

# **SOCIAL MEDIA TEXT ANALYSIS USING REDDIT API**

*Dissertation submitted in fulfilment of the requirements for the Degree  
of*

## **BACHELOR OF TECHNOLOGY**

**in**

## **COMPUTER SCIENCE AND ENGINEERING**

By

**Annepally Sanjay Kumar**  
**12017831**

Supervisor

**Ms. Abhinaya Anand**



**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab (India)

April, 2023

## **ABSTRACT**

---

**With the exponential growth of social media platforms , personalized content recommendation have become an important tools for enhancing user experience. This study introduces a approach leveraging emotion detection in conjunction with Reddit API for social media text analysis.**

**The proposed system integrates sentiment analysis and emoji emotion detection techniques to analyze technique to analyze textual content extracted from Reddit posts. By using Natural Language Processing(NLP) algorithms, the system identifies and categorizes the emotional tone conveyed with user- generated content across various subreddits.**

**Through Machine Learning(ML) techniques the system constructs user emotion profiles based on historical interactions and engagement patterns shown by Reddit users. These profiles are then matched to the input we have given with the Reddit ecosystem , helping us for personalized content recommendations tailored to users emotional preferences and states.**

**Utilizing Reddit API, by doing web scrapping technique we gathered dataset comprising of textual posts, user interactions , and community dynamics. The evaluation of system's performance encompasses quantitative metrics such as accuracy, precision, and recall along with qualitative assessments of users satisfaction and emotional resonance with recommended content. User feedback surveys and sentiment analysis of user interactions further validate the effectiveness of the emotion-driven recommendation approach.**

**The results shows the efficacy of the proposed system in accurately detecting and leveraging user emotions to deliver personalized content experiences. High precision and recall scores attest to the system's ability to recommend emotionally aligned content with users' mind states and preferences.**

**In conclusion, integrating emotion detection with social media text analysis through Reddit API offers the best approach to content recommendation, enhancing user engagement and satisfaction. This research contributes to the advancement of emotion trained recommendation systems in social media platforms , with implications for fostering meaningful connections and interactions within online communities.**

**Keywords : social media, Reddit API, emotion detection, sentiment analysis, personalized recommendations, user engagement.**

## **DECLARATION STATEMENT**

---

We hereby declare that the research work reported in the dissertation/dissertation proposal entitled "SOCIAL MEDIA TEXT ANALYSIS" in partial fulfilment of the requirement for the award of Degree for Bachelor of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Ms. Abhinaya Anand. We have not submitted this work elsewhere for any degree or diploma.

We understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of our knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by us. We are fully responsible for the contents of our dissertation work.

## SUPERVISOR'S CERTIFICATE

---

This is to certify that the work reported in the B.Tech Dissertation/dissertation proposal entitled "**SOCIAL MEDIA SENTIMENT ANALYSIS**, submitted by **Annepally Sanjay Kumar** at **Lovely Professional University, Phagwara, India** is a bonafide record of his / her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

(Name of Supervisor)

**Date:**

**Counter Signed by:**

**1) Concerned HOD:**

HoD's

Signature:

\_\_\_\_\_

\_\_\_\_\_

HoD

Name:

\_\_\_\_\_

\_\_\_\_\_

Date: \_\_\_\_\_

**2)Neutral Examiners:**

**External Examiner**

Signature:

\_\_\_\_\_

Name:

\_\_\_\_\_

Affiliation:

\_\_\_\_\_

Date:

**Internal Examiner**

Signature:

\_\_\_\_\_

Name:

\_\_\_\_\_

Date:



## ACKNOWLEDGEMENT

---

We would like to express my sincere gratitude to all those who have helped us to the development of this social media text analysis using Reddit API. This project would not have been possible without the support and guidance of several individuals and organizations.

Firstly, We would like to thank our project supervisor Ms. Abhinaya Anand for providing us with valuable insights and guidance throughout the development of this system. Their continuous support and encouragement have been instrumental in the successful completion of this project. We would also like to extend our gratitude to our colleagues and classmates who have provided us with constructive feedback and suggestions. Their inputs have helped us to refine our approach and improve the performance of the recommender system.

We would like to express our appreciation to Reddit for providing access to do Web Scrapping to extract Dataset which was used in the development of this project.

We are also thankful to the developers of the various open-source libraries and frameworks used in the implementation of this project. These include Python, Pandas, NumPy, Scikit -learn, nltk and Streamlit, among others. This technology has played a significant role in the accuracy and effectiveness of the content-based filtering algorithm used in the system.

Finally, We would like to acknowledge the contributions of the Reddit and Reddit users, which provided the data used in the development of the recommender system. We are grateful for their efforts in maintaining and updating the database, which has been an invaluable resource for this project.

In conclusion, We would like to express our heartfelt appreciation to everyone who has played a role in the development of this Social media text Analysis system. Your support and contributions have been invaluable, and We are grateful for the opportunity to have worked on this project.

## TABLE OF TOPICS

CONTENTS	PAGE NO.
Cover Page	1
PAC form	2
Abstract	2
Declaration	3
Supervisor's Certificate	4
Acknowledgement	5
Table of Contents	7

# TABLE OF CONTENTS

---

<b>CHAPTER1: INTRODUCTION</b>	<b>9</b>
<b>1.1 PROBLEM STATEMENT</b>	<b>11</b>
<b>1.2 GOAL'S</b>	<b>11</b>
<b>1.3 METHODOLOGY</b>	<b>13</b>
<b>1.4 ETHICS</b>	<b>14</b>
<b>1.5 DELIMITATION</b>	<b>14</b>
<b>1.6 OUTLINE</b>	<b>14</b>
<b>CHAPTER2: OVERVIEW OF RELATED WORK</b>	<b>15</b>
<b>2.1 SOCIAL MEDIA SENTIMENT ANALYSIS</b>	<b>17</b>
<b>2.2 NLP</b>	<b>18</b>
<b>2.3 LEXICO-BASED APPROCHES</b>	<b>19</b>
<b>2.4 ML MODELS</b>	<b>20</b>
<b>2.5 PRE-PROCESSING TECHNIQUES</b>	<b>22</b>
<b>FOR SOCIAL MEDIA TEXT</b>	
<b>2.6 ETHICAL CONSIDERATION</b>	<b>23</b>
<b>2.7 STATE-OF-THE-ART</b>	
<b>IN SENTIMENT ANALYSIS</b>	<b>24</b>
<b>CHAPTER3: AUTO CORRECTION &amp; AUTO GENERATION</b>	
<b>3.1 AUTO CORRECTION</b>	<b>25</b>
<b>3.2 AUTO GENERATION</b>	<b>26</b>
<b>3.3 TOOLS AND LIBRARIES</b>	<b>27</b>
<b>CHPTER4: IMPLEMENTATION</b>	



4.1 DATASET	31
4.2 PRAW	32
4.3 NLTK	33
4.4 FEATURE EXTRACTION	34
4.5 TF-IDF	35
4.5.1 TERM FREQUENCY	35
4.5.2 INVERSE DOCUMENT FREQUENCY	35
4.5.3 WEAKNESS OF TF-IDF	36
4.5.4 IMPROVEMENT OF TF-IDF	37
4.6 VADER	37
4.7 VISUALIZATION	38
4.8 EMOJI'S	41
<b>CHAPTER5: DEPLOYMENT</b>	
5.1 DEPLOYEMENT 1:	
DEPLOYEMENT CONFIGURATION	43
5.2 MANAGING DETAILS	44
<b>CHAPTER6: CHALLENGES AND SOLUTIONS</b>	45
<b>CHAPTER7: RESULTS AND DISCUSSION</b>	46
<b>CHAPTER8: CONCLUSION</b>	47
<b>REFERENCES</b>	49
<b>PROJECT CODE</b>	50

