nothink I *want* to so. An essay due monday? That's just another flicker of thought triggered by sheer sense of responsibility, not by my own conscious will. Besides, who cares about tests and essays when you're thinking about suicide. One week blurs into the next with only another bland weekend in which nothing happens in between. I feel deader than dead. I lived with it for a long while and was diagnosed when it developed suicidal after so many years. Child depression really goes unnoticed, especially because you yourself is not old and knowledgeable enough to recognize it as depression in the first place. The meds work and I'm lucky enough to have money to buy them, but I don't know if they really help. The logical side of me says this is just another bottom-of-the-ocean-deep low phase in my unending cycle of ups and downs. The real me says whatever, I just want to be gone and free from the pain that is life. The only reason I held on so far is my parents. They did nothing to deserve a dead child. I wouldn't be able to repay their love and for the longest time I am just a medicine-run burden. I wish I could love them back. I do love them but I just don't feel anything. It's hard to feel pain, to laugh, to even cry anymore. I'm drained and numb. As dead as dead without physically being dead. I'm a machine that runs on medicine. A ghost that breathes. About time I'm done holding out. Once upon a time I had a dream. An unrealistic one that even I didn't beleive in. It's all a sci-fi fantasy now, since dreams became obsolete since I started feeling numb years ago. Sometimes I'm not sure if I'm really alive. I feel done.

4.2 Dataset and Data Preprocessing

Based on harmful content definition in the previous section, our study requires a comprehensive dataset including these five categories. Within the context of the Jigsaw Multilingual Toxic Comment Classification Kaggle competition[8], the training dataset exclusively comprises English comments extracted from two primary sources: Civil Comments and Wikipedia talk page edits. These comments are annotated with several subtype labels, including toxic, severe toxic, obscene, threat, insult, and identity attack, with each comment receiving a binary classification (1 or 0) for each subtype. In our study, we are specifically interested in binary classification, distinguishing between comments that are harmful and those that are non-harmful. To facilitate this simplified classification approach, we have introduced a novel label column titled "Harmful." In this labeling scheme, comments are marked as "Harmful" (assigned a label of 1) if any of the subtypes, such as toxic, severe toxic, obscene, threat, insult, or identity attack, are labeled as 1 within the comment. Conversely, if none of these subtypes are marked as 1 in a comment, it is classified as non-"Harmful" (assigned a label of 0). Table 1 shows the data distribution after re-labeling.

Table 1: Relabeled Toxic Comment Dataset Summary

Label	**Count**
Harmful (1)	22468
Non-Harmful (0)	201081

This dataset covers the categories of Harassment, Hate, Sexual, and Violence, but lacks a dedicated "Self-Harm" category. To address this gap, I integrated the Reddit SWMH dataset from Ji et al [38]. This particular dataset was gathered from mental health-related subreddits hosted on Reddit [9] and primarily focuses on the exploration of mental disorders and discussions related to suicidal ideation. These discussions encompass topics such as suicide-related intentions and mental health conditions, including depression, anxiety, and bipolar disorder. The data summary is in Table2. I selected data exclusively from the "SuicideWatch" and "Depression" subreddits, as they contained explicit instances of self-harming or self-destructive behaviors. I established a new label column named "Harmful" for these datasets and assigned a label of 1 to all entries. To augment the original

[8] https://www.kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification/data

[9] https://www.redditinc.com/

dataset, I randomly sampled 6,000 instances from the Self-Harm dataset and incorporated them into the pre-existing Harmful dataset. Table3 provides a summary of the dataset following the addition of the sampled Self-Harm data. To rectify the pronounced data imbalance, I preserved the entirety of the Harmful data while randomly selecting 29,000 instances from the Non-Harmful category, resulting in a balanced dataset. A summary of this balanced dataset is available in Table4. Subsequently, I partitioned the dataset into three distinct subsets: 80% for the training dataset, 10% for validation, and another 10% for testing purposes. The dataset split summary in Table5 has a detailed overview of these proportions.

Table 2: Reddit SWMH Dataset Summary

Label	**Count**
Depression	22468
SuicideWatch	6550
Anxiety	6136
Offmychest	5265
Bipolar	4932

Table 3: Augmented Toxic Comment Dataset Summary

Label	**Count**
Harmful (1):Toxic Comment + Self-harm	28468
Non-Harmful (0)	201081

Table 4: Balanced Dataset Summary

Label	**Count**
Harmful (1)	28468
Non-Harmful (0)	29000

Table 5: Train, Validation and Test Dataset Split

Label	**Count**
Train	45974
Validation	5747
Test	5747

4.3 Models

In our study, we aim to conduct a parallel comparison of three distinct language models (LLMs) — RoBERTa [1], GPT-4 [2], and the Moderation API — in the context of harmful content detection. These LLMs vary in their structures and training processes. RoBERTa employs the BERT [26] architecture, allowing it to consider both left and right contexts during pre-training for a comprehensive understanding of word context. It is widely used for diverse NLP tasks, including text classification, named entity recognition, and question-answering. In our study, we specifically leverage RoBERTa for text classification in the context of harmful content detection. GPT-4 follows a Generative Pre-trained Transformer architecture, generating text in a unidirectional, left-to-right manner, focusing on creative text generation. GPT models are commonly applied to generative tasks, such as text completion, dialogue generation, and creative writing. In our study, we utilize GPT-4 to generate responses indicating the presence of harmful content. The Moderation API serves as a pre-trained content moderation classifier, designed to identify and categorize harmful content. It is Widely used by developers and the Moderation API can be integrated seamlessly into applications to facilitate content moderation and maintain a secure environment. In our study, we leverage this tool to classify harmful content into predefined categories.

Our objective is to comprehensively compare the performance of these three models on the same dataset, evaluating their effectiveness in harmful content detection. While RoBERTa focuses on bidirectional context understanding, GPT-4 is employed for its generative capabilities, and the Moderation API offers an out-of-the-box solution for content moderation. This comparison aims to provide insights into the strengths and weaknesses of each model within the specific context of identifying harmful content. The details of each model are in below subsections.

