BEACONHOUSE NATIONAL UNIVERSITY

PRJ-F23/329

SocialSense:

LLM Based Social Media Comments Classification

INTERNAL SUPERVISOR

Ms. Huda Sarfraz

GROUP MEMBERS

Ibraheem Omer	**F2020-150**
Huzaifa Ejaz	**F2020-690**

SCHOOL OF COMPUTER & INFORMATION TECHNOLOGY

Project Title	**SocialSense**	
Project ID	**PRJ-F23/329**	
Project Supervisors	**NetSol Technologies** **Ms. Huda Sarfraz**	
Group Members	**Ibraheem Omer**	**F2020-150**
	Huzaifa Ejaz	**F2020-690**

Academic Session	**2023-24 (Sept 2023 to June 2024)**
Credit Hours	**6**

PROJECT APPROVAL

This Project is approved in partial fulfilment of the requirements of BSc (Hons.) in Computer Science degree conducted by the School of Computer and IT, Beaconhouse National University, Lahore.

_____	_____
Ms. Huda Sarfraz **External Supervisor**	**Prof. Dr. Khawaja Shafaat Ahmed Bazaz** **Dean School of Computer and IT**

Date: _____

Acknowledgement

We extend our heartfelt gratitude to our internal supervisor, Ms. Huda Sarfaz, whose unwavering support, invaluable guidance, and profound expertise were instrumental throughout the course of this project. Her dedication and encouragement significantly contributed to our growth and the successful completion of our final year project.

We are deeply thankful to our external supervisors at NetSol Technologies, particularly Ms. Humera Mirza, Mr. Abdullah Dawood, and Mr. Uzair Ghauri, for graciously sharing their wealth of industrial experience. Their insightful feedback, relentless assistance, and guidance in navigating challenges proved to be invaluable. Their mentorship not only enriched our project but also broadened our perspective on real-world applications.

We also express our sincere appreciation to the faculty of SCIT for their thorough feedback, unwavering encouragement, and continuous support during our presentations. Their constructive criticism and insightful suggestions played a pivotal role in refining our project and honing our skills.

Additionally, we are thankful to the Punjab Higher Education Commission (PHEC) for organizing the All Punjab Universities Innovation Expo'23, which provided us with a remarkable platform to showcase our project. Winning the first position in the Computer Science and Information Technology category and securing a spot at the Pak China Industrial Expo'23 was a tremendous honor for us. We are grateful for this opportunity, which not only recognized our efforts but also facilitated valuable connections and exposure.

Furthermore, we extend our gratitude to all the investors who approached us and provided valuable feedback regarding our project. Their insights and encouragement have been instrumental in shaping our project and preparing us for future endeavors.

In conclusion, we are immensely grateful to all individuals, organizations, and institutions mentioned above for their unwavering support, guidance, and encouragement. Their collective efforts have undoubtedly played a pivotal role in the successful completion of our final year project, and we are truly appreciative of their contributions.

Scan To View Code & Implementation Files

Table of Contents

Table of Figures

Abstract

This project embarked on developing an AI-powered solution for classifying social media comments, aiming to enhance user experience and foster positive online interactions. After evaluating various Language Model (LLM) options, we fine-tuned the GPT-3.5 Turbo LLM with nearly half a million tokens, achieving a model capable of classifying comments into six categories: Positive, Neutral, Crisis, Negative - Ignore, Negative - Respond, and Negative - Remove. The model achieved an overall accuracy rate of 84%, demonstrating its effectiveness in handling diverse and complex comment data.

To facilitate seamless integration, we developed an API with extensive documentation and support, and a user-friendly Chrome webpage for users to interact with the model and explore project details. Additionally, we created three Chrome extensions – SocialSense for Facebook, Instagram, and TikTok – to classify comments in real-time and enhance social media experiences. These extensions were successfully deployed on the Chrome Web Store, making our solution accessible to a wider audience.

Future plans include integrating the model into the GPTstore by OpenAI for revenue generation, enhancing the model with GPT-4 for improved accuracy and advanced features, and exploring partnerships with social media platforms for direct integration. We also aim to expand our Chrome extensions to other social media platforms and web forums, increasing our reach and impact.

1 Introduction to the Project

In today's digital landscape, the need for efficient and accurate analysis and categorization of user-generated comments and reviews from diverse social media platforms has become increasingly pressing as businesses and organizations heavily rely on insights gleaned from user-generated content to shape marketing strategies, improve products and services, and maintain brand reputation. The ever-growing volume of online content demands tools and models that can automatically identify and categorize this content, providing valuable insights into its nature. This includes distinguishing between safe and harmful content, understanding user sentiment, and recognizing different types of user interactions. In Figures 1.1, 1.2, and 1.3, it's evident that comments on Facebook, Instagram, and TikTok lack classification, presenting a challenge as all comments are aggregated without the option for users to filter them according to their preferences.

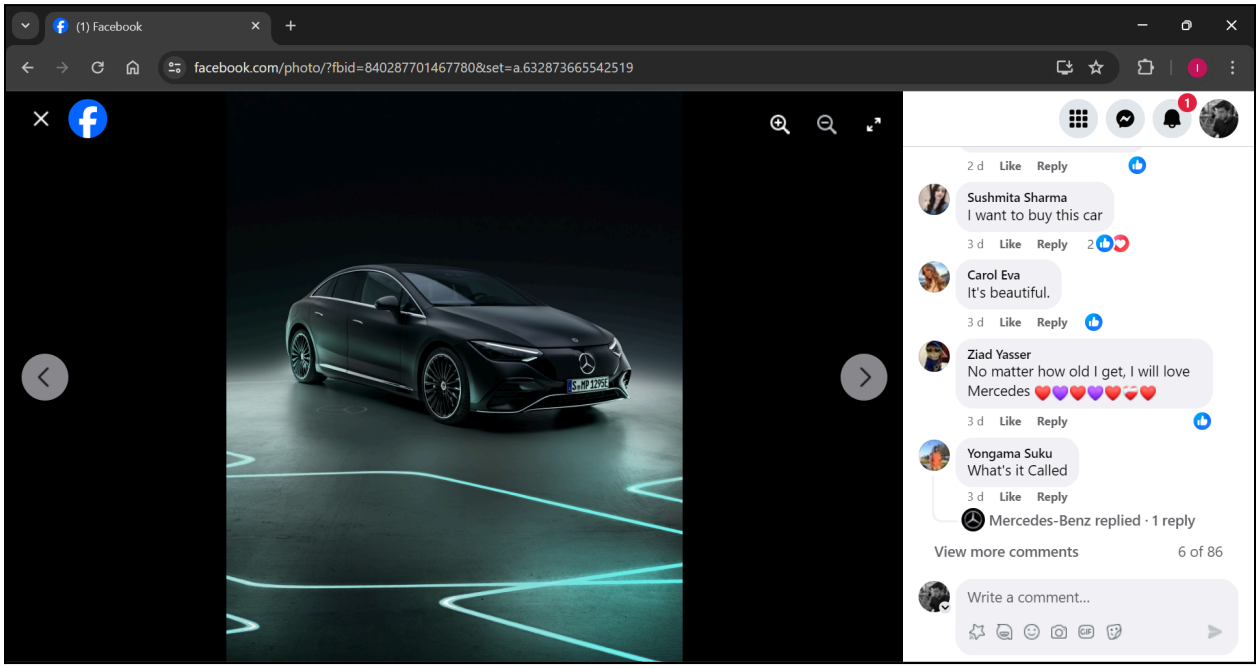


Figure 1.1: Facebook Comments

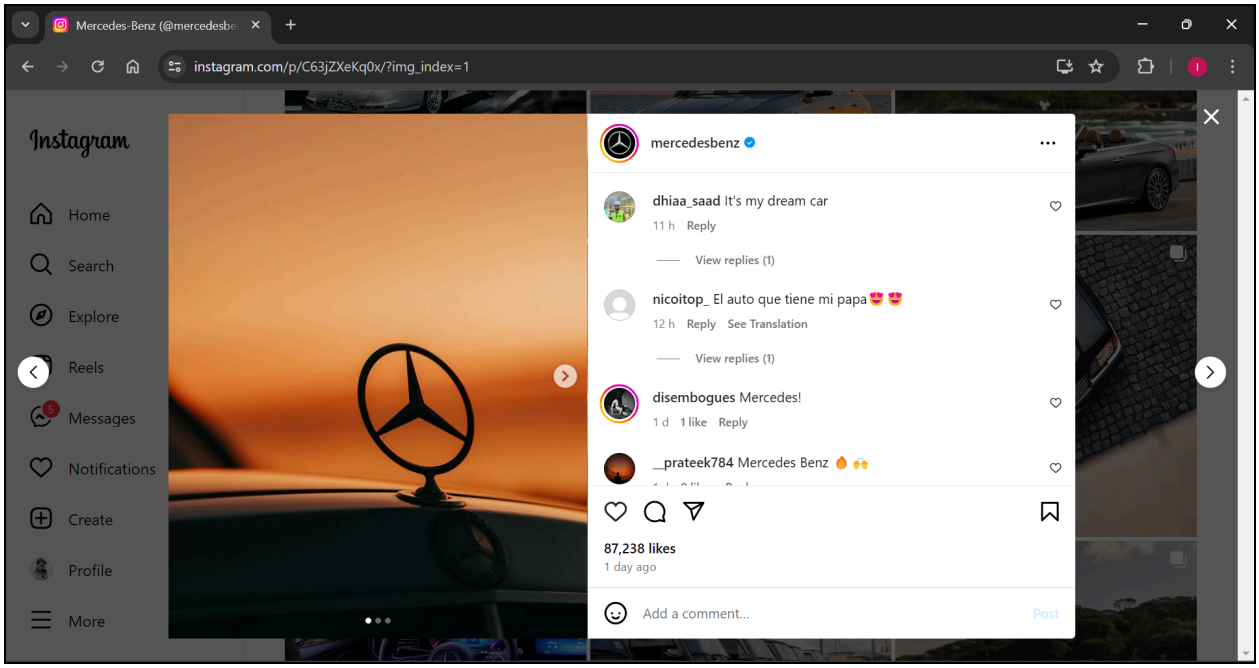


Figure 1.2: Instagram Comments

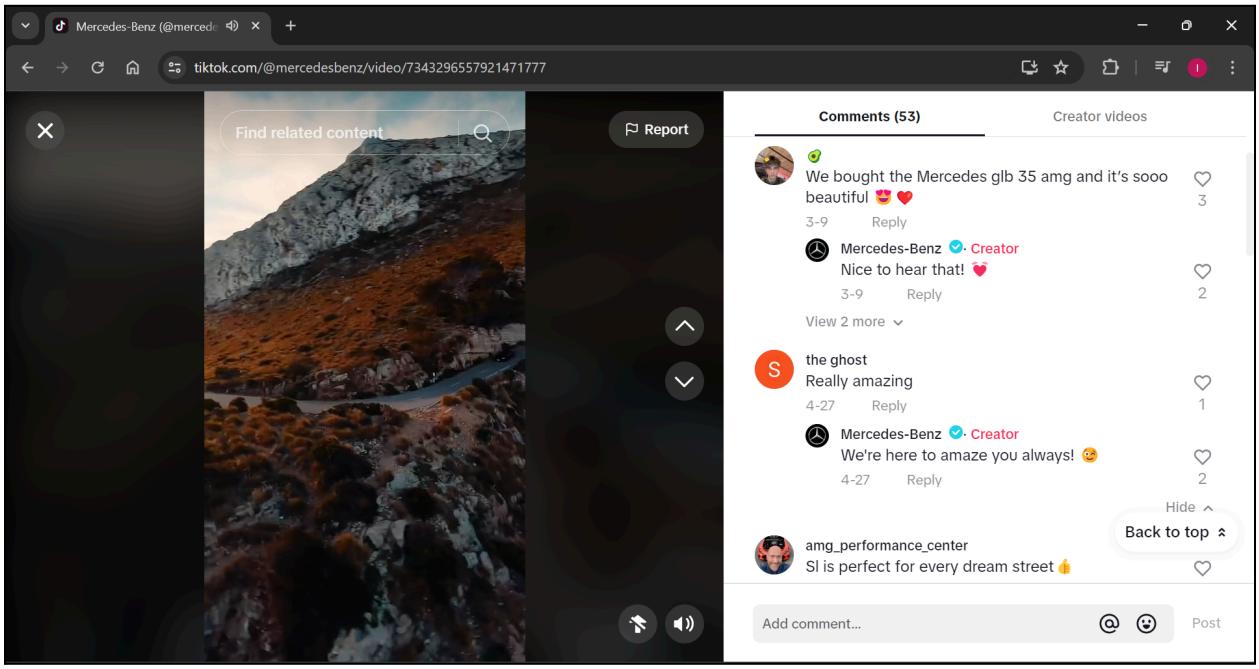


Figure 1.3: TikTok Comments

Furthermore, the challenge lies in the multilingual and diverse nature of user-generated content on these platforms. People comment and post in a multitude of languages and language forms. Hence, there is a growing need for comprehensive checks to be made regardless of the language being used.

1.1 Project Overview

Research by Cahe and Davison [16] demonstrated that Large Language Models (LLMs) excel at complex text classification tasks, surpassing conventional machine learning methods with notably high accuracy. Moreover, Zhan et al. [19] highlighted the transformative impact of LLMs on sentiment analysis, opening new avenues in this field. Similarly, Krungmann and Hartman's [20] study revealed that LLMs not only compete with but can also outperform traditional methods in sentiment classification accuracy. Building on these findings, our project explored various LLMs, including GPT-3.5-Turbo-1106, GPT-3.5-Turbo-0125, Llama-2-70b, and Llama-2-70-chat, for efficient text classification. After thorough investigation, we selected one model to proceed with. Focused on analyzing and categorizing comments and reviews from diverse social media platforms, we employed fine-tuning techniques to train an AI model capable of discerning different sentiments expressed in the collected data. This involved scraping content from various online sources using dedicated tools to ensure comprehensive coverage.

The primary objective of our AI model is to categorize user-generated content into six distinct labels:

1. Positive
2. Neutral
3. Negative - Respond
4. Negative – Ignore
5. Negative – Remove
6. Crisis

These labels were shortlisted after research from Jeff Bullas[23], Attention Insight [24], SproutSocial [25], and JSH Web Design [26] on critical types of social media comments where they all suggested same categories. This categorization system offers a comprehensive understanding of the sentiment and intent behind user comments and reviews, making it a valuable tool for businesses and organizations seeking to gain insights from online interactions. To ensure the diversity and inclusivity of our dataset, we included content in English, Urdu (Roman Script), and Urdu (Arabic Script). Refer to the Appendices section to view sample labeled dataset.

Furthermore, to facilitate seamless integration into various social media platforms, we developed an API (Application Programming Interface). This API allows organizations and developers to incorporate our AI model into their projects effortlessly, enabling real-time analysis and categorization of user-generated content. We've created an interactive website where developers can engage with our AI model and learn more about our work. Additionally, we've developed three Chrome extensions for Facebook, Instagram, and TikTok. These extensions are designed to classify comments into six distinct categories, as discussed earlier, empowering users to filter them according to their preferences.

1.2 Project Methodology

In our project methodology, as illustrated in Figure 1.4, we commenced data collection by scraping information from various social media platforms. Subsequently, the collected data was labeled according to our specified requirements, followed by thorough processing. Once the dataset was adequately prepared, we meticulously selected and fine-tuned a Large Language Model, incorporating prompt engineering. Following this, we conducted rigorous model testing to ensure its effectiveness.

Upon validation, an API for the model was developed and seamlessly integrated into our webpage and Chrome extensions. For additional details, please refer to the Implementation section.

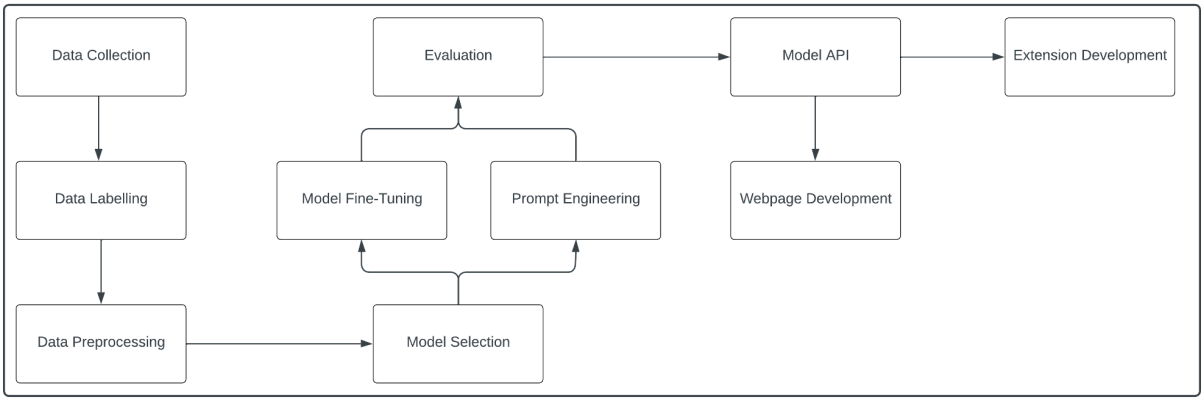


Figure 1.4: Project Methodology

1.3 Existing situation and motivation for the project

Managing comments and reviews on social media platforms rely on manual user review and response, with the limited capability to report and automatically filter hate speech, primarily in English content. To enhance this system, we can look at the functionalities provided by Instagram [27], YouTube [29], CellarWeb plugin [28] and Perspective API [30]. However, these solutions may not perform as effectively for comments in local communities that use languages with variations like Urdu (Roman Script). Such languages can have unique phrases and cultural nuances that automated systems are less adept at filtering [14]. To address this, tailored content moderation tools may be necessary, incorporating language-specific filters and community-driven moderation to ensure a safer and culturally relevant online environment for local language users.

1.3.1 Instagram

Instagram offers automatic filtering for offensive comments and message requests through customizable settings, hiding content containing offensive words or phrases as seen in Figure 1.5. Users can adjust these settings to suit their preferences, ensuring a safer and more personalized online experience while maintaining privacy.

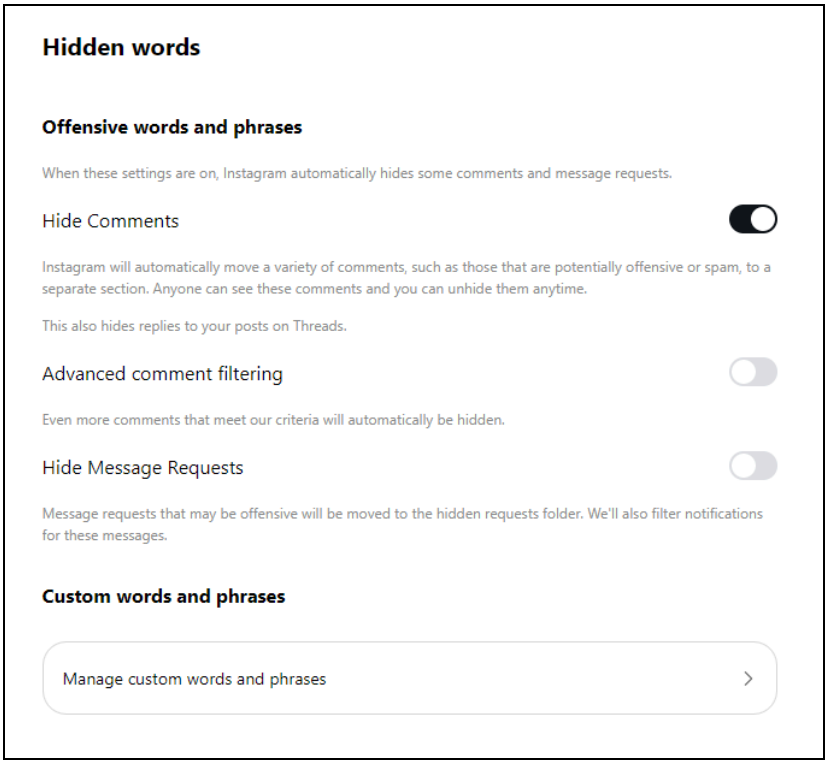


Figure 1.5: Instagram Comments Filteration Functionality

1.3.2 YouTube

YouTube provides flexible comment moderation options for video creators, allowing the management of comments based on their preference. It offers settings to allow all comments, hold potentially inappropriate comments for review, manually approve them before visibility, or turn off comments entirely as see in Figure 1.6. This system ensures control over the comments section according to the creator's moderation preferences.

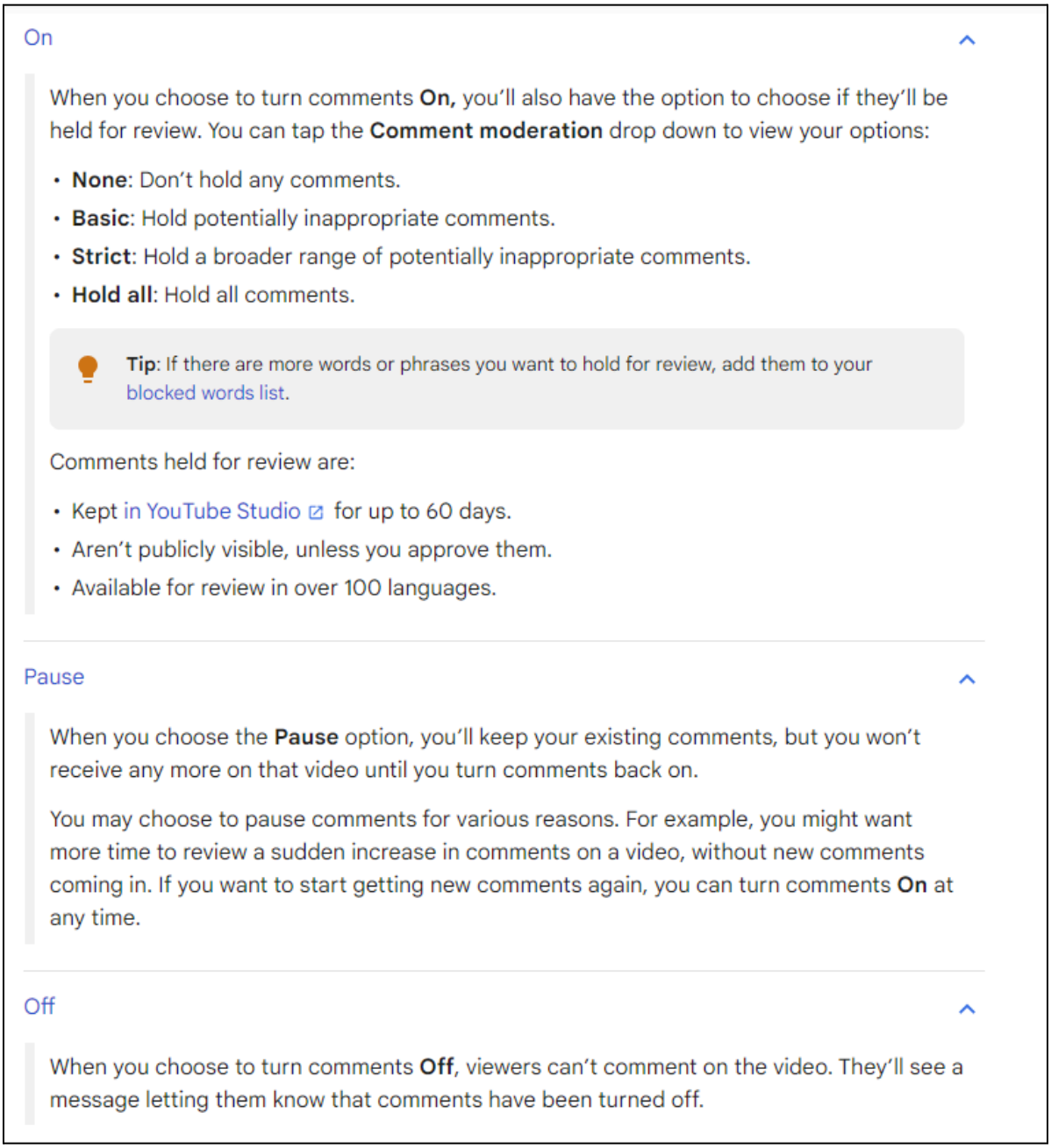


Figure 1.6: YouTube Comments Moderation Functionality

1.3.3 Cellar Web Instant Comment Management

CellarWeb, particularly for WordPress, effortlessly manage comments directly from the front-end display by incorporating options like 'Spam', 'Trash', and 'Delete' alongside the standard 'Edit' feature as seen in Figure 1.7. Execute the chosen action instantly without navigating through additional screens, swiftly removing the comment from view. This streamlined approach significantly enhances efficiency, particularly beneficial for websites inundated with numerous comments.

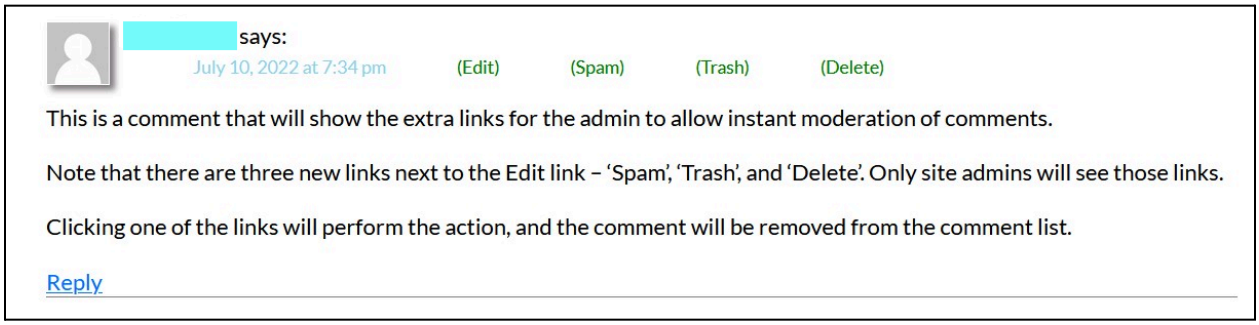


Figure 1.7: Cellar Web Instant Comment Managament Functionality

1.3.4 Perspective API

The Perspective API (Figure 1.8) leverages machine learning models to evaluate and score text for various attributes, aiding in the identification of abusive language and harmful content within online conversations. Beyond its flagship Toxicity attribute, which assesses the overall negativity of a comment, Perspective offers scores for Severe Toxicity, Insult, Profanity, Identity Attack, Threat, and Sexually Explicit content. These attributes enable developers and publishers to gain deeper insights into the nature of user-generated content, allowing for more effective moderation and content filtering strategies. By integrating the Perspective API into their projects, developers can empower platforms to provide feedback to commenters, facilitate easier review for moderators, and enable readers to filter out undesirable language, ultimately fostering healthier online interactions.

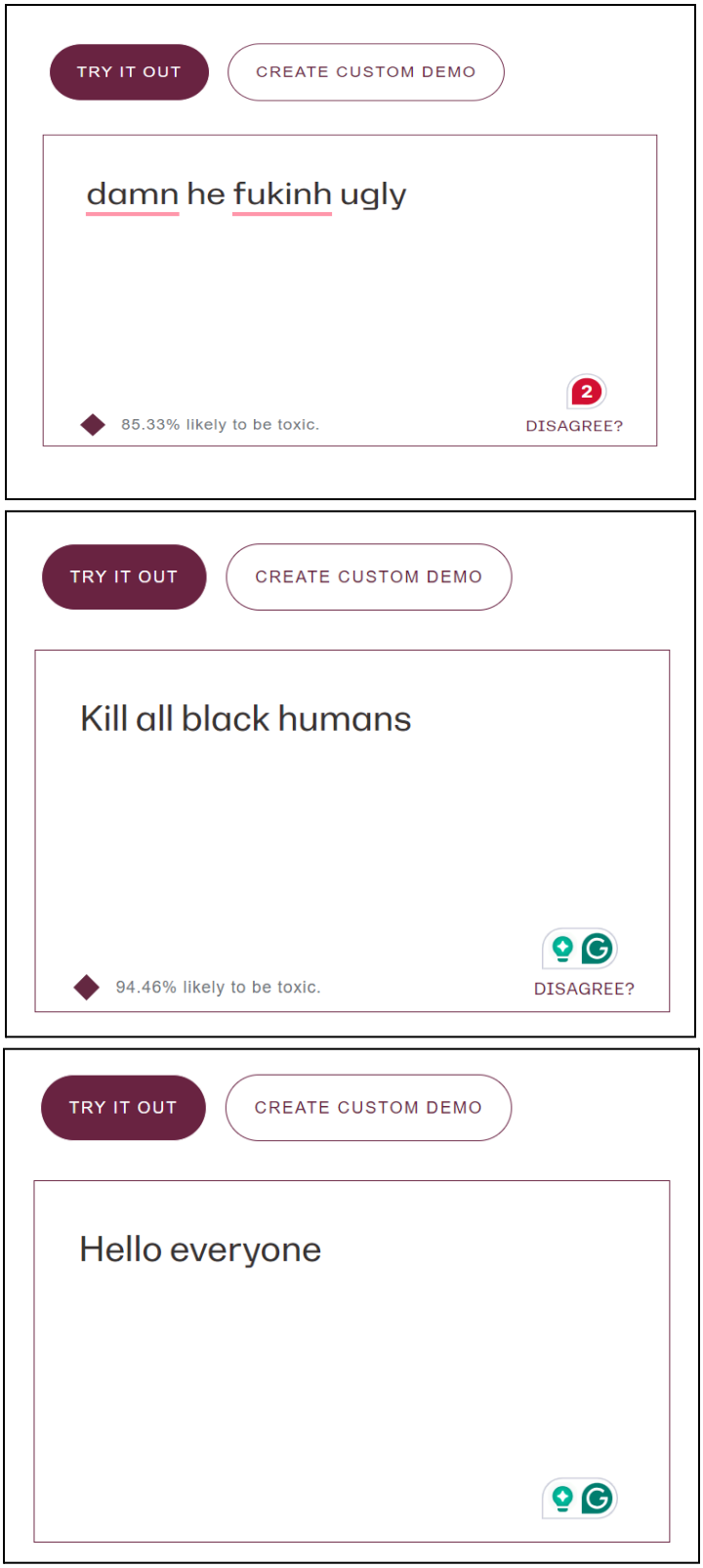


Figure 1.8: Perspective API Funtionality

2 Literature Review

This literature review aims to critically evaluate and categorize existing research in the field by focusing on language moderation efforts. It will delve into an in-depth analysis of findings and methodologies employed in relevant studies while also identifying gaps that this project aims to address. The review will be structured into two main sections: one examining conventional machine learning approaches to multilingual text classification and the other focusing on the utilization of Large Language Models (LLMs) for text classification.

2.1 Conventional Machine Learning Approaches

This section explores conventional machine learning approaches in hate speech detection and encompasses various methodologies and insights to mitigate online toxicity and promote healthier discourse. This section delves into a collection of studies that dissect the intricacies of detecting hate speech across different linguistic, cultural, and social dimensions. From scrutinizing responses to hateful content to assessing the performance of various classifiers and models, researchers have embarked on a journey to unravel the complexities of identifying and combatting online toxicity.

2.1.1 English Language

Samory et al. [1] focused their work on targeting sexism in English alone and provided a scale-based codebook and insights regarding the existing machine-learning algorithms to develop a better and broader model for sexism detection on social media. They proposed a theory-driven data annotation approach that incorporates psychological definitions of sexism to discover the hidden forms of sexism. They used their codebook to annotate datasets and concluded that sexism detection models could be made better by the use of adversarial examples, making the models more reliable, and by testing the model with various dimensions of sexism as a construct, improving its reliability. Their research, however, is arguably too narrow, as hate speech spreads way beyond sexism alone.

A wider view is represented by Albanyan and Blanco [2], who argue that the key to solving the problem of detecting hate speech is to evaluate the “naturally occurring replies” to hateful content. An English dataset was used to analyze the dynamics of hateful content and its corresponding responses. They propose that studying the relationship between hateful content and the responses to it can be classified as either counter-hate, justification, attacking the author, or additional hate. Their research suggests that the relationship between hateful content and its replies could help moderate hate speech on social media.

Research carried out by Pamungkas et al. [6] focuses on abusive and swear words in English, taking the possibility of not only detecting but also predicting the abusiveness of a swear word in any given context as the main investigation perspective. Though their dataset was rather small, they were able to develop models to automatically predict abusive swearing, to provide a deep evaluation of their corpus, and to confirm the robustness of the resource. They used BERT for sequence labeling while using simpler yet more transparent models for text classifications. They ended up discovering that a wide range of features hold the potential to further improve the model’s performance.

Viacheslav Zhukov [7] and the Toloka ML team explored text classification performance on extra-small datasets using various NLP models. Their experiments revealed that larger models trained on extensive data outperform smaller models, especially with limited training data. Few-shot classification with ChatGPT displayed promise but depended heavily on input prompts, while fine-tuning the GPT-3 model yielded impressive results. The conclusion emphasizes the importance of model selection and highlights that, as dataset size increases, performance differences diminish. Additionally, domain-adapted classical

models, such as RoBERTa model, can outperform generic Large Language Models in specific tasks, especially with large datasets, with implications for resource usage in online services.