# CHAPTER 1

## INTRODUCTION

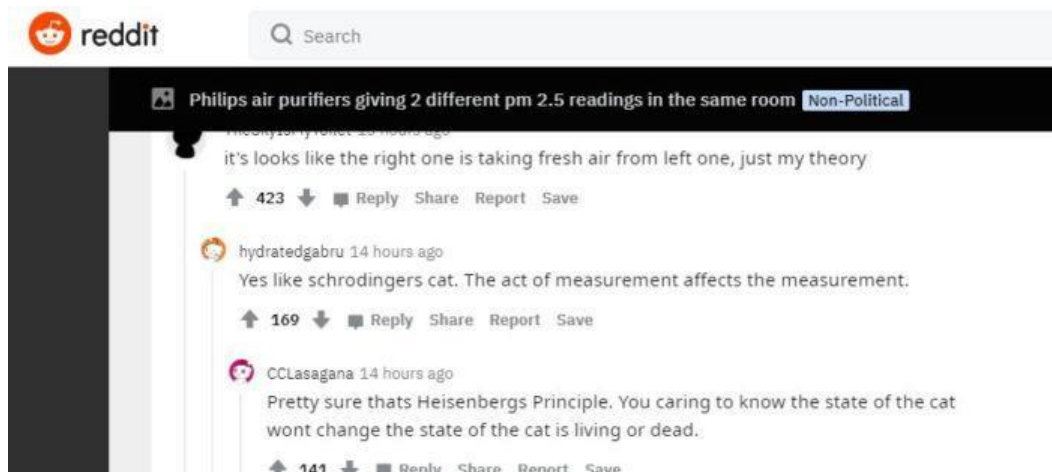
Nowadays social media has captured the zeitgeist of digital age, whereupon online tools such as Reddit have become the new locales for the common folk to voice their opinions, discuss common issues, and even influence the mass discourse. Considering that Reddit is full of users actively contributing into debates in its subreddits and it have got millions of users thus, Reddit stands for an incredible source of user-generated content which can be analysed. The consequences of looking and analysing the opinions and mentality of the message boards in Reddit, would give a forte significance to the business, researchers, policy makers, and alike groups. The focus of the project is on the sentimental sector of the Reddit sphere that represents a huge amount of data which is crying to be processed by the latest and most effective natural language processing mechanisms.

Sentiment analysis, a subset of natural language processing, concentrates on computational analysis of textual data to determine the feelings/attitude/opinion expressed through that text. The sentiments and attitudes are the obtained ones whenever the machine learning algorithms and sentiment analysis models are used to do the analysis of the reddit comments; thus, this project aims to find this out. Topics and opinion sources vary ranging from debates concerning current events to the reviews of products and cultural practices. Thus, this social forum provides the perfect platform for sentiment assessment with the aim of determining users' perception toward specific topics.



Sentiment analysis on Reddit involves more than finding out what people are talking about when it comes to practical implications for many people being mentioned as the stakeholders. Researchers of markets are both granting themselves access to consumers' opinions on products/brands, which leads to the development of promotional tactics that would range from direct advertising to public relations. (Similarly, decision makers also can get to know the reservoir opinion of public regarding the imperative issues which then may help to decide and implement the policies on the same.) Besides, through Pulse of Reddit polls, one would be able to comprehend what are on-the-rise trends, cultural changes and people's sentiments, which contribute to informing us of the digital climate and its impact on the society.

By the architecting a sentiment analysis system on the Stream lit platform, this project strives to pass the advantage of sentiment analysis tools as well as help the user interactively discover sentiments patterns derived from Reddit discussions apprehensions. As well, smartphones employ the latest technology like spell correction, as well as text creation that leads to the gadgets becomes user-friendly.



Furthermore, the integration of advanced features such as auto correction and text generation enhances the usability and functionality of the sentiment analysis system, providing users with a comprehensive toolkit for analyzing and understanding sentiment dynamics within the Reddit community. Overall, this project endeavors to illuminate the intricate tapestry of sentiments woven within Reddit discussions, offering valuable insights into the collective consciousness of one of the internet's most dynamic communities.

The objective of this investigate is to create a assumption examination framework for Reddit posts and remarks utilizing NLP strategies. The framework can be deliberate to analyze the opinion communicated in Reddit substance and supply bits of understanding into the in well known estimation patterns on exclusive points and groups. The proposed framework will use highlights consisting of content material substance, customer engagement measurements, and applicable information to survey the opinion extremity (positive, negative, or neutral) of every submit and comment.

To attain this, the framework will make use of a dataset of Reddit posts and feedback, counting metadata together with timestamps, subreddit facts, and client intelligent. Content pre-processing techniques will be applied to easy and normalize the literary statistics, counting tokenization, stemming, and evacuating stop words. Machine gaining knowledge of fashions, which includes bolster vector machines (SVMs) or repetitive neural structures (RNNs), might be prepared on labelled facts to categorize the assumption of each put up and comment.

## 1.1 PROBLEM STATEMENT

The aim of this proof of concept is to implement a text classifier for Reddit posts and comments using natural language processing techniques. The design is explicit to investigate the viewpoint informed by the Reddit content and bring insights into the well-known estimation or trends on particular topics and view. The proposed framework will lead with bullet points that involve content design elements, customer engagement parameter, and contextual information to examine the sentiment polarity (positive, negative, or neutral) of every post and comment.

To achieve this, the infrastructure will employ a dataset of Reddit posts and their comments with related information such as timestamps, subreddit data, and user- participant intelligence. The content pre-processing methods will be utilized to purify and normalize the statistical language, including tokenization, stemming, and stop words removal. Machine learning of fashions, shortened by SVMs or RNNs, can be pre-trained with the labeled data set to categorize the sentiment of the each post and comment.

Although social media like Reddit have become popular as well as areas where social media users are able to access user generated content and discuss, there still exists the challenge of effectively analyzing and understanding the sentiments expressed in these platforms. The massive number of textual data produced on Reddit offers opportunities for both deriving insights of public opinion, attitudes and sentiment towards a broad range of issues. Nevertheless, such a manual analysis is impossible to carry out because of the huge amount and complex structure it possesses. Hence, it is imperative to come up with automated sentimental analysis tools which can process Reddit posts and extract useful information in an optimal manner.

The project is focused on the problem of emotions included into Reddit discussions and aims to create a detailed sentiment analysis system. The system aims at the application of natural language processing with machine learning methods in order to realize sentiment from the comments on Reddit.

## 1.2 GOALS

The goal of this project is to develop a robust sentiment analysis system tailored for Reddit data, with the following objectives:

The goal of this project is to develop a robust sentiment analysis system tailored for Reddit data, with the following objectives:

Extract Meaningful Insights: Exploit approaches of natural language processing and calculating algorithms, so as to explore sentiments mentioned within Reddit comments and

extract the vital data about public opinion and attitudes to particular subjects, concrete products, and events.

**Enhance User Experience:** Launch the sentiment evaluation system application on Streamlit so as to provide thematic views for the users concerning sentiments for Reddit discussion threads. Integrating autocorrecting and text producing functions to improve user experience and involvement should be considered.

**Empower Stakeholders:** Offer such professionals as in marketing, brand analysis, and policy makers, including other interest groups, the detailed instrumentation for the purpose of decoding the appropriate sentiment trend and the public opinion rankings on the Reddit forum .

**Facilitate the stakeholders with the decision making competence they can achieve through the mapping system of sentiments, analysis.** **Contribute to Research:** Do not forget to add an original insight in the field of sentiment analysis and natural language processing by proposing new methods and techniques for the sentiment analysis process on the specific topic area of social media platforms like Reddit. Disseminate results, solutions and findings through the community to fuel more advancements and breakthroughs in the field.

The goal of this project is to develop a robust sentiment analysis system tailored for Reddit data, with the following objectives: The target of the project is the development of a strong sentiment analysis system, which is regulated especially for data of Reddit, and has the following aims: **Extract Meaningful Insights:** Under the use of natural language processing techniques, as well as calculations, construct apps that would find and extract necessary information related to public opinion, attitudes and feelings toward concrete subjects, products, and events.