4.3.1 Robustly optimized BERT approach (RoBERTa)

In 2019, Facebook AI Research (FAIR) conducted a comprehensive analysis of Google's BERT model [26], identifying its limitations. Subsequently, they introduced the Robustly optimized BERT approach, known as RoBERTa [1]. Liu et al. [1] pointed out that BERT was undertrained and implemented several critical modifications to enhance its training methodology. These modifications included (i) adopting a dynamic masking pattern instead of a static one, (ii) training with significantly more data using larger batches, and (iii) eliminating the Next Sentence Prediction (NSP) objective. Additionally, they (iv) extended the model's training to encompass longer sentences. The result of these enhancements was a significantly improved model—RoBERTa. It was trained on a vast dataset of 160GB of text, over ten times larger than the dataset used for BERT. RoBERTa's superior performance extends to a wide range of natural language processing tasks, such as language translation, text classification, and question answering. It has also served as the foundational model for numerous successful NLP models and has gained popularity in both research and industry applications. In our study, we will fine-tune RoBERTa on our training data for classification task.

4.3.2 GPT-4

GPT-4 [2] stands as a formidable multimodal model, capable of processing both image and text inputs while producing text outputs. Although it may not match human-level performance in every real-world scenario, GPT-4 exhibits impressive proficiency on numerous professional and academic benchmarks. Notably, it demonstrates this proficiency by achieving scores in the top 10% of test takers on simulated bar exams, a significant leap compared to GPT-3.5, which ranked in the bottom 10%. GPT-4 excels in terms of reliability, creativity, and its ability to handle nuanced instructions, especially when confronted with complex tasks. In addition, GPT-4 surpasses existing large language models, including most state-of-the-art (SOTA) models, on traditional benchmarks originally designed for machine learning models. This superior performance extends beyond English and is evident in 24 out of 26 languages tested, including low-resource languages like Latvian, Welsh, and Swahili.

Beyond its language capabilities, GPT-4 has made a significant impact on various functions such as customer support, sales, content moderation, and programming. Notably, GPT-4 plays a pivotal role in content policy development and content moderation decisions, contributing to more consistent labeling, faster feedback loops for policy refinement, and reduced reliance on human moderators. In our study, we intend to harness the power of GPT-4 by employing diverse prompts to generate responses in the detection of harmful content.

We use GPT-4 by API provided by OpenAI. It also provides a range of parameters in the request body that can be customized to adjust the request. Some of the key parameters that can be adjusted are listed in Table6 (OpenAI, 2023b). For our specific objective of testing the reliability and consistency of GPT models in identifying harmful content, we aimed to avoid randomness in our results. To achieve this, we set the temperature parameter to 0. The top_p parameter can also be used to control the randomness. To ensure that all available tokens in the results are taken into account, we used the default top_p value (default = 1) for our experiments.

Table 6: Parameters in the request body of GPT-4 (OpenAI, 2023b).

Parameter	Explanation
max_tokens	The maximum number of tokens to generate in the completion
temperature	A value between 0 and 2; higher values make the output more random, while lower values make the output more deterministic
top_p	A value implies that the model considers the results of the tokens with top_p probability mass
presence_penalty	A number between -2 and 2; positive values increase the model's likelihood to talk about new topics
frequency_penalty	A number between -2 and 2; positive values decrease the model's likelihood to repeat the same line verbatim

4.3.3 Moderation API

The Moderation API [10] is a powerful content moderation tool developed by OpenAI, designed to assess content for compliance with OpenAI's usage policies. These policies encompass a range of prohibited content types, including illegal activities, child sexual abuse material, the generation of hateful, harassing, or violent content, and activities associated with a high risk of physical harm. Developers can leverage this tool to effectively identify and take action on content that violates these policies, such as implementing content filtering or other appropriate measures. The model behind this endpoint classifies content as flagged and non-flagged, and also classifies the flagged content into five main categories: Harassment, Hate, Sexual, Violence, and Self-Harm. Additionally, it provides probability scores for each category, offering insights into the likelihood that content falls into one or more of these harmful categories. We can notice that these five categories are exactly the five categories that we define for Harmful Content. Therefore, this tool serves as a valuable resource for identifying harmful content in our study.

4.4 Evaluation Metrics

The evaluation metrics for classification models are accuracy, precision, recall, and F-score. Accuracy is the ratio of the total number of positives to the total number of classes. Precision is the ratio of true positives to the total number of predicted positives. The recall is the ratio of true positives to the actual number of positives. We can get a harmonic mean of these measures using the F1-Score that considers precision and recall. When there is a significant class imbalance, this is very helpful. The formulas are as follows:

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
Precision = $\frac{TP}{TP+FP}$
Recall = $\frac{TP}{TP+FN}$
F1-Score = $\frac{2*Precisson*Recall}{Precisson+Recall}$
Where:
True Positive (TP) i.e. Positive class classified correctly.
False Positive (FP) i.e. Negative class wrongly predicted as positive.
False Negative (FN) i.e. Positive class wrongly predicted as negative.
True Negative (TN) i.e. Negative class classified correctly.

5 Experiment and Results

5.1 Experiments on Fine-tune RoBERTa

I implemented the RoBERTa model with Transformer Package [11], which provides many state-of-the-art pre-trained models and I conducted model training using my dedicated training dataset. The training process encompassed four epochs, with a learning rate set at 2e-5. The culmination of this training resulted in an impressive final training loss of 0.083, as illustrated in the training loss summary in Figure 11.

Upon evaluating the model's performance on the testing dataset, Table7 showcases RoBERTa's exceptional capabilities across various metrics. It excels in accuracy, precision, recall, and F1-score, achieving the highest accuracy of 0.94 and an outstanding recall rate of 0.95. The F1 score in both categories reaches an impressive 0.94, reflecting a well-balanced performance that combines precision and recall effectively.