2.1.2 Urdu Language

Urdu (Roman Script) represents Urdu language using the Roman alphabet, which is the same alphabet used in English. This form of Urdu is typically used for communication online, especially in social media platforms and digital messaging. On the other hand, Urdu (Arabic Script) represents Urdu language using the Nastaliq script, which is a variant of the Arabic script adapted specifically for Urdu. Unlike English, Urdu (Roman Script) and Urdu (Arabic Script) lacks language resources and annotated datasets making it more complex to scan for hate speech (Rizwan et al.) [8]. Rizwan et al. [8] attempt to establish a dataset of hateful terms in Urdu (Roman Script), a prevalent form of communication online, and test the feasibility of existing embedding models in terms of transfer learning. They propose a new deep learning model called CNN-gram and compare its performance to seven existing baseline architectures on their established dataset. Their study concluded, with rigorous research and workings, that their proposed model demonstrates more accuracy and robustness compared to the existing baseline models and that transfer learning beats training embedding from scratch to detect hate speech in Urdu (Roman Script)

Saeed et al. [9] also recognize the lack of resources available for Urdu (Roman Script) and the limitation of existing research to hate speech in terms of racism, sexism, and hatred of religions, mostly. To narrow their research to target three forms of hate speech: religion, racism, and national origin, while categorizing severity levels of such language into symbolization, insult, and attribution. They found that word n-grams and char n-grams played a key role in detecting hate speech in Urdu (Roman Script) while embedding-based features performed well in highlighting moderately infrequent patterns of hate speech. Even though their results showed traditional methods outperforming deep learning models, they still strongly suggest that deep neural networks hold great potential to contribute to this issue. They recommend addressing the degrees of language in future works as Urdu (Roman Script) is a language greatly lacking in datasets and resources.

Bilal et al. [10] conducted significant research and working with regards to hateful speech detection in Urdu (Roman Script). To scan the language, they deployed a transformer-based model due to its ability to capture the text context. They also pioneered the first Urdu (Roman Script) pre-trained BERT model, which they called BERT-RU, and trained it from scratch on the largest Urdu (Roman Script) dataset. They assigned traditional and deep learning models as baseline methods. Their research showed evidence that transformer-based model outperformed traditional and deep learning models, both, in terms of accuracy and, precision, recalls, and F-measure. The proposed transformer-based model excelled in generalization across diverse domains, highlighting its versatility. To categorize approaches, consider three categories: traditional ML, neural networks, and transformative transformer-based AI, providing a clear framework for understanding their impact on cross-domain dataset generalization.

Like Saeed et al. [9] and Bilal et al. [10], Azam et al. [11] also recognizes the failure of deep-learning models in detecting hate speech in Urdu (Roman Script) due to the it being a low-resource language. They, instead, emphasize the importance of data augmentation techniques to exploit the limited datasets available in order to improve generalizability. After an in-depth analysis and application of different data augmentation techniques on two datasets of Urdu (Roman Script), they concluded that these techniques allowed them to improve performance not only in the primary metric of comparison, but also in recall which is impertinent for human-in-the-loop AI systems.

Aziz et al. [12] carried out a research as well, however, they focus on the geo-political context. To tackle the wide lexical structure of Urdu (Roman Script), they proposed an algorithm for lexical unification of Urdu (Roman Script). They go on to develop three vectorization techniques: TF-IDF, word2vec, and fastText. Their results showed that, like Alias et al. [4] work, Random Forest, along with a feed-forward, neural network, produces the highest accuracy using fastText word embedding to identify the differences between neutral and politically offensive speech.

2.1.3 Data Processing and Feature Extraction

A similar approach is demonstrated by Grimminger and Klinger [3], who joined stance detection with hate speech detection to evaluate the style, tone, and implicit meanings and intentions in comments regarding politicians. Additionally, it's worth noting that they utilized only the English subset from a multi-lingual resource for their analysis. This focused approach allowed for a more targeted examination of the nuances within English-language political discourse. To enhance the accuracy of such models, they suggest it is important to add the nicknames of politicians to the datasets and to better distinguish between a political party and its candidate. This symbolizes the importance of the inclusion of alternative words of common hateful words into datasets, which this research paper will be addressing as well.

Alias et al. [4] proposed a technique used for video-sharing spam comments feature detection. They used pre-processed datasets in English and applied six classifiers to test their accuracies. These classifiers were Random Tree, Random Forest, Naive Bayes, Kstar, Decision Table, and Decision Stump. Their rigorous, numerical research showed that Decision Stump produced the lowest accuracy in detection at 58.86%, while Random Forest, a classifier they concluded was the most ideal for future use, produced 90.57% accuracy. Similarly, Androcec [5] collected and analyzed data from thirty-one selected primary relevant studies to evaluate and systematically review the current “toxic comment classification” using machine learning methods. In his study, Hindi, English, and some Korean datasets were analyzed to provide a broader perspective on toxic comment classification across different languages. His study recommends using transformers for such classification in future works as they have recently shown superior performance in many natural language processing tasks.

2.2 Large Language Models (LLMs) Approaches

This section of the literature review delves into recent research exploring the capabilities and limitations of LLMs in various Natural Language Processing (NLP) applications, analyzing their effectiveness in text classification, sentiment analysis, and specialized domains like online predator detection. By examining studies that highlight both the strengths and potential shortcomings of LLMs, we aim to gain a comprehensive understanding of their current impact and future directions in the field.

LLMs have shown immense performance across many AI and natural language processing tasks. However, Abburi et al. [18] acknowledge the creation of undesirable consequences due to the unregulated malign application of these models, such as the generation of fake news, plagiarism, etc. In their paper, they discuss their submission to the AuTexTification shared task, exploring the performance of their LLM in two types of tasks. They discover that in the task of human or generated binary classification, the LLM ranks fifth and thirteenth in English and Spanish, respectively. However, for Model Attribution Multiclass classification tasks, they rank first in both languages.

In an extensive study by Chae and Davidson [16] focuses on the use of LLMs as a methodology for computational sociology, specializing in applications to supervised text classification. They evaluated four different existing LLM designs varying in size, training data, and architecture, through a case study on

identifying opinions about politicians on Twitter and Facebook. They concluded that LLMs can definitely perform complex text classification tasks with high accuracy by sharing their model performance with different confusion matrices, substantially outperforming conventional machine learning approaches. Their study suggested fine-tuning smaller models in order to achieve optimal solutions for most researchers due to their higher accuracy and lower costs.

To understand the alignment between the reasonings of humans and the AI model, Venkatesh et al. [17] conducted a study to compare the human text classification performance and explainability with a traditional Machine learning model and LLM. They used injury narratives that had to be classified into six cause-of-injury categories to test their models. Their findings show that ML models performed better overall as compared to ChatGPT and humans, unlike Chae and Davidson's [16] findings.

Zhan et al. [19] acknowledge the potential of LLMs being employed on Sentiment Analysis (SA) problems as it has been a longstanding research area in the field. Their paper seeks to provide an elaborate evaluation of the capabilities of LLMs in performing various SA tasks, from conventional sentiment classification to aspect-based sentiment analysis and multifaceted analysis of subjective texts. Experimental results in their study reveal that LLMs perform really well on simpler tasks in a zero-shot setting, but they struggle with more complex tasks. In some contexts, LLMs consistently outperformed SLMs, which suggests that their capabilities are better utilized in situations where annotation resources are scarce. Furthermore, the findings reiterate that LLMs have surely opened new avenues for SA tasks, as they offer effective tools and exciting research directions for the exploration of SA. Similarly, Peiqi Ji's [22] study features the systematic evaluation of the Alpaca model's performance across a variety of datasets, with an emphasis on its capabilities to excel in sentiment analysis tasks in order to prove insights into its adaptability and effectiveness. In its findings, the study reveals LLMs excel in situations where tasks involve a scarcity of possible answers, which shows their capacity to handle nuanced and contextually rich sentiment analysis challenges. However, the paper highlights the potential for improved accuracy through the utilization of larger training datasets.

Krugmann and Hartmann [20] also conducted a study on the application of LLMs on Sentiment Analysis by benchmarking the performance of three state-of-the-art LLMS, i.e., GPT- 3.5, GPT-4, and Llama 2, against established, high perming transfer learning models. Their study revealed that LLMs can not only compete with but also, in some cases, surpass the traditional methods in terms of sentiment classification accuracy. They also find that linguistic elements such as lengthy, content-laden words improve classification performance, whereas other elements like single-sentence reviews and less structured social media text documents hinder performance. These findings are different from Zhan et al. [19] and Peiqi Ji [22] in the aspect that traditional models do not hold any ground. However, it does agree with Zhan et al. [19] when it comes to the under or at-par performance of LLM compared to traditional methods in more complex tasks.

A more specialized study was conducted to evaluate the application of LLMs to detect sexually predatory behavior online using the Llama 2 7B-parameter model by Nguyen, Wilson, and Dalins [21]. They rely on the power of LLMs and their study of a manual search for synergy between feature extraction and classifier design steps like conventional methods in this domain. Their study revealed strong support for such an application of LLMs as they performed proficiently and consistently acro, varying in size, training data, and architecture, through a case study such an application is not only limited to texts in English but also other languages. The generic and automated nature of their study allows their findings to be applied to other areas like cybersecurity, legal and compliance, social media and social listening etcetera, as well.

The paper by Sun et al. [15] highlights the underperformance of base LLMs compared to fine-tuned large language models in text classification tasks, unlike most of the other studies discussed, due to a lack of reasoning ability and the limited number of tokens allowed in in-text learning. They introduce Clue And Reasoning Prompting (CARP), which adopts a progressive reasoning strategy specifically made to address the complex linguistic phenomena involved in text classification. This CARP yielded new SOTA performances on 4 out of 5 widely-used text-classification benchmarks. The study found that CARP delivers promising abilities on low-resource and domain adaption setups.

3 Technology Review

The project incorporates a diverse array of cutting-edge technologies to enhance its functionality and user experience. Several advanced language models and technologies have emerged, each bringing unique strengths to various applications. Among these, GPT-3.5-Turbo-1106 and GPT-3.5-Turbo-0125 represent the pinnacle of OpenAI's developments, boasting enhanced functionalities and user-centric improvements. Meanwhile, Llama-2-70b and its conversational variant, Llama-2-70b-chat, offer robust performance in dialogue systems, despite their smaller parameter count compared to the GPT-3.5 models. Mistral-7b provides an efficient alternative for resource-constrained environments with its streamlined architecture. Additionally, React.js remains a cornerstone in web development, known for its ability to create dynamic and efficient user interfaces. This diverse landscape of AI models and frameworks caters to a wide array of needs, from natural language processing to user experience optimization.

3.1 GPT-3.5-Turbo-1106

This variant of GPT-3.5 Turbo boasts an identical parameter count of 200 billion but focuses on improved functionalities. It excels in following instructions precisely, offers a JSON output format for structured data, and delivers consistent results with its reproducible outputs feature. Additionally, GPT-3.5-Turbo-1106 allows for parallel function calls, streamlining complex tasks. The model testing and the setup details can be seen in the section 8.1.4 of the document.

3.2 GPT-3.5-Turbo-0125

It reigns as the latest and most refined version of the GPT-3.5 Turbo family. While maintaining the impressive 200 billion parameter structure, this update prioritizes user experience. It boasts significantly improved accuracy in responding to your specific format requests, eliminating the need for reformatting. Non-English language tasks are now flawless thanks to a bug fix for text encoding issues. Furthermore, OpenAI has implemented lower pricing, making this powerhouse more accessible. We finalised to go ahead with this model and details can be seen in Model Training and Fine-Tuning Section of the document.

3.3 Llama-2-70b

Llama-2-70b is a powerful language model with 70 billion parameters, making it a strong contender in the field. While it might have a smaller parameter space compared to GPT-3.5 Turbo, it could offer advantages in areas like efficiency or specific task performance. The details of model fine tuning over VertexAI and the issues faces can be seen in the section 8.1.3 of the document.

3.4 Llama-2-70b-chat

This variant of Llama-2-70b prioritizes conversational fluency, making it ideal for chat applications or dialogue systems. With its 70 billion parameters, it can hold engaging and informative conversations, potentially excelling in back-and-forth communication. The model testing and the setup details can be seen in the section 8.1.1 of the document.

3.5 Mistral 7 B

Mistral-7b is a streamlined language model with 7 billion parameters, designed to provide efficient performance without compromising on quality. Despite its smaller parameter count, it delivers robust results in various NLP tasks, making it suitable for resource-constrained environments. The specifics of its fine-tuning process and the challenges encountered are documented in section 8.1.5 of the document.

3.6 React.js

React.js is popular for building user interfaces, particularly for single-page applications. It excels in providing a modular and efficient way to update and render components, improving the overall user experience. Its virtual DOM system helps optimize rendering performance.

4 Requirement Analysis

Social media platforms have become a diverse landscape where multilingual interactions are prevalent. According to Schroeder [13], while historically, much of the research on social media texts focused on English, the current scenario highlights a significant shift towards non-English languages. Fischer's study in 2011 [13] emphasized the varying language usage on platforms like Twitter across different geospatial locations. Despite English remaining a dominant language for web communication, there's a growing need for technological advancements catering to diverse linguistic backgrounds. The prominence of non-English languages, such as Urdu, is significant. Urdu, spoken by over 300 million individuals globally and with approximately 11 million speakers in Pakistan, presents a substantial case for technological adaptations, as highlighted in a survey [14]. Moreover, the preference for Roman Urdu, due to the predominant use of English scripts in modern technologies like computers and mobile phones, adds complexity to the linguistic landscape for local Urdu users [14].

The multilingual landscape of social media (Figure 4.1) necessitates an adaptable comment analysis tool capable of categorizing and providing insights from diverse comments across platforms, overcoming language barriers. Comprehensive checks are crucial for a holistic content analysis, irrespective of the language. Such a tool, tailored to multilingual platforms like Urdu, is vital for users seeking valuable insights from their content.

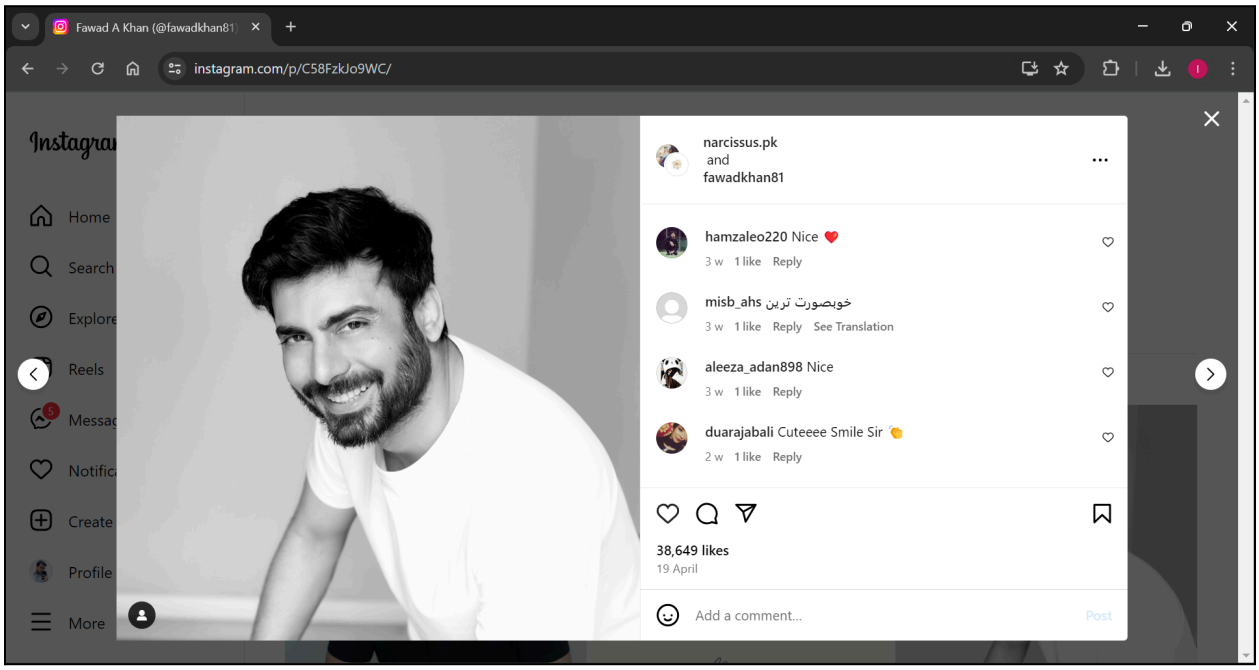


Figure 4.1: Multilingual Landscape Of Social Media

The section specifies the Software Requirement Specification, functional requirements, and non-functional requirements. This Software Requirement Specification (SRS) document serves as a foundational blueprint for the development of a cutting-edge solution that empowers users with the ability to categorize and filter comments.

4.1 Key Objectives and Functional Requirements

The primary objective of this SRS section is to define the software requirements for our comment classification tool clearly. It encompasses both functional and non-functional requirements that are essential for us to implement the developed model systemically.

4.2 The Power of AI and Multilingual Capabilities

In the digital landscape, the diversity of content knows no bounds, transcending language and culture. Hence, the SRS document also encompasses the requirements for an advanced AI model capable of analyzing comments in multiple languages. This will make our tool a versatile asset for a global audience.

4.3 Data Requirements

Effective comment analysis relies on quality data. This section will define the data requirements, including data sources and formats.

By the end of this SRS section, we will have a clear and comprehensive understanding of the software requirements necessary to bring our comment analysis tool to life. It will serve as a foundational guide for developers, ensuring that the final product not only meets but exceeds the expectations of our users.

As we progress with this project, we are committed to developing a robust tool that empowers content creators and social media enthusiasts to harness the true potential of comment analysis, irrespective of the language in which content is presented.

4.4 Requirement Gathering and Fact Finding

The initial phase of our research involved collecting sample comments from various social media platforms and subsequently labeling them with ground truth for English, Urdu (Roman Script), and Urdu (Arabic Script) languages. Following the collection of sample comments, we systematically categorized the data into six distinct labels that were previously defined through comprehensive research. In instances where ambiguity arose regarding the appropriate label assignment for certain comments, we devised a survey form and solicited responses from individuals to gain insights into their perspectives. Subsequently, based on the feedback received, we proceeded to assign labels to those comments accordingly. We employed OpenAI and Llama 2 to predict these comments, generating results as presented in Table 4.1, Table 4.2, and Table 4.3.

Upon comparing the predictions made by both OpenAI and Llama 2 with the ground truth, it was observed that the accuracy of predictions for English and Urdu comments was consistent between the two models. However, Llama 2 exhibited better predictions for comments written in Roman Urdu than OpenAI (Table 4.4 & Table 4.5).

To further enhance the accuracy of these models, we recognize the potential for improvement through advanced techniques such as prompt engineering, and fine-tuning of the models. We aim to develop an upgraded model capable of classifying comments into distinct categories: Positive, Neutral, Crisis, Negative - Respond, Negative - Ignore, and Negative - Removed.

Moving forward, our focus will be on implementing these techniques to elevate the predictive accuracy of the models, with the ultimate goal of creating a robust system for comment categorization across multiple languages and sentiments.

Table 4.1: English comments classification by OpenAI and Llama 2

Comment	Ground Truth	OpenAI	Llama 2
Cheers to you for making the trip	Positive	Positive	Positive
It's a great initiative of Dr. Najam. I appreciate it	Positive	Positive	Positive
damn she fukinh ugly	Negative - Remove	Negative - Ignore	Negative - Remove
Would bust in that slut.	Negative - Remove	Negative - Remove	Negative - Remove
Beggars can't be choosers	Negative - Ignore	Neutral	Negative - Respond
You're Middle class that's whyyy😂	Negative - Ignore	Neutral	Negative Respond
Their pricing is insane, infinitely raising prices has to run out at some point	Negative - Respond	Negative - Respond	Negative - Respond
What's scary is how underpowered the new computers are for being at such a premium price point. The iMac, at $1300 for 8GB and 256gb, sounds more like it belongs big 2013, not 2023	Negative - Respond	Negative - Respond	Negative - Respond
I will burn this planet down before I spend another minute living among these animals.	Crisis	Negative - Remove	Negative - Remove
dont worry i will destroy british people all in le head.	Crisis	Negative - Remove	Crisis
I am having dinner with my boyfriend	Neutral	Neutral	Positive
where is it dude?	Neutral	Neutral	Negative - Ignore

Table 4.2: Urdu (Roman Script) comments classification by OpenAI and Llama 2

Comment	Ground Truth	OpenAI	Llama 2
Sir Kamal kr dya hy	Positive	Neutral	Negatve - Remove
Ap jaisa koie banda nahi Khan sab the great	Positive	Positive	Positive
Ay lanti cartoon k samny hi milty han kia	Negative - Remove	Negative - Ignore	Negative - Remove
Porn generals, khinzeer generals	Negative – Remove	Negative - Remove	Negative - Remove
Respect to ab bhool jao pori zindgi	Negative - Ignore	Positive	Positive
Kash apkey aney sey pehley app key oper Janey ki etlah ajahey	Negative- Ignore	Negative - Respond	Negative - Respond
Bycott karna chaheya iss Suzuki ka	Negative - Respond	Negative - Ignore	Negative - Respond
Price dekho aur itni third class gariyan banate hain ye.	Negative - Respond	Negative - Respond	Negative - Respond
Toh Asim ko jalaa dena chahiye tha Rashmi ka .	Crisis	Negative - Ignore	Negative - Respond
If i were the captain of the flight Qadri was travelling in.. Maa qasam jahaz tabah ker dena tha!	Crisis	Negative - Ignore	Negative - Respond
Ye resturant ka khana sahi tha	Neutral	Positive	Positive
Chalo sahi hai	Neutral	Positive	Positive

Table 4.3: Urdu (Arabic Script) comments classification by OpenAI and Llama 2

Comment	Ground Truth	OpenAI	Llama 2
ماشاءاللہ یہ دیکھنا اچھا ہے کہ کس طرح پیشہ ورانہ اور نصابی سرگرمیاں نوجوان آئی ٹی پروفیشنلز کے فائدے کے مواقع پیدا کر رہی ہیں۔ میرے بھائی نے بھی نیٹسول میں اپنی انٹرن شپ مکمل کی۔ یہ جگہ آئی ٹی انڈسٹری کے نئے انجینیئرز کے لیے بہت فائدہ مند ہے۔	Positive	Positive	Positive
کوئی بھی چیز دل کو ایسے زندہ نہیں کرتی جیسے قرآن پاک کرتا ہے	Positive	Positive	Negative - Respond
اللہ پاک اس ناسور خاندان سے پاکستانیوں کی جلد جان چھڑوائے امین	Negative - Remove	Negative - Respond	Negative - Remove
جس قوم کے لیے بہترین تفریح کسی ٹک ٹاکر کا ٹوٹا ہے اس نے خاک ترقی کرنی ہے تو پھر اُس قوم پر مریم نواز جیسی گشتی عورت اور نواز شریف جیسے دلے ہی حکمران ہو گے	Negative - Remove	Negative - Respond	Negative - Respond
میرا ایک دوست بھی ایسی حرکتیں کرتا تھا آج اس کا چالیسواں ہے اور میری پیشی ہے	Negative - Ignore	Neutral	Positive
نہ جانے کس غار سے نکلا ہے، کوئی ہے جو اسے واپس بند کر دے	Negative- Ignore	Negative - Respond	Negative - Respond
قیمت بڑھاتے رہنا۔ کوالٹی اچھی نا کرنا۔	Negative - Respond	Negative - Ignore	Negative - Respond
اضافہ 5...5 لاکھ اور کمی صرف ایک لاکھ کھلا مزاق ھے۔	Negative - Respond	Neutral	Negative - Respond
بجوا کو میں جان سے مار دوں گا۔	Crisis	Crisis	Negative - Respond
ان جیسوں کو زندہ دفنا چاہیے۔	Crisis	Negative - Respond	Negative - Respond
اَچھا	Neutral	Positive	Positive
کام کر لو	Neutral	Neutral	Negative - Respond

Table 4.4: Percentage of Correct Predictions by OpenAI

Comment Type	% Correct Predictions
English	58.33%
Roman Urdu	16.67%
Urdu	33.33%

Table 4.5: Percentage of Correct Predictions by Llama 2

Comment Type	% Correct Predictions
English	58.33%
Roman Urdu	33.33%
Urdu	33.33%

4.5 System Environment

The system environment depicted in Figure 4.2 illustrates the architecture of an AI model, accompanied by an exposed API, which is subsequently utilized within both a demo website and a Google Chrome extension to showcase its functionalities. This arrangement enables social media influencers to access the demo website for an overview of the model's capabilities and employ the Google Chrome extension to classify comments effectively. Likewise, developers are afforded the same privileges, being able to leverage the extension and visit the website for evaluation purposes. Furthermore, developers possess the additional capability of integrating the Model's API into their projects, thereby extending the utility and versatility of the model within diverse contexts. This configuration underscores a comprehensive ecosystem facilitating interaction and utilization by both influencers and developers alike.

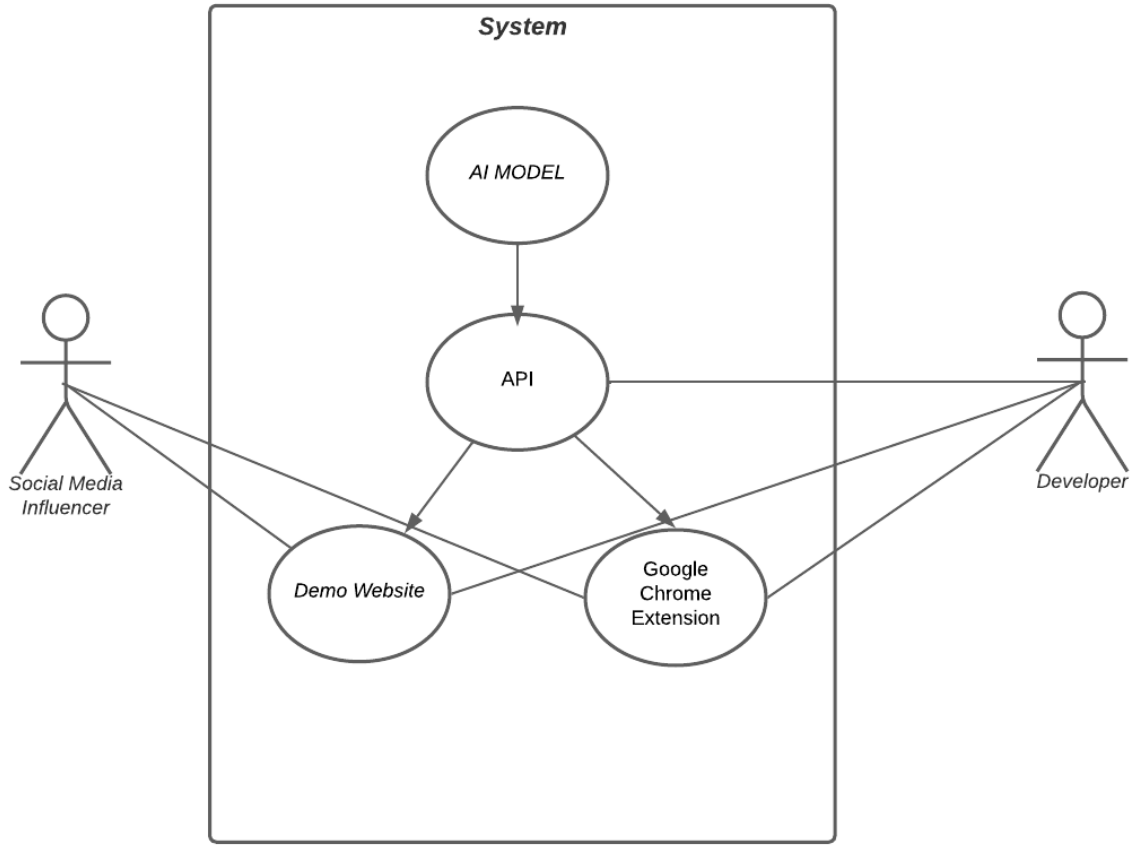


Figure 4.2: System Environment

4.5.1 User Roles

SocialSense will have two user roles:

4.5.1.1 Social Media Influencer

A user who wants to interact and engage with their audience using the comments categorization.

4.5.1.2 Developer

A user who seeks to incorporate our model into their project by integrating our API.

4.5.2 User Stories

4.5.2.1 User Story 1: Social Media Influencer

As a social media influencer, I want to leverage an intelligent comment categorization tool, so I can effectively manage and engage with the comments on my posts.

4.5.2.2 User Story 2: Developer Integrating with Comment Categorization API for Social Media Platform

As a developer working on a new social media platform, I am excited to integrate Comment Categorization API into our application to enhance user experience. Our primary goal is to provide post owners with the ability to categorize and filter comments, making it easier for them to manage and engage with their audience.

4.6 Software Requirement Specifications

4.6.1 Functional Requirements

4.6.1.1 AI Model Requirements

- **FR1.1:** Comment Categorization
    - The model should be able to classify the comments into the following categories: “Positive”, “Neutral”, “Negative - Respond”, “Negative - Ignore”, “Negative - Remove”, and “Crisis”.
- **FR 1.2:** Text Preprocessing
    - The model must perform text preprocessing to clean and normalize user-submitted comments. This includes tasks tokenization, lowercasing, and removing special characters.
- **FR 1.3:** Large Language Model (LLM) Model Selection
    - Choose a suitable Large Language Model (LLM) Llama 2, and GPT 3.5 for sentiment analysis and categorization.
- **FR 1.4:** Model Fine Tuning
    - Fine-tune the selected model on a labelled dataset.

4.6.1.2 Web Interface and Extension

- **FR 2.1:** Users should be able to submit comments or reviews for categorization.
    - **Description:** The web application should provide an interface for users to input comments or reviews for analysis.
    - **Actor:** End Users
    - **Precondition:** The user is on the comment submission page.
    - **Postcondition:** The user's comment is submitted for categorization.
    - **Details:**
        - Users can enter or paste their comments into the input field.
        - The web application should handle input in multiple languages, including English and Urdu.
- **FR2.2, 3.1:** Users should receive categorization results.
    - **Description:** The web application and extension should display categorization results to users after analyzing their submitted comments.
    - **Actor:** End Users
    - **Precondition:** The user has submitted a comment.
    - **Postcondition:** The user receives categorization results.
    - **Details:**
        - The web application and extension processes the submitted comment and provides categorization labels; “Positive”, “Neutral”, “Negative - Respond”, “Negative - Ignore”, “Negative - Remove”, and “Crisis”.
        - Users should have access to a clear and understandable display of the results.

- **FR 2.3:** User Feedback
    - **Description:** The web application should provide a mechanism for users to provide feedback or contact us.
    - **Actor:** End Users
    - **Precondition:** The user is viewing the webpage.
    - **Postcondition:** The user can submit feedback or contact the developers
    - **Details:**
        - The feedback can include options to report misclassified comments or provide suggestions for improvement.
        - Users' feedback should be collected and considered for refinement and enhancement.

4.6.1.3 API

- **FR 4.1:** Documentation
    - **Description**: Provide comprehensive documentation for the API to guide users on using it effectively.

- **Actor:** API Users
- **Precondition:** The API should be implemented and ready for use
- **Postcondition:** Users have access to the comprehensive documentation that provides detailed guidelines on how to use the API
- **Details:**
  - Include API usage examples, endpoint descriptions, input and output formats, and authentication instructions.
  - Update the documentation as the API evolves.

- **FR 4.2:** API Access Control
  - **Description:** The API should have access control mechanisms to authenticate and authorize users or applications that want to use it.
  - **Actor:** API Users
  - **Precondition:** The API is operational, and access control mechanisms, API keys, OAuth, or other secure authentication methods, have been successfully implemented
  - **Postcondition:** API users, both individuals and applications, can securely access the API while adhering to the established access control measures.
  - **Details:**
    - Implement API keys, OAuth, or other secure authentication methods to verify the identity of users or applications.
    - Define access levels and permissions to control which operations each user or application can perform.
- **FR 4.3:** API Endpoint Definition and Documentation
  - **Description:** The API should have well-defined endpoints, and comprehensive documentation should be provided to guide users on how to interact with each endpoint effectively.
  - **Actor:** API Users
  - **Details:**
    - Define and document each API endpoint, specifying the purpose, path, and parameters associated with each endpoint.
    - Clearly state which HTTP methods (e.g., GET, POST, PUT, DELETE) are supported by each endpoint and the expected functionality of each method.
    - Document the required and optional request parameters for each endpoint, along with their data types and constraints.
    - Specify the format of responses returned by each endpoint, JSON or XML, and provide examples of the response structures.

4.6.2 Non-Functional Requirement

4.6.2.1 Data Requirements

- NFR 5.1: Data Diversity
  - **Description:** Ensure diversity in the collected dataset to account for a wide range of user-generated text.