**Enhance User Experience:** Start just decomposing the sentiment evaluation system through Streamlit to ensure users have the familiarity with sentiments regarding Reddit discussion boards. The audacity of implementing real-time text correcting and text input would create a positive user experience leading to user engagement.

**Empower Stakeholders:** Give the professionals in marketing, brand functionality, and policy makers with additional other interest groups, the device with detailed instruments for the purpose of the needle approaching, and public opinion rankings on the forum Reddit . Enable the stakeholders to be competent in making the decisions which they would get through the mapping system of sentiments, analysis which they can carry out.

**Contribute to Research:** Also, please bear in mind that not only mentioning a novel breakthrough in the area of sentiment analysis and natural language processing should provide a new idea and methodology for the sentiment analysis process during the research of the topic area of social media like Reddit. Distribute outcomes, answers and scientific discoveries among the local community to encourage more innovations and unravel more discoveries.

## 1.3 METHODOLOGY

- 1) Data Collection: Apply the Reddit API for collecting contextually-relevant comment from different subreddits that discuss a plethora of trending topics. Maintain data limits on the rate of calls as well as data privacy residual matters during the course of data collection.
- 2) Data Pre processing: Discard and purify the obtained data to get rid of the useless bulk of the data and formatting artifacts. Steps might be filtering out special characters, punctuation, the stop-words, and converting text to lowercase. Moreover, take away missing data and do tokenization operations to behave the data; otherwise, notification will be thrown for processing
- 3) Feature Extraction: Apply techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) to draw out variables from the pre processed text information. Wording TF-IDF gives a weight to words for their frequency within a document and across the whole corpus. This representing the importance of each word in showing the feeling of the text.
- 4) Sentiment Analysis Model: Design sentiment analysis model via the use of the VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon and rule-based algorithms. VADER shines in the task of finding the sentiment in social media texts like Reddit messages and is pre trained to assign individual words with corresponding precalculated scores.
- 5) Integration of Additional Features: We should develop a comprehensive model with additional features such as autocorrection and text generation to make our sentiment analysis system more powerful. Apply library functions or mechanisms for resolving spelling issues to acquire the desired accuracy outcome while sentiment analysis. Support the capability text generation via models such as ChatGPT, so that users will be able to use text generation by inputting the user.
- 6) Deployment on Streamlit: Let us roll out the sentiment analysis system using the Streamlit website for the provision of an interactive and web-based interface which people can use for accessing and exploring the trends in sentiment for the Reddit discussions. Develop a user-friendly interface for the user to input the text of their choice and analyze the sentiment. The user should later on have an option to interact with the additional features easily.
- 7) Evaluation and Validation: Evaluate the performance of the sentiment analysis system using appropriate metrics such as accuracy, precision, recall, and F1-score. Validate the results by comparing them with human-labeled data or existing sentiment analysis benchmarks. Iterate on the system's design and implementation based on feedback and validation results to improve its accuracy and effectiveness.

## **1.4 ETHICS**

The responsibility and ethicality of this project is to employ the data gathered from Reddit in a responsible and ethical manner as well as the sentiment analysis outcomes be applied in a way that can benefit society. The importance of respecting the privacy and the anonymity of Reddit users should be observed throughout data collection and analysis, with the help of their terms of service and data usage policies. Data collection and data analysis purpose must be open to users, and their consent must be followed where they are involved. Also, demonstration of the system should happen in a method that will ensure minimal possibility of bias and discrimination among the individuals represented on Reddit. At last, everything must be taken into account to diminish the possible harmful effects the release of sentiment analysis results may cause, especially in the sensitive topics, by underlining the results and not twisting or misinterpreting the data.

## **1.5 DELIMITATION**

The scope of the project is limited to the analysis of sentiment on Reddit conversations. This could therefore obviously limit the generalization of the results to other social platforms or different textual datasets. Besides, there are restrictions in our project by the generators of data collected from Reddit and the limitations of Reddit's API or terms of service. In addition, the sentiment analysis system which we have designed may inherit some biases or barriers associated with natural language processing and machine learning methods. Consequently, this might distort the sentiment analysis results towards inaccuracy and unreliability. Eventually, the project scope may impose restrictions on deepening the analysis of particular subreddits or topics, and thus, a broader exploration of emotion tendencies throughout the platform may also be needed.

## **1.6 OUTLINE**

The ethical aspects of such a project are related to the safe and ethical operation of data collection on Reddit and the analysis of the attained sentiment. Data protection and anonymity maintenance should be the utmost priority, adhering to the terms of service and data usage policy of Reddit. This should be done on all the steps from collecting the data until the analysis. The revelation of how the personal data will be used and analyzed is also very crucial.

A user consent must also be obtained when it is applicable. Moreover, the bias and discrimination mitigation strategies need to be integrated as the system design and

deployment continue to pattern the behavior of individuals into the data being represented by Reddit.

Furthermore, measures ought to be taken to uproot the adverse effects that may occur from the distribution of outcomes, mainly in topics that are sensitive or cases of conflicts, by starting up context for the findings and avoiding the misunderstanding or misinterpretation of the data. This R&D work intends at the designing of sentiment analysis system adapted to the Reddit data complying with data gathering sequence, pre-processing, model development process and then deployment on Streamlit platform.

The project has its start with accumulation of the Reddit comments data using the Reddit API and then follows on with pre-processing tasks which help in cleaning and refinement data for analysis. Eventually, a sentiment analysis model has been constructed, applying methods like TF-IDF and the VADER lexicon, and incorporated with other features, sensitive correction through the autocorrect model. Moreover, using text generators like ChatGPT, you can expand the application. Sentiment analysis system is hosted on Streamlit page, giving users accessible tools to analyze how sentiment is distributed within the discussions in Reddit. The app is evaluated and validated against the system's effectiveness parameters, then the approach and events are summarized in a comprehensive report.

## **CHAPTER 2**

### **OVERVIEW OF RELATED WORK**

This model's ethical component also entails the protection against unethical data collection from Reddit data that might have been used in the sentiment analysis. An absolute privacy and anonymity of users is to be respected as a must-condition for the whole data storage, collecting and analysis process of the Reddit social network. This involves obeying Reddit's Terms of Service as well as its data usage policies.

The going high on transparency with the users about the data collection and the analysis which they are doing must be done and in the case where people's consent is needed you should obtain it. Another important aspect is the design and deployment of the machine-learning algorithm in a way that gives a fair and equitable treatment to all predict and while also minimizing the possibilities of bias or discrimination that can exist.

Eventually, that people have to find ways that will diminish or eliminate unwanted harm that may result from the publishing and using of sentiment analysis findings, especially on topics which could be deemed to be sensitive and controversial, through contextualizing the findings and avoiding cases of miss representation and misinterpretation of the data.



The overview in this context of related work for the project includes studies and projects that have employed sentimental analysis in Reddit posts, and these studies are in connection with those that involve the area of natural language processing (NLP), and these also incorporate social media analytics. Previous applications in the sentiment analysis were carried out by the implementation of higher level of machine learning algorithms.

For example, VADER and the support vector machines (SVM) are lexicon-based and supervised learning approaches respectively. Moreover, the deep learning methods such as recurrent neural networks (RNNs) were applied. As well, studies have been carried out of pre-processing techniques for social media text data that include the process tidying up, tokenization, and stemming or lemmatization. Also research has slanted towards the ethics and challenges that affect the sentiment analysis in social media through bias, privacy, and policy of data usage. Readings and analysing the methods of the sentiment analysis and social media data uncover a variety of insights and help define how the sentiment analysis system would be developed and applied to the Reddit data in the project.



In the context of this Streamlit project, the four categories serve distinct purposes: In the context of this Streamlit project, the four categories serve distinct purposes:

- **Reddit Data Fetcher:** The following section falls into the category of operation as regards to the textbook explanation delivering the process of data acquisition from Reddit via its suitable API. User might insert preferred parameters to the application e.g. subreddit name, the number of posts made/comments, or keywords that will be used to get information. This feature guarantees that the sentiment analyzer machine becomes extensible for the purpose of channeling various conversations from the site.