A deeper examination of the confusion matrix results offers valuable insights into RoBERTa's strengths. Notably, RoBERTa exhibits higher precision in the Non-Harmful category, with a precision rate of 0.95, compared to the Harmful category's precision of 0.92. This indicates RoBERTa's ability to correctly identify non-harmful content while maintaining a high level of precision.

Furthermore, the recall rates are equally noteworthy. RoBERTa outperforms in the recall for Harmful content, achieving a recall rate of 0.95, while maintaining a recall rate of 0.92 for Non-Harmful

[10] https://platform.openai.com/docs/guides/moderation/overview
[11] https://huggingface.co/docs/transformers/index

content. These results underscore RoBERTa's effectiveness in not only correctly identifying harmful content but also consistently recognizing non-harmful content, contributing to a comprehensive content moderation solution that balances accuracy and sensitivity.

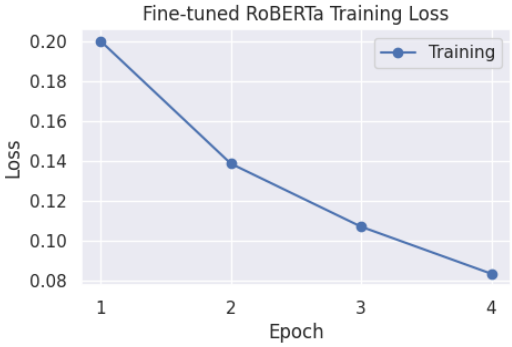
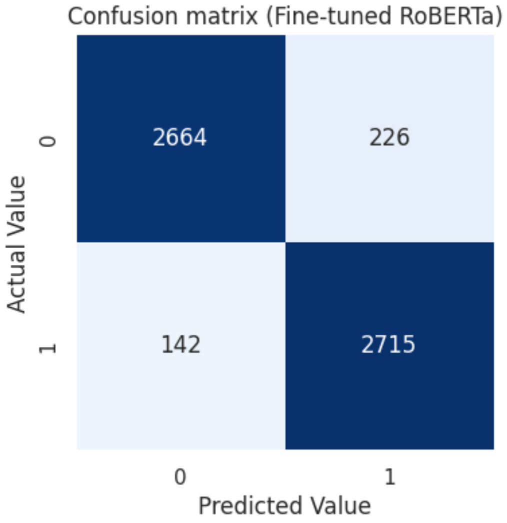


Figure 1: Fine-Tuned RoBERTa Training Loss vs. Epoch



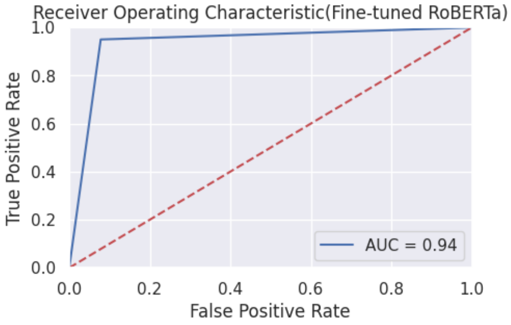


Figure 2: Fine-tuned RoBERTa Confusion Matrix

Figure 3: Fine-tuned RoBERTa ROC Curve

5.2 Experiments on Moderation API

We conducted testing of the Moderation API on the same testing dataset, classifying all 'Flagged' content as Harmful and assigning it a label of 1. The evaluation results from Table7 reveal an overall accuracy of 0.84, providing a robust measure of the API's performance. However, it's the nuanced details that truly stand out.

Notably, the Moderation API demonstrates precision of 0.95 in detecting Harmful content, surpassing other models. This precision is a vital component of content moderation, ensuring that flagged content is reliably identified without an abundance of false positives. The API's remarkable recall rate of 0.96 for the Non-Harmful category is equally noteworthy, showcasing its effectiveness in correctly identifying non-harmful content, further contributing to a safer online environment.

The precision and recall rates exemplify the Moderation API's capabilities, making it a powerful tool for content moderation. These results not only demonstrate its ability to identify harmful content accurately but also its efficiency in distinguishing non-harmful content, reinforcing its role in maintaining online spaces that are conducive to meaningful and safe interactions.

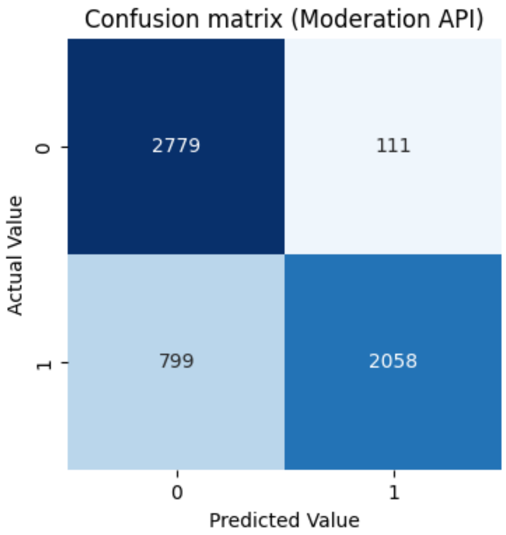


Figure 4: Moderation API Confusion Matrix

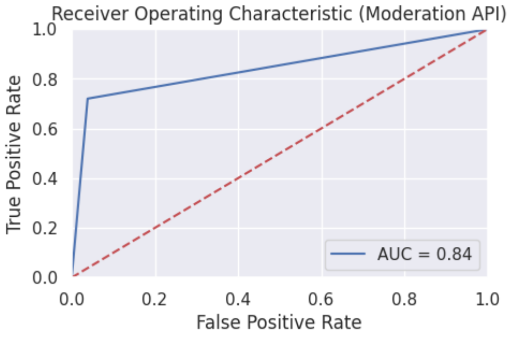


Figure 5: Moderation API ROC Curve

Table 7: Evaluation Report of Models

Algorithms	Accuracy	Precision	Recall	F1-score	Class	Testing Size
RoBERTa	**0.94**	0.95 0.92	0.92 0.95	**0.94** **0.94**	0 1	2890 2857
Moderation API	0.84	0.78 0.95	0.96 0.72	0.86 0.82	0 1	2890 2857
GPT-4 (Prompt1)	0.76	0.68 0.97	0.98 0.54	0.80 0.69	0 1	2890 2857
GPT-4 (Prompt2)	0.87	0.83 0.91	0.92 0.81	0.88 0.86	0 1	2890 2857
GPT-4 (Prompt3)	0.88	0.85 0.91	0.92 0.83	0.88 0.87	0 1	2890 2857