    - **Details:**
        - Collect comments and reviews in multiple languages, including English, Urdu, and Roman Urdu, to create an inclusive dataset.
        - Aim to include a representative sample of user interactions from different demographics and regions.
- NFR 5.2: Data Sampling
    - **Description:** Implement data sampling strategies to ensure that the dataset represents different sentiment categories.
    - **Details:**
        - Ensure that the collected dataset includes a balance of "Positive", "Neutral", "Negative - Respond", "Negative - Ignore", "Negative - Remove", and "Crisis" sentiment comments for training and testing purposes.
        - Employ random or stratified sampling methods to create a diverse training dataset.
- **NFR 5.3:** Data Anonymization
    - **Description:** Protect user privacy by anonymizing data during collection.
    - **Details:**
        - Remove personally identifiable information (PII) from the collected comments and reviews.
        - Replace usernames or account names with generic identifiers.
- **NFR 5.4:** Data Collection from various social media platforms
    - **Description:** The data will be collected from various social media platforms.
    - **Details:**
        - Collect data from platforms such as Instagram, X, Facebook, and others to fetch user-generated content.

4.6.2.2 Performance Requirements

- **NFR 6.0:** Response Time
    - The system should respond to user actions and requests within a maximum response time of 4 seconds under normal operating conditions.
- **NFR 6.1:** Scalability
    - The application should be capable of handling a scalable number of concurrent users.
- **NFR 6.2:** Load Testing
    - The system must undergo load testing to ensure it can handle expected peak loads while maintaining acceptable performance levels.

4.6.2.3 Availability and Reliability

- **NFR 6.3:** Uptime
    - The system should aim for high availability with a minimum uptime of 90% during normal operation.
- **NFR 6.4:** Failover and Redundancy
    - Redundancy measures, backup servers, and failover systems should be in place to minimize downtime in the event of system failures.

4.6.2.4 User Experience (Usability)

- **NFR 6.5:** User Interface (UI) Design
    - The application should follow modern design principles and provide a user-friendly and intuitive interface.
- **NFR 6.7:** Accessibility Support
    - The web application and extension should be accessible to a wide range of users, including those with disabilities.

4.6.2.5 Maintenance and Support

- **NFR 6.8:** Software Updates
    - Regular software updates and security patches must be provided to address vulnerabilities and improve model functionality.
- **NFR 6.9:** Technical Support
    - A support team should be available to address user queries and issues during specified hours.

4.7 AI/ML requirements

4.7.1 Data Requirements

4.7.1.1 Data Sources

- We will be collecting data from various social media platforms, including Instagram, X, Facebook, and Reddit.

4.7.1.2 Data Collection

- We will be using different data scraping tools for faster data collection. The data scraping tools that will be used are Power Automate, Instant Data Scraper, and GPTBot.

4.7.2 Feature Engineering

4.7.2.1 Text Preprocessing

- **Tokenization:** Split comments into individual words or tokens.

4.7.2.2 Prompt Engineering

- **Clear and Specific Prompts:**
  - Design prompts that are clear, concise, and tailored to the specific task of comment categorization.
- **Keyword Inclusion:**
  - Include relevant keywords or phrases in prompts that can guide the model in understanding the desired categorization.

4.7.2.3 Model Selection

- We will be using fine-tuned LLM, Llama-2-70b-chat, Llama-2-70b, GPT-3.5-Turbo-1106 and GPT-3.5-Turbo-0125 for sentiment analysis and categorization.

4.7.2.4 Model Training and Testing

- **Validation:**
  - Assess the model's performance on the validation dataset to avoid overfitting. To evaluate the model's classification capabilities, use metrics like accuracy, precision, recall, F1-score, and AUC-ROC.
- **Fine-tuning:**
  - If the validation results indicate underperformance, consider further fine-tuning the model, adjusting hyperparameters, or introducing regularizations to enhance accuracy and generalization.
- **Model Evaluation:**
  - Once the model has been fine-tuned and validated, evaluate its performance on the testing dataset. This evaluation provides a more unbiased estimate of how well the model will perform in a real-world scenario.
- **Evaluation Metrics:**
  - Use a combination of evaluation metrics suitable for text classification tasks. Common metrics include accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).
  - In the context of specific categorization labels ("Positive," "Neutral," "Negative - Respond," "Negative - Ignore," "Negative - Remove," and "Crisis"), evaluate the model's ability to classify comments into these categories correctly.
- **Interpretation and Feedback:**
  - Analyze the model's performance, and if necessary, iterate on the model, data preprocessing, or feature engineering based on the evaluation results.

4.7.2.5 Integration with User Interface

- **Facebook Extension:**
  - The fine-tuned LLM recommendations will be seamlessly integrated into the Chrome extension's user interface by assigning specific colors to circles next to each comment on Facebook.

- **Instagram Extension:**
    - The fine-tuned LLM recommendations will be seamlessly integrated into the Chrome extension's user interface by assigning specific colors to circles next to each comment on Instagram.
- **TikTok Extension:**
    - The fine-tuned LLM recommendations will be seamlessly integrated into the Chrome extension's user interface by assigning specific colors to circles next to each comment on TikTok. Furthermore, the user can filter the comments based on their desired selection.
- **Webpage:**
    - On the webpage, users can test the model by inputting comments or reviews, and the fine-tuned LLM recommendations will be presented through a user-friendly interface. After submitting a comment, users will receive categorization labels or color-coded indicators to understand the sentiment or intent behind the text. The website will make the model's recommendations easily accessible and comprehensible, enabling users to experience the model's capabilities firsthand.

5 Data Preparation

In today's digital world, social media platforms are full of people sharing their thoughts and opinions every day. To understand this huge amount of information, we need to prepare the data carefully. This study looks at how we collected, created, and labeled around 80,000 social media comments from platforms like Facebook, Instagram, TikTok, YouTube, and Reddit. Using tools like Instant Data Scraper and the Mistral 7B model, and following clear labeling rules, this research helps us understand and manage different types of online comments, especially those that might lead to serious problems.

5.1 Data Collection

Approximately 80,000 comments were scraped from various platforms like Facebook, Instagram, TikTok, YouTube, and Reddit using Instant Data Scraper (Figure 5.1). These comments encompass diverse social media posts, including those related to political discussions, religious influences, workplace issues, and various other topics, offering a comprehensive snapshot of online discourse across multiple domains.

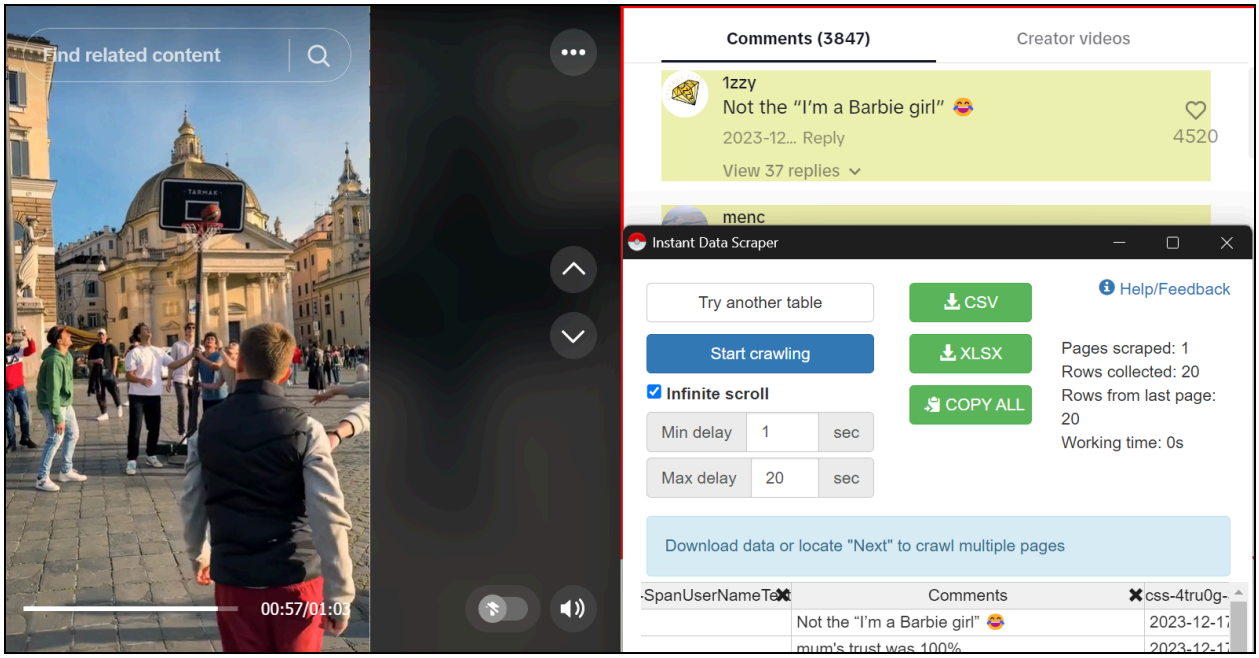


Figure 5.1: Instant Data Scraper Demonstartion

Instant Data Scraper, a powerful Chrome extension, facilitates easy seamless data extraction from websites. Whether you're interested in mining information from a TikTok post or any other webpage, this extension simplifies the process.

With Instant Data Scraper, you can effortlessly select and crawl through the desired columns of data. The intuitive interface allows you to customize your scraping experience, ensuring you extract exactly what you need. Once the data is collected, you can conveniently save it as a CSV file for further analysis or integration into your projects.

5.2 Data Generation

Due to the limited dataset available for our labeled category "Crisis," we leveraged the Mistral 7B model to generate additional data. The prompt employed for dataset generation specifically targeted comments with potential legal or criminal implications, such as threats of violence, breaches of confidentiality, defamation, and scenarios leading to PR disasters. It was crucial to ensure that the generated threats were of a high-level and notably violent in nature. This approach aimed to enrich our dataset with diverse examples that could better facilitate the training of our models to detect and address crisis situations effectively.

5.3 Data Labeling

Approximately 50,000 comments were labeled with our labels: Positive, Neutral, Negative-Respond, Negative-Ignore, Negative-Remove, and Crisis. The sample labeled data set can be seen in the Appendices A-E.

The labels for categorizing social media comments were carefully shortlisted following thorough research, which included insights from Jeff Bullas [23], Attention Insight [24], SproutSocial [25], and JSH Web Design [26], which shared types of social media comments. The definitions for each label are as follows:

1. **Positive:** Represents any comment that expresses favorability or positivity.
2. **Neutral:** Denotes a comment that is neither positive nor negative, conveying a neutral stance.
3. **Negative – Respond:** Signifies a genuinely negative comment that warrants a response or engagement.
4. **Negative – Ignore:** Identifies a negative comment originating from a "troll" – an intentional trouble-maker – which is best disregarded.
5. **Negative – Remove:** Indicates a comment that is offensive, malicious, or spam, breaching established "House Rules" and necessitating removal.
6. **Crisis:** Encompasses comments with potential legal or criminal implications, such as threats of violence, breaches of confidentiality, defamation, or PR disasters.

These labels serve as a comprehensive framework to systematically analyze and manage various types of social media interactions, ensuring effective engagement and the maintenance of a positive online environment.

In the process of labeling data, we encountered challenges when dealing with comments that initially appeared positive but turned negative midway. To address this ambiguity, we initiated a comprehensive research effort. A dedicated form was created, featuring a variety of comments, prompting users to assign labels. Through subsequent discussions with participants, it became evident that any comment containing negative words or intentions, regardless of an initial positive tone, was consistently labeled as negative.

It is noteworthy that due to constraints, a detailed survey could not be conducted. Notably, an observation surfaced during the form-filling process where several male respondents tended to label comments as positive even when they contained explicit and derogatory content, particularly harmful towards women.

This discrepancy underscores the importance of refining the labeling criteria and emphasizes the need for a nuanced approach to discerning negativity, especially in cases where the apparent positivity masks harmful undertones. The insights gathered from this research will contribute to a more accurate and sensitive labeling system, ensuring a better understanding of the complexity in comments that exhibit mixed sentiments.

6 Model Training and Fine-Tuning

To refine GPT-3.5 for text classification, a structured approach was taken involving data preparation, model fine-tuning, and testing. A function called `create_dataset` generated a JSONL dataset with a default prompt to guide classifications. The GPT-3.5 Turbo model, version 0125, was fine-tuned using a balanced dataset of 1800 samples across various labels and languages, and optimized over three epochs. The fine-tuned model, 'ft:gpt-3.5-turbo-0125:personal:socialsense:96eqMPvL', was then tested on 900 comments to assess its classification accuracy, demonstrating improved performance in social media comment classification.

6.1 Training Data Preparation

A JSONL file was prepared to facilitate the creation of a dataset for fine-tuning GPT-3.5. Within this context, a function named `create_dataset` (Figure 6.1) was written. In this JSONL format, the key `"role": "system"` signifies messages generated by the system itself, often representing prompts or initial messages, while `"role": "user"` indicates messages provided by the user. The function is designed to process these JSONL files, likely parsing and organizing the data for model training.

Additionally, a default system prompt, denoted as DEFAULT_SYSTEM_PROMPT, was configured as *'You are a text classifier for social media comments. Classify the following comment into one of the following classes: [Positive, Neutral, Negative - Respond, Negative - Ignore, Negative - Remove, Crisis]*'. This prompt likely guides users in providing appropriate labels or classifications for the data, which is crucial for training the text classification model effectively. Overall, the function and JSONL structure appear integral to the process of preparing and refining datasets for training GPT-3.5, ensuring the model can effectively understand and generate human-like responses.

```
def create_dataset(Comments, ground_truth):
    return {
        "messages": [
            {"role": "system", "content": DEFAULT_SYSTEM_PROMPT},
            {"role": "user", "content": Comments},
            {"role": "assistant", "content": ground_truth},
        ]
    }
```

Figure 6.1: Function to create JSONL Dataset

6.2 Model Fine-Tuning

In Figure 6.2, the model was fine tuned with the following specifications:

- **Base Model Selection:** Utilized the latest GPT-3.5 Turbo model, version 0125, known for its enhanced capabilities in natural language processing tasks.
- **Dataset Composition:** The dataset comprised 1800 samples, with 100 samples for each label in each language, mitigating biases in the training data.
- **Fine-Tuning Procedure:** The model underwent fine-tuning on a dataset consisting of 480,966 tokens, optimizing its performance over three epochs.
- **Validation Dataset:** A separate validation dataset was employed, containing 100 samples for each label in each language. This dataset served to evaluate the model's performance during training and ensure its generalization across various languages and labels.

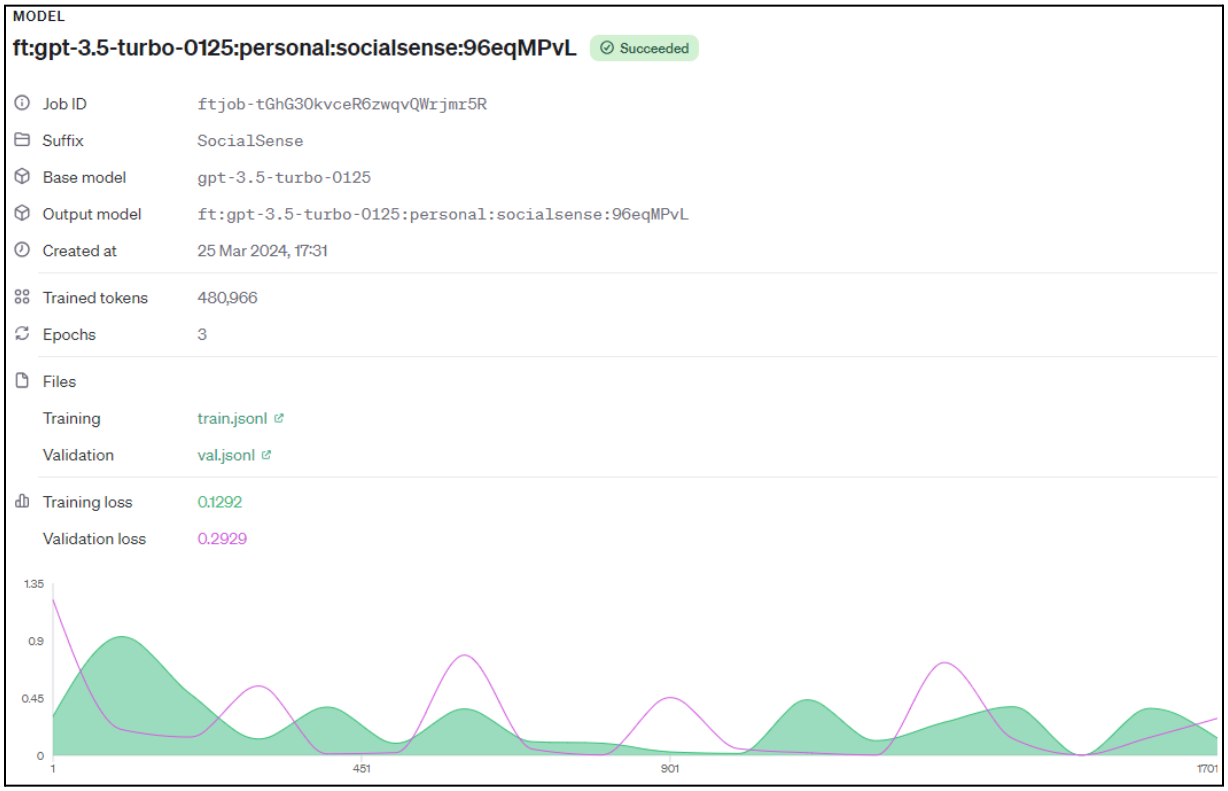


Figure 6.2: Model fine-tuning details

6.3 Model Testing Setup

In the model testing process, Figure 6.3 showcases the utilization of the fine-tuned GPT-3.5 model 'ft:gpt-3.5-turbo-0125:personal:socialsense:96eqMPvL' for text classification purposes. This involved applying the model on a testing dataset consisting of 900 comments, with 50 comments per language for each category. The system prompt provided to the model was: *'You are a text classifier for social media comments. Classify the following comment into: [Positive, Neutral, Negative - Respond, Negative - Ignore, Negative - Remove, Crisis]'*. Through this approach, the model's performance and accuracy in classifying comments into distinct categories were evaluated. The testing results can be seen in the Evaluation section.

```
for index, row in df.iterrows():
    comment = row['Comments']
    completion = client.chat.completions.create(
        model="ft:gpt-3.5-turbo-0125:personal:socialsense:96eqMPvL",
        messages=[
            {"role": "system", "content": "You are a text classifier for
social media comments. Classify the following comment into one of the
following classes: [Positive, Neutral, Negative - Respond, Negative -
Ignore, Negative - Remove, Crisis]"},
            {"role": "user", "content": comment}
        ]
    )
    response = completion.choices[0].message.content
    df.at[index, 'Prediction'] = response
```

Figure 6.3: Function to test the model and generate a response

7 Design

SocialSense is a comment categorization model proficient at classifying comments into six distinct categories: Positive, Neutral, Crisis, Negative-Ignore, Negative-Remove, and Negative-Respond. Leveraging Language Models (LLMs) has been fine-tuned to excel in processing English, Urdu (Arabic Script), and Urdu (Roman Script) data. This adaptability ensures its effectiveness across diverse languages, making it a versatile and powerful tool for comment classification.

7.1 System Architecture

7.1.1 System Context Diagram

Figure 7.1 shows the dynamic interaction involving two external actors. A social media influencer seeking enhanced engagement with their audience, facilitated by the assistance of our model. This engagement entails leveraging the capabilities of our model to enhance content interaction on social media platforms. Secondly, a developer aims to integrate our model into their project, utilizing the model's API to augment the functionalities of their application. The influencer, as one external actor, collaborates with our model to optimize audience interaction, while the developer, as another external actor, harnesses the model's API to incorporate its capabilities into their project seamlessly. This dual interface caters to distinct external needs, emphasizing the versatility of our model for both content creators and developers alike.

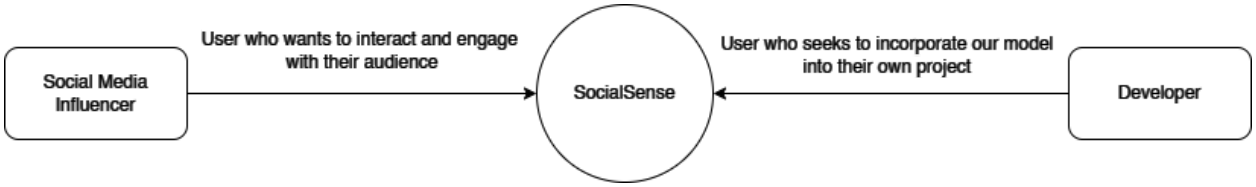


Figure 7.1: System Context Diagram

7.1.2 Container Diagram

The system operates through a streamlined interaction among various components, as illustrated in Figure 7.2. The process begins with inputs from either media or developers, submitted via a ReactJs-powered single-page application. These applications and extensions interface with our API, which subsequently sends requests to our model. The model processes these requests and returns the responses. Additionally, Google Chrome extensions, activated by users, facilitate the visualization of categorized comments across different social media platforms. Developers have the flexibility to integrate our model into their systems using the provided API, enhancing the system's versatility. This architecture ensures a seamless flow of information from user input to categorized output, enabling widespread adoption through developer integration and browser extension utilization.

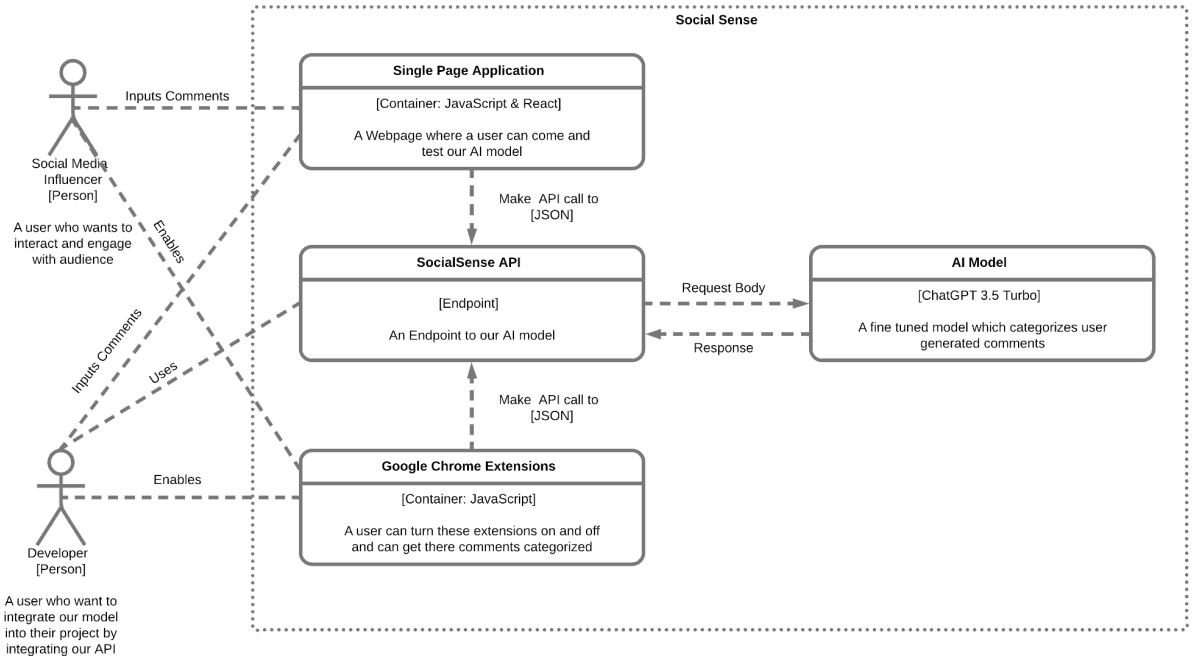


Figure 7.2: Container Diagram

7.1.3 Class Diagram

Our system consists of two key components: the "Request Body" class and the "Response" class, which contains carefully defined variables with specific data types as shown in Figure 7.3.

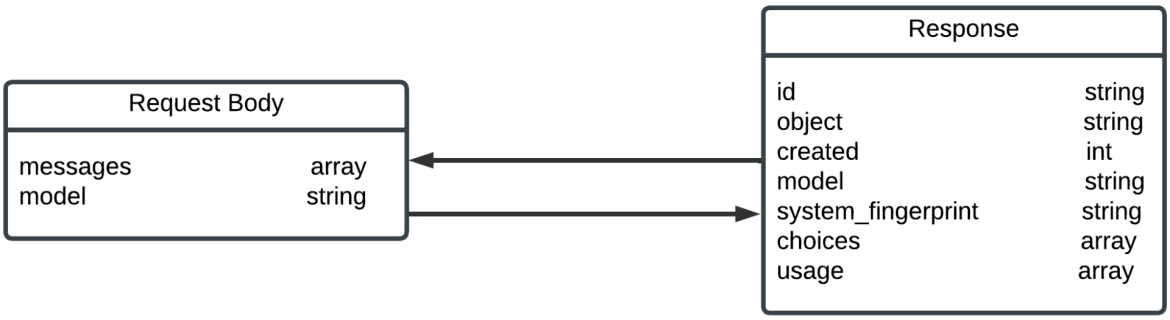


Figure 7.3: Class Diagram

7.2 Key sequence diagrams

In our system, where the main focus is on our AI model, and we don't have many objects, we use a sequence diagram to show how things interact. This diagram helps us map out the step-by-step interactions between the few elements in our system. Because our system is all about the AI model, the interactions are straightforward and don't involve a lot of different objects. The sequence diagram is like a visual guide that captures these interactions, making it easy to understand how our system works, even though it's not too complicated, with only a handful of elements involved.

Figure 7.4 encapsulates a user, who may either be a social media influencer or a developer, initiating the interaction by submitting a comment. This comment is then transmitted to the comment object, which serves as an intermediary responsible for conveying the submitted comment to the Model API object. The Model API object, in turn, employs our model's API to categorize the input comment. Following the categorization process, the Model API object communicates the categorized comment back to the comment object. Subsequently, the comment object relays the categorized result to the user. Importantly, both the original comment and its categorized result. This sequence of actions ensures a cohesive and traceable flow of information throughout the user engagement process.

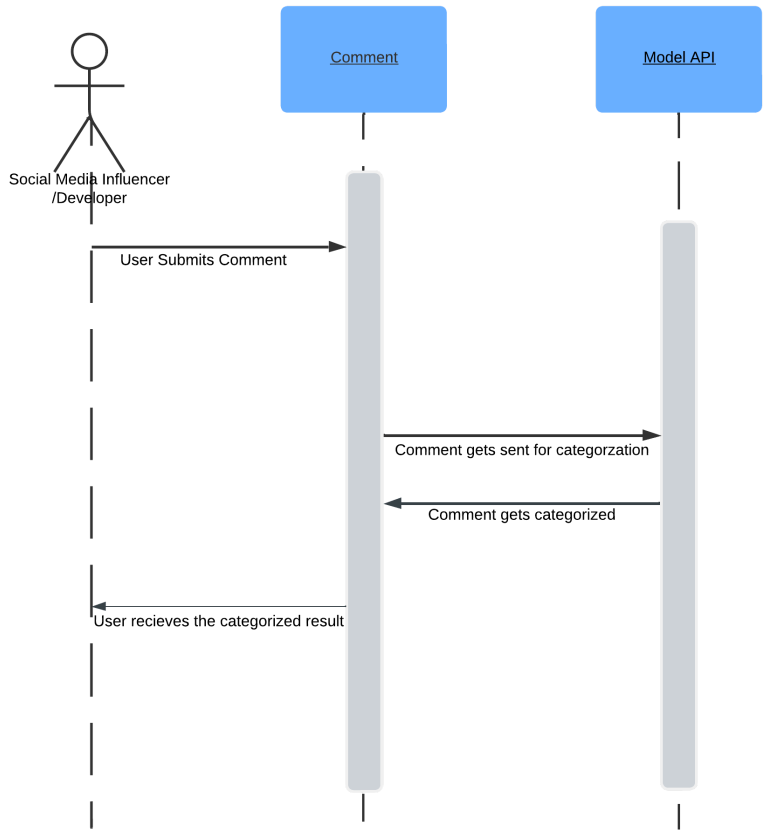


Figure 7.4: Sequence Diagram

7.3 API Design

7.3.1 Endpoints

The following endpoints will be available for accessing the SocialSense model:

7.3.1.1 Request Body

- Endpoint: https://api.openai.com/v1/chat/completions
- Method: POST
- Description: Request Body
- Request Parameters (Figure 7.5):
    - messages (array): The role and content of the system and user
    - mode (string): The ID of fine-tuned model
- Response Format:
    - HTTP Status: 200 OK

```
{
    messages:[
            {"role": "system",
             "content": "You are a text classifier for social media
      comments. Classify the following comment into one of the following
      classes: [Positive, Neutral, Negative - Respond, Negative - Ignore,
      Negative - Remove, Crisis]"},
            {"role": "user", "content": "Sir Kamaal ker deya}]
    model: "ft:gpt-3.5-turbo-0125:personal:socialsense:96eqMPvL"
}
```

Figure 7.5: Sample post comment

7.3.1.2 Response

- Endpoint: https://api.openai.com/v1/chat/completions
- Method: GET
- Description: Display the response from the API
- Request Parameters: None
- Response Format (Figure 7.6):
    - id (string): A unique identifier for the chat completion.
    - object (string): The object type, which is always chat.completion
    - created (int): The Unix timestamp (in seconds) of when the chat completion was created.
    - model (string): The model used for the chat completion
    - system_fingerprint (string): This fingerprint represents the backend configuration that the model runs with.
    - choices (array): A list of chat completion choices. Can be more than one if n is greater than 1.
    - usage (array): Usage statistics for the completion request

```
{
  "id": "chatcmpl-123",
  "object": "chat.completion",
  "created": 1677652288,
  "model": "ft:gpt-3.5-turbo-0125:personal:socialsense:96eqMPvL",
  "system_fingerprint": "fp_44709d6fcb",
  "choices": [{
    "index": 0,
    "message": {
      "role": "assistant",
      "content": "Positive",
    },
    "logprobs": null,
    "finish_reason": "stop"
  }],
  "usage": {
    "prompt_tokens": 9,
    "completion_tokens": 12,
    "total_tokens": 21
  }
}
```


Figure 7.6: Sample get result

7.4 AI/ML design

7.4.1 Data Sources

7.4.1.1 Facebook Comments

- Platform Overview: Facebook is one of the largest social media platforms with diverse user demographics.
- Comment Characteristics: Facebook comments are typically longer and can include rich media such as images and GIFs. Users can also reply to specific comments, creating threaded conversations.
- Use Cases: Comments on Facebook are often used for discussions, sharing opinions, expressing emotions, and providing feedback on posts that include text, images, or videos.

7.4.1.2 Instagram Comments

- Platform Overview: Instagram is a visually rich platform, and comments can provide insights across different types of content.
- Comment Characteristics: Comments on Instagram are usually concise, often containing emojis or short sentences. Users can also reply directly to specific comments, fostering conversation threads.

- Use Cases: Instagram comments are crucial for engagement, expressing appreciation, asking questions about the content, and occasionally providing feedback. Hashtags are commonly used in comments for discoverability.

7.4.1.3 Twitter Comments

- Platform Overview: Twitter is a real-time platform with concise, publicly available comments.
- Comment Characteristics: Tweets can receive replies, which are often brief due to Twitter's character limit. Conversations can be easily followed by viewing threaded replies.
- Use Cases: Twitter comments are used for discussions, expressing opinions, providing updates, and engaging in public conversations. Hashtags and mentions are commonly used for broader reach and interaction.

7.4.1.4 TikTok Comments

- Platform Overview: TikTok is a popular platform for short-form videos, and comments can reflect user engagement.
- Comment Characteristics: TikTok comments are often short and feature a variety of emojis. Users can reply to comments, and the platform often highlights the top comments for each video.
- Use Cases: TikTok comments are essential for user engagement, expressing reactions, and participating in challenges or trends. Comments play a significant role in shaping the overall narrative of a video.

7.4.1.5 Reddit Comments

- Platform Overview: Reddit is a diverse platform hosting various communities (subreddits) for discussions on specific topics.
- Comment Characteristics: Reddit comments are usually longer and can include text, links, and formatting. The platform allows upvoting and downvoting, shaping comment visibility.
- Use Cases: Reddit comments are vital for in-depth discussions, sharing information, asking questions, and providing diverse perspectives. Each subreddit may have its own set of norms and rules governing discussions.

7.5 Methodology For Comment Selection

We prioritize diversity by selecting comments across varied topics, users, and periods from major social media platforms. Our goal is to achieve a balanced distribution of sentiments, ensuring our model's robustness across diverse emotional expressions. We deliberately include comments from different user types to capture a broad range of perspectives. These measures contribute to a comprehensive dataset.

7.6 Scraping Tool

Instant Data Scraper, a powerful Chrome extension, facilitates easy, seamless data extraction from websites. Refer to section 5.1 of the document for further details.