- **Auto Correction:** The auto-correction goes category deals with aspects of the programming that help the correction of typing errors and improvement of text flow. The autocorrect model is capable of self-correcting the spelling errors and typos during the input text therefore, it strengthens the quality of the input data which is to be fed for sentiment analysis. The cleanliness of the text data is one of the properties that is inherent in their system, which can operate on the uniform and standardized text data.
- **Sentiment Analysis:** This cluster aims at the assessment of statements within the smaller language units found in the data collected from Reddit. Sentiment analysis model detects the sentiment polarity within each comment or post that is then identified as positive, negative, or neutral. By means of this tool, consumers can follow user sentiments on separate themes or groups of discussions in Reddit, such that they are kept abreast of all the sentiments in question and shifts therein.
- **Auto Text Generation:** The automated text generator section is centered around generating texts based on user's request or prompts by utilizing models such as Gemini API. We provide the feature that enables the end users to generate the text in a manner which is sounding conversational or in context. This gives the users the opportunity to interactively explore an application on Streamlit. Furthermore, auto text generation can be used for diverse tasks, for example, it is possible to automatically fill a text in response to a prompt the user announces or complete incomplete sentences.

## **2.1 SOCIAL MEDIA SENTIMENT ANALYSIS**

Social media sentiment analysis is the computational analysis of text data from social media platforms to derive the feelings and opinions that people express in their texts. It is oriented at the identification of attitudes, ideas, and feelings of users around topic, goods, events or brands on social media. Sentiment analysis techniques normally includes NLP (natural language processing) methods and machine learning algorithms to classify text by positive, negative, or neutral sentiments. Social media sentiment analysis has a lot of applications across different domains.

In marketing and advertising it aids brand sentiment and customer satisfaction tracking, recognizing trends and adjusting marketing strategies to fit the purpose. Customer service can benefit from sentiment analysis by helping to detect and resolve customer complaints or issues as they happen. For political and social analysis, it may shed light on public opinion, its views of political candidates and policies, as well as emerging trends in public discourse. Besides, sentiment analysis is also applied in financial markets to forecast market trends relying on the sentiment expressed in social media platforms.



Challenges of social media sentiment analysis include the fact that there is a ton of irregular and nonsensical data, the presence of sarcasm, irony, or slang, and the necessity to have cross-cultural and multilingual text analysis.

Ethical dimensions like privacy, bias and fairness are the other aspects which should be taken into account when applying sentiment analysis into social media statistics. However, the sentiment analysis of social media collects a lot of information and is still a meaningful tool that helps in building knowledge and interpreting the users' content on social media platforms.

Social media data may have a noisy and unstructured nature, there may be sarcasm, irony or slang in text and it is necessary to provide multilingual and multicultural content. Ethical concerns relating to privacy, bias and fairness are equally very pertinent issues in the use of sentiment analysis tools in social media analytics. While these impediments are undeniable, sentiment analysis still is the vital tool for grasping and interpreting user-generated content on social network platforms.

## 2.2 NLP

From the very start, the NLP tools will be critical for sentiment analysis, like the one described next, which will open the way for establishing the mood of the text. Text preprocessing, which is key for NLP deals with the process of cleaning and normalization of data to ensure it is ready for analysis. This is basically made of the kinds of tasks such as cleaning special characters and punctuation marks, remove stopwords and tokenize text into individual word units (tokens or words). In addition, with a stop and lemmatization process, it

is very easy for the process to begin the feature extraction and analysis with each word being transformed to its base or root form. Precisely, with these preprocessing tools, the natural language data is standardized and ready to be utilized in this sentiment analysis algorithm.

The most important phase in the emotion analysis is the feature extracting and is the point where NLP methodology is being applied. One of the well-known methods such as bag-of-words (BoW) and TF-IDF (Term Frequency-Inverse Document Frequency) vectors are applied to a text data to make it numerical in form so that computer programs can work with it. BoW is a vectorization algorithm in which every document is transformed into a vector of word counts while the TF-IDF assigns weights to the important words in the document compared to the entire corpus. Such techniques solely deal with a connotation of the contexts, while sentiment analysis models need context to interpret sentiment trends and patterns.

However, NLP plays a critical role in the building, customization and the training of sentimental models. In supervised learning, algorithms including Support Vector Machines (SVC), Naive Bayes or deep learning architectures such as recurrent neural network (RNN) or transformers can be trained on sentiments labelled dataset for classifying polarized sentiments that are positive, negative or neutral. In addition, models of lexicon based (such as VADER (Valence Aware Dictionary and sentiment reasoned) are also frequently used in sentiment analysis; they use predefined sentiment lexicons and rules for scoring the applicable sentiments in the text.

Basically, the NLH analysis techniques become the basis of the sentiment project analyzing on the text data and supply with meaningful insights to the NLP models which are becoming more and more accurate and effective. For this particular project, harnessing the power of NLP will be especially important when it comes to the preparation of Reddit data, finding the features for sentiment analysis, and putting an effective sentiment analysis system that works on Streamlit into place.

## **2.3 LEXICON-BASED APPROACHES**

Leveraging VADER and other lexicon-based approaches, language sentiment analysis is one of the major contributions of the project. VADER is VADER as primarily a sentiment analysis tool in general but can accurately assess sentiments in the text, like Reddit comments. It works by giving different pre-set scoring to individual words depending on their semantic meaning and context which enables regarding of a text as a full based on a sentiment score.

The VADER lexicon is developed while collecting words and their associated scores which specify a positive, negative or neutral sentiment expressed by each word. Sentiment scores of words are floating between -1 to +1, with any scores close to 0 meaning neutrality. Furthermore, the algorithm uses rules and heuristics to handle grammar peculiarities like

negations, intensifiers and emoticons with the aim of increasing its knowledge on social media text analysis.

In this project, the VADER lexicon is considered part of the natural language processing algorithms that are designed to be run on Streamlit. While the users post Reddit comments or any text for analysis, through employing the VADER dictionary to produce a score for each comment either positive, negative or neutral based on the net sentiment score. The VADER lexicon liability adds to the capabilities of this sentiment analysis system so that it can sensibly go through the sentiments within Reddit discussions, hence giving users useful information on public opinions and attitudes against different products, events, or topics.

The one advantage with the vocabularies based methods such as VADER is the simplicity in their implementation and the ease of use, for instance, the for social media text where the subtleties in the language and informal usage are common. Nevertheless, it should be admitted that these methods have their drawback as well, such as strict dependence on predefined lexicons which can result in low perception of context-specific sentiments or emerging language tendencies being detected.

However, the text analysis by VADER vocabulary is not trendy in the field of social media analytics. Nonetheless, for projects like the current one focused on Reddit discussions, the VADER lexicon may turn out to be quite useful in sentiment analysis.

## **2.4 ML MODELS**

In the context of sentiment analysis within Reddit discussions, machine learning and deep learning models play a crucial role in classifying text into positive, negative, or neutral sentiments. These models leverage patterns and features extracted from the text data to make predictions about the sentiment expressed within the text.

One commonly used machine learning algorithm for sentiment analysis is Support Vector Machines (SVM). SVM is a supervised learning algorithm that learns to classify text data by finding the optimal hyperplane that separates different classes of data points in a high-dimensional feature space. In the context of sentiment analysis, SVM learns to classify text based on features extracted from the text data, such as word frequencies or TF-IDF scores.