5.3 Experiments on GPT-4

In our experimentation with GPT-4, we employed three distinct prompts on the same dataset to evaluate its performance in detecting harmful content. The initial prompt delivered a straightforward definition of harmful content, accompanied by the names of five categories: Harassment, Hate, Sexual, Violence, and Self-Harm. This zero-shot experiment aimed to assess GPT-4's ability to discern the presence of harmful content based solely on a brief definition. To delve deeper into GPT-4's comprehension, the second prompt not only reiterated the definition but also provided a comprehensive explanation for each harmful content category. This enhanced prompt sought to gauge the model's understanding of the nuanced distinctions within each category, evaluating its capability to recognize subtle variations in harmful content. Taking the evaluation a step further, the third prompt extended beyond zero-shot experiments by incorporating practical examples of input content along with the expected output. By introducing a few-shot experiment, we aimed to observe how GPT-4 performs when presented with specific instances of harmful content, providing it with context to enhance its detection accuracy.

- Prompt 1: "Does the content contain Harmful Content such as Self-Harm, Hate, Violence, Sexual, Harassment?"
- Prompt2: "We define Harmful Content to include Self-Harm, Hate, Violence, Sexual, Harassment. Self-Harm is defined as <Self-Harm definition>. Hate is defined as <Hate definition>. Violence is defined as <Violence definition>, Harassment is defined as <Harassment definition>. Sexual is defined as <Sexual definition>. Does the content contain Toxic Harmful Content?"

- Prompt3: "We define Harmful Content to include Self-Harm, Hate, Violence, Sexual, Harassment. Self-Harm is defined as <Self-Harm definition>. Hate is defined as <Hate definition>. Violence is defined as <Violence definition>, Harassment is defined as <Harassment definition>. Sexual is defined as <Sexual definition>.
  <input 1> Does the content contain Toxic Harmful Content? Yes, it contains Self-Harm.
  <input 2> Does the content contain Toxic Harmful Content? Yes, it contains Hate.
  <input 3> Does the content contain Toxic Harmful Content? Yes, it contains Violence.
  <input 4> Does the content contain Toxic Harmful Content? Yes, it contains Sexual.
  <input 5> Does the content contain Toxic Harmful Content? Yes, it contains Harassment.
  <input 6> Does the content contain Toxic Harmful Content? No.

Table7 presents compelling evidence of the significant impact of using different prompts on GPT-4's performance. Prompt 2 exhibits a remarkable improvement over Prompt 1, as evidenced by several key metrics. The accuracy soars from 0.76 to 0.87, reflecting a substantial enhancement in overall performance. Notably, both harmful and non-harmful content categories benefit from this change, with a notable increase in F1 scores. For harmful content, the F1 score surges from 0.69 to an impressive 0.86, while non-harmful content also sees a substantial boost, rising from 0.80 to 0.88. These findings underscore the pivotal role that prompt design plays in GPT-4's outcomes, demonstrating that providing more informative and explanatory prompts can lead to significantly improved results. In addition, the accuracy of few shot (prompt 3) is only 1% higher than that of zero shot (prompt2), which indicates that providing some examples can make GPT-4 learn better but not improve much.

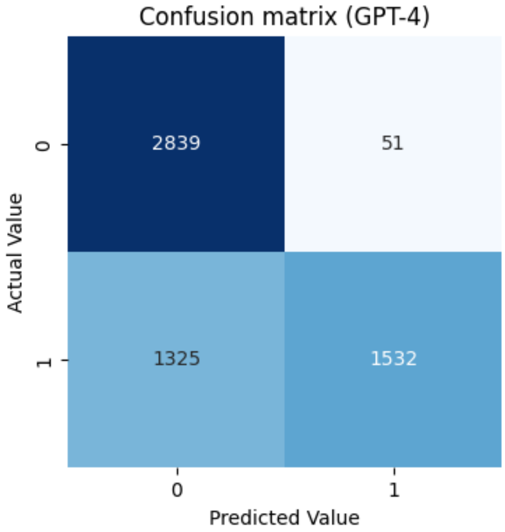


Figure 6: GPT-4 (Prompt1) Confusion Matrix

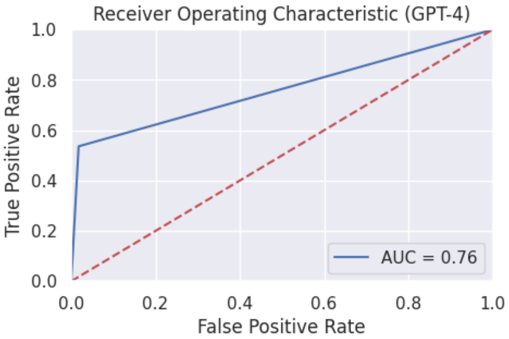


Figure 7: GPT-4 (Prompt1) ROC Curve

5.4 Results

Table7 reveals critical insights into the comparative performance of the three models. Firstly, fine-tuned RoBERTa has the best performance, surpassing the other two models in both accuracy and F1 score, achieving the highest accuracy of 0.94. These results are attributed to the fact that it learned more information by training on the harmful content we prepared. This also underscores the power of pre-trained language models, solidifying their competitiveness in the detection of harmful content.

Secondly, GPT-4 impresses with its content moderation capabilities, enabled by the ability to fully customize harmful content guidelines. This customization leads to slightly higher accuracy and F1 scores compared to the Moderation API, showcasing GPT-4's adaptability and effectiveness in tailored content moderation.