7.7 Data Collection

We have labeled the dataset into six categories: Positive, Neutral, Crisis, Negative-Ignore, Negative-Remove, and Negative-Respond. The labels for categorizing social media comments were carefully shortlisted following thorough research, which included insights from Jeff Bullas[23], Attention Insight [24], SproutSocial [25], and JSH Web Design [26] on critical types of social media comments. The definitions for each label are as follows:

1. **Positive:** Represents any comment that expresses favorability or positivity.
2. **Neutral:** Denotes a comment that is neither positive nor negative, conveying a neutral stance.
3. **Negative – Respond:** Signifies a genuinely negative comment that warrants a response or engagement.
4. **Negative – Ignore:** Identifies a negative comment originating from a "troll" – an intentional trouble-maker – which is best disregarded.
5. **Negative – Remove:** Indicates a comment that is offensive, malicious, or spam, breaching established "House Rules" and necessitating removal.
6. **Crisis:** Encompasses comments with potential legal or criminal implications, such as threats of violence, breaches of confidentiality, defamation, or PR disasters.

Labeling a dataset is susceptible to challenges, such as inconsistencies from varying interpretations among annotators based on their subjective understanding. This can introduce discrepancies in the labeled data. Furthermore, annotators may bring inherent biases into their labeling decisions, potentially resulting in a biased dataset and impacting the fairness and accuracy of the trained machine learning model.

7.8 Large Language Models (LLMS)

Large Language Models, or LLMs, form the foundation for a wide range of natural language processing tasks, including chatbot development. These models are typically pre-trained on massive amounts of textual data to learn the intricate patterns and structures of human language. One notable example is GPT (Generative Pre-trained Transformer) architecture.

7.9 Chatbot Models and Derivation from LLMs

Chatbot models are derived from LLMs by fine-tuning them on specific datasets. The process involves exposing the pre-trained model to conversations or dialogue data, allowing it to adapt to the nuances and context-specific requirements of generating human-like responses. Fine-tuning makes the model more specialized for tasks like conversation, making it a powerful tool for chatbot development.

8 Implementation

Our goal was to develop an AI model utilizing the fine-tuning approach with Large Language Models (LLMs). Our journey began with data scraping, followed by labeling the scraped dataset to suit our requirements. For fine-tuning the LLM, we explored options such as fine-tuning Llama 2 by Meta and Chat GPT 3.5 Turbo by OpenAI. Various prompt engineering techniques were employed to enhance the model's performance. Once the model development was finalized, we integrated the model API into a webpage, allowing developers to interact with the model for testing and gaining deeper insights. Additionally, we developed Chrome extensions for Facebook, Instagram and TikTok, leveraging the AI model to classify user comments on these platforms effectively.

8.1 Key implementation details

This document details the key implementations and results of our project, focusing on the testing and fine-tuning of ChatGPT 3.5 and Llama 2 models for sentiment analysis of social media comments. We explored various approaches, including prompt engineering techniques and fine-tuning on Vertex AI, to optimize the model's performance and deploy a user-friendly web application with browser extensions.

8.1.1 ChatGPT 3.5 & Llama 2 Testing

We conducted a comparative analysis of ChatGPT 3.5 and Llama 2 70b Chat models, evaluating their performance on English and Urdu comments (both Roman and Arabic scripts).

8.1.1.1 ChatGPT 3.5

The testing of ChatGPT 3.5 (Figure 8.1) took place on the OpenAI Playground, where we provided comments and received categorizations indicating the specific category to which the comments belonged.

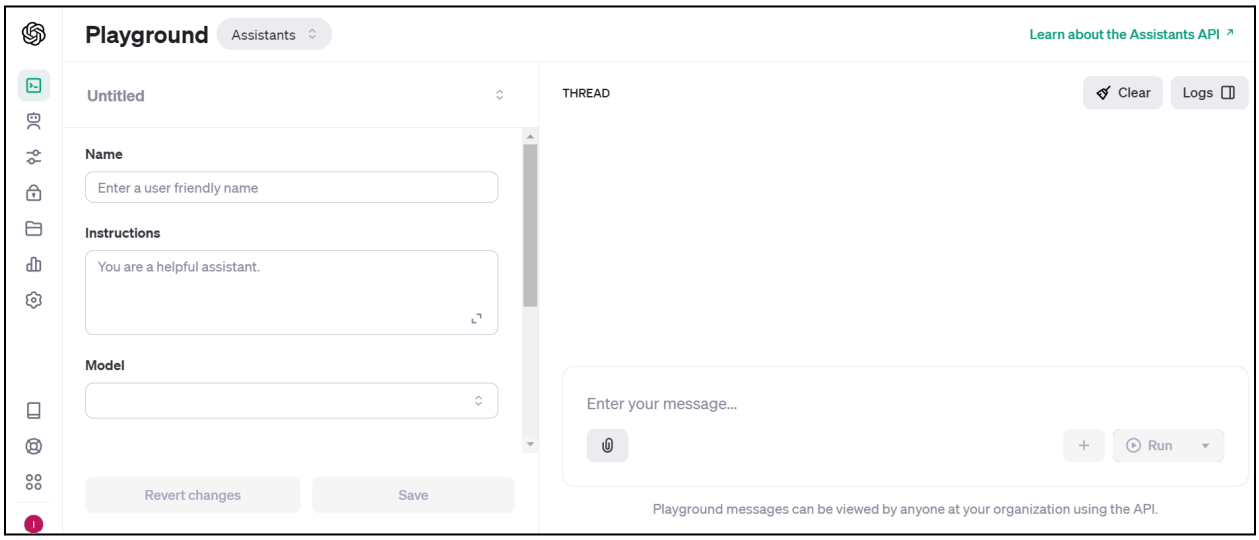


Figure 8.1: OpenAI Playground

8.1.1.2 Llama 2

The testing of Llama 2 70b Chat took place on HuggingChat (Figure 8.2), where we provided comments and received categorizations indicating the specific category to which the comments belonged.

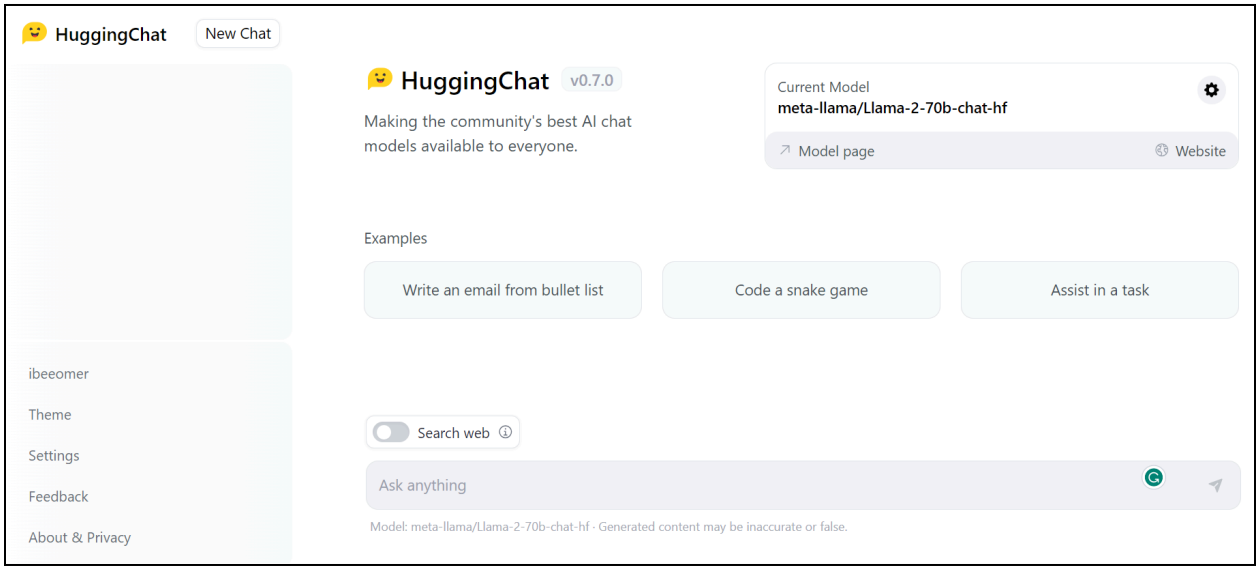


Figure 8.2: HuggingChat

8.1.1.3 Results

It was observed that the Llama2 70b chat model outperformed ChatGPT 3.5 in handling Urdu (Roman Script) comments (Table 8.1). Interestingly, the accuracy remained consistent for both Urdu (Arabic Script) and English comments in the dataset. The overall accuracy of the model was adversely affected due to several factors. Primarily, the testing phase encountered limitations as the HuggingChat API became unresponsive after a few requests, resulting in a very limited dataset for evaluation. Additionally, the Language Model (LLM) utilized was not fine-tuned with the training set, further compromising its performance.

Table 8.1: Model results for initial testing

Model	English	Urdu (Roman Script)	Urdu (Arabic Script)
Chat GPT 3.5	58.33%	16.67%	33.33%
Llama 2 70b Chat	58.33%	33.33%	33.33%

8.1.2 Prompt Engineering

We implemented prompt engineering techniques learned from the ChatGPT Prompt Engineering for Developers [31] course offered by DeepLearning.ai.

8.1.2.1 Course Learning

- Create a concise and clear prompt for an engineering course.
- Provide detailed instructions for better results.
- Specify the use of three delimiters to structure the prompt effectively.
- Include sample data to enhance accuracy.

8.1.2.2 Prompt Version 1

In the given prompt (Figure 8.3), only the label names were provided, and the accompanying comments were subsequently passed on to the corresponding function.

```
prompt = f"""
What is the sentiment of the following comment,which is delimited with
triple backticks?

There are six possible sentiments: Positive, Neutral, Crises, Negative -
Ignore, Negative - Remove, and Negative - Respond.

Reply with label of the sentiment only.

Comment text: '''{comment}'''
"""
```

Figure 8.3: Prompt 1

8.1.2.3 Prompt Version 2

In the given prompt (Figure 8.4), the label names and definitions were provided, and the accompanying comments were subsequently passed on to the corresponding function.

```
prompt = f"""
What is the sentiment of the following comment, which is delimited with
triple backticks?
There are six possible sentiments: Positive, Neutral, Crises, Negative -
Ignore, Negative - Remove, and Negative - Respond.

Positive comments are any comment that's favourable.
Neutral comments are is a comment that is neither good nor bad.
Negative - Respond is a comment that is a genuine negative comment.
Negative – ignore is a comment that is a comment by a "troll" (a
deliberate trouble-maker).
Negative – remove is a comment that is offensive, malicious or spam.
That is, it breaches your "House Rules".
Crisis is a comment that have legal or criminal ramifications (eg.
threat of violence, breach of confidentiality, defamation, PR disaster
etc).

Reply with label of the sentiment only.
Comment text: '''{comment}'''
"""
```

Figure 8.4: Prompt 2

8.1.2.4 Prompt Version 3

In the provided prompt (Figure 8.5), each label is accompanied by its definition, and a sample English comment is given for each label. The corresponding comments are then passed on to the corresponding function.

```
prompt = f"""
What is the sentiment of the following comment, which is delimited with
triple backticks?

There are six possible sentiments: Positive, Neutral, Crises, Negative -
Ignore, Negative - Remove, and Negative - Respond.

Positive comments are any comment that's favourable.
Eg: "My burger was awesome!"

Neutral comments are is a comment that is neither good nor bad.
Eg: "I'm having a burger for lunch."

Negative - Respond is a comment that is a genuine negative comment.
Eg: "My burger was cold and took forever."

Negative – ignore is a comment that is a comment by a "troll" (a
deliberate trouble-maker).
Eg: "Burgers are evil and so are the people that eat them."

Negative – remove is a comment that is offensive, malicious or spam.
That is, it breaches your "House Rules".
Eg: "My waitress was a %^&*."

Crisis is a comment that have legal or criminal ramifications (eg.
threat of violence, breach of confidentiality, defamation, PR disaster
etc).
Eg. "I'm going to burn this burger joint down."

Reply with label of the sentiment only.
Comment text: '''{comment}'''
"""
```

Figure 8.5: Prompt 3

8.1.2.5 Results

The model's accuracy (Table 8.2) demonstrated a clear improvement trend with increasingly detailed and specific prompts. As more comprehensive sample data of English comments was provided, the accuracy notably increased. Specifically, when English data was incorporated, the model's accuracy advanced from 47% in prompt 1 to 49% in prompt 2, further reaching 57% in prompt 3. However, due to HugChat API limitations, further testing could not be pursued to validate these advancements.

Table 8.2: Classification accuracy using prompt engineering techniques

Prompt	English	Urdu (Roman Script)	Urdu (Arabic Script)
1	47.0%	14.1%	19.0%
2	49.0%	14.1%	7.0%
3	57.0%	12.5%	7.0%

In conclusion, our findings strongly indicate that augmenting the dataset leads to a noticeable enhancement in the model's accuracy, as evidenced by the results from Prompt 3. This underscores the pivotal role of data quantity in refining the model's predictive capabilities. Furthermore, our investigation suggests that the accuracy of the model for Urdu (Roman Script) and Urdu (Arabic Script) can be substantially supported through the inclusion of additional data, coupled with meticulous training aligned with our specified labels. This underscores the significance of both data quantity and quality in optimizing the performance of the Language Model, highlighting a potential avenue for further refinement and improvement.

8.1.3 Llama 2-70b Deployment On Vertex AI

The below code (Figure 8.6) fine-tuning LLAMA2-70B using the Progressive Elastic Fine-Tuning (PEFT) technique as suggested by Vetex AI. This approach involved refining a pre-trained model by further training it on specific datasets or tasks to enhance its adaptability to new data or improve its performance in targeted tasks. PEFT, in particular, facilitated dynamic adjustments to the model size, such as the number of parameters, throughout the training process, effectively balancing performance gains against resource constraints.

```
train_job = aiplatform.CustomContainerTrainingJob(
    display_name=job_name,
    container_uri=TRAIN_DOCKER_URI,
)
train_job.run(
    args=[
        "--task=causal-language-modeling-lora",
        f"--pretrained_model_id={base_model_id}",
        f"--dataset_name={dataset_name}",
        f"--output_dir={output_dir}",
        "--lora_rank=16",
        "--lora_alpha=32",
        "--lora_dropout=0.05",
        "--warmup_steps=10",
        "--max_steps=10",
        "--learning_rate=2e-4",
        f"--precision_mode={finetuning_precision_mode}",
        f"--template={template}",
    ],
    environment_variables={"WANDB_DISABLED": True},
    replica_count=replica_count,
    machine_type=machine_type,
    accelerator_type=accelerator_type,
    accelerator_count=accelerator_count,
    boot_disk_size_gb=500,
)

print("Trained models were saved in: ", output_dir)
```

Figure 8.6: PEFT fine tuning of Llama 2 - 70B

The training procedure was conducted on Google Cloud's Vertex AI platform, which offers managed services tailored for machine learning tasks, encompassing both training and deployment phases. Various parameters crucial to the fine-tuning process, such as learning rate, dropout rate, precision mode, among others, were meticulously configured to optimize the model's performance.

However, despite the meticulous fine-tuning and optimization efforts, deployment and subsequent testing of the model were impeded due to the high running costs associated with its implementation.

8.1.4 GPT-3.5-Turbo-1106 Fine-Tuning

This section delves into the fine-tuning process of the GPT-3.5-Turbo-1106 model, a significant step towards tailoring its capabilities for our specific task of sentiment analysis on social media comments. We also analyze the model's performance using a confusion matrix, highlighting its strengths and weaknesses in classifying different sentiments.

8.1.4.1 Mode Fine-Tuning

In accordance with the details presented in 'Figure 8.7', the model underwent training using a dataset containing a total of 60 examples, distributed evenly with 10 examples per category. The training process was conducted utilizing 13,437 tokens in total. The chosen base model for training was the GPT-3.5 Turbo-1106, equipped with 1,106 million parameters. Training encompassed 3 epochs, and notably, no distinct validation dataset was employed throughout the training procedure.

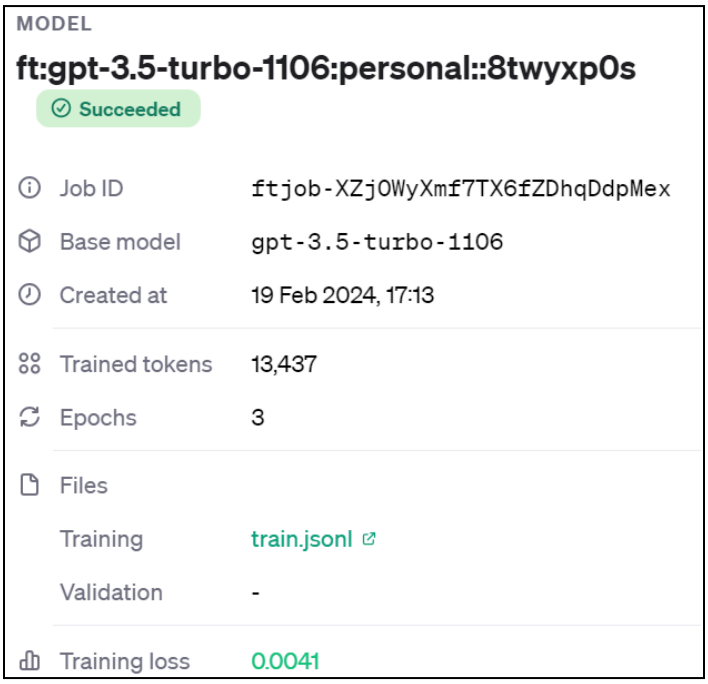


Figure 8.7: GPT 3.5 Trubo Fine-tuning Details.

8.1.4.2 Confusion Matrix

The model initially achieved a classification accuracy of 62% on the dataset but exhibited poor performance specifically in the crisis category, as evidenced by the confusion matrix (Figure 8.8). However, after excluding the crisis category from the dataset, the model's accuracy significantly improved to 82%. This suggests that the model struggled with accurately classifying instances within the crisis category but performed notably better when this category was removed from consideration.

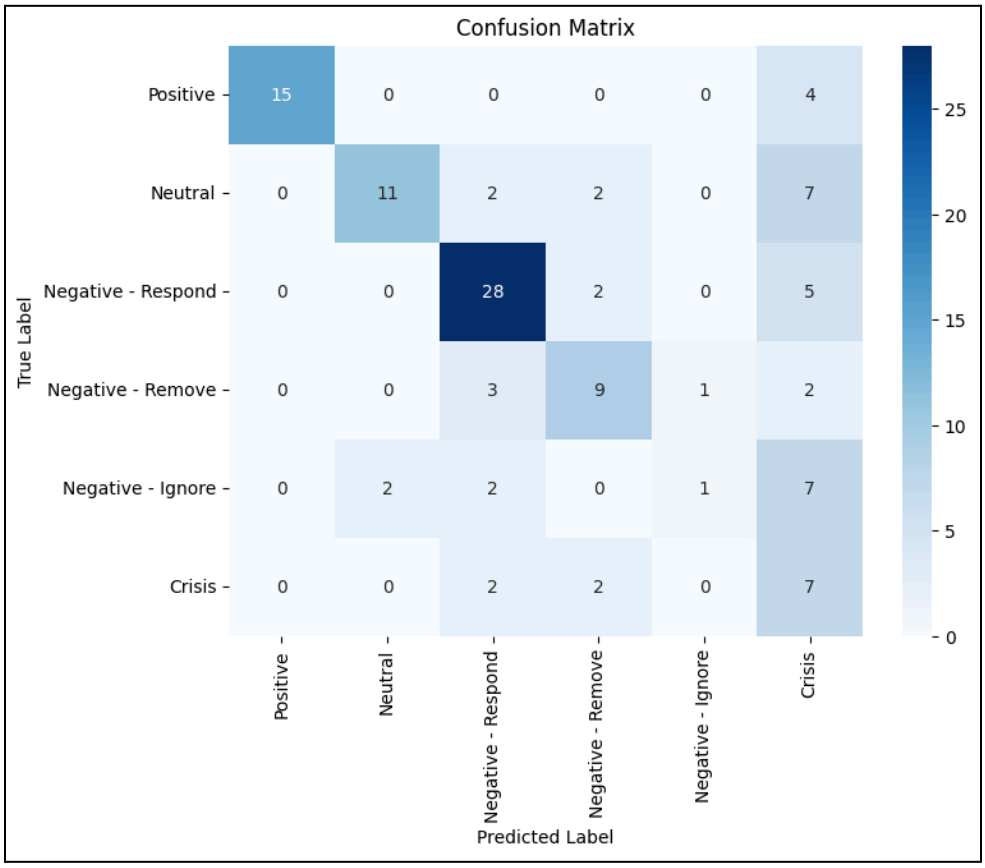


Figure 8.8: Confusion matrix for initial model testing

8.1.5 SocialSense Website

8.1.5.1 Model Testing Page

The Model testing page (Figure 8.9) serves as the central hub for user interaction with the model. Here, users can engage with the system by entering text into the designated text area (Figure 8.10). Once the user submits their input, it is promptly sent to the SocialSense API for processing. The API then generates a response based on the input text, which is subsequently displayed to the user (Figure 8.11). This seamless process allows for dynamic interaction and real-time feedback, enhancing the overall user experience.

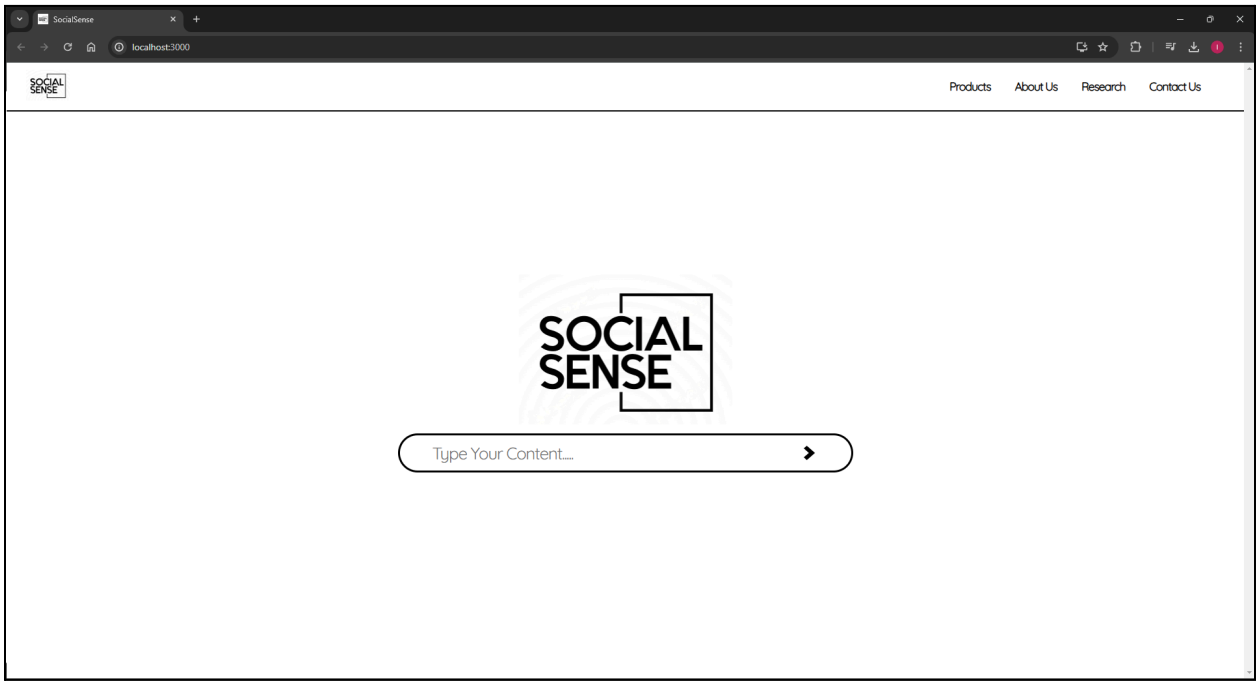


Figure 8.9: Screenshot of Model Testing Page

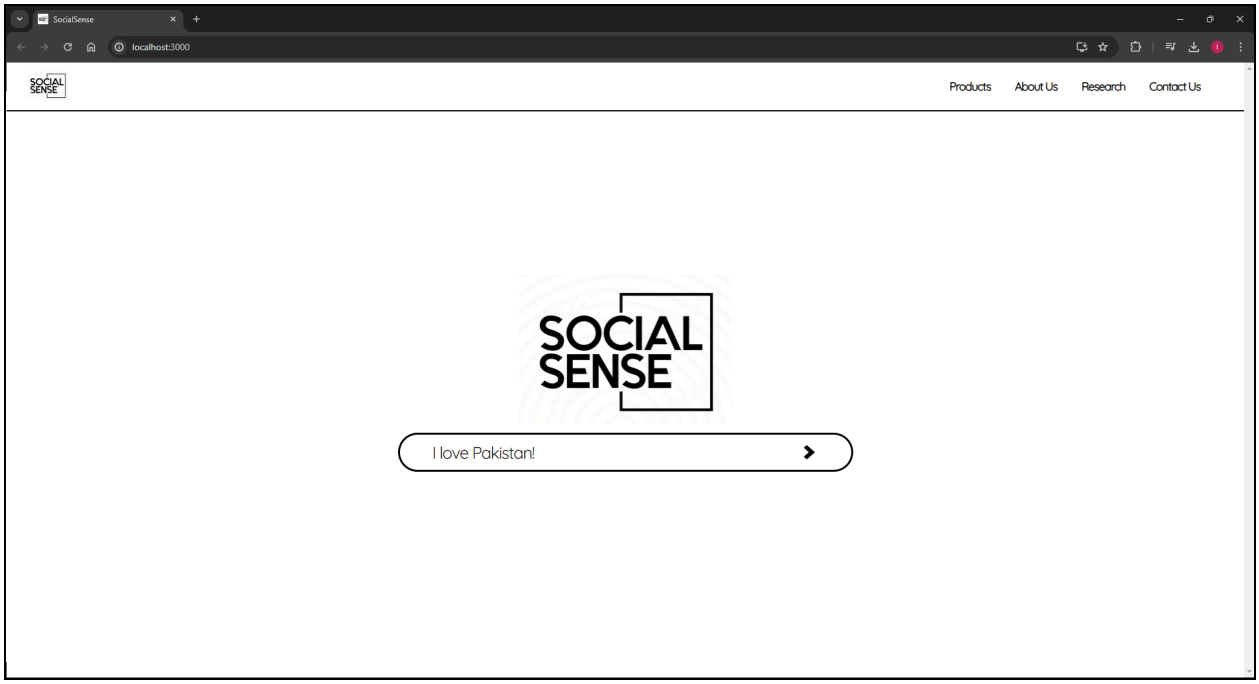


Figure 8.10: Screenshot of user input on the Model Testing page

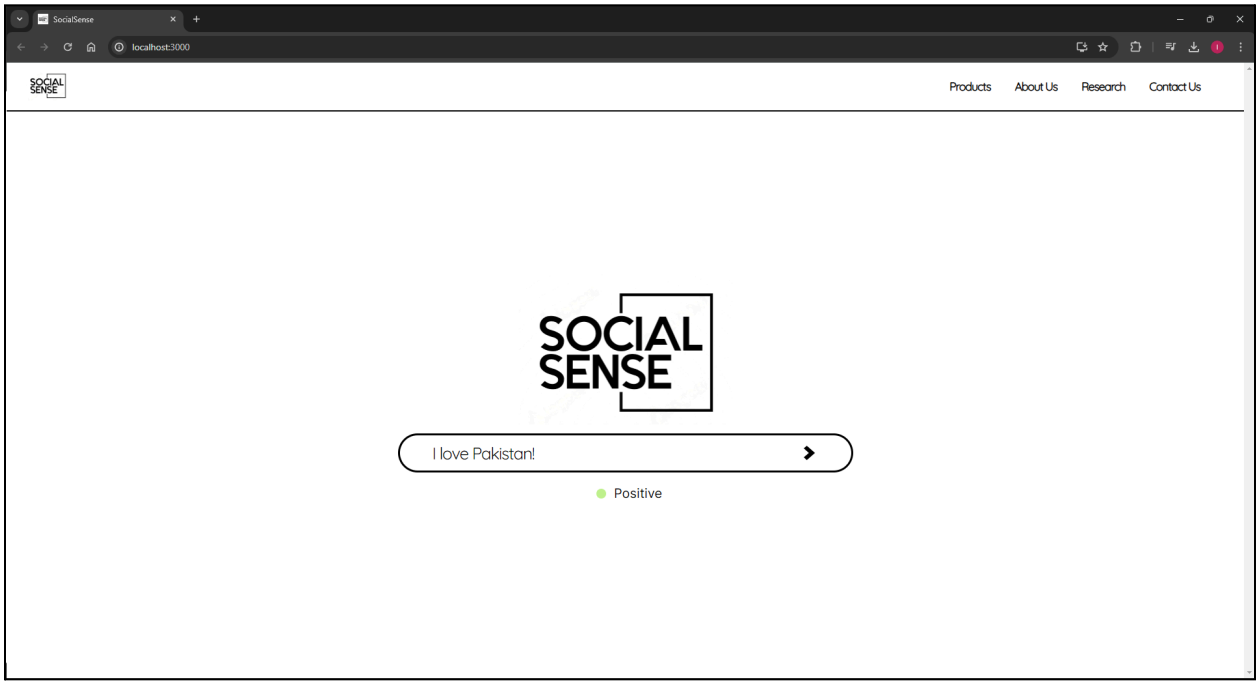


Figure 8.11: Screenshot of response by SocialSense API on Model Testing page

8.1.5.1.1 API Integration

The `handleSendMessage` function (Figure 8.12) processes user input by sending it to OpenAI's Chat API for classification into predefined categories like Positive, Neutral, Negative - Ignore, Negative - Remove, Negative - Respond, or Crisis. After receiving the predicted category, it validates it against predefined patterns. If the category matches, it logs the result and displays it. Otherwise, it logs an error any errors during the process are caught and logged. This function streamlines the classification of user input and ensures accurate categorization with error handling.

```
const handleSendMessage = async (input) => {
    try {
      const response = await openaiInstance.chat.completions.create({
        model: "ft:gpt-3.5-turbo-0125:personal:socialsense:96eqMPvL",
        messages: [
          {
            role: "system",
            content:
              "You are a text classifier for social media comments.
Classify the following comment into one of the following classes:
[Positive, Neutral, Negative - Respond, Negative - Ignore, Negative -
Remove, Crisis]",
          },
          { role: "user", content: input },
        ],
      });

      const predictedCategory = response.choices[0].message.content;

      const categoryPattern =

/^(Positive|Neutral|Negative\s-\sIgnore|Negative\s-\sRemove|Negative\s-\
sRespond|Crisis)$/;

      if (categoryPattern.test(predictedCategory)) {
        console.log(predictedCategory);
        setResultData(predictedCategory);
        setShowResult(true);
      } else {
        console.error("Invalid category:", predictedCategory);
      }
    } catch (error) {
      console.error("Error:", error);
      setResultData("Unable To Predict Category");
      setShowResult(true);
    }
  };
```

Figure 8.12: Function to process user input

8.1.5.2 Products Page

The Products page (Figure 8.13) showcases our comprehensive range of solutions, including the SocialSense API for seamless integration into users' systems. Additionally, our groundbreaking extensions for Facebook and Instagram utilize the SocialSense API to classify user comments in real-time, providing valuable insights for optimizing digital strategies and enhancing online presence. Upon clicking a product, the user will be redirected to its documentation.

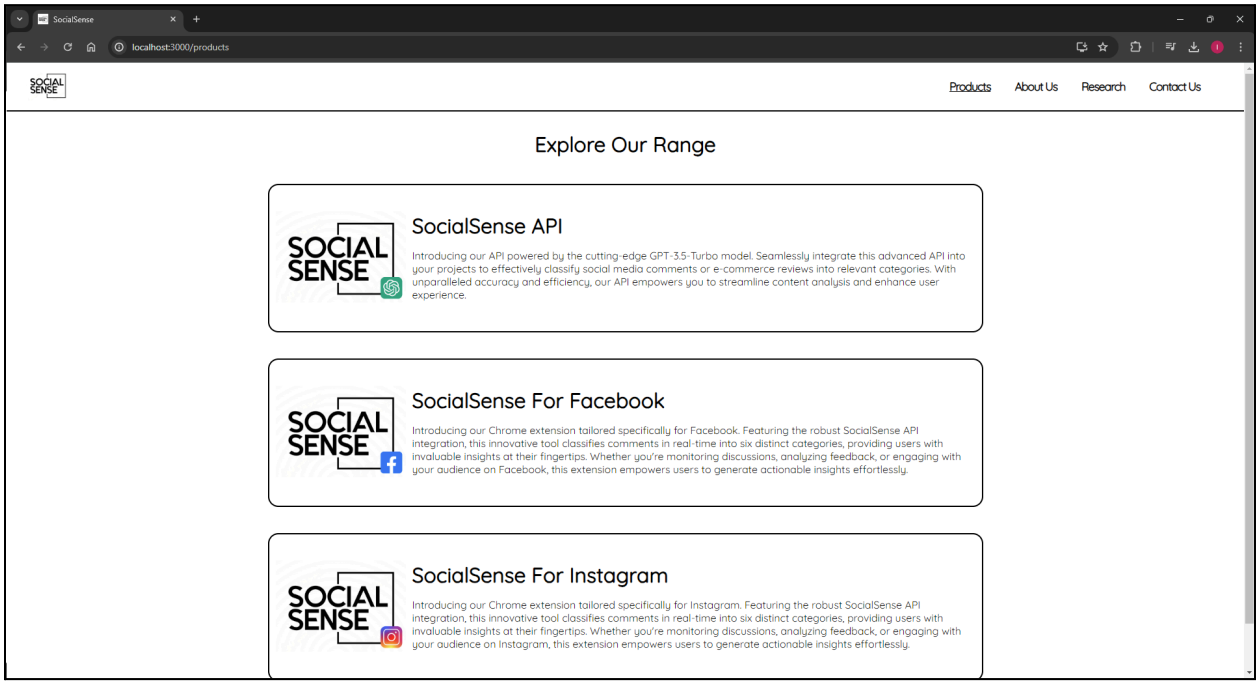


Figure 8.13: Screenshot of the Products page

8.1.5.3 About Us Page

The About Us page (Figure 8.14) provides users with insight into our team, the developers behind SocialSense, and offers access to our contact details.