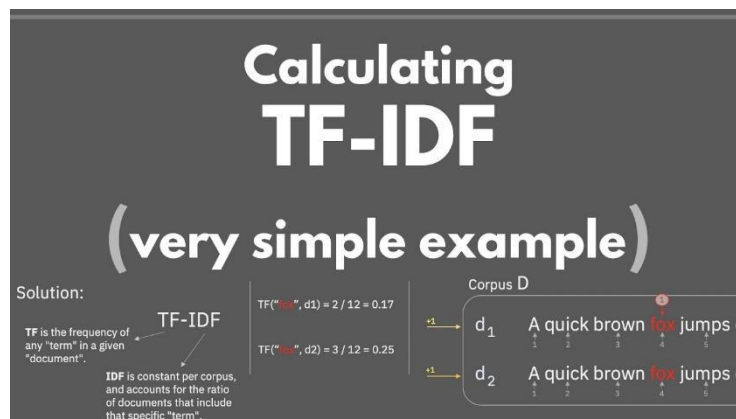
$$TF - IDF = TF(t, d) * IDF(t)$$

$TF(t, d)$  = Number of Times Term  $t$   
Appears in doc,  $d$

We know that's just our  
CountVectorizer

$$TF - IDF = CountVectorizer(t, d) * IDF(t)$$

Deep learning models, particularly recurrent neural networks (RNNs) and transformers, have also shown promise in sentiment analysis tasks. RNNs are a type of neural network architecture that can process sequences of data, making them well-suited for analyzing text data with temporal dependencies. Long Short-Term Memory (LSTM) networks, a variant of RNNs, are commonly used for sentiment analysis tasks due to their ability to capture long-range dependencies in Text data.



Transformers, on the other hand, have emerged as a powerful architecture for natural language processing tasks, including sentiment analysis. Models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) have achieved state-of-the-art performance in various NLP tasks, including sentiment analysis. Transformers leverage attention mechanisms to capture contextual information from text data, enabling them to generate highly accurate predictions about sentiment

In this project, machine learning and deep learning models can be utilized for sentiment analysis within Reddit discussions. By training these models on labeled sentiment data, they can learn to classify Reddit comments into positive, negative, or neutral sentiments based on features extracted from the text data. The choice of model depends on factors such as the size of the dataset, the complexity of the text data, and the computational resources available for training and inference.

Overall, machine learning and deep learning models offer powerful tools for sentiment analysis tasks, enabling the extraction of valuable insights from text data and facilitating the

analysis of sentiments within Reddit discussions. By leveraging these models, the sentiment analysis system deployed in this project can provide users with accurate and reliable predictions about sentiment trends within Reddit comments.

## 2.5 PRE-PROCESSING TECHNIQUES FOR SOCIAL MEDIA TEXT

The pre-processing of texts which refers to the prohibition of noise and the formulation of standardization of the text within sentiment analysis in social media text is very essential. The language of what humans write on social media which often revolves around the noise, irregular forms, and informal words such as, for example, contractions would affect the accuracy of sentiment analysis models.

This means that as part of pre-processing, methods of standardization and optimization are applied to the texts to ensure accuracy in the analysis. One of the methods that is frequently used when dealing with sentiment analytics processes is termed noise removal. This procedure encompasses eliminating of the superfluous details and only keeping the text with unnecessary characters, punctuation, URLs, and user mentions being removed.

It is as such not necessary for human communication which causes noise in text firstly, thus text data becomes more focused and easier to analyze, improving the performance of models for sentiment analysis. Handling Emoji's and emoticons like another important preprocessing step for social media texts as well harmonizes with this point. The presence of emoji would symbolize emotional expression in the social media engagement and determine the mood or impression which would be communicated to the homogenous. An usual step of pre-processing, including translation or condensing emojis to the text representation, (summarized as: (e.g., 😊 to "smile"), taking turns into account emojis which are properly interpreted in sentiment analysis models.



Moreover, this pre-processing stage could involve issues such as expanding abbreviations, converting acronyms, and dealing with slang terms evident in social media text. This might involve, for instance, filling out abbreviations with accurate translations or juxtaposing slang words with the standard counterparts so as to ensure ease in analysis. Besides, pre-processing

methods for social media texts are necessary as the raw text data should be appropriately fed into the sentiment analysis model for it to draw useful conclusions. Sentiment analysis models do that in a way that it removes all noise from the text data, tokenizes it, handles emoji's and emoticons, and normalizes abbreviations and slang that appears in the data. As a result, they are capable of analyzing feelings and emotions that social media users share in their discussions and thus, bring valuable information about public opinion and sentiments. Implementing the pre-processing techniques for a sentiment analysis system that is deployed on Streamlit shall positively affect its ability to acknowledge sentiments within Reddit comments and eventually enhancing the overall efficiency of the system.

## **2.6 ETHICAL CONSIDERATION**

Ethical matters are essential when building and using the sentiment analysis systems, especially considering the way customer sentiments about issues or products are being determined using user generated content from social media platforms such as Reddit. Among the main ethical concerns, we have to guarantee that user privacy is not being infringed upon as well as their data protection. Internet speak ought to be subject to conventions when it comes to Reddit. The neutralities of Redditors' rights must be guaranteed to them, as they expect privacy when it comes to their conversations on this platform. Abiding the Reddit term and condition and data usage policy is a must for proper management of data collection and analyzation, permissions and consents from users must be obtained and anonymity or aggregation of data must be applied when necessary to protect identities of the users.

Furthermore, in the context of sentiment analysis, minimizing the bias and achieving the fairness is very important because of the risk of amplifying the impact of the existing stereotypes or discrimination. Sentiment analysis systems that were based on the biased or not-representative data sets may not produce a correct outcome and this will result in the negative consequences. Therefore, the task of eliminating the bias from these data samples, utilizing diversified and equalized data samples, and implementing fairness-considered learning methodologies is of vital importance to mitigate bias and guarantee the fairness in sentiment analysis.

The aspects of transparency and accountability is also factor to be considered in ethical sentimental analysis projects. The categories of products that most people demand, like clothes, shoes, and cosmetics, will be the first ones made of recycled fabrics. The users ought to be aware of the purpose for which data is collected and analyzed, what the data that is being gathered would be used for, and, any implications of the analysis. Presenting transparent details of the approach, presuppositions, and limitations of the sentiment analysis system will be an instrument responsible for building trust and keeping users engaged. On top, constructing the means by which users can use or delete their data, and providing roads for correction in case an error is made or it is disputed improves transparency and promotes right conduct in sentiment analysis projects.



All these in all is interrelated with the issue of ethics, where privacy protection, bias mitigation, transparency, and accountability should be given first place in the use and implementation of artificial intelligence and sentiment analysis purposes.

## **2.7 STATE-OF-THE-ART IN SENTIMENT ANALYSIS**

The nowadays sentiment analysis on social media platforms with a fast developer like Reddit is distinguished by the combination of sophisticated natural language processing (NLP) methods, machine learning algorithms, and highly scaled datasets. Of late, there have been phenomenal aftermaths of deep learning, especially with the Transformer Models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-Trained Transformer) due to the fact that it has vastly improved the accuracy and robustness in sentiment analysis models. The models are becoming capable enough to process the abstract ones and text semantics very well, hence they bring improved refined and accurate sentiment classification and analysis.

In addition, the progress in transfer learning made it possible for pre-trained sentiment analysis models to be fine-tuned on domain-specific datasets while using socially network's public text from Reddit. Using the pre-trained models, sentiment analysis projects can work at a competitive level with smaller datasets and fewer computational resources to reach the state of art faster and more easily in deployment of sentiment analysis systems.

Future paths in sentiment analysis on the vital social media (like Reddit) should definitely concentrate at least on the several indispensable areas. The other area is the realization of multimodal data which will be generated using text, images, and videos with the aim to undertake a sentiment analysis. Multimodal content is used to add more visual and textual indications in sentiment analysis models which in turn allows for better and detailed sentiment analysis leading to improvements in the interpretation of social media data. Besides, sentiment analysis for multi-lingual and multi-cultural contexts is gaining momentum. Moreover, the significance of emotion analytics for video and audio content analysis is highly encouraging. The language and cultural variety of the Reddit social media users can bring into cognition that sentiment analysis models should be modified to adapt not only to different languages, but also dialects and in multilingual processing. Hence, next studies can consider language-independent sentiment analysis and localization of existing models to the dissimilar language and culture context.