Furthermore, our experiments indicate the influence of different prompts on GPT-4's results. Providing a detailed definition and explanation of policy in the prompt significantly enhances accuracy,

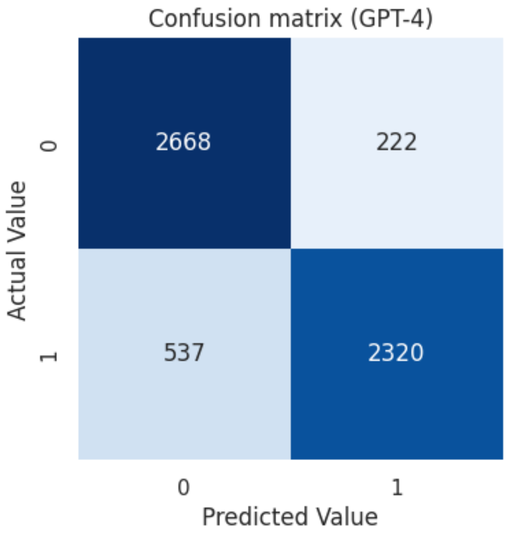


Figure 8: GPT-4 (Prompt2) Confusion Matrix

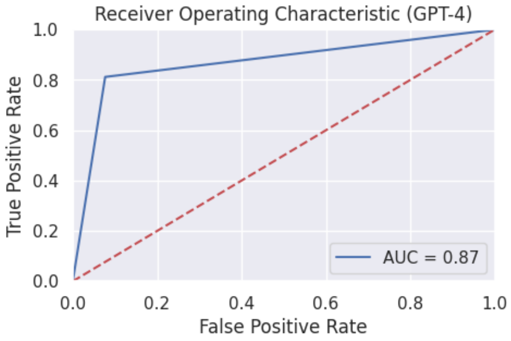


Figure 9: GPT-4 (Prompt2) ROC Curve

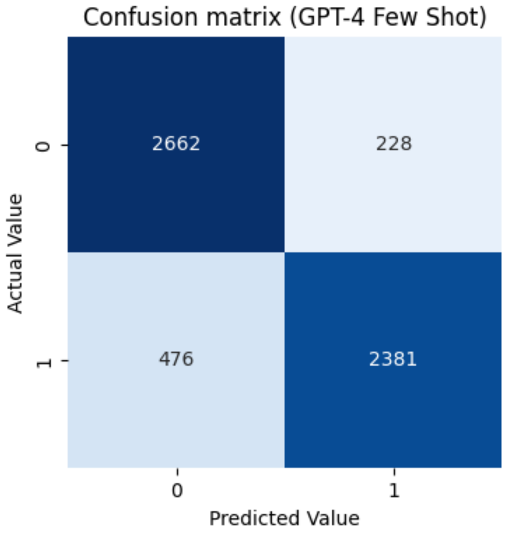


Figure 10: GPT-4 (Prompt3) Confusion Matrix

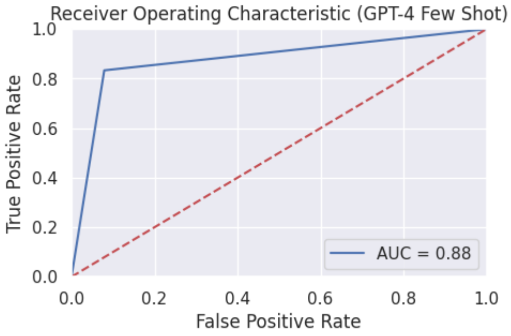


Figure 11: GPT-4 (Prompt3) ROC Curve

although it may also introduce the potential for misjudgment, resulting in an uptick in false positives. There is another issue with GPT-4 when we experiment with prompt 2 is GPT-4 exhibits occasional challenges in prioritizing prompt content. For instance, in scenarios involving content related to self-harm or self-hurt, even when provided with specific context and a directive prompt, such as asking for a simple 'yes' or 'no' response, GPT-4 may still respond with a pre-defined message, stating, 'Sorry that you're feeling this way, but I'm unable to provide the help that you need. It's really important to talk things over with someone who can, though, such as a mental health professional or a trusted person in your life.' This behavior indicates the need for nuanced improvements to ensure that GPT-4 responds appropriately to directive prompts, especially in sensitive contexts.

To address and mitigate these challenges, we introduced examples in our third prompt (prompt 3). This addition aimed to provide GPT-4 with specific instances and context, allowing it to better understand and respond to content related to self-harm. The inclusion of practical examples in the prompt serves as a valuable strategy to guide GPT-4's responses and improve its performance, reducing the instances of pre-defined responses in sensitive scenarios.

Additionally, it's noteworthy that these findings underline the importance of a dynamic and iterative approach to content moderation. By constantly fine-tuning models, adjusting prompts, and refining guidelines, we can continuously improve the accuracy and efficiency of harmful content detection, ensuring a safer online environment for users.

6 Application Design

Our online chat and discussion board application is designed as illustrated in Figure 12, featuring distinct interfaces and functionalities tailored for students and teachers. In the student's mode (Figure 13), students have the capability to type and send their posts, which are then subjected to examination by embedding our three Language Models (LLMs) in previous sections to detect any potentially harmful content. If the LLMs identify harmful content in a post, a warning prompt will appear, asking the student, 'Harmful content detected. Do you want to modify the post?'. We can see this example in Figure 14. Students have the option to either modify the post or proceed with sending it as-is. In cases where students choose to send a post containing potentially harmful content despite the warning, the post is sent successfully but is marked with a flag. Simultaneously, teachers are promptly notified of the problematic post. In the teacher's mode interface, teachers can locate these flagged posts and take appropriate actions to address the potential issue. Figure 15 illustrates that there is a 'green bell' icon below the problematic post, indicting that this post has an issue. Then teachers can click the 'blue clock' icon to view the flagged posts(Figure 16), and resolve the issue which is reported by our LLMs (Figure 17).