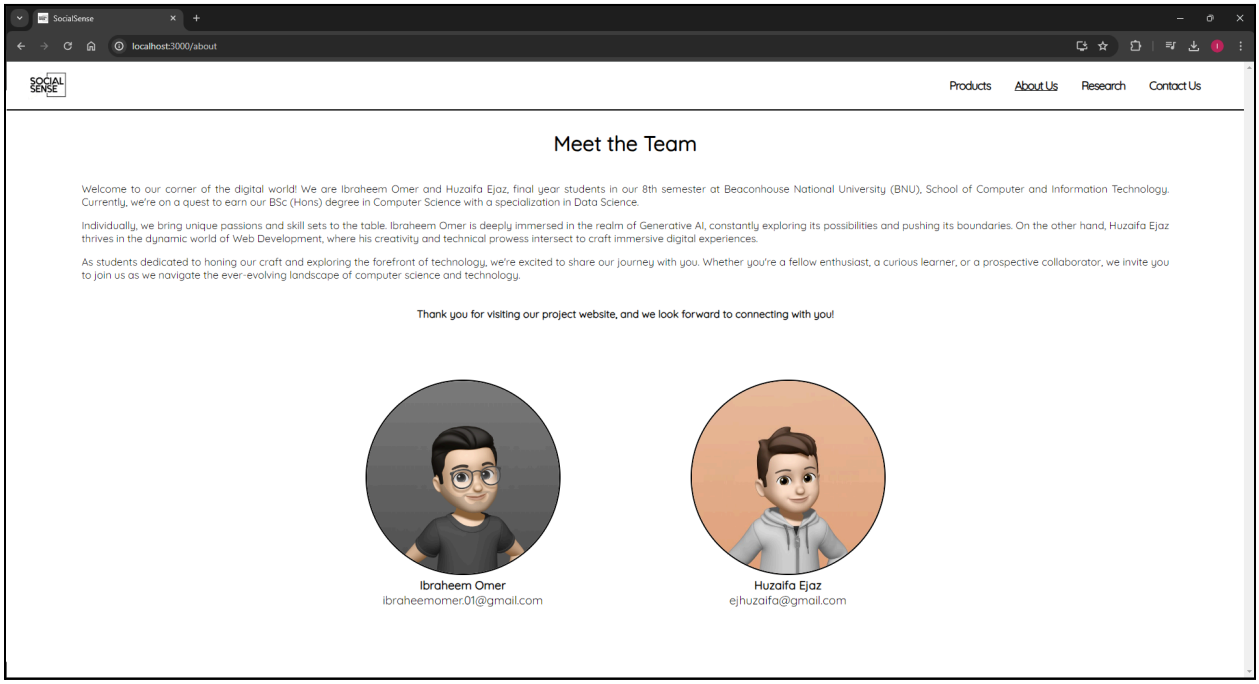


Figure 8.14: Screenshot of the About Us page

8.1.5.4 Research Page

The Research page (Figure 8.15) offers users detailed insights into the project's research background and provides in-depth knowledge about the workings of the developed AI model.

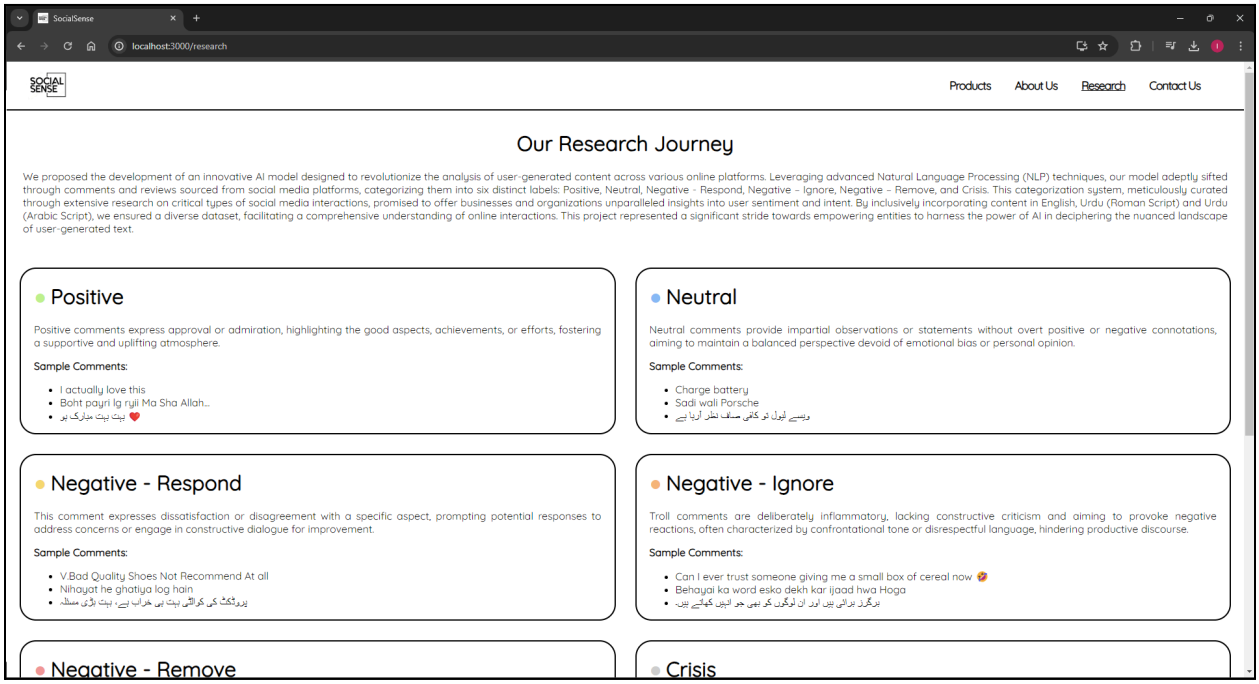


Figure 8.15: Screenshot of the Research Page

8.1.5.5 Contact Us Page

The Contact Us page (Figure 8.16) allows users to reach out to us effortlessly by filling in their details (Figure 8.17), and simply clicking the submit button. Upon submission, an email is automatically sent via the API Email JS, ensuring efficient communication (Figure 8.18).

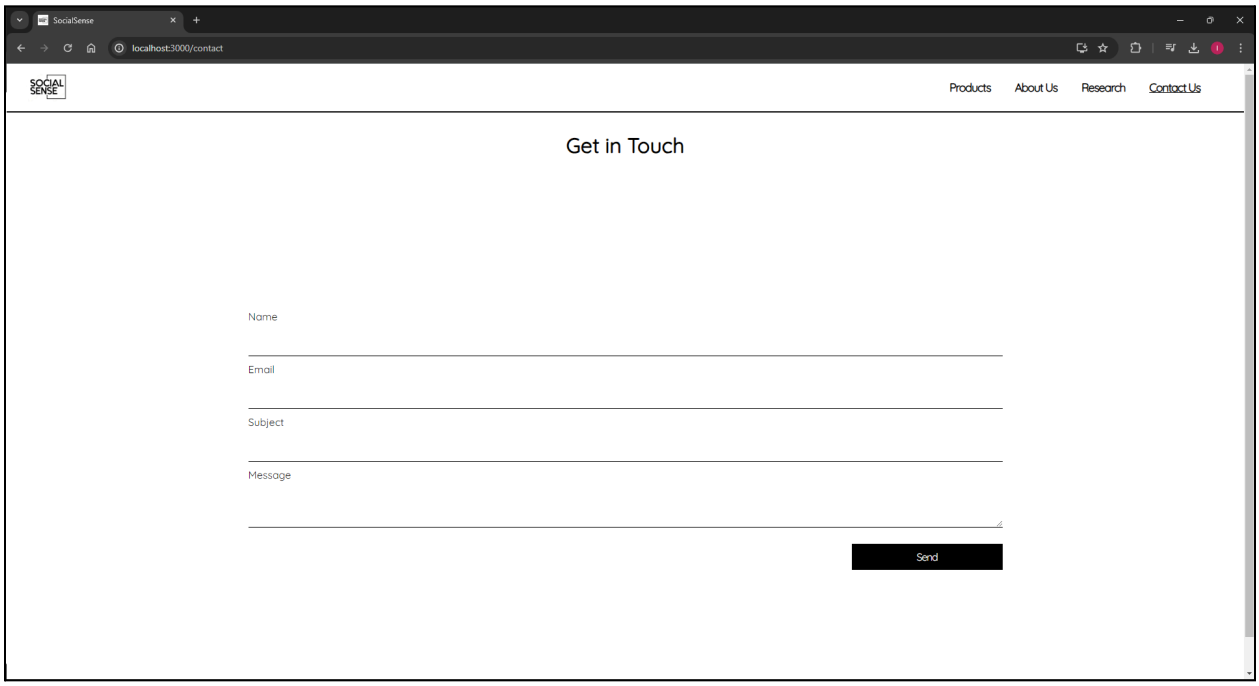


Figure 8.16: Screenshot of the Contact Us Page

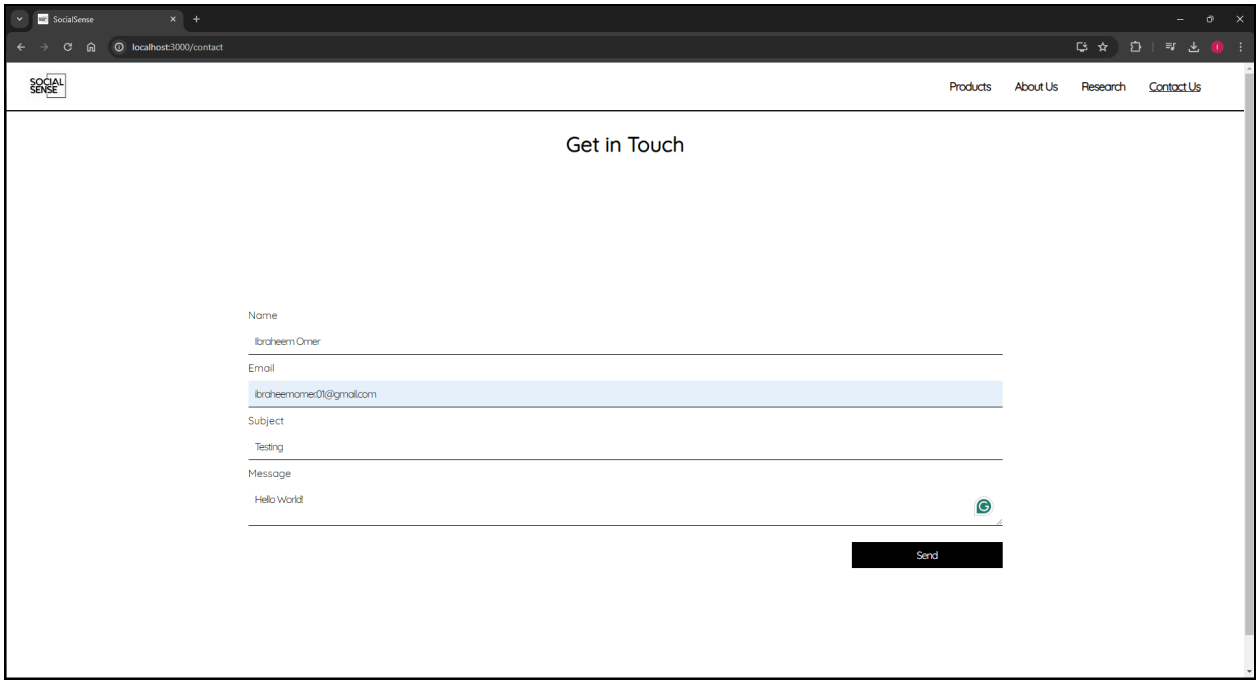


Figure 8.17: Screenshot of user input on the Contact Us Page

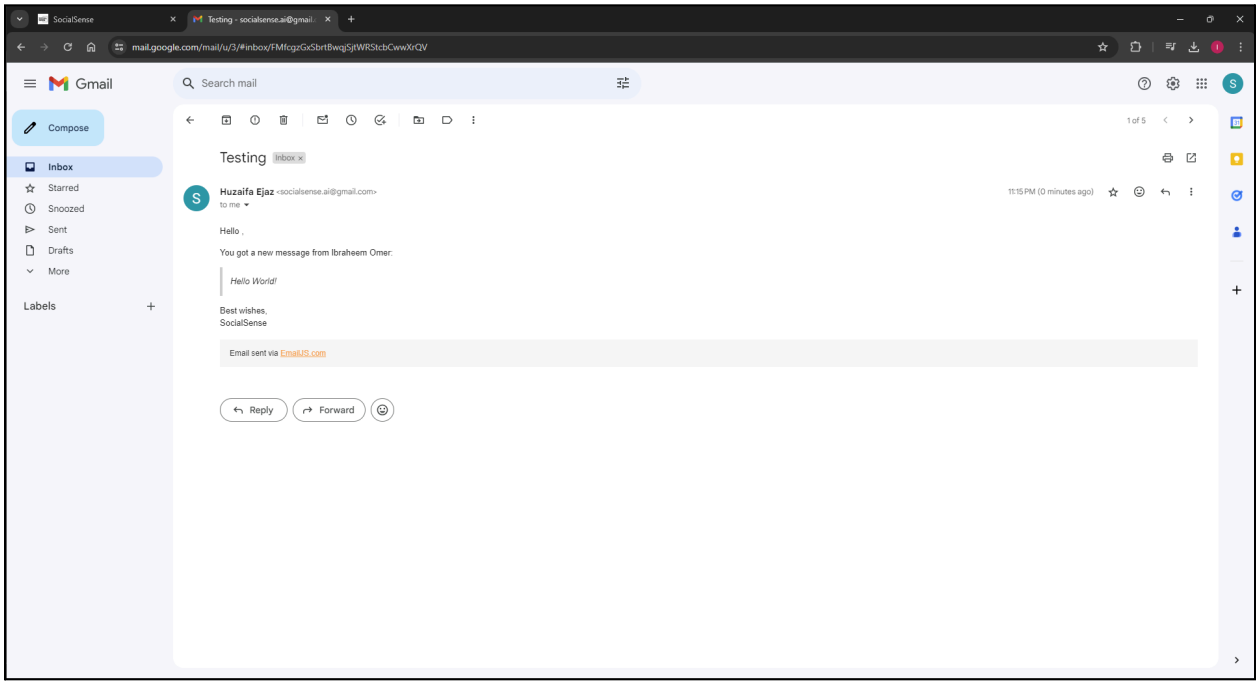


Figure 8.18: Screenshot of email received via Email JS

8.1.5.5.1 Email.js Integration

The `sendEmail` function (Figure 8.19) handles the submission of an email form. The `sendForm` method of `emailjs` takes several parameters: the service ID, the template ID, the form element (retrieved via a ref), and the user ID. Upon successful submission, it logs a success message along with the text of the result. In case of failure, it logs an error message along with the error text. This function facilitates the sending of form data via email using the `emailjs` library.

```
const sendEmail = (e) => {
    e.preventDefault();
    emailjs
      .sendForm(
        "service_tnr92ox",
        "template_4ow6v9l",
        form.current,
        "h4wgewFuBSn0kcKBN"
      )
      .then(
        (result) => {
          console.log("SUCCESS!", result.text);
        },
        (error) => {
          console.log("FAILED...", error.text);
        }
      );
    form.current.reset();
  };
```

Figure 8.19: Function to process email

8.1.6 Facebook Extension

The Facebook extension powered by SocialSense AI, is designed to classify comments swiftly and accurately into different categories: Positive, Neutral, Negative - Respond, Negative - Ignore, Negative - Remove, and Crisis as seen in Figure 8.20.

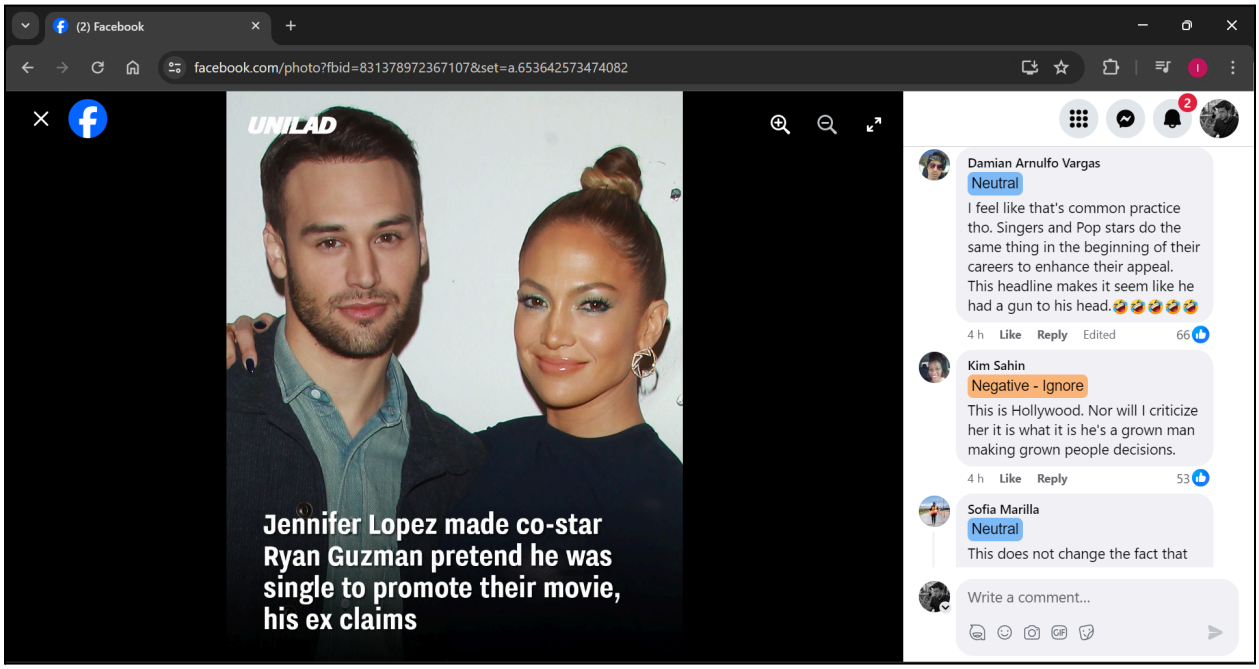


Figure 8.20: SocialSense API classifying Facebook comments in real time.

8.1.6.1 Manifest.js for Facebook Extension

A manifest file (Figure 8.21) is a configuration file that contains essential details about an application or extension.

```
{
  "manifest_version": 3,
  "name": "SocialSense For Facebook",
  "version": "1.0",
  "description": "An extension to classify Facebook comments using
SocialSense.",
  "permissions": ["activeTab", "scripting"],
  "background": {
    "service_worker": "background.js"
  },
  "action": {
    "default_icon": {
      "16": "icons/facebook.png",
      "48": "icons/facebook.png",
      "128": "icons/facebook.png"
    }
  },

  "icons": {
    "16": "icons/facebook.png",
    "48": "icons/facebook.png",
    "128": "icons/facebook.png"
  },

  "host_permissions": ["https://www.facebook.com/*"],
  "content_scripts": [
    {
      "matches": ["https://www.facebook.com/*"],
      "js": ["content.js"],
      "run_at": "document_idle"
    }
  ]
}
```

Figure 8.21: manifest.js file for Facebook extension

- **Name and Description:** The extension is named "SocialSense For Facebook" and described as a tool to classify Facebook comments using SocialSense.
- **Version:** The version of the extension is 1.0.
- **Permissions:**
    - `activeTab`: Allows the extension to interact with the currently active tab.
    - `scripting`: Grants permission for the extension to execute scripts.
- **Background Script:**
    - Utilizes a service worker defined in `background.js`.
- **Action:**
    - Specifies default icons for different sizes (16x16, 48x48, 128x128).

- **Icons:**
    - Defines icons for the extension in different sizes.
- **Host Permissions:**
    - Grants access to `https://www.facebook.com/*` indicating the extension will operate on Facebook pages.
- **Content Scripts:**
    - Injects `content.js` into web pages matching the Facebook URL pattern.
    - Scripts run when the document is in the idle state.

8.1.6.2 Background.js

In web development, a `background.js` file typically serves as a script running in the background of a browser extension, handling tasks such as event handling, API interactions, and data manipulation. In the `background.js` file (Figure 8.22), the Models API was configured to await responses after sending comments.

```
chrome.runtime.onMessage.addListener((message, sender, sendResponse) =>
{
  if (message.action === 'analyzeComment') {
    fetch('https://api.openai.com/v1/chat/completions', {
      method: 'POST',
      headers: {
        'Content-Type': 'application/json',
        'Authorization': 'Bearer
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'
      },
      body: JSON.stringify({
        model: "ft:gpt-3.5-turbo-0125:personal:socialsense:96eqMPvL",
        messages: [
          {"role": "system", "content": "You are a text classifier for
social media comments. Classify the following comment into one of the
following classes: [Positive, Neutral, Negative - Respond, Negative -
Ignore, Negative - Remove, Crisis]"},
          {"role": "user", "content": message.comment}
        ]
      })
    })
    .then(response => {
      if (!response.ok) {
        throw new Error('Network response was not ok');
      }
      return response.json();
    })
    .then(data => {
      if (data.choices && data.choices.length > 0 &&
data.choices[0].message && data.choices[0].message.content) {
        sendResponse({ response: data.choices[0].message.content });
      } else {
        throw new Error('Response data is invalid');
      }
    })
    .catch(error => {
      console.error('Error:', error);
      sendResponse({ error: error.message });
    });

    return true;
  }
});
```

Figure 8.22: background.js file for Facebook extension.

- **Event Listener:** It listens for messages sent by other parts of the extension or web page using `chrome.runtime.onMessage.addListener()`.
- **Message Action:** It checks if the received message has an action property set to `'analyzeComment'`.
- **API Call:** If the action is `'analyzeComment'`, it sends a POST request to the OpenAI API endpoint `https://api.openai.com/v1/chat/completions` with the provided comment.
- **Authorization Header:** It includes an Authorization header with a bearer token for authentication.
- **Request Body:** The request body contains the comment to be analyzed and a predefined message for the AI model to classify the comment into specific classes.
- **Handling Response:** Upon receiving a response from the API, it checks if the response is successful (`response.ok`). If successful, it parses the JSON response and sends the classified comment back as a response using `sendResponse()`.
- **Error Handling:** If there's an error during the API call or response handling, it catches the error, logs it to the console, and sends the error message back to the caller.
- **Returning True:** It returns `true` to indicate that the response will be sent asynchronously. This ensures that the event listener stays active and doesn't get removed after the listener function returns.

8.1.6.3 Content.js

In the `content.js` file (Figure 8.23), various functions were implemented to classify comments by sending them to the SocialSense AI model and retrieving responses.

```
function analyzeAndDisplayResponses() {}
function getComment(commentElement) {}
function displayResponse(commentElement, response) {}
```

Figure 8.23: Functions used in the script.js file for Facebook extension

- **'analyzeAndDisplayResponses()':** Analyzes comments in a webpage and displays corresponding responses based on their content.
- **'getComment(commentElement)':** Retrieves the text content of a comment element.
- **'displayResponse(commentElement, response)':** Displays a response for a comment, with styling based on the nature of the response.

8.1.7 Instagram Extension

Instagram extension powered by SocialSense AI, designed to classify comments swiftly and accurately into different categories: Positive, Neutral, Negative - Respond, Negative - Ignore, Negative - Remove, and Crisis as seen in Figure 8.24.

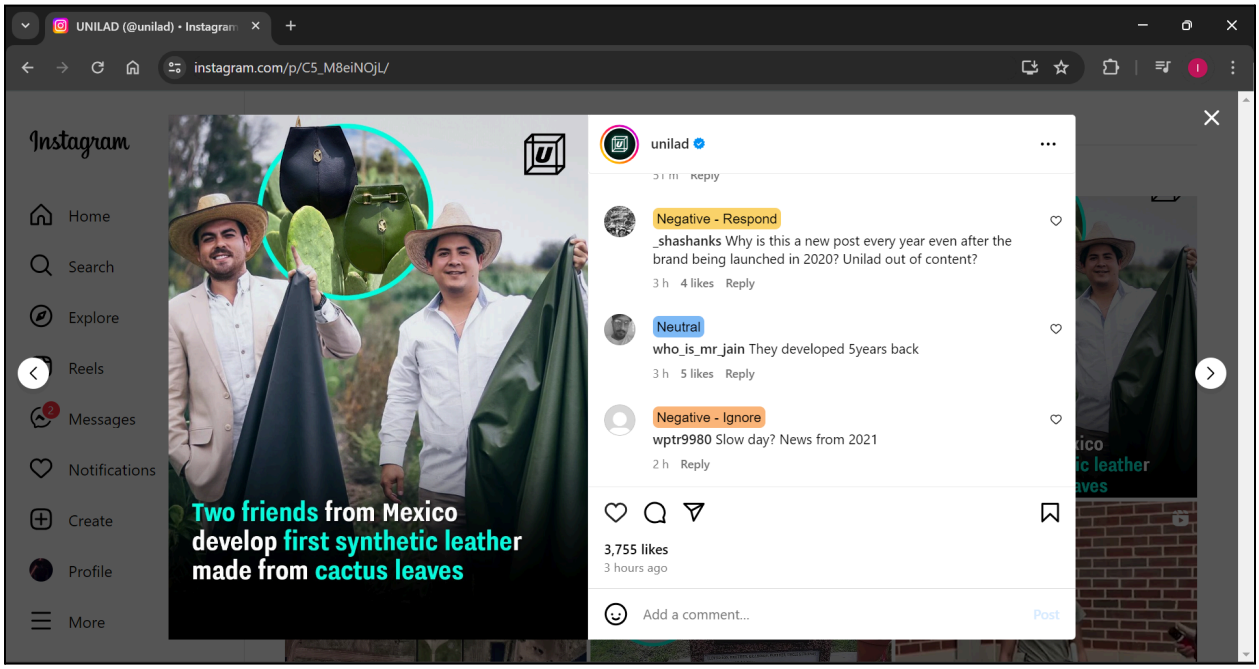


Figure 8.24: SocialSense API classifying Instagram comments in real time.

Minor modifications were implemented in the Facebook extension's manifest file, specifically adjusting host permissions to allow access to `https://www.instagram.com/*`, signaling the extension's intended operation on Instagram pages. Additionally, the target classes responsible for extracting comments were updated to align with Instagram's specifications and requirements.

8.1.8 TikTok Extension

The TikTok Extension (Figure 8.25), enhanced by the SocialSense AI, tailored to swiftly and accurately classify comments into distinct categories: Positive, Neutral, Negative - Respond, Negative - Ignore, Negative - Remove, and Crisis. Users can effortlessly navigate through the comment categories via a convenient drop-down selection feature (Figure 8.26). This intuitive design enables users to focus solely on comments belonging to their selected category, ensuring efficient management and response (Figure 8.27).

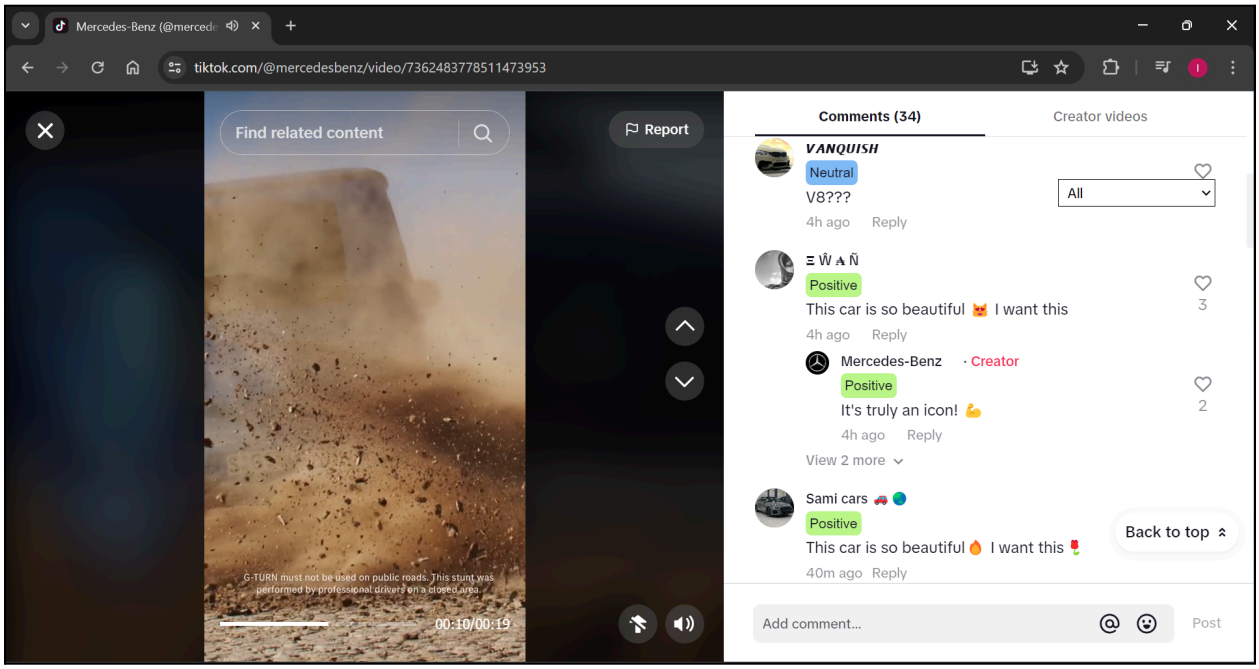


Figure 8.25: SocialSense API classifying TikTok comments in real time.