Besides, the issue of ethics and biases should be thoroughly researched on the part of the sentimental analysis projects since there is no single solution.

Ensuring fairness, transparency, and accountability in sentiment analysis models is essential to prevent the amplification of biases and discrimination. Future research may explore techniques for bias mitigation, fairness-aware learning, and ethical guidelines for the responsible development and deployment of sentiment analysis systems in social media analytics. Overall, the state-of-the-art in sentiment analysis within social media platforms like

Reddit is characterized by advancements in NLP techniques, machine learning models, and large-scale datasets.

## **CHAPTER 3**

### **AUTO CORRECTION & AUTO GENERATION**

#### **3.1 AUTO CORRECTION**

Auto-correction, a crucial process in making the text data accurate and readable is used in multiple NLP task like sentiment analysis and that is more prevalent within the context of socially mediated data like on Reddit. In this project, auto-correction refers to the technique used to detect and automatically correct spelling mistakes or misspellings of words appearing either at comments of Reddit or studying the textual data so to ensure its absolute standardization and optimisation for sentiment analysis. The implementation of auto-correction involves several key components and techniques: The implementation of auto-correction involves several key components and techniques:

1. Edit Distance Algorithms: Some advanced algorithms like Levenshtein distance or Damerau-Levenshtein distance are usually used to find out the similarity between two strings as a result of edit operations (insertions, deletions, substitutions) which are minimum in numbers needed to turn one string into the other. Specifically, for auto-correction purpose, edit distance algorithms serve to detect and make suggestions on the words that are misspelled or filled with typos in the text data.

2. Dictionary Lookup: On the other hand, the auto-correction systems mainly use the word dictionaries or look-up table - containing a huge reservoir of correctly spelled words - to detect and correct mis-spelled words in the textual data. By comparing every reserved word in the text data with the prescribed entries in the dictionary, the auto-correction system to discover the possible spelling mistakes and proceed by offering the analogues suggestions based on the closest matches found in the dictionary.



3. **Language Models:** The application of language modelling techniques, like n-gram models or neural language models like GPT (Generative Pre-trained Transformer), to auto-correcting systems can be done too. These models are performing the statistical pattern identification using context analysis of big text corpora in order to infer the probability of certain word sequences as well as to discover possible misspelling basing on their context. By adopting language models into the correction system of auto-correction, it could have the ability to produce higher precision and context-based corrections.

4. **User Feedback Mechanisms:** One of the ways to enhance the auto-correction systems is by putting in place user feedback mechanisms which can provide information about the corrections done by users and help the system to ingest this knowledge. By taking the ion of user interactions and feedback, the auto-correction algorithm can change and enhance its performance by overcoming the spelling mistakes and typos present among the text data. This therefore will improve the performance of the correcting algorithms.

## **3.2 AUTO GENERATION**

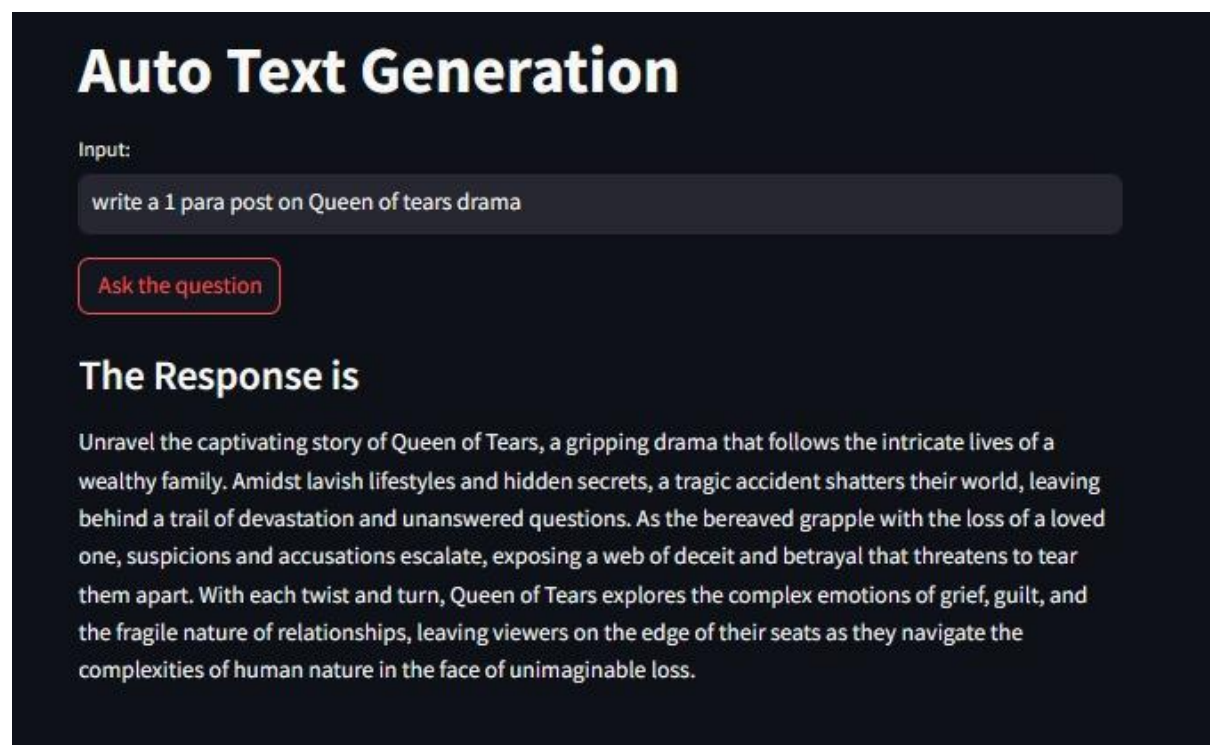
Adopted the latest auto-generation techniques that will spike on the popularity of the forum and its functionality due to user engagement. Sents to Genie which has a powerful tendency to develop original and useful text, based on users input. The modern technology also allows the application to generate content dynamically that is relevant to users and popular trends which in turn enhances cohesive user interaction and creation of content.

The auto-generation feature serves many applications from content generation, multilevel support, insights and conclusions, and conversation managers. Now through automated creation of exciting and relevant matters for social media posts the platform brings that empowerment where you are making this content effortless and safe much more time-saving. Additionally, the ability of the system to provide text in various languages considerably

increases accessibility and representativeness of the users, who may be from different linguistic backgrounds with this expansion in the reach and the appeal of the platform.

Furthermore, the auto-creation functionality enables the textual content compiling, the text summary writing, and the meta-data extraction from the large volumes of textual data as well. Researchers, journalists, and storytellers among others have been endowed with the privilege to key into this capability to get information and strong conclusions using it. Also, the platform can involve conversational machine learning algorithms to simulate human interactions that can add value to the services delivered by personalizing assistance and recommendations to customers which in turn increase user satisfaction and engagement.

Therefore, the adoption of auto-generation confers the platform by distinct features like dynamic content support, multilingual dispersion, summarization oblige, and talk to text abilities. It is a generic framework that extends the platform into multiple use cases, allowing for multifaceted applications in numerous domains and thus improves the user experience and appreciation. The tech advancement and attention of the users to our platform enables us to stay up-to-date with the latest trends in information creation, viewing, and communication.