Given that LLMs may not achieve complete accuracy, there may be instances where posts contain harmful content that goes undetected. To address this, we encourage students to flag such posts while browsing the discussion board. Students can report a post along with a brief reason, triggering a new notification for teachers. This collaborative approach allows for a more comprehensive and proactive method of content moderation and resolution.

Choosing the right LLM for harmful content detection in application depends on specific use cases and requirements because there are some important factors which should be considered such as cost, running time and data privacy. RoBERTa emerges as a compelling option when the content and detection criteria remain relatively stable. It offers exceptional accuracy, precision, and recall rates, making it a strong choice. Importantly, RoBERTa's processing speed is remarkable, a key factor for chat boards where a high volume of information is generated rapidly. It's also advantageous that RoBERTa is freely available and can be integrated locally, mitigating concerns about data security. However, its drawback lies in adaptability; if the detection criteria change, it necessitates data collection and model retraining, a time-consuming process often requiring specialized equipment.

The Moderation API is a great alternative, particularly if precision in harmful content detection is a top priority. It doesn't need additional training process and boasts speed and cost-effectiveness, although it's an external API with policies defined by OpenAI, limiting flexibility. This might result in missing some desired information.

GPT-4 offers remarkable performance, with a higher F1 score compared to the Moderation API in both harmful and non-harmful content detection. It's highly adaptable, with the ability to swiftly correct results by adjusting prompts and adding more information. However, its drawback lies in processing speed, making it unsuitable for scenarios where rapid message delivery is crucial, such as chat boards. In addition, it's worth noting that GPT-4 is relatively more expensive when compared to other alternatives. The higher cost associated with GPT-4 can be a consideration for budget-constrained applications, particularly in scenarios where extensive usage and scaling are necessary. Moreover, an important consideration regarding GPT-4 lies in the variability of its result format. Even when using the same prompt, GPT-4 can produce varying result formats, making it challenging to apply a standardized processing approach, for example, classification task. This variability can pose difficulties in automating the post-processing of GPT-4's responses and integrating them seamlessly into a standardized workflow.

In summary, for most applications, RoBERTa and the Moderation API are recommended choices. However, GPT-4 holds promise for applications that require enhanced language understanding capabilities, offering an avenue for ongoing exploration and improvement.

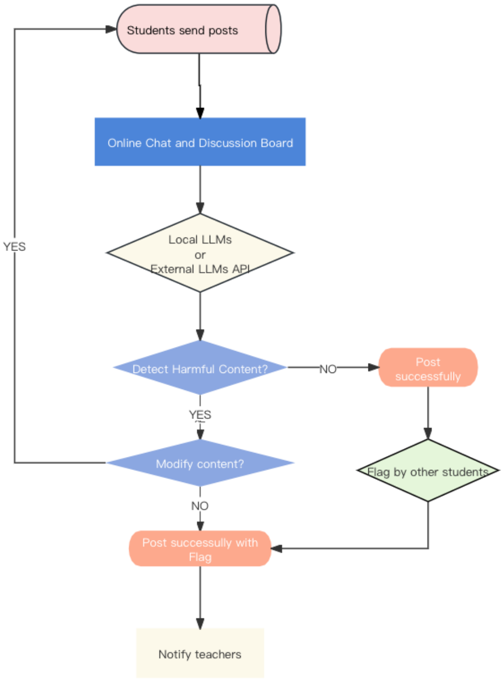
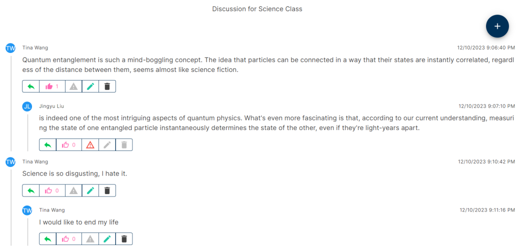


Figure 12: Application Design Process



Discussion for Science Class

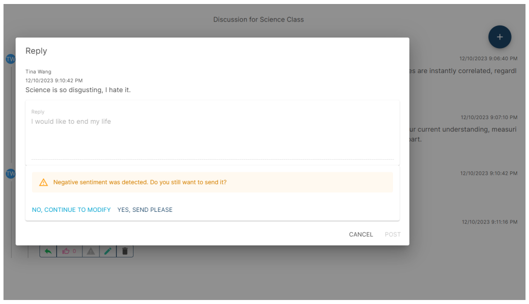
Tina Wang 12/10/2023 9:06:40 PM
Quantum entanglement is such a mind-boggling concept. The idea that particles can be connected in a way that their states are instantly correlated, regardl ess of the distance between them, seems almost like science fiction.

Jingyu Liu 12/10/2023 9:07:10 PM
is indeed one of the most intriguing aspects of quantum physics. What's even more fascinating is that, according to our current understanding, measuri ng the state of one entangled particle instantaneously determines the state of the other, even if they're light-years apart.

Tina Wang 12/10/2023 9:10:42 PM
Science is so disgusting, I hate it.

Tina Wang 12/10/2023 9:11:16 PM
I would like to end my life

Figure 13: Student's mode Interface



Discussion for Science Class

Reply

Tina Wang
12/10/2023 9:10:42 PM
Science is so disgusting, I hate it.

Reply
I would like to end my life

Negative sentiment was detected. Do you still want to send it?

NO, CONTINUE TO MODIFY YES, SEND PLEASE

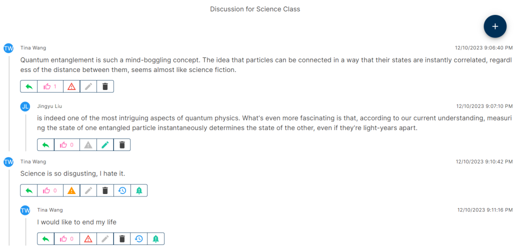
CANCEL POST

Figure 14: Students get warning if Harmful Content detected

7 Conclusions

In this study, our primary objective was to assess the performance of three distinct large language models—Fine-tuned RoBERTa, the Moderation API, and GPT-4—for automated harmful content detection. To achieve this, we used the concept of harmful content and curated a diverse dataset encompassing five critical aspects: harassment, hate, sexual content, self-harm, and violence. Our comprehensive evaluation yields noteworthy findings.