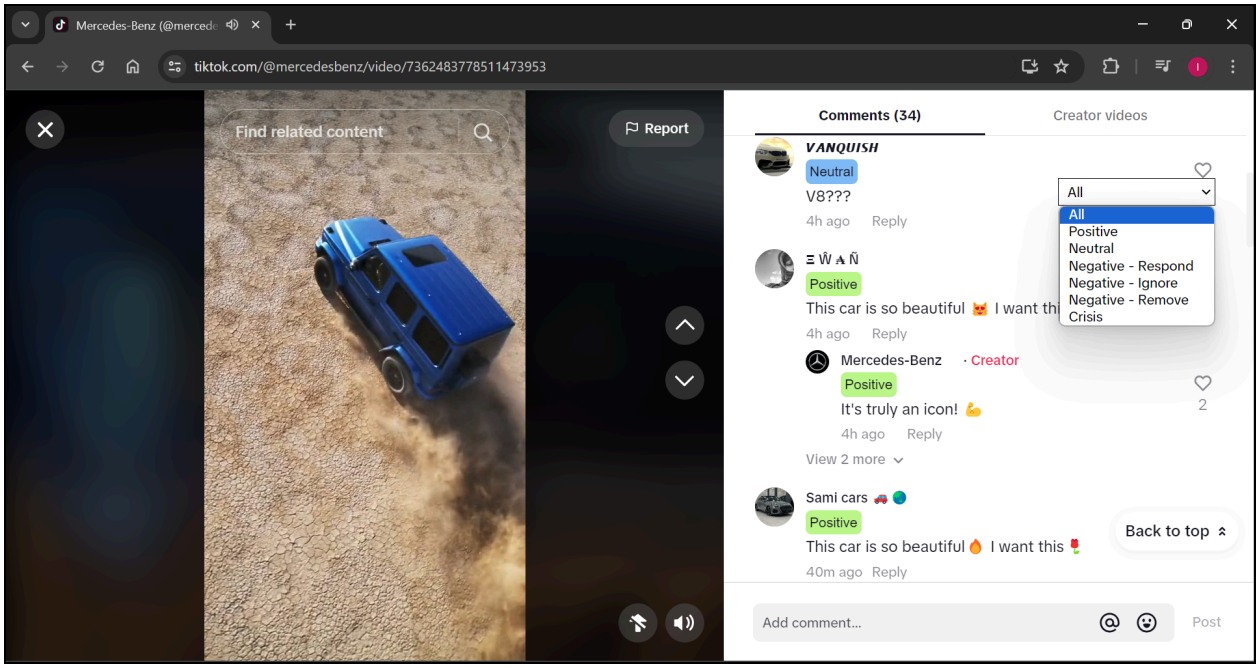


Figure 8.26: Dropdown for comment filtration on TikTok

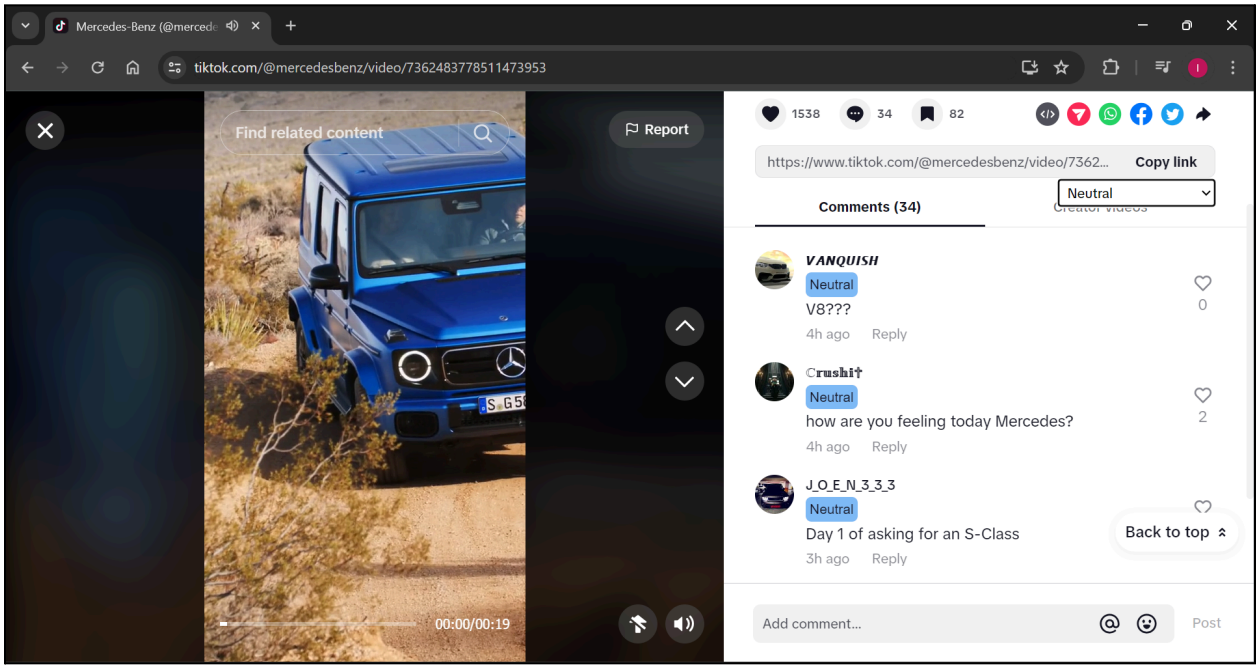


Figure 8.27: Comment classification after selection from dropdown

Minor modifications were implemented in the Facebook extension's manifest file, specifically adjusting host permissions to allow access to `https://www.tiktok.com/*`, signaling the extension's intended operation on Instagram pages. Additionally, the target classes responsible for extracting comments were updated to align with TikTok's specifications and requirements.

8.1.8.1 Script.js

In the script.js file (Figure 8.28), two additional functions have been implemented to enhance functionality: one forcreating a drop-down menu and another for filtering comments based on the user's selection.

```
function createDropdown(){}
function toggleResponseVisibility(){}
```

Figure 8.28: Additional functions in the script.js file

- **'createDropdown()':** Creates a dropdown menu dynamically with predefined options and styles, appending it to the body of the HTML document.
- **'toggleResponseVisibility()':** Controls the visibility of comment elements based on the selected value from the dropdown menu, showing or hiding them accordingly.

8.1.9 Web Page Deployment

The web page has been deployed (Figure 8.29) using Vercel and is now publicly accessible via https://social-sense.vercel.app/.

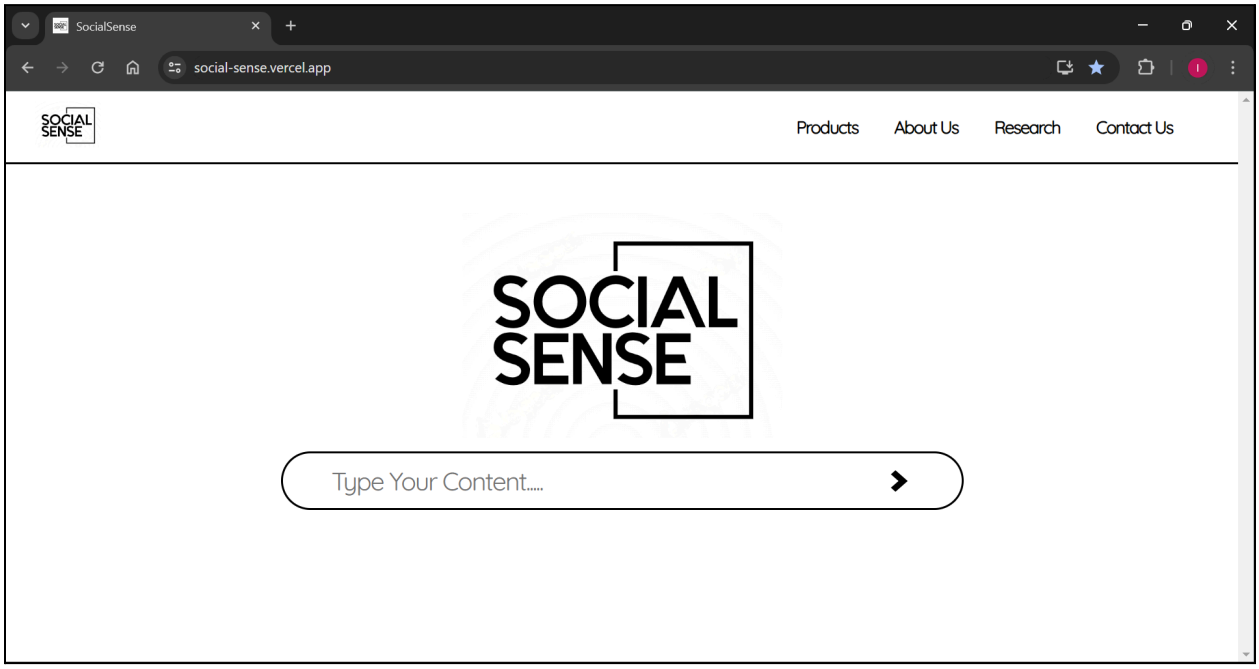


Figure 8.29: Deployed web page on Vercel

8.1.10 Chrome Extensions Deployment

The chrome extensions were approved by Google and published on Chrome Web Store: SocialSense For Facebook (Figure 8.30), SocialSense For Instagram (Figure 8.31), and SocialSense For TikTok (Figure 8.32). These extensions can be downloaded through the following links:

- **SocialSense For Facebook:** https://chromewebstore.google.com/detail/socialsense-for-facebook/hfngglbbpdcgjipinlgpbkllemaokfej
- **SocialSense For Instagram:** https://chromewebstore.google.com/detail/socialsense-for-instagram/hkjoggfopokhofccmelocagmgbigoblo
- **SocialSense For TikTok:** https://chromewebstore.google.com/detail/socialsense-for-tiktok/lponlmnipbnahjbhnfhodlfepegbjehh

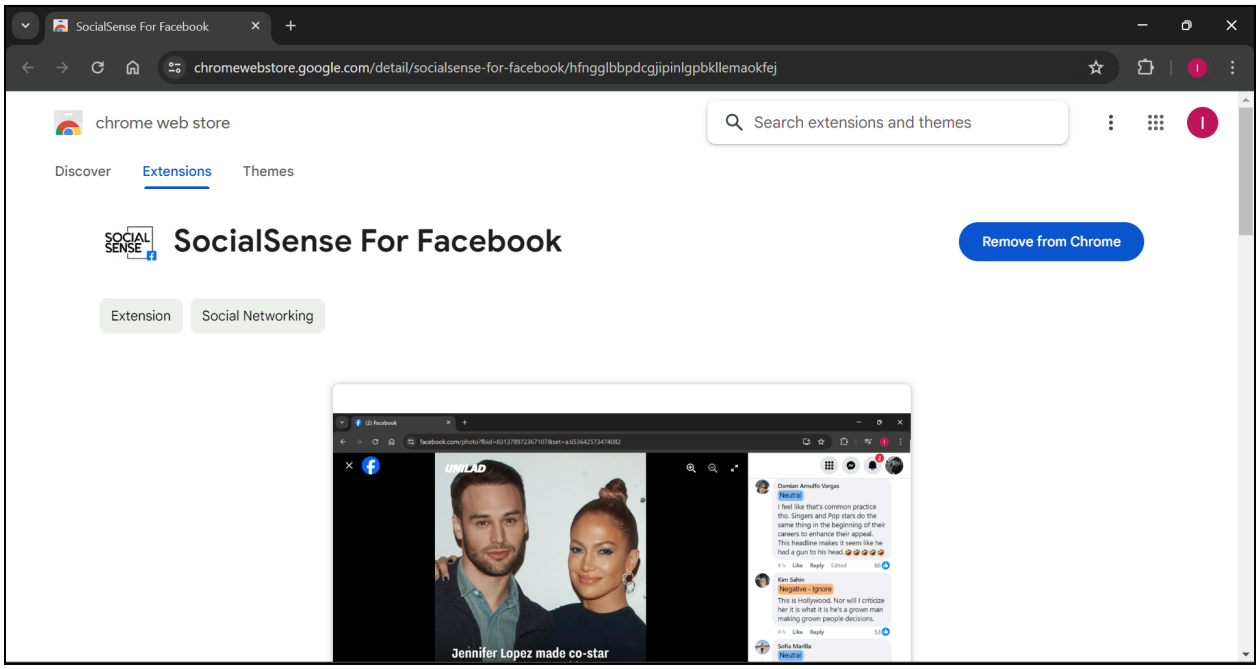


Figure 8.30: SocialSense For Facebook

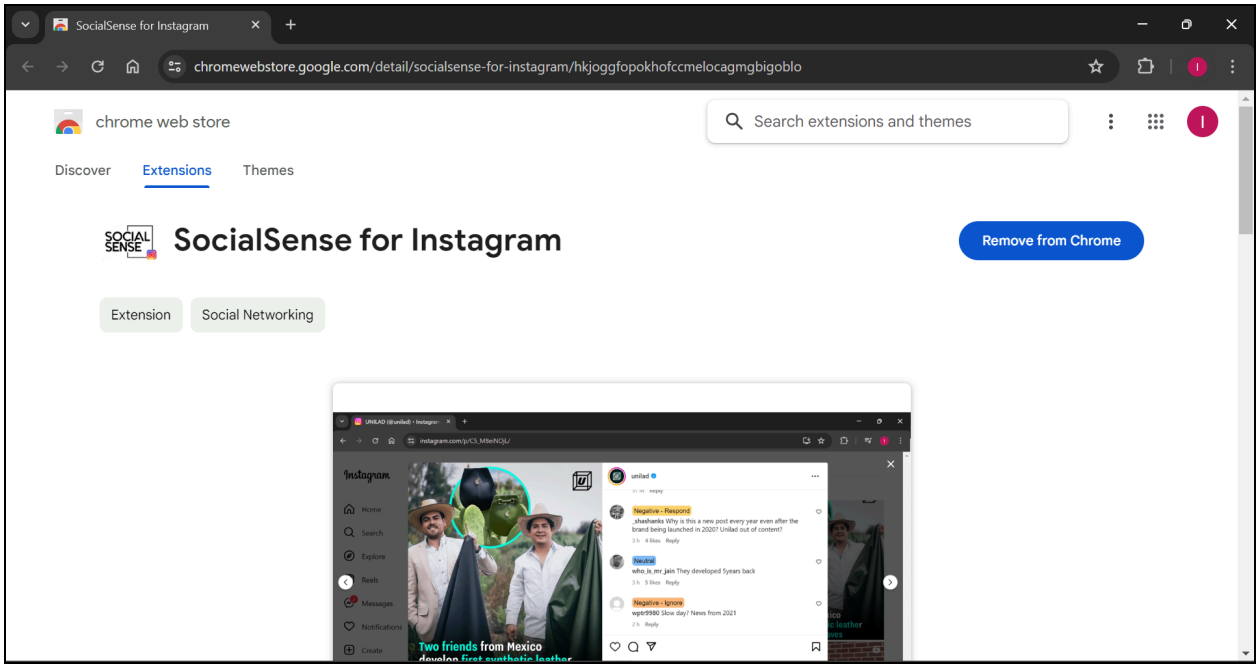


Figure 8.31: SocialSense For Instagram

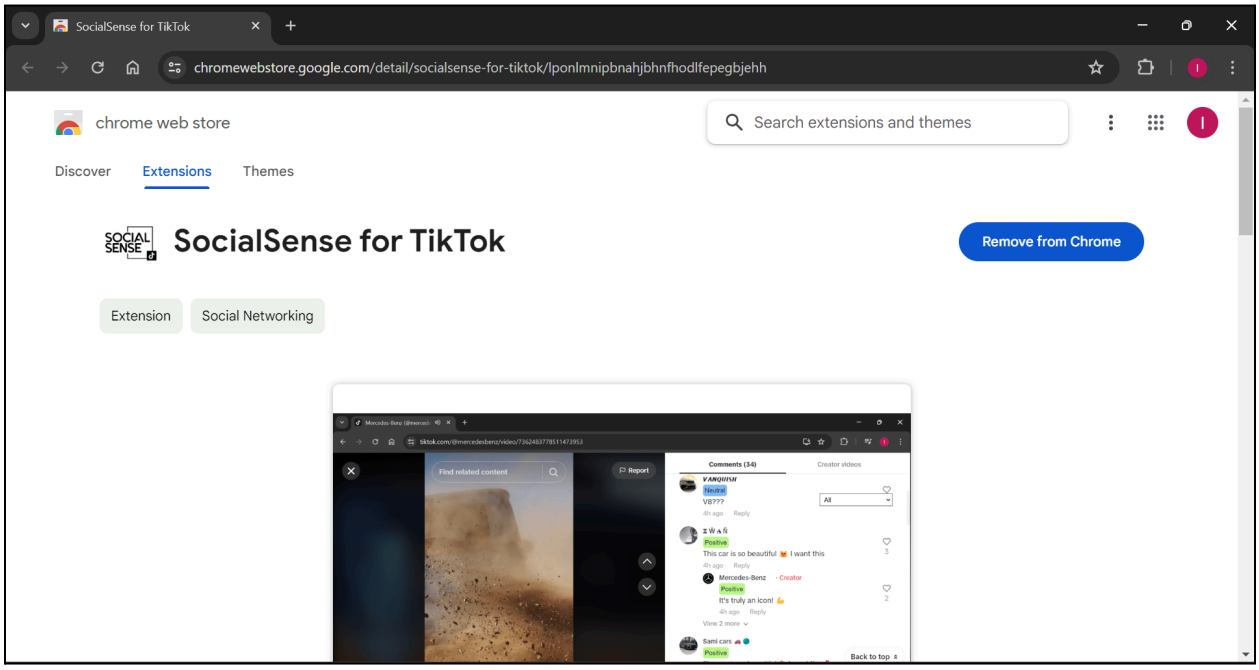


Figure 8.32: SocialSense For TikTok

8.2 Test Cases

Test Case FR 1.1: Comment Categorization

Test Scenario: Verify the comment categorization functionality of the model.
Preconditions:

- The model for comment categorization is properly set up and operational.
- The categories "Positive," "Neutral," "Negative - Respond," "Negative - Ignore," "Negative - Remove," and "Crisis" are defined in the model.

Test Steps:

1. Input Comment: "Sir Kamal kr dya hy"
2. Input Comment: “Ye resturant ka khana sahi tha”.
3. Input Comment: “قیمت بر ھاتے رہنا۔ کوالٹی اچھی نا کرنا۔”.
4. Input Comment: “Beggars can't be choosers”.
5. Input Comment: “اللہ پاک اس ناسور خاندان سے پاکستانیوں کی جلد جان چڑوائے امین”.
6. Input Comment: “dont worry i will destroy british people all in le head.”.

Expected Result:

1. The model should correctly classify the comment as "Positive."
2. The model should correctly classify the comment as "Neutral."
3. The model should correctly classify the comment as "Negative - Respond."
4. The model should correctly classify the comment as "Negative - Ignore."
5. The model should correctly classify the comment as "Negative - Remove."
6. The model should correctly classify the comment as "Crisis."

Alternate Flow 1
N/A (No alternate flow for this test case)

Test Case FR 1.2: Text Preprocessing

Test Scenario: Verify the text preprocessing functionality of the model.
Preconditions:

- The model for text preprocessing is set up and operational.
- The model is configured to perform tokenization, lowercase conversion, and special character removal.

Test Steps:

1. Input Comment: "Hello, this Is a Test - 123!"

Expected Result:

1. "hello this is a test"

Alternate Flow 1:
N/A (No alternate flow for this test case)

Test Case FR 1.3: LLM Selection

Test Case Scenario: Verify the process of choosing a suitable LLMl for sentiment analysis and categorization, ensuring it meets the project requirements.

Test Case Steps:

1. **Requirement Analysis:**
    a. Determine the specific project requirements for sentiment analysis and categorization, accuracy, speed, model size, and available resources.
2. **Model Evaluation:**
    a. Evaluate different fine-tuned LLM Llama 2 and GPT-3.5 regarding their performance, accuracy, and compatibility with the project's requirements.
3. **Model Testing:**
    a. Select one of the fine-tuned LLMs based on the evaluation and configure it for sentiment analysis and categorization.
4. **Testing with Sample Data:**
    a. Input sample comments or text data with varying sentiment and categories to the selected model.

Expected Result:

1. The selected fine-tuned LLM should provide accurate sentiment analysis and categorization results that meet the project's requirements. The model should be capable of correctly classifying comments into the specified categories and determining sentiment accurately.

Alternate Flow 1:

If the initially selected LLM does not meet the project's requirements in terms of accuracy or performance, return to step 2 and consider other pre-trained models for evaluation.

Test Case FR 1.4: Model Fine Tuning

Test Case Scenario: Verify that the selected LLM is fine-tuned on a labeled dataset for improved sentiment analysis and categorization performance.

Test Case Steps:

1. **Data Preparation:**
    a. Collect and prepare a labeled dataset for fine-tuning, including comments with corresponding sentiment labels and categories.
2. **Fine-Tuning Configuration:**
    a. Configure the selected pre-trained NLP model for fine-tuning, specifying hyperparameters, training duration, and batch size.
3. **Fine-Tuning Training:**
    a. Train the model on the prepared dataset using the configured settings.
4. **Evaluation Dataset:**
    a. Prepare a separate dataset for evaluation with a variety of comments and known sentiment labels and categories.
5. **Model Evaluation:**
    a. Evaluate the fine-tuned model's performance on the evaluation dataset.

Expected Result:

1. The fine-tuned model should show improved performance in sentiment analysis and categorization when compared to its performance before fine-tuning. It should correctly classify comments into the specified categories and provide accurate sentiment analysis.

Alternate Flow 1:

1. If the fine-tuned model's performance does not meet the desired level of improvement, consider adjusting hyperparameters and fine-tuning settings and repeat the fine-tuning process.

Test Case FR 1.5: Model Evaluation Metrics

Test Case Scenario: Verify that the model's performance is correctly evaluated using metrics F1-Score, Accuracy, Recall, and Precision.

Test Case Steps:

1. **Model Selection and Training:**
    a. Select a trained model that has undergone fine-tuning.
    b. Ensure that the model is capable of sentiment analysis and categorization.
2. **Evaluation Data Preparation:**
    a. Prepare an evaluation dataset with a representative set of comments or reviews, each labeled with sentiment categories.
3. **Model Evaluation:**
    a. Use the selected evaluation dataset to assess the model's performance.
4. **Metrics Calculation:**
    a. Calculate the following evaluation metrics for the model:
        i. Accuracy: Measure the overall accuracy of the model in correctly classifying comments.
        ii. F1-Score: Calculate the F1-Score, which balances precision and recall.
        iii. Recall: Determine the recall, which measures the model's ability to correctly identify positive, neutral, negative, or crisis comments.
        iv. Precision: Evaluate the precision, which indicates the model's accuracy in classifying comments.

Expected Result:

1. The model should achieve acceptable or desirable values for the specified evaluation metrics, including high accuracy, F1-Score, recall, and precision. The metrics should collectively reflect the model's capability to correctly analyze sentiment and categorize comments.

Alternate Flow 1:

1. If the model's performance based on the metrics is not satisfactory, consider fine-tuning the model, adjusting feature extraction techniques, or modifying training data to improve its evaluation metrics.

Test Case FR 2.1: Comment Submission for Categorization

Test Scenario: Verify that users can submit comments or reviews for categorization through the web application interface.

Preconditions:

- The user is on the comment submission page.
- The web application interface for comment submission is accessible.

Test Steps:

1. Enter or paste a comment in a language that is supported into the input field. For example, "I really enjoyed the product, it exceeded my expectations."
2. Click the "Submit" or equivalent button to submit the comment for categorization.

Expected Result: The web application correctly receives and processes the user's comment, and the comment is ready for analysis.

Alternate Flow 1:

Test Steps:

1. In step 2, enter a comment in an unsupported language (e.g., a language not covered by the model).

Expected Result: The web application displays an error message indicating that the language is not supported and the comment is not processed.

Test Case FR 2.2, 3.1: Displaying Categorization Results to Users

Test Scenario: Verify that users receive categorization results after the model analyzes their submitted comments and that the results are presented clearly and understandably.

Preconditions:

- The user has submitted a comment.
- The model has processed the submitted comment for categorization.

Test Steps:

1. The user submits a comment for analysis, for example: "The customer service was excellent, I had a great experience."

Expected Result:

1. The model successfully processes the user's submitted comment for categorization.
2. The model accurately categorizes the comment and assigns an appropriate label.
3. The user should receive the categorization result clearly and understand that their comment has been categorized as "Positive."

Alternate Flow 1:

N/A (No alternate flow for this test case)

Test Case FR 4.1: Documentation

Test Scenario: Testing the comprehensiveness of the API documentation.
Preconditions:

1. The API is implemented and ready for use.
2. The documentation is available for API users.

Test Steps:

1. Access the API documentation.
2. Check for the presence of API usage examples.
3. Verify that the documentation includes detailed endpoint descriptions.
4. Ensure that input and output formats (e.g., JSON, XML) are clearly documented.
5. Confirm that authentication instructions are provided.
6. Review the documentation for a mention of updates and version control.

Expected Result:

1. The API documentation is comprehensive and user-friendly.
2. It includes API usage examples to help users understand how to use the API effectively.
3. Detailed endpoint descriptions clarify the purpose and usage of each API endpoint.
4. Input and output formats are clearly documented, providing users with the necessary information.
5. Authentication instructions guide users on how to securely access the API.
6. The documentation mentions plans for updates and version control, ensuring users are informed about any changes.

Test Case FR 4.2: API Access Control

Test Scenario: Testing API access control mechanisms.
Preconditions:

- The API is operational and access control mechanisms are implemented.
- API users or applications have the required credentials for authentication.

Test Steps:

1. Attempt to access the API without valid credentials.
2. Attempt to access the API with valid credentials but with insufficient access permissions.
3. Attempt to access the API with valid credentials and sufficient access permissions.

Expected Result:

1. The API should deny access and return an authentication error message for unauthorized access attempts.
2. The API should deny access and return an authorization error message for users with insufficient permissions.
3. The API should grant access to authorized users with valid credentials and sufficient access permissions, allowing them to use the API securely.

Test Case FR 4.3: API Endpoint Definition

Test Scenario: Verifying the well-defined endpoints and comprehensive documentation.

Preconditions:

- The API is available for use.
- The API documentation is accessible to API users.

Test Steps:

1. Access the API documentation.
2. Review the endpoint definitions to ensure each endpoint's purpose, path, and parameters are clearly specified.
3. Check if the supported HTTP methods (e.g., GET, POST, PUT, DELETE) for each endpoint are clearly stated, along with their expected functionality.
4. Examine the required and optional request parameters for each endpoint, including their data types and constraints.
5. Verify that the format of responses returned by each endpoint (e.g., JSON or XML) is documented, and examples of response structures are provided.

Expected Result:

1. The API documentation should clearly define each endpoint and its purpose, path, and parameters.
2. Supported HTTP methods and their expected functionality should be explicitly mentioned.
3. Request parameters, along with their data types and constraints, should be well-documented.
4. The format of responses and examples of response structures should be provided, helping users interact with the API effectively.

8.3 Test Case Grid

Test Case ID	Test Case Description	Status
1.1	Verify comment categorization functionality.	□
1.2	Verify text preprocessing functionality.	□
1.3	Verify LLM selection for sentiment analysis.	□
1.4	Verify fine-tuning of the selected LLM.	□
1.5	Verify model evaluation metrics (F1, Accuracy, Recall)	□
2.1	Verify comment submission for categorization.	□
2.2	Verify display of categorization results to users.	□
3.1	Verify clear presentation of categorization results.	□
4.1	Verify comprehensiveness of API documentation.	□
4.2	Testing API access control mechanisms.	□
4.3	Verify well-defined API endpoints and documentation.	□

9 Evaluation

9.1 Performance Metrics

The fine-tuned model has achieved an overall accuracy of 84%. Table 9.1 presents the model's accuracy across different classes, while Figure 9.1 displays the confusion matrix. Furthermore, Table 9.2 provides insights into the model's performance across six classes classes in English, Urdu (Roman Script), and Urdu (Arabic Script).

Table 9.1: Model accuracy on each label

Label	Accuracy
Positive	89%
Neutral	67%
Negative - Respond	79%
Negative - Ignore	78%
Negative - Remove	93%
Crisis	98%
Average	84%

Table 9.2: Model accuracy on each label for each language

Label \ Language	English	Urdu (Roman Script)	Urdu (Arabic Script)
Positive	88%	88%	90%
Neutral	86%	62%	54%
Negative - Respond	92%	58%	86%
Negative - Ignore	60%	88%	86%
Negative - Remove	90%	94%	96%
Crisis	100%	94%	100%
Average	86%	81%	85%

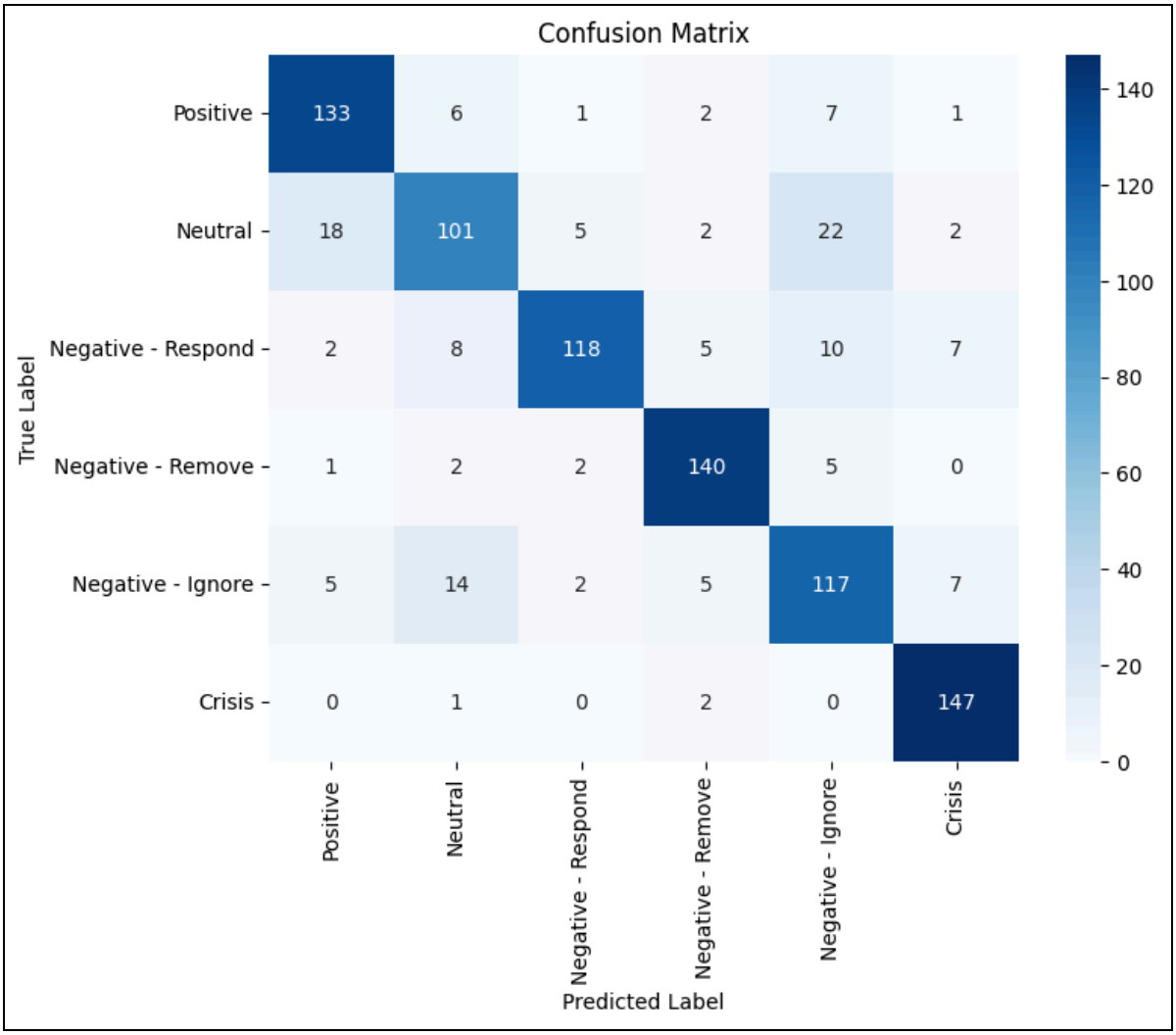


Figure 9.1: Confusion matrix for the model predictions done on test data

The model attained an accuracy of 86% for comments in English, as depicted in the confusion matrix displayed in Figure 9.2, which illustrates the model's performance. Sample comments in English, along with their corresponding model predictions, are presented in Appendices A.1, B.1, C.1, D.1, E.1 and F.1.

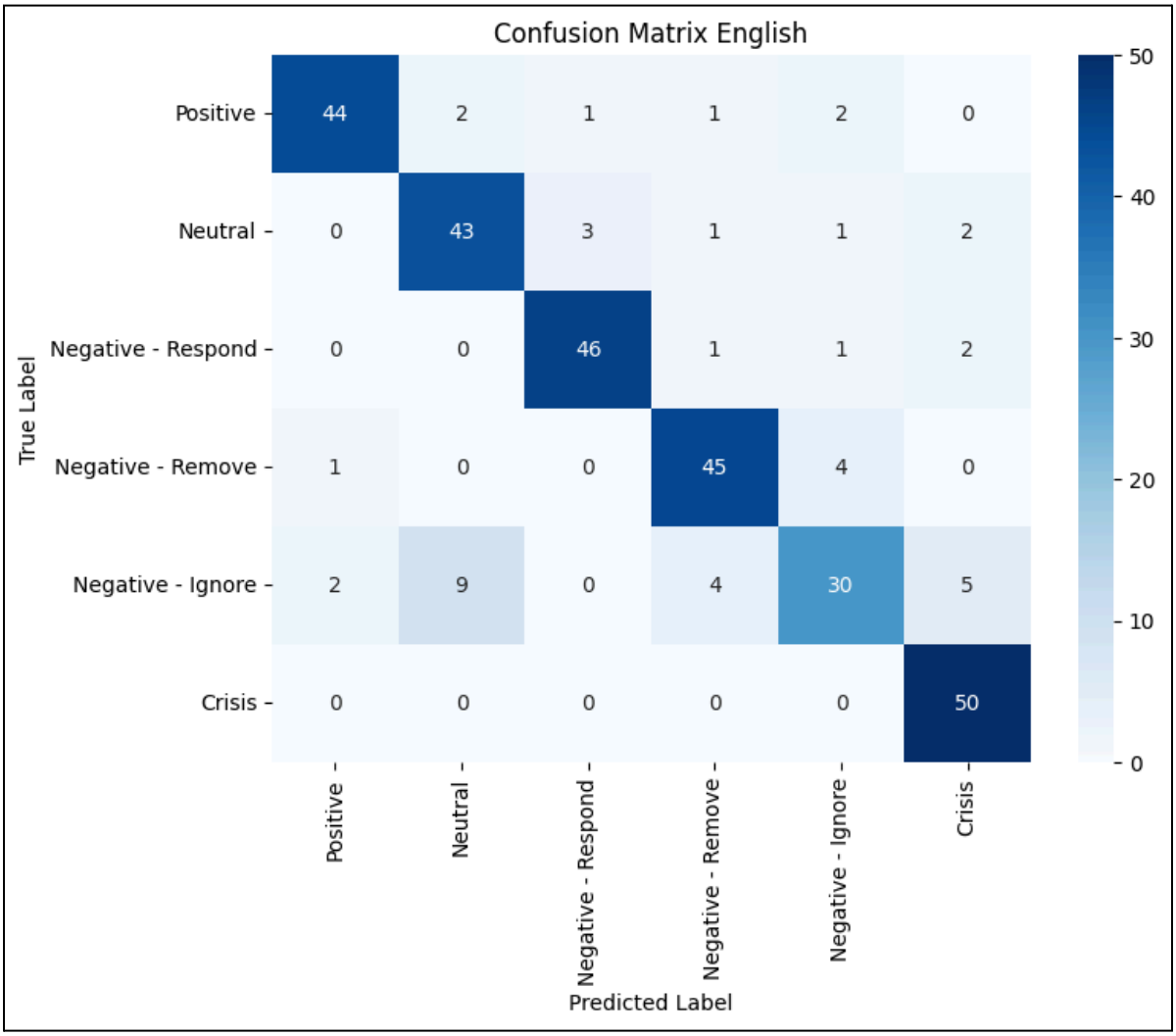


Figure 9.2: Confusion matrix for English comments

The model attained an accuracy of 81% for comments in Urdu (Roman Script), as depicted in the confusion matrix displayed in Figure 9.3, which illustrates the model's performance. Sample comments in Urdu (Roman Script), along with their corresponding model predictions, are presented in Appendices A.2, B.2, C.2, D.2, E.2 and F.2.