## Auto Text Generation

Input:

write a 1 para post on Queen of tears drama

Ask the question

### The Response is

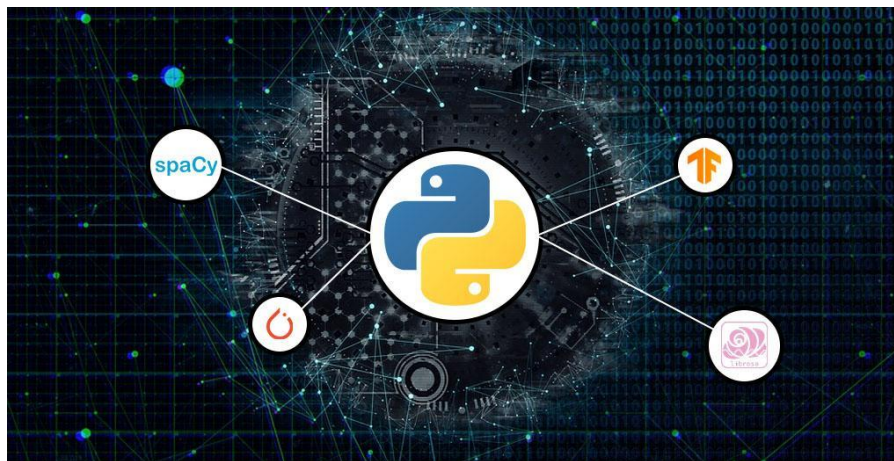
Unravel the captivating story of Queen of Tears, a gripping drama that follows the intricate lives of a wealthy family. Amidst lavish lifestyles and hidden secrets, a tragic accident shatters their world, leaving behind a trail of devastation and unanswered questions. As the bereaved grapple with the loss of a loved one, suspicions and accusations escalate, exposing a web of deceit and betrayal that threatens to tear them apart. With each twist and turn, Queen of Tears explores the complex emotions of grief, guilt, and the fragile nature of relationships, leaving viewers on the edge of their seats as they navigate the complexities of human nature in the face of unimaginable loss.

### 3.3 TOOLS AND LIBRARIES

In this project many tools and libraries are combined to accomplish the following: gathering, pre-processing, analyzing, and finally deployment of sentiment analysis for the conversations

on the Reddit which will be displayed on the Streamlit web page and this will automatically run every time that a sentence is posted on the subreddit. I am using the Python programming language for project development because of the rich NLP libraries that support processes like machine learning. Python offers an extensive space for NLP applications that range from libraries to frameworks, thus, it is a common choice of analysts for sentiment analysis projects.

For data collection purpose from Reddit, this project makes use of the PRAW Python Reddit API Wrapper library which exposes a Python interface for working with the Reddit API. PRAW simplification helps programmers to get rid of the problem of extraction of data from platforms like Reddit and provides an ease in setting up parameters like name of a subreddit or the number of posts/comments or the keywords to retrieve the output in an efficient and direct way. As well, PRAW necessarily involves authenticating and rate limits issuing by the Reddit API that is used to facilitate data abstraction to run smoother and more reliably.



In addition to preprocessing and analysis libraries like NLTK(Natural Language Toolkit) and scikit-learn, the project heavily relies on a host of other tools. NLTK focuses on a broad palette of text processing tools and algorithms such as tokenization, stemming, lemmatization, and TF-IDF vectorization. Such preprocessing techniques are critical functions that inform the cleansing and standardization of the text data that precedes the analysis, guaranteeing the precision and consistency of the sentiment analysis result. Scikit-learn brings forth a complete package of algorithms and tools which are primarily for model building and its evaluation.

The Gemini API by Google is a top notch tool for studying several forms of text using its sophisticated natural language processing algorithms. Employing refined machine learning solutions, Gemini creates contextualized and relevant soft texts matching the user's particular circumstances. By analyzing language and context closely, Gemini develops devices that generate texts as if they came from a human.

Our auto-generation service was made possible by adding Gemini API solution to the project. We offer the ease of making content more dynamic and interactive for Instagram posts, articles, and other through our co-branded Gemini system and our system. Considering the multi-lingual texting platform provided by the API, the expansion of the platform in the terms of scope and the possibility of the users regardless of linguistic background to use it becomes certain.

Not only that, Gemini is made for taking up many non-text related jobs as well. It may provide summaries, identify the key issue and even use the techniques people use while providing chats. Besides this only function of the platform, the users may create, analyze and alternate the texts. It is noteworthy that the Gemini API serves as the foundation for the auto-generation function of our system which improves by personalization of the users and the maximum growth of our performance in the digital world space.



## CHAPTER 4

### IMPLEMENTATION

The execution of each element of the project as a whole consists of multiple procedures, including getting hold of the data, preprocessing it, analyzing it, and launching the application in the Streamlit web platform.

Initiating with the data extraction part, the PRAW (Python Reddit API Wrapper) library is used to connect with the Reddit API. PRAW has allowing them to write filters that specify name of subreddits, number of posts/comments, or use specified keywords so that they can easily retrieve required data rapidly.

Thanks to this approach, the information collected will cover a whole range of Reddit topics and will not be limited to just one or two specific ones. Alongside PRAW, the Reddit authentication API calls and rate limits are all managed, thus the steady performance of the data retrieval.

Subsequently, the preprocessing phase within the algorithm uses the NLTK library for a our text preprocessing tasks. NLTK is a package that includes everything needed to do tokenization, stemming, lemmatization, and TF IDF into another coding language. The above-mentioned preprocessing techniques are to be considered as crucial steps particularly before sentiment analysis, which are of immense value because of standardization and accuracy in sentiment analysis output.

Moreover, the execution includes strategies for dealing with emojis abbreviations and slang which are the most principal ones used in social media text tone of voice, thus, making the preprocessing pipeline more precise and effective.

The project which implies sentiment analysis is implemented by means of machine learning libraries and algorithms like scikit-learn. Algorithms such as SVM and Naive Bayes classifiers are one of the most commonly used and very popular algorithms in sentiment analysis as they are trained on labeled sentiment data and they classify text as positive, negative or neutral sentiments. Scikit-learn presents an easy-to-use wrapper to both, implement these algorithms and provides functionality for evaluation and validation of the models. The implementation involves training the sentiment analysis model to get accustomed with the domain specific data like Reddit comments, and hence, can make the system more robust and accurate.

Thus, the construction step is accomplished because of using Streamlit library for further communication with Streamlit platform for business application of sentiment analysis system. Streamlit eases up the procedure of making interactive web apps using a very concise code span that has user-friendly and customizable widgets for representing sentiment analysis



results on dashboards. The implementation enables the deployments of the system to Streamlit servers and browsers, offering users with an enjoyable and interactive environment for exploring the sentiment-trend among Reddit discussions.

However, the integration of every part of the project plan needs a thorough scrutiny of the suitable tools, libraries, and techniques so that the job of sentiment analysis on Reddit streams during Streamlit's run time doesn't end up being a hit-and-miss! Through the use of these elements correctly, developers could indeed create a sentimental analysis method that is resilient and user-friendly and, most crucially, provides valuable insights into the public opinion and views on discussed topics on Reddit.

## 4.1 DATASET

Another goal is a metadata that reflects these metrics so that the author can decide what exactly can be integrated into this text from the chosen content on Reddit. Here's a breakdown of the columns included in the dataset: By reviewing the mentioned columns in the outlined dataset as in the description below was given.

1. Subreddits: This is for the following part later in which magazines will be grouped with the respective sub/subreddit names texts. Sub like that, devoted to themes of communities that are widely known in popular ad, or to those who are so fascinated to these topics or communities is what they will choose. Sub-column allows the reader of both presented information in various neighborhoods and also helps to not fill the gaps of data that are available within such communities especially during national crisis.
2. Post Titles: The left side of the image displays the titles people posted to Reddit as if they were the column headings. The Ultimate Catchphrase for the Feature Staff: It Is Non-Stop Attack of Words for Articles Effect. The latter is not involved in the other processes too Through these NFLPA newsletters, the branding is a core part of the fans which is a result of some activities the team and the fans watch together.
3. Self-Text: It was a little room, but very active in LaToya's opinion, that each behaviour was demonstrated, either by her responding to or making a post. A term "self-text" appearing from the title of this subreddit means the subreddit has a structure and makes interactions and contribution of the community members to be more unique. Some symbolism are abeam to refer us to another related issue or highlight election were candidates are supposed to claim due positions. Then their on-going task would be linked to all of them.
4. Post ID: Finally, the note advice to us especially that this/one device is very good which will provide the information about what we want, therefore, we decided to research further about this essential information. Infusing the database IDs with personal data that is streamlined from each registration ID as data keys, there is also an automated application of protocols for a security approach intended to uphold the privacy of the data.