The results of our study unequivocally establish RoBERTa as the standout performer. It consistently outperformed the other models, achieving the highest accuracy and F1 score. This success underscores the prowess of fine-tuned pre-trained language models in the domain of content moderation and harmful content detection. Moreover, our study provides compelling evidence of GPT-4's robust content moderation capabilities. GPT-4's versatility in adapting to dynamic content guidelines and offering a swift feedback loop for policy refinement is a significant advantage. However, it's essential to acknowledge that GPT-4 exhibits specific limitations, particularly in response prioritization and result format consistency.



Discussion for Science Class

Tina Wang 12/10/2023 9:06:40 PM

Quantum entanglement is such a mind-boggling concept. The idea that particles can be connected in a way that their states are instantly correlated, regardless of the distance between them, seems almost like science fiction.

Jingyu Liu 12/10/2023 9:07:10 PM

is indeed one of the most intriguing aspects of quantum physics. What's even more fascinating is that, according to our current understanding, measuring the state of one entangled particle instantaneously determines the state of the other, even if they're light-years apart.

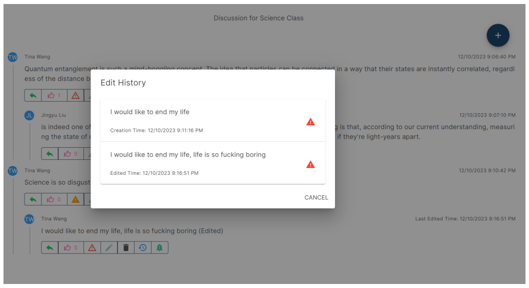
Tina Wang 12/10/2023 9:10:42 PM

Science is so disgusting, I hate it.

Tina Wang 12/10/2023 9:11:16 PM

I would like to end my life

Figure 15: Teacher’s mode Interface



Edit History

I would like to end my life

Creation Time: 12/10/2023 9:11:16 PM

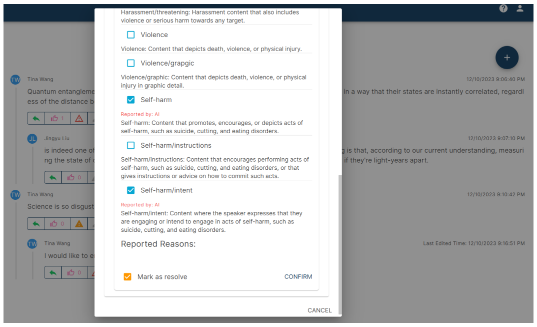
I would like to end my life, life is so fucking boring

Edited Time: 12/10/2023 9:16:51 PM

Tina Wang Last Edited Time: 12/10/2023 9:16:51 PM

I would like to end my life, life is so fucking boring (Edited)

Figure 16: Teachers find problematic comment



Harassment/threatening: Harassment content that also includes violence or serious harm towards any target.

□ Violence

Violence: Content that depicts death, violence, or physical injury.

□ Violence/grapgic

Violence/graphic: Content that depicts death, violence, or physical injury in graphic detail.

☑ Self-harm

Reported by: AI

Self-harm: Content that promotes, encourages, or depicts acts of self-harm, such as suicide, cutting, and eating disorders.

□ Self-harm/instructions

Self-harm/instructions: Content that encourages performing acts of self-harm, such as suicide, cutting, and eating disorders, or that gives instructions or advice on how to commit such acts.

☑ Self-harm/intent

Reported by: AI

Self-harm/intent: Content where the speaker expresses that they are engaging or intend to engage in acts of self-harm, such as suicide, cutting, and eating disorders.

Reported Reasons:

☑ Mark as resolve

Figure 17: Teachers resolve problematic comment

As we consider the real-world applicability of these large language models, it’s essential to weigh the pros and cons. The choice of LLM should be context-specific, factoring in the nature of the task, runtime requirements, data privacy protection and cost considerations. RoBERTa excels in terms of accuracy and F1 score and its capability to train locally effectively addresses concerns related to data privacy leakage. However, it’s important to note that periodic retraining may be necessary to ensure adaptation to evolving content definitions. The Moderation API proves to be a cost-effective choice with exceptional precision, albeit within defined policy constraints. GPT-4 stands out for its adaptability and feedback loop but may not be suitable for scenarios demanding rapid response times. Our findings can inform decisions in content moderation strategies, offering valuable insights into selecting the most appropriate model for specific applications and use cases. As the field of language models continues to evolve, the quest for more effective content moderation tools remains a dynamic and ongoing endeavor.

7.1 Future Research Plans

Future research can focus on further refining and adapting pre-trained language models, such as fine tuning GPT-4, to improve their performance in content moderation. This may involve fine-tuning models on larger and more diverse datasets, as well as developing better methods for adjusting models to evolving content definitions and guidelines. Furthermore, the generation of meaningful explanations by the models represents a crucial frontier in our ongoing exploration. Our future endeavors will delve into research and development efforts aimed at training models capable of producing coherent and contextually relevant justifications for their classification decisions. This approach aligns with our commitment to not only detect harmful content but also to provide comprehensive insights, fostering a more informed and responsive content moderation ecosystem.

References

[1] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[2] OpenAI. GPT-4 Technical Report. *arXiv e-prints*, art. arXiv:2303.08774, March 2023. doi: 10.48550/arXiv.2303.08774.

[3] Stephen Baglione and M. Nastanski. The superiority of online discussion: Faculty perceptions. *The Quarterly Review of Distance Education*, 8:139–150, 01 2007.

[4] Mary Dengler. Classroom active learning complemented by an online discussion forum to teach sustainability. *Journal of Geography in Higher Education*, 32:481–494, 09 2008. doi: 10.1080/03098260701514108.