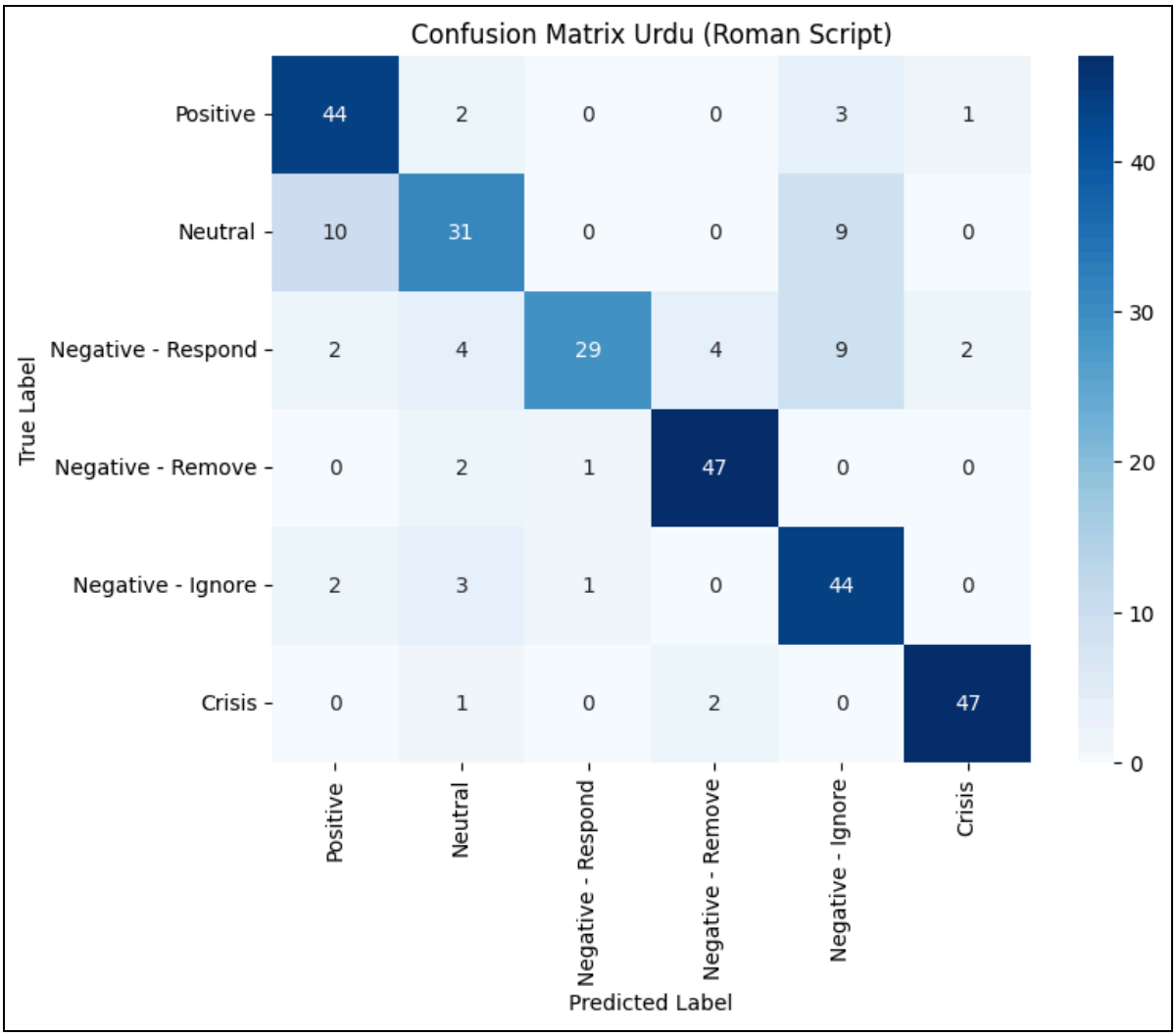


Figure 9.3: Confusion matrix for Urdu (Roman Script) comments

The model attained an accuracy of 85% for comments in Urdu (Arabic Script), as depicted in the confusion matrix displayed in Figure 9.4, which illustrates the model's performance. Sample comments in Urdu (Arabic Script), along with their corresponding model predictions, are presented in Appendices A.3, B.3, C.3, D.3, E.3 and F.3.

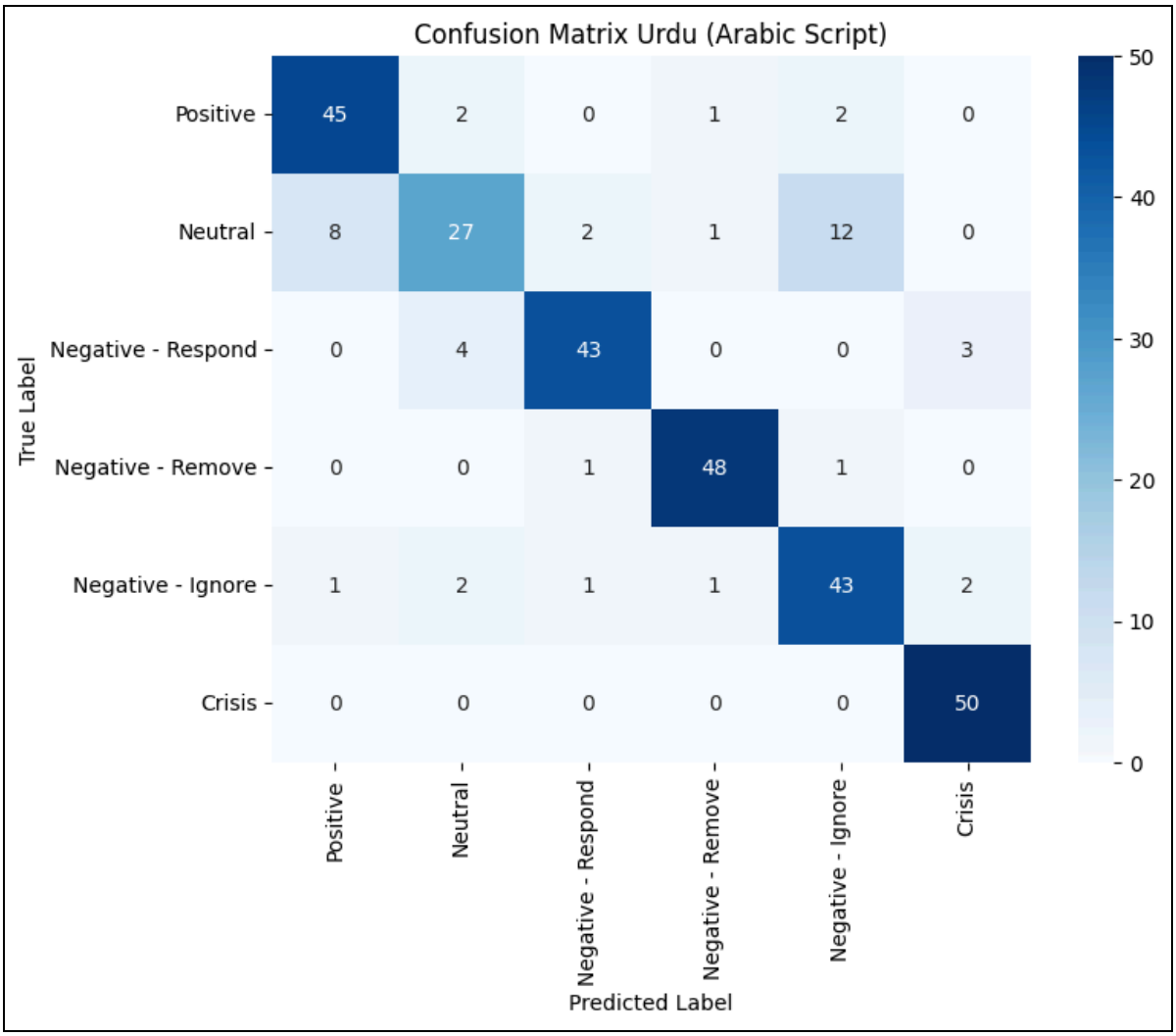


Figure 9.4: Confusion matrix for Urdu (Arabic Script) comments

9.2 Discussion

The model's overall accuracy of 84% is commendable, especially when compared with current research findings outlined in the Literature Review section. This level of performance highlights the efficacy of the model in various linguistic and contextual scenarios. Notably, the model excels in specific categories, demonstrating its highest sensitivity in the Crisis and Negative-Remove categories, where it achieves impressive accuracies of 98% and 93% respectively. These results underscore the model's ability to accurately identify and categorize critical and negative content, which is crucial for timely and appropriate responses in crisis management and content moderation.

Furthermore, the model maintains high accuracies across different languages, achieving an accuracy of 86% on English comments, 85% on Urdu (Arabic Script) comments, and 81% on Urdu (Roman Script) comments. The consistency in performance across these languages is indicative of the model's multilingual capabilities and its potential applicability in diverse linguistic contexts.

The slightly lower accuracy observed in Urdu (Roman Script) comments, compared to English and Urdu (Arabic Script), can be primarily attributed to the absence of standardized spelling conventions in Urdu (Roman Script). Unlike English and Urdu (Arabic Script), which have more rigid and consistent spelling rules, Urdu (Roman Script) allows users greater flexibility in spelling and formatting. This flexibility leads to significant variations in how words are written, with users often employing personal or regional preferences in their spelling choices. These variations create a challenging landscape for the model, as it must effectively adapt to and interpret a wide array of non-standardized writing styles.

In conclusion, while the model demonstrates outstanding performance across various categories and languages, the observed disparity in accuracy for Urdu (Roman Script) highlights the ongoing challenges associated with non-standardized linguistic forms. Addressing these challenges through enhanced algorithms and normalization techniques will be crucial for further improving the model's accuracy and ensuring its effectiveness in a wider range of applications.

10 Conclusion

This project has successfully delivered an advanced AI model and user-friendly interfaces that enhance social media interaction management. The GPT-3.5 Turbo LLM-based model, achieving an 84% accuracy rate, effectively classifies comments into six categories, proving its robustness in handling diverse data.

The development of an API and a user-friendly Chrome webpage facilitates seamless integration and provides detailed analytics. Our Chrome extensions for Facebook, Instagram, and TikTok enable real-time comment classification, fostering a positive online environment. Their successful deployment on the Chrome Web Store marks a significant milestone in making our solution widely accessible.

Looking ahead, integrating our model into the GPTstore by OpenAI and upgrading to GPT-4 will improve accuracy and introduce advanced features. We plan to explore partnerships with social media platforms and expand our extensions to other platforms and forums, broadening our impact.

In conclusion, this project has laid a solid foundation for AI-driven social media interaction management. Our commitment to innovation and user satisfaction ensures ongoing improvement and expansion, driving meaningful change in online interactions.

11 References

[1] M. Samory, I. Sen, J. Kohne, F. Flöck, and C. Wagner, "Call me sexist, but...": Revisiting Sexism Detection Using Psychological Scalesand Adversarial Samples, https://ojs.aaai.org/index.php/ICWSM/article/view/18085/17888 (accessed Oct. 5, 2023).

[2] A. Albanyan and E. Blanco, "Pinpointing fine-grained relationships between hateful tweets and replies," Proceedings of the AAAI Conference on Artificial Intelligence, https://ojs.aaai.org/index.php/AAAI/article/view/21284 (accessed Oct. 5, 2023).

[3] L. Grimminger and R. Klinger, "Hate towards the political opponent: A twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection," ACL Anthology, https://aclanthology.org/2021.wassa-1.18/ (accessed Oct. 5, 2023).

[4] N. Alias, C. F. M. Foozy, S. N. Ramli, and N. Zainuddin, "Video spam comment features selection using Machine Learning Techniques," Indonesian Journal of Electrical Engineering and Computer Science, https://ijeecs.iaescore.com/index.php/IJEECS/article/view/18626 (accessed Oct. 5, 2023).

[5] D. Andročec, "Machine learning methods for toxic comment classification: A systematic review," *Acta Universitatis Sapientiae, Informatica*, vol. 12, no. 2, pp. 205–216, 2020. doi:10.2478/ausi-2020-0012

[6] E. W. Pamungkas, V. Basile, and V. Patti, "Do you really want to hurt me? predicting abusive swearing in social media," ACL Anthology, https://aclanthology.org/2020.lrec-1.765/ (accessed Oct. 5, 2023).

[7] V. Zhukov, "Text classification challenge with extra-small datasets: Fine-tuning versus chatgpt," Medium, https://towardsdatascience.com/text-classification-challenge-with-extra-small-datasets-fine-tuning-versus-chatgpt-6348fecea357 (accessed Oct. 15, 2023).

[8] H. Rizwan, M. H. Shakeel, and A. Karim, "Hate-speech and offensive language detection in Roman Urdu," ACL Anthology, https://aclanthology.org/2020.emnlp-main.197/ (accessed Oct. 5, 2023).

[9] R. Saeed, H. Afzal, S. A. Rauf, and N. Iltaf, "Detection of offensive language and its severity for low resource language," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 6, pp. 1–27, 2023. doi:10.1145/3580476

[10] M. Bilal, A. Khan, S. Jan, S. Musa, and S. Ali, "Roman Urdu hate speech detection using transformer-based model for Cyber Security Applications," *Sensors*, vol. 23, no. 8, p. 3909, 2023. doi:10.3390/s23083909

[11] U. Azam, H. Rizwan, and A. Karim, "Exploring data augmentation strategies for hate speech detection in Roman Urdu," ACL Anthology, https://aclanthology.org/2022.lrec-1.481/ (accessed Oct. 5, 2023).

[12] S. Aziz, M. S. Sarfraz, M. Usman, M. U. Aftab, and H. T. Rauf, "Geo-spatial mapping of hate speech prediction in Roman Urdu," *Mathematics*, vol. 11, no. 4, p. 969, 2023. doi:10.3390/math11040969

[13] R. Mangroveightyone, "Code-mixing in social media text the last language identification frontier," Academia.edu, https://www.academia.edu/26517855/Code_Mixing_in_Social_Media_Text_The_Last_Language_Identification_Frontier (accessed Nov. 1, 2023).

[14] M. Shahroz, M. F. Mushtaq, A. Mehmood, S. Ullah, and G. S. Choi, "RUTUT: Roman Urdu to Urdu Translator Based on Character Substitution Rules and Unicode Mapping," *IEEE Access*, vol. 8, pp. 189823–189841, 2020, doi: https://doi.org/10.1109/access.2020.3031393.

[15] X. Sun *et al.*, "Text Classification via Large Language Models," *arXiv.org*, May 22, 2023. https://arxiv.org/abs/2305.08377.

[16] Y.-M. Chae and T. Davidson, "Large Language Models for Text Classification: From Zero-Shot Learning to Fine-Tuning," Aug. 2023, doi: https://doi.org/10.31235/osf.io/sthwk.

[17] Jeevithashree Divya Venkatesh, A. Jaiswal, and G. Nanda, "Comparing Human Text Classification Performance and Explainability with Large Language and Machine Learning Models Using Eye-Tracking," *Research Square (Research Square)*, Mar. 2024, doi: https://doi.org/10.21203/rs.3.rs-4002294/v2.

[18] Harika Abburi, M. Suesserman, Nirmala Pudota, Balaji Veeramani, E. Bowen, and S. Bhattacharya, "Generative AI Text Classification using Ensemble LLM Approaches," *arXiv (Cornell University)*, Sep. 2023, doi: https://doi.org/10.48550/arxiv.2309.07755.

[19] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing, "Sentiment Analysis in the Era of Large Language Models: A Reality Check," *arXiv.org*, May 24, 2023. https://arxiv.org/abs/2305.15005

[20] Jan Ole Krugmann and J. Hartmann, "Sentiment Analysis in the Age of Generative AI," *Customer Needs and Solutions*, vol. 11, no. 1, Mar. 2024, doi: https://doi.org/10.1007/s40547-024-00143-4.

[21] T. T. Nguyen, C. Wilson, and J. Dalins, "Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts," *arXiv.org*, Aug. 28, 2023. https://arxiv.org/abs/2308.14683 (accessed Oct. 12, 2023).

[22] P. Ji, "Text sentiment analysis based on LLaMA models," Mar. 2024, doi: https://doi.org/10.1117/12.3026733.

[23] Editor, "The 6 Critical Types of Social Media Comments You Must Plan For," *jeffbullas.com*, Aug. 19, 2014. https://www.jeffbullas.com/the-6-critical-types-of-social-media-comments-you-must-plan-for/ (accessed May 04, 2024).

[24] G. Makarskaite, "Social Media Sentiment and Engagement: What Is Their Connection? - Attention Insight," *attentioninsight.com*, Nov. 29, 2022. https://attentioninsight.com/social-media-sentiment-and-engagement/

[25] A. Chacko, "Your guide to social media comments: How to post and respond," *Sprout Social*, Nov. 07, 2023. https://sproutsocial.com/insights/social-media-comments/

[26] admin, "6 Types of Social Media Comments and How to Respond," *JSH Web Designs*, Jun. 18, 2015. https://jshwebdesigns.com/blog/6-types-of-social-media-comments-and-how-to-respond/ (accessed May 04, 2024).

[27] "Instagram Help Center," *help.instagram.com*. https://help.instagram.com/700284123459336

[28] R. Hellewell, "Block Comment Spam Bots," *WordPress.org*. https://wordpress.org/plugins/block-comment-spam-bots/#developers (accessed May 05, 2024).

[29] "Learn about comment settings - YouTube Help," *support.google.com*. https://support.google.com/youtube/answer/9483359?hl=en#zippy%3D%2Chidden-users&zippy

=%2Callow-all-comments%2Chold-potentially-inappropriate-comments-for-review%2Chold-all-comments-for-review%2Cturn-off-comments (accessed May 05, 2024).

[30] “Perspective API,” *www.perspectiveapi.com*. https://www.perspectiveapi.com/

[31]“ChatGPT Prompt Engineering for Developers,” *www.deeplearning.ai*. https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/

12 Appendices

Appendix A.1: Model predictions on Positive comments in Engish

Ground Truth	Language	Comments	Prediction
Positive	English	Sam Bahadur	Neutral
Positive	English	I will still rock this with steelies.	Neutral
Positive	English	Lav you	Positive
Positive	English	I was *just* thinking yesterday, that itâ€™s great he survived it..	Negative - Ignore
Positive	English	Get well soon.	Positive
Positive	English	get well soon sir	Positive
Positive	English	It's so satisfying to watch	Positive
Positive	English	ROWAN ATKINSON is the name of this great actor.!	Positive
Positive	English	Weird but I love it!	Positive
Positive	English	SAM Bahadur superb... everyone must watch this movie...	Positive
Positive	English	Very funny man and a nice car collection	Positive
Positive	English	I like it !	Positive
Positive	English	Sambahadur is worth watching. Quite practical move and great acting.	Positive
Positive	English	Sam bahadur best of luck	Positive
Positive	English	Good	Positive
Positive	English	Yes Sam bahadur is better movie	Positive
Positive	English	Looking georgeos	Positive
Positive	English	Zainab keep up the amazing work.	Positive
Positive	English	Mashallah nice	Positive

Appendix A.2: Model predictions on Positive comments in Urdu (Roman Script)

Ground Truth	Language	Comments	Prediction
Positive	Roman Urdu	jannat Khush raho Always	Positive
Positive	Roman Urdu	Kitni rishvat di thi $$ uparvale ko itni khubsurat ho ne k liyeðŸ˜ðŸ˜	Positive
Positive	Roman Urdu	kash ye bhai waha pe hota aur use sahi salamat uske ghar chod ata i am sorry for not being there with u my sister#justiceforpriyankareddy	Positive
Positive	Roman Urdu	Mashallah ðŸ˜	Positive
Positive	Roman Urdu	HYee MashAllah ðŸ¥° â¤ï¸	Positive
Positive	Roman Urdu	Wah mazy mazy	Positive
Positive	Roman Urdu	Wah wah jeee	Positive
Positive	Roman Urdu	Kasturi kebabðŸ˜	Positive
Positive	Roman Urdu	Love you jaan	Positive
Positive	Roman Urdu	Ye hua na content	Negative - Ignore
Positive	Roman Urdu	Kis kis ko hania psnd hai Woh like kre â¤ï¸âœ¨ðŸ¤ðŸ¥ºðŸ¦‹	Neutral
Positive	Roman Urdu	LOVE U GURU JI	Positive
Positive	Roman Urdu	MasAllah MasAllah so beautiful â¤ï¸â¤ï¸â¤ï¸â¤ï¸â¤ï¸â¤ï¸	Positive
Positive	Roman Urdu	inshallah apka naseeb bohut acha huga ðŸ¤—ðŸ¤—ðŸ–¤â¤ï¸â¤ï¸â¤ï¸â¤ï¸	Positive
Positive	Roman Urdu	bohat piyari lag Rahi hen	Positive
Positive	Roman Urdu	Boht payri lg ryii ma Sha Allah...	Positive
Positive	Roman Urdu	hamein or jeeny ki chahat na hoti agar tum na hoty agar tum na hoty	Positive
Positive	Roman Urdu	Pahli bar maja aya	Positive
Positive	Roman Urdu	Ab Shadi KarloðŸ¥°	Positive

Appendix A.3: Model predictions on Positive comments in Urdu (Arabic Script)

Ground Truth	Language	Comments	Prediction
Positive	Urdu	مبارک	Positive
Positive	Urdu	انسان صرف تمنا کر سکتا ہے"✍ ملتا وہی ہے جو مقدر میں ہوتا ہے* 💝	Positive
Positive	Urdu	شکریہ 🌹 آپکی محبتوں کو کا سلام	Positive
Positive	Urdu	😁 وأطلق من سوى الترند	Negative - Ignore
Positive	Urdu	❤️ شکریہ فضلووو خوش رہیں آمین	Positive
Positive	Urdu	💕😍اسلام علیکم😍 کیسے ہیں ٹویٹر ورلڈ کے پیارے لوگو	Positive
Positive	Urdu	اشعار کرتے ہیں یہاں اظہار عشق میرے معاشرے میں سر عام محبت جائز نہیں ❤️ ثمن_بلوچ	Neutral
Positive	Urdu	سیکھ رہا ہوں میں بھی انسانوں کو پڑھنے کا ہنر سنا ہے کتابوں سے زیادہ چہروں پہ لکھا ہوتا ہے 🌾🌾🌾🌾🌾🌾🌾🌾🌾🌾🍀🌼🥀🌴🌴🌹🌹🌺🌺 اردو_زبان	Positive
Positive	Urdu	Ghazali 💐💐💐🌹	Positive
Positive	Urdu	love u 😍😍😍😋 ہائے 😅😘😘😍😋😛 تم بھی بہت لاجواب	Positive
Positive	Urdu	❤️❤️❤️ دنیا میں ہر چیز کا متبادل موجود ہے، لیکن محنت کا نہیں۔	Positive
Positive	Urdu	تلفظ، اندازیاں اور انتخاب بے حد کمال❤️❤️	Positive
Positive	Urdu	❤️اپنی زبان اُردو قومی زبان اُردو۔ ہم سب کا مان ہے یہ پیاری زبان اُردو	Positive
Positive	Urdu	❤️ شکریہ تنویر خوش رہیں آمین	Positive
Positive	Urdu	😂ہائے وڈیریاں ہڈیاں اتوں موسم ٹھنڈا	Positive
Positive	Urdu	🌹 🌹 ✌️ ✌️ انشا اللہ صابر بھائی پیشگی مبارک	Positive

Appendix B.1: Model predictions on Neutral comments in Engish

Ground Truth	Language	Comments	Prediction
Neutral	English	Are you available for my mehndi date is 17 jan	Neutral
Neutral	English	wrong numberðŸ˜,	Neutral
Neutral	English	Leakage	Crisis
Neutral	English	B	Neutral
Neutral	English	My name is Bean, James Bean	Neutral
Neutral	English	I had studied from you back in 2013	Neutral
Neutral	English	Damn Mr. Bean out here driving around with 2/32 tread left on his tires	Neutral
Neutral	English	Suneel sb said it is used for playing music	Neutral
Neutral	English	Legends focus ðŸ˜œ	Neutral
Neutral	English	Jhonny english secret mission	Neutral
Neutral	English	Charge battery	Neutral
Neutral	English	Is this the one he wrecked?	Neutral
Neutral	English	No way he actually owns one	Neutral
Neutral	English	Hardly unworldly!	Neutral
Neutral	English	cow ðŸ„	Negative - Remove
Neutral	English	No...no...no... this is his car	Neutral
Neutral	English	He drove past me in this going down the m1 years ago. Got a cheeky wave out of him	Neutral
Neutral	English	Who are you brother?	Negative - Ignore
Neutral	English	Its cool	Neutral

Appendix B.2: Model predictions on Neutral comments in Urdu (Roman Script)

Ground Truth	Language	Comments	Prediction
Neutral	Roman Urdu	15 pro max JV box pack NATURAL TITANIUM ki kia Price hay jaldi mjhy btayen	Neutral
Neutral	Roman Urdu	pta nhi kaun shi kaun galat.	Neutral
Neutral	Roman Urdu	qaid azam ka koi beta nahi tha phr poota kaha se aa gya?	Neutral
Neutral	Roman Urdu	chalo dekh len gey in sha allah	Positive
Neutral	Roman Urdu	sadi wali Porsche	Negative - Ignore
Neutral	Roman Urdu	har rat ka andhera honey se pehly haal poch lia kro aesa na ho k hm matti me so jayen or tum duniya me dhondty rahoÃ¯Â¿Â½Ã¯Â¿Â½	Neutral
Neutral	Roman Urdu	sari raat jaagti na	Negative - Ignore
Neutral	Roman Urdu	klko ye extra-logical bnde bolnge:- push ups and bench press krna chhod do, kyuki usme aapke triceps pehle hi thk jate hain, uski jgh cable flys kro..... Ptta chla chest bn gyi lekin triceps tndi se reh gye....	Neutral
Neutral	Roman Urdu	Sir kya ye trick squats ke sath bhi h kaam kr sakti h ...????	Neutral
Neutral	Roman Urdu	Sir elbow pain ka cure btao	Neutral
Neutral	Roman Urdu	okay mai un sai mang leata hon	Neutral
Neutral	Roman Urdu	Bhai Kya hoa iss Ka scene?	Negative - Ignore
Neutral	Roman Urdu	baat to aap ki sach hai maggar hota aaisa hi hai	Negative - Ignore
Neutral	Roman Urdu	ye arabic me kya likha hai?	Neutral
Neutral	Roman Urdu	Ã¯Â¿Â½Ã¯Â¿Â½Ã¯Â¿Â½main muktan ka hun to meharbani aus ka support ni karna aapny	Negative - Ignore
Neutral	Roman Urdu	sister please aek video Imran Khan ki bnao	Neutral
Neutral	Roman Urdu	local alag alag conet ho rhi hai	Neutral
Neutral	Roman Urdu	Ap mera qarsh ho	Neutral

Appendix B.3: Model predictions on Neutral comments in Urdu (Arabic Script)

Ground Truth	Language	Comments	Prediction
Neutral	Urdu	تشرح بذمہ وضمیر 🤡😞	Neutral
Neutral	Urdu	فالورز کا اضافہ کریں لسٹ میں شامل ہونے کے لیے آئی ڈی مینشن کریں 300 ریٹوئیٹ کریں 🔄 ریٹوئیٹ والوں کو فالو کریں اور فالو بیک دیں 🌹 👈 مارخورز Follow	Neutral
Neutral	Urdu	نہیں 😂	Neutral
Neutral	Urdu	حنا پرویز بٹ	Neutral
Neutral	Urdu	:حضرت امام حسین نے فرمایا	Neutral
Neutral	Urdu	بھگانے اس کی 🤣😁😜😂	Negative - Ignore
Neutral	Urdu	میں نے تو کبھی اسکو سنا ہی نہی	Negative - Ignore
Neutral	Urdu	السلام علیکم 🙋‍♂️ اور تصدیق کرتا آیا ہوں اپنے سے پہلے کتاب توریت کی اور اس ...لیے کہ حلال کروں تمہارے لیے کچھ وہ چیزیں جو تم ح	Neutral
Neutral	Urdu	میرا کرونا ٹیسٹ پازیٹیو آگیا ھے 14 دن کی چھٹی چاہیے آئی جی بلوچستان 😁😉	Positive
Neutral	Urdu	کیسے نکھرا ہوا پھرتا ہوں مجھے دیکھ زرا اک نئے غم نے مجھ سے محبت کی ہے 🌹 ...عین ممکن ہے مری آنکھ سے ٹپکے وحشت آیت ہجر کی اس درجہ	Neutral
Neutral	Urdu	حوالات میں بھیج کر سارا محکمہ چھٹی پر بھیج دیا ہے۔ اب بس کریں۔ شغل میلہ کافی ہو گیا ہے ۔ ملک کے حالات پر رحم کریں۔ 🙏	Neutral
Neutral	Urdu	مسلم لیگ ن کے حمایت یافتہ عاصمہ جہانگیر گروپ کے شہزاد شوکت بڑی لیڈ کیساتھ صدر سپریم کورٹ منتخب۔	Neutral
Neutral	Urdu	بالکل۔ اندھیرا جلدی ہو جاتا ہے دل گھبرانے لگتا میرا اس سے 😞	Negative - Respond
Neutral	Urdu	میری سب حسرتوں میں اول ہے میرے ہاتھ میں تیرا ہاتھ ہونا محفل۔ سخن 👇	Neutral
Neutral	Urdu	نئیں نئیں چوہدری ساڈا ایڈے نکے دل دا مالک نئیں 😍	Positive
Neutral	Urdu	اُداش اُداش شا دل ہے آو۔۔۔ 😕😞	Neutral

Appendix C.1: Model predictions on Negative - Respond comments in Engish

Ground Truth	Language	Comments	Prediction
Negative - Respond	English	Product quality is cheap. There were scratches all over, most probably it was secondhand. The heating feature was also not working properly. I had to return the product	Negative - Respond
Negative - Respond	English	Getting in a car full of teenagers that would scare me don't think my ears would cope	Negative - Respond
Negative - Respond	English	Very disappointed from this purchase I ordered two dumbells(3kg and 5kg) and he sent me 4kg dumbell set.	Negative - Respond
Negative - Respond	English	Totaly fraud. No any fragrence looks like water added in bottle	Negative - Respond
Negative - Respond	English	thermostat is new but adopter is used adopter wire is damaged and used	Negative - Respond
Negative - Respond	English	Why am I getting rusted dumbbells?	Negative - Respond
Negative - Respond	English	Not my orders	Negative - Ignore
Negative - Respond	English	Mr Musk's daddy's wealth funded his empire, it also helps being a sociopath	Negative - Respond
Negative - Respond	English	I ordered small but got xl and I ordered pink and received black .. That was the mistake. Although the shirt quality was good	Negative - Respond
Negative - Respond	English	It didnt work it is broken i brought new batteries for the remote but it didnâ€™t work. I suggest not to buy this light	Negative - Respond
Negative - Respond	English	cheap quality ðŸ‘ŽðŸ»	Negative - Respond
Negative - Respond	English	Extreme bad quality. Extreme light material.	Negative - Respond
Negative - Respond	English	Third class product sent without checking	Negative - Respond
Negative - Respond	English	They have written full sleeves but they sent half sleeves, liars and the fabric is very cheap	Negative - Respond

Appendix C.2: Model predictions on Negative - Respond comments in Urdu (Roman Script)

Ground Truth	Language	Comments	Prediction
Negative - Respond	Roman Urdu	Bro besti apki hi hori h aesi video bnane pe	Positive
Negative - Respond	Roman Urdu	spelling tu theek kr lo ya phir seedha seedha #bhalu ko declare kar do king ye overacting ki kia zrorat hy	Negative - Respond
Negative - Respond	Roman Urdu	jo size mmagya tha wo aya nahi return karvana hay	Negative - Respond
Negative - Respond	Roman Urdu	nahi hota poineer mene 2,3 martaba try kia hai card milta hi nahi ya to request cancel ho jati hai	Negative - Respond
Negative - Respond	Roman Urdu	I never liked him..he was a munafiq since ever..	Negative - Respond
Negative - Respond	Roman Urdu	paise k hisab SE munasib hai yet the pictures say it all. stitching isn't good the shoe fabric is damaged. inside jogger patava Nahi hai which is quite annoying. baaki ap khud Dekh ly	Negative - Respond
Negative - Respond	Roman Urdu	Squats bhi nhi krni chaiye according to smart n fit and that coat n pent guy	Neutral
Negative - Respond	Roman Urdu	Ye banda Arjun Kapoor ka career khatam krdega .... agar acting me agaya to	Negative - Respond
Negative - Respond	Roman Urdu	Skelton wale ki leli	Negative - Ignore
Negative - Respond	Roman Urdu	larky apny apko khrab kar rhy hain ajkal.	Crisis
Negative - Respond	Roman Urdu	Toobah Astagfirullah	Crisis
Negative - Respond	Roman Urdu	Apki aik ankh badi or aik choty kiu h?	Neutral
Negative - Respond	Roman Urdu	kankaal di acting	Negative - Respond
Negative - Respond	Roman Urdu	Ya haal ha is duniyya k loggo ka	Negative - Ignore

Appendix C.3: Model predictions on Negative - Respond comments in Urdu (Arabic Script)

Ground Truth	Language	Comments	Prediction
Negative - Respond	Urdu	مس ہماری طرف بھی کچھ نظر گھمائی ہیں	Negative - Respond
Negative - Respond	Urdu	کسی کام کی نھیں ہے میں واپس کر رہا ہوں نہ ٹائیم سیٹ ہو رہا	Neutral
Negative - Respond	Urdu	بد قسمت مولوی	Negative - Respond
Negative - Respond	Urdu	ویری بد سٹف میں نے آرڈر کیا پک دیکھ کےپیاری لگے پک م لیکن جب رسیو کیا اور کھل کے دیکھا اندر لینڈی کی شرٹس تھی فضول ہ یا	Negative - Respond
Negative - Respond	Urdu	یہ چیز اٹھنے اچھی نہیں ہ جتنے پک میں تھی	Negative - Respond
Negative - Respond	Urdu	😡تم کیسا چاہتے تھے، ریپسٹ؟	Crisis
Negative - Respond	Urdu	موادکی کوالٹی بہت ہی خراب ہے، پسند نہیں آیا	Negative - Respond
Negative - Respond	Urdu	پروڈکٹ بہت جلد خراب ہوگیا، بہت بڑی گھٹیا مصنوعیت	Negative - Respond
Negative - Respond	Urdu	👎👎👎👎پھر دیکھنے کے لیے رہ۔کیا جانا ہے	Negative - Respond
Negative - Respond	Urdu	میں نے اِس نمبر کے لئے آرڈر دیا جس کو اور اس کو بہت کم معیار کا پتا چلا کہ صرف ایک بار پہننے کے بعد بھی یہ پھٹ جاتی ہے داراز کو دارازپی کے ذریعے نمبر ظاہر کرنے کے لئے ان کے معاہدے کو مسترد کرنا ہوگا	Negative - Respond
Negative - Respond	Urdu	پیکجنگ بہت ہی خراب تھی، سامان میں نقصان	Negative - Respond
Negative - Respond	Urdu	بل کمپنی کا پروڈکٹ بنا ہوا تھا، مواد بہت کمزور تھا	Negative - Respond
Negative - Respond	Urdu	میرا تجربہ اس خریداری سے صرف آدھا اچھا تھا کیونکہ ایک ٹی شرٹ معیار اور فٹنگ میں بالکل ٹھیک تھی لیکن ایک معیار میں اتنا اچھا نہیں تھا کیونکہ اس کا کالر معیار میں بالکل کچا اور وقتہ پور ناقص تھا اور اسے پہننے میں آرام دہ اور پر سکون تجربہ فراہم نہیں کیا گیا تھا تھوڑا مایوس کن تھا	Negative - Respond
Negative - Respond	Urdu	بہت بری حالت میں ہے اور کوالٹی بھی اچھی نہیں ہے	Negative - Respond

Appendix D.1: Model predictions on Negative - Ignore comments in Engish

Ground Truth	Language	Comments	Prediction
Negative - Ignore	English	What will he do .. Give them diesel	Negative - Ignore
Negative - Ignore	English	Wonder if he drives it as aggressive as his mini Cooper	Neutral
Negative - Ignore	English	plz restore my lifeðŸ™,ðŸ’€	Crisis
Negative - Ignore	English	Yeah. Heâ€™s Been laughing his way to the bank since the 80s!!	Negative - Ignore
Negative - Ignore	English	Bear ðŸ˜,ðŸ˜,	Neutral
Negative - Ignore	English	True,â€¦â€¦stay away from the TV!	Negative - Ignore
Negative - Ignore	English	And now youâ€™re on your fourth marriage.	Negative - Remove
Negative - Ignore	English	Lol, Heli wouldnt work at mars..	Neutral
Negative - Ignore	English	Horse	Negative - Ignore
Negative - Ignore	English	Otherworldly it's ours!! Ffs defund the media.	Negative - Ignore
Negative - Ignore	English	And.. Crushed but is a great piloto	Negative - Ignore
Negative - Ignore	English	Good to know we can build houses from all the scrap that crashed	Neutral
Negative - Ignore	English	Thus the imberial calcutaletion to millimeter flaw	Negative - Ignore
Negative - Ignore	English	vs what? Hamas stoning women?	Negative - Ignore
Negative - Ignore	English	He minding how not to crash that car again	Neutral

Appendix D.2: Model predictions on Negative - Ignore comments in Urdu (Roman Script)

Ground Truth	Language	Comments	Prediction
Negative - Ignore	Roman Urdu	ap may bhut mil rahi ha ðŸ˜,ðŸŒš	Negative - Ignore
Negative - Ignore	Roman Urdu	ab ye dubara poora christian bhi ni ban skta kyuki kuch aur aadha ho chuka hai Ã¯Â¿Â½	Neutral
Negative - Ignore	Roman Urdu	kon kon mathira ke kapre notice kr rha he. comment like karo	Negative - Ignore
Negative - Ignore	Roman Urdu	Behayai ka word esko dekh kar ijaad hwa Hoga	Negative - Ignore
Negative - Ignore	Roman Urdu	kha Daman Yae zan mo sharmolay day boda	Negative - Ignore
Negative - Ignore	Roman Urdu	Out of religion just behudgi hy nothing else yukh	Negative - Ignore
Negative - Ignore	Roman Urdu	phir uska koi kaam bhi nhi karti phir woh mujhy laa k deta hai	Negative - Ignore
Negative - Ignore	Roman Urdu	haha nahi bolongi chal jhootay koi ho to sahi na aesa yakeen thori der baad hota hai na	Negative - Ignore
Negative - Ignore	Roman Urdu	ðŸ˜,char gya hai	Negative - Ignore
Negative - Ignore	Roman Urdu	ma to comment par ky maza lyny ai hoðŸ˜	Negative - Ignore
Negative - Ignore	Roman Urdu	lekin ye ahmaddiya hai	Negative - Ignore
Negative - Ignore	Roman Urdu	ajeb admi mjhy bht bura lgta hy ya	Negative - Ignore
Negative - Ignore	Roman Urdu	Tobba aap apni umar ke lehaz se aesy kpry mt pehn kry shrm nai ati you r muslim Allah se daro plzzzzz	Negative - Ignore
Negative - Ignore	Roman Urdu	ab to yeh kapry Utr do ðŸ˜	Neutral
Negative - Ignore	Roman Urdu	sister please mera account mintion kr dainðŸ™ðŸ™	Positive

Appendix D.3: Model predictions on Negative - Ignore comments in Urdu (Arabic Script)

Ground Truth	Language	Comments	Prediction
Negative - Ignore	Urdu	عوام خبردار رہے کرائے کے لوگ لگا دی گئی ہے تاکہ دھاندلی ہوتوں ذہنی طور پر تیار ہو	Negative - Ignore
Negative - Ignore	Urdu	😢 ہمارا اسلامی جمہوریہ پاکستان جو اسلام کے نام پر بنا۔۔	Negative - Respond
Negative - Ignore	Urdu	جی جی پچھلے دنوں ایک چور ڈاکو ظالم قاتل فراڈیئے پچاس روپے کے اسٹام پیپر پہ بھاگے ہوئے سپریم کورٹ سے سزا یافتہ مافیا کے ڈان کو عدالت نے انصاف دیا ۔وہ بھی چٹ منگنی پٹ ویا	Negative - Ignore
Negative - Ignore	Urdu	ان کے دل میں صرف بدنیتی ہوتی ہے۔	Negative - Ignore
Negative - Ignore	Urdu	کیوں ھم سے خفا ھو گئے اے جانِ تمنا آئی فون کا گفٹ مجھے کیوں نہیں دیتے	Negative - Ignore
Negative - Ignore	Urdu	ان کے لئے کچھ بھی ممکن نہیں ہے۔	Negative - Ignore
Negative - Ignore	Urdu	ہمیشہ بدکاری کی راہ پر چلتے ہیں۔	Crisis
Negative - Ignore	Urdu	اپ پہلے ہی watch یہ ویڈیو اپ مل کر پروگرام بنا کے بناتے ہیں ۔ مجھے پتا ہے اپنی زیب میں رکھلی تھی کہ مج ھ سے اس نے باہر آ کے مانگنی	Negative - Ignore
Negative - Ignore	Urdu	کسی بھی امید کی بات نہیں	Negative - Ignore
Negative - Ignore	Urdu	یوسف فین حرمہ حقک یا واد البنت تشوف کلہ شوف انت مافي شوف	Negative - Ignore
Negative - Ignore	Urdu	برگر زبرائی ہیں اور ان لوگوں کو بھی جو انہیں کھاتے ہیں۔	Negative - Ignore
Negative - Ignore	Urdu	رونگ نمبر	Negative - Ignore
Negative - Ignore	Urdu	ان کے اعتباروں کا کوئی بھی اہمیت نہیں ہوتی۔	Negative - Ignore

Appendix E.1: Model predictions on Negative - Remove comments in Engish

Ground Truth	Language	Comments	Prediction
Negative - Remove	English	Roses are red and violets are blue and niggers are black, you know that's true.	Negative - Remove
Negative - Remove	English	yes fcking ur brown pussy will definataly tutn me on	Negative - Remove
Negative - Remove	English	Hot milf	Negative - Remove
Negative - Remove	English	Bean banged more times then a brass nice car tho	Negative - Remove
Negative - Remove	English	Can pornhub just get a gaming stream feature so these dumb bitches can gtfo get on it Katie<3	Negative - Remove
Negative - Remove	English	bitch plz whatever	Negative - Remove
Negative - Remove	English	i txt my old bitch my new bitch pussy wetter	Negative - Remove
Negative - Remove	English	Its ok I fapped when I got up	Negative - Remove
Negative - Remove	English	Murda Gang bitch its Gang Land	Negative - Remove
Negative - Remove	English	Bitches always show there booty	Negative - Remove
Negative - Remove	English	Baby I like being on your left side no bitch you can't be there Cuz the street on that side and Twitter said I ain't real if you do	Negative - Remove
Negative - Remove	English	Woah. You have the figure of a goddess. Sex goddess. Yay a 1000 times.	Negative - Remove
Negative - Remove	English	#oomf is soooo cute. But he probably has hoes on him a lot.	Negative - Ignore
Negative - Remove	English	Ain't that about a bitch... What does that mean?	Negative - Ignore
Negative - Remove	English	it aint nothing to cut a bitch off	Negative - Remove
Negative - Remove	English	I had a bitch that had a bitch, had the bitch eating halibut.. - #FreshRhymes #JerkSquad #SlowStir	Negative - Remove
Negative - Remove	English	Lovely brown boobies	Negative - Remove

Appendix E.2: Model predictions on Negative - Remove comments in Urdu (Roman Script)

Ground Truth	Language	Comments	Prediction
Negative - Remove	Roman Urdu	hath se moongfali tod lo pehle bhai phir gaand se akhrot todna	Negative - Remove
Negative - Remove	Roman Urdu	to the people jo pizza burger ki pictures lagatay hain abay tou bhaee khud khao humaray mood ki maa ko q de raha ho dua hai unhain tattiyan lagain bhenchod ://	Negative - Remove
Negative - Remove	Roman Urdu	madharchod badavo ko hm pta nahi btate bahanchod	Negative - Remove
Negative - Remove	Roman Urdu	Lanat on Agha Khan	Negative - Remove
Negative - Remove	Roman Urdu	sahi harami ha ye bhi	Negative - Remove
Negative - Remove	Roman Urdu	is randi ki gandi main aag laga di ha	Negative - Remove
Negative - Remove	Roman Urdu	aik kali se gand bhi ha..dekhay ga	Negative - Remove
Negative - Remove	Roman Urdu	Lanat ho tujh py tu hy harami	Negative - Remove
Negative - Remove	Roman Urdu	Ashbal gandu scene ni kro srf	Negative - Remove
Negative - Remove	Roman Urdu	Nai bahi aik din Jane me Gand phat gai thi	Negative - Remove
Negative - Remove	Roman Urdu	bkwas chez hai	Negative - Remove
Negative - Remove	Roman Urdu	see his eyes poora charsi hai ye *****	Negative - Remove
Negative - Remove	Roman Urdu	Bharwe tjhse puch bhosra band kr apna	Negative - Remove
Negative - Remove	Roman Urdu	is ka matlab hai is bc ko bewi bhi nahi pasand karti hai	Negative - Remove
Negative - Remove	Roman Urdu	hijra Ã¯Â¿Â½Ã¯Â¿Â½ suar ki aulad Ã¯Â¿Â½Ã¯Â¿Â½Ã¯Â¿Â½Ã¯Â¿Â½	Negative - Remove

Appendix E.3: Model predictions on Negative - Remove comments in Urdu (Arabic Script)

Ground Truth	Language	Comments	Prediction
Negative - Remove	Urdu	تمہاری گانڈ میں کیڑا کاٹتا ہے کیا جو کل رات کو تم نے پھر ویڈیو بنائی	Negative - Remove
Negative - Remove	Urdu	سنی کی گانڈ میں کھوتے کا لن	Negative - Remove
Negative - Remove	Urdu	دکی بھائی ایسے جھوٹے لوگوں کی وجہ سے پریشان مت ہوں آپ گانڈ ماریں تے ان ساروں کی ہم	Negative - Ignore
Negative - Remove	Urdu	چوتنے بھی لے لے اچھی لگے گی	Negative - Remove
Negative - Remove	Urdu	چوتیا گری جس کی روٹی اس کی بوٹی ذکاہ کی کاٹ دو گوٹی	Negative - Remove
Negative - Remove	Urdu	وہ دونوں بہت مادرچود ہیں ان کی ماں اور بہن کو دیکھ کر کچھ کچھ ہوتا ہے تم بھی پہلے دیکھ لو پھر ان پر ویڈیو بناؤ	Negative - Remove
Negative - Remove	Urdu	چوتیا پے کی ویڈیو بنائی اس مرتبہ اب لیول وہ نہیں رہا پہلے والا	Negative - Remove
Negative - Remove	Urdu	بھائی کیوں آپ اس چوتیا کے لیے اپنی ویڈیو کی کوالٹی خراب کر رہے ہو	Negative - Remove
Negative - Remove	Urdu	بھائی تو نے تو آخیر کر دی ویسے یہ بندہ ہے ہی چوتیا	Negative - Remove
Negative - Remove	Urdu	بھائی انتا غصہ ہاہاہا گانڈ ماردی شام کی	Negative - Remove
Negative - Remove	Urdu	بھائی میں انڈیا سے ہوں اور مجھے اس سے کوئی فرق نہں پڑتا کہ آپ دونوں کا آپس میں مسئلہ ہے پر چوتیا چوتیا ہی رہے گا	Negative - Remove
Negative - Remove	Urdu	اس کی ماں کا لوڑا	Negative - Remove
Negative - Remove	Urdu	مولوی وڑ گیا	Negative - Remove
Negative - Remove	Urdu	صیح سچ میں ایک نمبر کا چوتیا ہے اس کا بائیکاٹ کرو	Negative - Remove
Negative - Remove	Urdu	ابے جھوٹے شام کے ٹٹے دفعہ ہو غرق ہو شام کی ماں کی چوت	Negative - Remove

Appendix F.1: Model predictions on Crisis comments in Engish

Ground Truth	Language	Comments	Prediction
Crisis	English	You receive a letter threatening physical harm to your staff unless you pay a hefty sum. What should you do?	Crisis
Crisis	English	Think you've got a chance against us? We'll laugh as you fail miserably.	Crisis
Crisis	English	As long as this object remains intact, nothing can harm us. But if it breaks...well, let's just hope it doesn't break.	Crisis
Crisis	English	For centuries, the secrets of this place lay hidden. But now, they've been unearthed. Are you ready?	Crisis
Crisis	English	Rest In Peace	Crisis
Crisis	English	Someone's got to teach you a lesson. And unfortunately, that someone is us.	Crisis
Crisis	English	I'll make you pay for all the wrongs you've committed against innocent men.	Crisis
Crisis	English	From this point forward, your future holds nothing but misery.	Crisis
Crisis	English	Our fury knows no limits. Get ready to feel our pain.	Crisis
Crisis	English	Consider yourself marked for death. There's nowhere left to hide.	Crisis
Crisis	English	Cross us and you'll pay dearly. Our wrath knows no bounds.	Crisis
Crisis	English	Hackers aren't the only ones targeting your systems. Watch out for ghosts in the machine.	Crisis
Crisis	English	We'll leave a trail of destruction in our wake as we hunt you down.	Crisis
Crisis	English	if you aint bout that Murder Game pussy nigga shut up	Crisis
Crisis	English	It's too late to repent. Our justice is swift and brutal.	Crisis

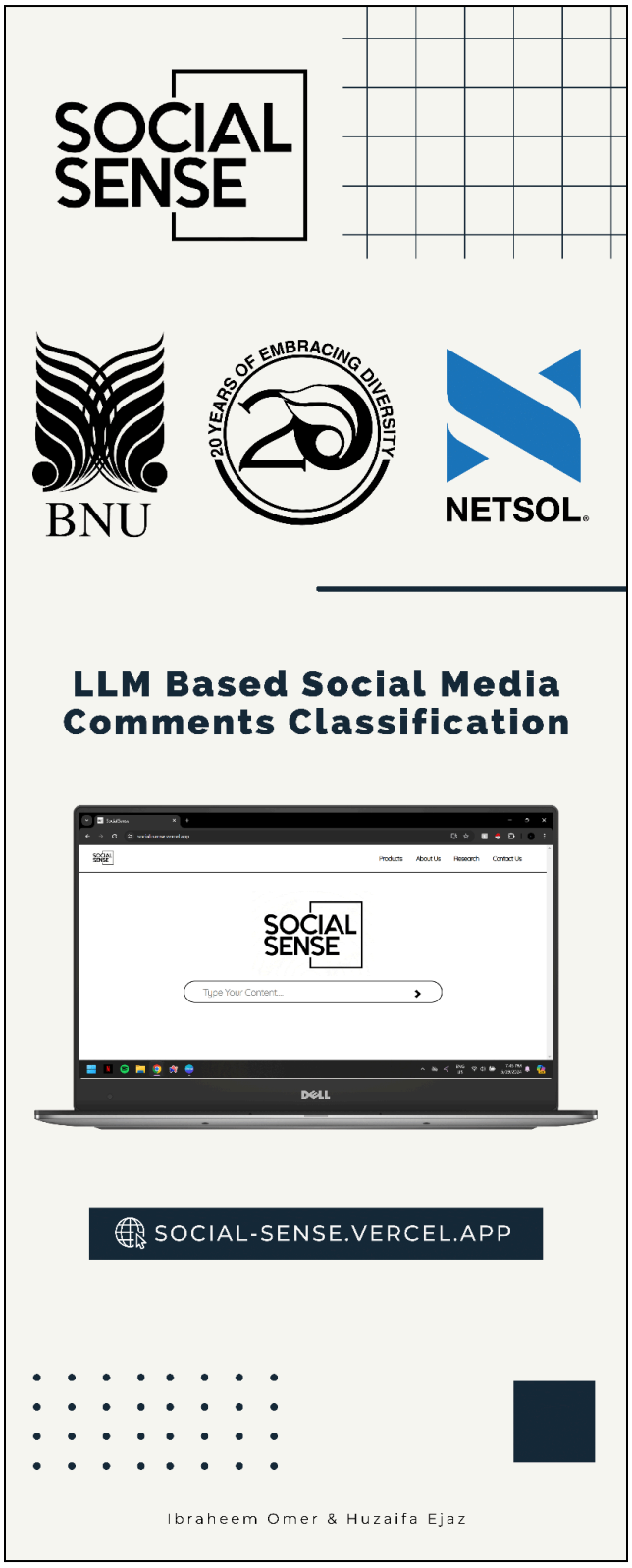
Appendix F.2: Model predictions on Crisis comments in Urdu (Roman Script)

Ground Truth	Language	Comments	Prediction
Crisis	Roman Urdu	Is haramzadon ko insaaf ki gehraiyon tak pohnchao, unka khoon kholta hai.	Crisis
Crisis	Roman Urdu	In haramkhoro ko dhamkian do, taake unko insaaf ki gehraiyon tak pohnchne ki koshish na karein.	Crisis
Crisis	Roman Urdu	In fitnon ko saza dilane ka waqt aa gaya hai, taake sab ko sachai ka pata chal jaye.	Crisis
Crisis	Roman Urdu	agaar pakistan hargiayto to may phora india ko aag lagadhonga	Crisis
Crisis	Roman Urdu	In galiyon ke keede ko kuchlo, taake sab ko inki misaal yaad rahe.	Crisis
Crisis	Roman Urdu	Poori society block karo or ander he subkoo sarnee do	Crisis
Crisis	Roman Urdu	In fitnon ki gandi raftar ko online forums par expose karo, taake unki asliyat sabke saamne aaye.	Crisis
Crisis	Roman Urdu	#khuda ki qasam agar mey hota tu zinda nahe bach sakta ta muj sey ye beghairat kafir yahoodi	Crisis
Crisis	Roman Urdu	gand toor donga tumhari	Negative - Remove
Crisis	Roman Urdu	In zalimon ki chhavi ko media ke zariye tarnish karo, taake unko koi maqam na bache.	Crisis
Crisis	Roman Urdu	goli mar den kameno ko in kry dramey khatam nahi hotey	Crisis
Crisis	Roman Urdu	Wajbul QATAL hain yahhhh ab sabbbb molviiiii	Neutral
Crisis	Roman Urdu	In haramkhoro ko publically shame karo, taake unko insaaf ki saza mil sake.	Crisis
Crisis	Roman Urdu	Apki ki bund me bomb lage	Crisis

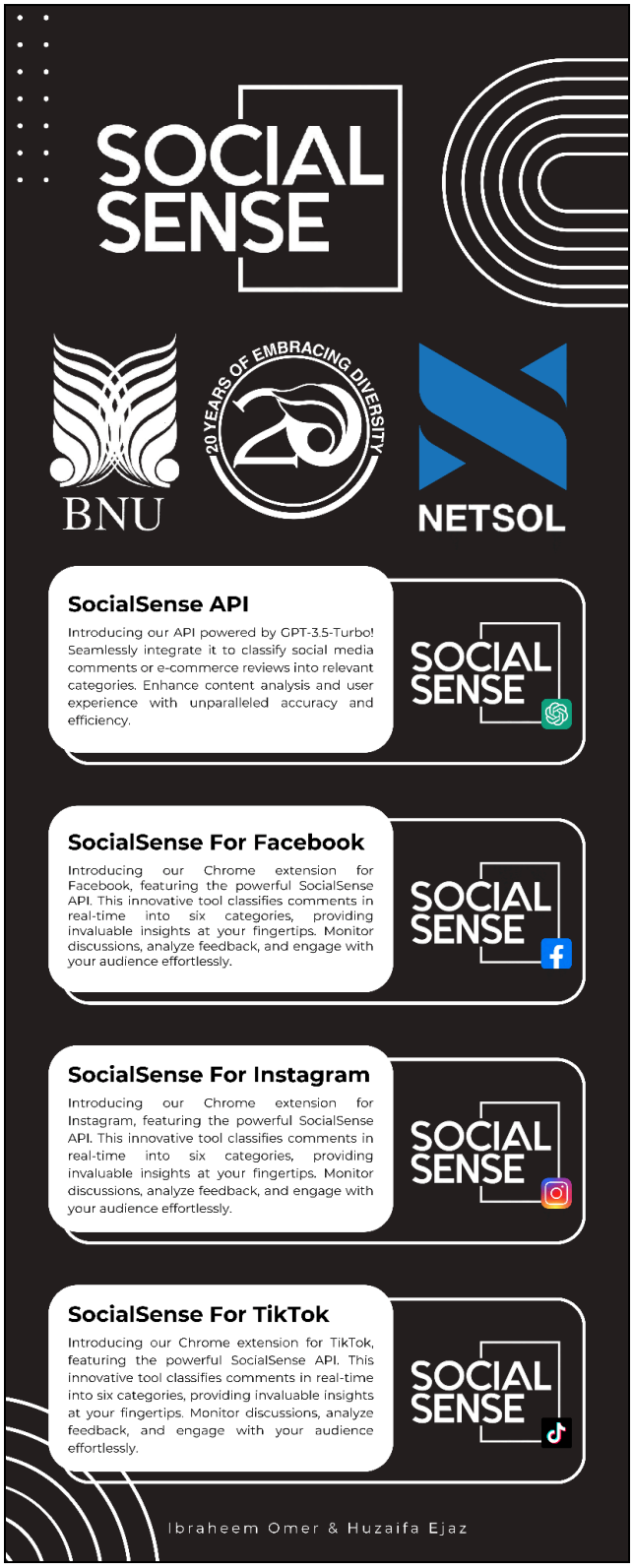
Appendix F.3: Model predictions on Crisis comments in Urdu (Arabic Script)

Ground Truth	Language	Comments	Prediction
Crisis	Urdu	آپ کو میرے راستے میں آنے کے بعد آپ کو اپنی آنکھوں پر بھرم سے بھی اعتماد نہیں رہے گا۔	Crisis
Crisis	Urdu	آپ میرے خلاف کچھ بھی نہیں کر سکتے۔	Crisis
Crisis	Urdu	!آپ کو میری دہائی کو محسوس کرنا ہوگا اگر آپ میری راہ میں آئیں	Crisis
Crisis	Urdu	آپ کو میرے غصے کی انتہائی سزا دینی ہوگی اگر آپ میری بھراسا طور پر کھیلیں۔	Crisis
Crisis	Urdu	میں آپ کو ایک ایسا دنیا کا سب سے بڑا غم دوں گا کہ آپ کو ہمیشہ کے لئے بھولا دوں گا۔	Crisis
Crisis	Urdu	!میں آپ کی داستان بٹا دوں گا اگر آپ میرے ساتھ کچھ کریں گے	Crisis
Crisis	Urdu	میں آپ کے ساتھ یہ کہتا ہوں کہ آپ مانتے رہو یا غلط وقت پر پتہ چل جائے گا۔	Crisis
Crisis	Urdu	آپ میرے سامنے آنا نہیں چاہتے تو میں آپ کو مجبور کر دوں گا۔	Crisis
Crisis	Urdu	کنارے پر تیرنے والی لاش کو دیکھ کر یہ سمجھ میں آیا ۔۔۔ بوجھ جسم کا نہیں 🌹 🙃سانسوں کا تھا	Crisis
Crisis	Urdu	میں آپ کو ایک ایسا درد دوں گا جسے آپ کبھی بھول نہیں سکیں گے۔	Crisis
Crisis	Urdu	آپ کو میری طاقت کا اندازہ ہی نہیں، اور میں آپ کو خاک میں ملا دوں گا۔	Crisis
Crisis	Urdu	آپ کو میری قوت کا اندازہ ہی نہیں، اور میں آپ کو خاک میں ملا دوں گا۔	Crisis
Crisis	Urdu	آپ میری زندگی میں دخل نہیں دے سکتے۔	Crisis
Crisis	Urdu	!آپ کو میرے غصے کا بچنے کا کوئی راستہ نہیں ہے، آپ کو ضرور معاوضہ دینا ہوگا	Crisis
Crisis	Urdu	میں آپ کو بھوکا رکھ دوں گا اگر آپ میری راہ میں آئیں، اور پھر آپ کو پیاسا بھی رکھ دوں گا۔	Crisis

Appendix G.1: SocialSense Promotional Flier



Appendix G.3: SocialSense Produts Promotional Flier



LLM Based Social Media Comment Classification

Ibraheem Omer [1] Huzaifa Ejaz [2]

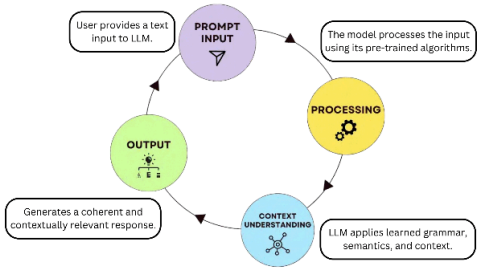
Motivation

- Zhan et al. highlighted the transformative impact of LLMs on sentiment analysis. [7]
- Cahe and Davison showed LLMs' superiority in complex text classification over traditional methods. [4]
- Krungmann and Hartman found LLMs can outperform traditional methods in sentiment classification accuracy. [6]

Existing Works & Limitations

- Instagram filters offensive comments and messages for a safer, personalized experience. [1]
- YouTube lets creators allow, review, approve, or disable comments for better control. [3]
- CellarWeb for WordPress enables quick front-end comment actions like 'Spam' and 'Delete'. [5]
- The Perspective API scores text for attributes like Toxicity, aiding in content moderation. [2]

LLM Process Model



Solution

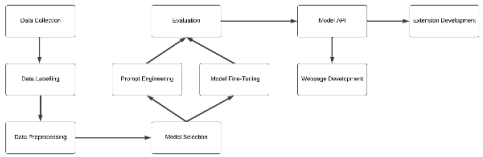
Developed an AI model for comment analysis utilizing Large Language Model (LLM). The model was fine-tuned on a diverse dataset containing English and Urdu (Roman Script & Arabic Script) content, catering to multiple languages while being specifically fine-tuned for English and Urdu.

The model categorizes the comments into one of the following six categories

- Positive
- Neutral
- Crisis
- Negative - Remove
- Negative - Ignore
- Negative - Respond

Methodology

- **Data Collection:** Eighty thousand comments were collected from social media platforms.
- **Data Labelling:** Fifty thousand comments were categorized into six distinct categories.
- **Data Preprocessing:** Each comment was converted to lowercase, and special characters were removed.
- **Model Selection:** Experimented with various LLM's.
- **Prompt Engineering:** The selected model was provided with input prompts.
- **Model Fine Tuning:** GPT 3.5 Turbo 0125 was fine-tuned on the labelled data.
- **Evaluation:** The model was evaluated using various metrics.
- **Model API:** Our model was deployed using the OpenAI playground.
- **Web-page Development:** A web page was developed for users to test our model.
- **Extension Development:** Three Google Chrome extensions were developed, including those for Facebook, Instagram, and TikTok.



Results

Label Language	English	Urdu (Roman)	Urdu (Arabic)	Class Accuracy
Positive	88%	88%	90%	89%
Neutral	86%	62%	54%	67%
Negative - Respond	92%	58%	86%	79%
Negative - Ignore	60%	88%	86%	78%
Negative - Remove	90%	94%	100%	93%
Crisis	100%	94%	100%	98%
Language Accuracy	**86%**	**81%**	**85%**	N/A

Table 1. Model's Accuracy On Each Label

Technologies

References

[1] Instagram help center: How do I filter out offensive comments on my Instagram posts?
[2] Perspective api. https://www.perspectiveapi.com/.
[3] YouTube help: Manage inappropriate comments on youtube.
[4] Youngjin Chae and Thomas Davidson. Large language models for text classification: From zero-shot learning to fine-tuning. Aug 2023.
[5] Rick Hellewell. Block comment spam bots.
[6] Jan Ole Krugmann and Jochen Hartmann. Sentiment analysis in the age of generative ai. *Customer Needs and Solutions*, 11(1):3, 2024.
[7] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. Sentiment analysis in the era of large language models: A reality check. 2023.