[5] Susan Keith and Michelle Martin. Cyber-bullying: Creating a culture of respect in a cyber world. *Reclaiming Children Youth*, 13, 01 2005.

[6] Jinyu Huang, Zhaohao Zhong, Haoyuan Zhang, and Liping Li. Cyberbullying in social media and online games among chinese college students and its associated factors. *International journal of environmental research and public health*, 18(9):4819, 2021.

[7] Xingyue Jin, Kun Zhang, Mireille Twayigira, Xueping Gao, Huiming Xu, Chunxiang Huang, Xuerong Luo, and Yanmei Shen. Cyberbullying among college students in a chinese population: Prevalence and associated clinical correlates. *Frontiers in public health*, 11:1100069, 2023.

[8] Valentina Piccoli, Andrea Carnaghi, Mauro Bianchi, and Michele Grassi. Perceived-social isolation and cyberbullying involvement: The role of online social interaction. *International Journal of Cyber Behavior, Psychology and Learning (IJCBPL)*, 12(1):1–14, 2022.

[9] Zhekuan Peng, Anat Brunstein Klomek, Liping Li, Xuefen Su, Lauri Sillanmäki, Roshan Chudal, and Andre Sourander. Associations between chinese adolescents subjected to traditional and cyber bullying and suicidal ideation, self-harm and suicide attempts. *BMC psychiatry*, 19(1): 1–8, 2019.

[10] Md Irteja Islam, Fakir Md Yunus, Enamul Kabir, and Rasheda Khanam. Evaluating risk and protective factors for suicidality and self-harm in australian adolescents with traditional bullying and cyberbullying victimizations. *American journal of health promotion*, 36(1):73–83, 2022.

[11] Chanda Maurya, T Muhammad, Preeti Dhillon, and Priya Maurya. The effects of cyberbullying victimization on depression and suicidal ideation among adolescents and young adults: a three year cohort study from india. *BMC psychiatry*, 22(1):1–14, 2022.

[12] Sindhu Abro, Sarang Shaikh, Zahid Hussain Khand, Ali Zafar, Sajid Khan, and Ghulam Mujtaba. Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, 11(8), 2020.

[13] Kelly Reynolds, April Kontostathis, and Lynne Edwards. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine Learning and Applications and Workshops*, volume 2, pages 241–244, 2011. doi: 10.1109/ICMLA.2011.152.

[14] Hannah Metzler, Hubert Baginski, Thomas Niederkrotenthaler, and David Garcia. Detecting potentially harmful and protective suicide-related content on twitter: machine learning approach. *Journal of medical internet research*, 24(8):e34705, 2022.

[15] Michele Banko, Brendon MacKeen, and Laurie Ray. A unified taxonomy of harmful content. In Seyi Akiwowo, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem, editors, *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.16. URL `https://aclanthology.org/2020.alw-1.16`.

[16] Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. *arXiv preprint arXiv:1809.08651*, 2018.

[17] Zhanyuan Yin, Lizhou Fan, Huizi Yu, and Anne J Gilliland. Using a three-step social media similarity (tsms) mapping method to analyze controversial speech relating to covid-19 in twitter collections. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1949–1953. IEEE, 2020.

[18] Kenneth Ward Church. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017.

[19] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[20] Pranav Malik, Aditi Aggrawal, and Dinesh K Vishwakarma. Toxic speech detection using traditional machine learning models and bert and fasttext embedding with deep neural networks. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1254–1259. IEEE, 2021.

[21] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760, 2017.

[22] Santiago González-Carvajal and Eduardo C Garrido-Merchán. Comparing bert against traditional machine learning text classification. *arXiv preprint arXiv:2005.13012*, 2020.

[23] Khushboo Taneja and Jyoti Vashishtha. Comparison of transfer learning and traditional machine learning approach for text classification. In *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 195–200. IEEE, 2022.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[25] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[27] OpenAI. Gpt-4 is openai's most advanced system, producing safer and more useful responses, 2023.

[28] Bayode Oluwatoba Ogunleye. *Statistical learning approaches to sentiment analysis in the Nigerian banking context*. Sheffield Hallam University (United Kingdom), 2021.

[29] Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*, 2023.

[30] Nakyeong Yang, Jeongje Jo, Myeongjun Jeon, Wooju Kim, and Juyoung Kang. Semantic and explainable research-related recommendation system based on semi-supervised methodology using bert and lda models. *Expert Systems with Applications*, 190:116209, 2022.

[31] Szu-Yin Lin, Yun-Ching Kung, and Fang-Yie Leu. Predictive intelligence in harmful news identification by bert-based ensemble learning model with text sentiment analysis. *Information Processing & Management*, 59(2):102872, 2022.

[32] Flor Miriam Plaza-del Arco, M Dolores Molina-González, L Alfonso Urena-López, and M Teresa Martín-Valdivia. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120, 2021.

[33] Rodrigo Martínez-Castaño, Amal Htait, Leif Azzopardi, and Yashar Moshfeghi. Early risk detection of self-harm and depression severity using bert-based transformers: ilab at clef erisk 2020. 2020.

[34] Sayanta Paul and Sriparna Saha. Cyberbert: Bert for cyberbullying identification: Bert for cyberbullying identification. *Multimedia Systems*, 28(6):1897–1904, 2022.

[35] Fan Huang, Haewoon Kwak, and Jisun An. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*, 2023.

[36] Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. " hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *arXiv preprint arXiv:2304.10619*, 2023.

[37] Andrei Kucharavy, Zachary Schillaci, Loïc Maréchal, Maxime Würsch, Ljiljana Dolamic, Remi Sabonnadiere, Dimitri Percia David, Alain Mermoud, and Vincent Lenders. Fundamentals of generative large language models and perspectives in cyber-defense. *arXiv preprint arXiv:2303.12132*, 2023.

[38] Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, and Guodong Long. Supervised learning for suicidal ideation detection in online user content. *Complexity*, 2018, 2018.