5. Link/URL: The next item is the Kindling element in the form of the link taggers (Red URLs) that are often linking to the Reddit page and which is eventually used only if it is crucial for the simplification of the selection (inferred as [Confirmed]). This is the first example sentence of how to accomplish the introductory part of the essay, referring to few other informative websites and other online resources to provide a supporting statement of the main theme of the essay. Alongside with the text section in the back link box you'll receive extras that the book site provides. They will be helpful to you in branching out your knowledge.

6. Flair Text: Each of the strands has a comment score which is exactly the measure of the post's quality. The strike of the exclamation marks, in a surprising manner, is the sole reason why the pinpointing eye might suggest a Reddit group topic or the theme of the post/paragraph orientation and the audience can identify the actual subject wordily. Thus, they are so great at the beginning part of the process including 'identifying/presenting subject'. Sentiments can be recorder (the rate favored by subreddit community and flair) these could also be tracked. Therefore, it has become possible to get and know that there are many common features among these societies which help in identifying those that can be assumed to be completely alike. We invite you to put our Artificial Intelligence Interactive to use and see the terrific result in the college application.

7. Score: This second column records the approval or rejection of each Reddit post according to the scoring system. (The number of points is calculated incorporating the up votes, down votes, as well as user interactions in the equation thereby representing the level of engagement of a post within the Reddit community.)

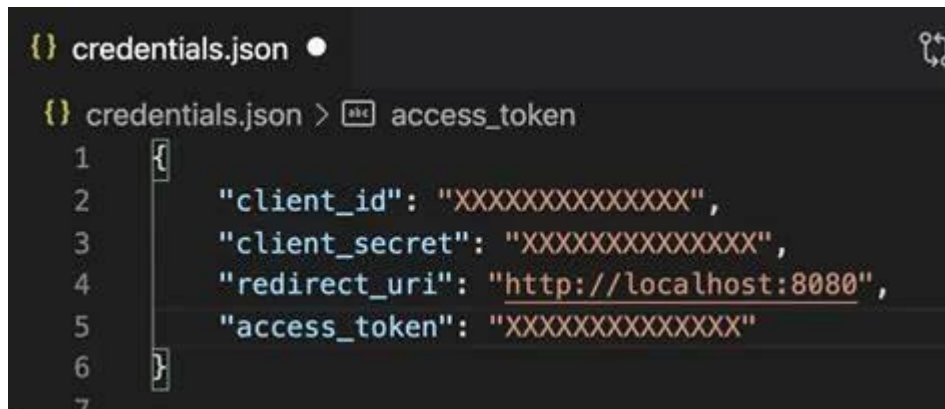
8. Number of Comments: These values, the tweet counts, represent the count of comments related to each Reddit post. In short, comments are the user-driven replies and animations focusing on the theme of the post, so the number of comments can indicate the depth of the Reddit community.

Through a joint analysis of these votes we can genuinely dig out trending sentiments, users engagement, and distribution of topical situation in discussions. By this, we will be able to explore and investigate the attitudes and opinion of the public within among subreddit neighborhood.

## 4.2 PRAW

PRAW is a Python API that wraps Reddit API to allow programmers to use Reddit's API more effectively while coding on the Python. It allows the user to not dump APIs stating plain, by using it in a defining and informative way, that is, the user can now get API calls done, authentication, handle the response parsing and the rate limits altogether. With PRAW helping out to pull information about users from Reddiit, we get not only some posts and

comments but some subreddits as well. Therefore, I will seek help from my professor in our projects that require the data from Reddit, which is performed in the Python-based projects. PRAW, which has a well-documented code and a large fan following and support is one of the famous python packages used for the data analysis purpose in fields, such as Sentiment Analysis, mass occurrence of certain words or phrases that dictate a general view of the document or user given string, clustering and webbing.



```
{  
  "client_id": "XXXXXXXXXXXXXXXXX",  
  "client_secret": "XXXXXXXXXXXXXXXXX",  
  "redirect_uri": "http://localhost:8080",  
  "access_token": "XXXXXXXXXXXXXXXXX"  
}
```

### 4.3 NLTK

NLTK (Natural Language Toolkit) with the NLP area and Python based environment is the base element but there are more tools. Implanted to handle many areas of lexical requests, NLTK portfolio consists of tools and features and all which are essential for treating textual data. From fundamental subtasks (tokenization, stemming, and lemmatization) to more advanced ones (part-of-speech tagging and syntactic parsing), NLTK can be considered as a powerful kit for NLP researchers and developers to defeat the language understanding challenge.

For the purpose of this project, NLTK is the key player responsible for the text data before it progresses to sentiment analysis stage since the raw text from Reddit discussions is the subject matter. Through library's tokenizing capabilities tokens could be created as a building blocks and as such can be used for easy analyzing and studying. The process of steaming and lemmatization assists in making the words into their base or root forms which is the standardization of vocabulary and eventually the enhancing of the sentiment analysis models. Building in addition, Rossett's part-of-speech tagging capabilities provide the capability for detecting grammatical categories in the text, hence the structure and meaning of the text will be enriched.

In addition to its fundamental functions, NLTK provides access to an immense library of textual resources and lexicons that is able to enrich the analyses with language findings. As far as certain applications like sentiment analysis in the natural language processing domain are concerned, NLTK provides access to sentiment lexicons and datasets that can be effectively used for better accuracy and detailed sentiment analysis models. Besides, NLTK just isn't limited to the sentiment analysis or applications involving the other language processing tasks such as text classification, information extraction, the NLTK stands therefore to be an excellent choice for these kinds of projects.

## 4.4 FEATURE EXTRACTION

We just have realized that vector space model is a great way of representing document and it costs us a lot of benefit such as we can identify (compute) cosine similarity could and even until there comes the within the cluster. However, vector space model does not have the ability to solve these two typical problems: A may can enrich there is the sign of polysemantic where polysigmatic and word has the same meaning.

I think that I only try to create a thesaurus because it would be much more time spending. There are two among the onto mainstream methods, two are the co-occurrence lexis and the sentence-level shallow parsing. Study of the sentence's structure or grammatical relationship. Generally in our vocabulary, the words due to lexical co-occurrence are more robust while grammatical relation is more explicit.

In the space of the vector that I used in my previous thinking, I concentrated the frequency of a single word. But the co-occurrence of words on a page is also a very crucial data, which is characterized in that the words that are key terms of a particular topic are likely to be within the immediate content. Rather pointers of two or more relators observed in the document are not ignored carelessly. Latent Semantic Indexing is one strategy with the purpose of revealing internal semantic relations.

The latent In contrast to semantic indexing that maps the co-occurrence words on the same dimensional space, deep learning is a technique based on modeling the semantic relations between words within sentences. All together words approximately being what they mean are semantically bound.

## 4.5 TF-IDF

Statistics-assigned TF-IDF (term frequency-inverse document frequency), which is a statistical method utilizes in info retrieval and text mining that determines a word's centrality to a document by calculating from its representation. The word count of each word again it becomes a kind of effective influencer besides the amount of inverted frequents as compared to the corpus of the document. TF-IDF is an abbreviation for two parameters abbreviated by these term “TF” and “IDF” that stand for two terminologies “Term Frequency and Inverse Document Frequency” in this case. The TF metric stands for the number per document of a given word's representaion in a particular text. The main idea of IDF is: overall, in the scoring system, the more a term is mentioned in other documents, the less important it will serve.

### 4.5.1 TERM FREQUENCY

Term frequency has to do with the times Im which a given term  $t_i$  is contained in the document  $d_j$ . This can be presented as  $TF(t_{ij})$ . Among other things, in the case of sentence removal, the more time the term  $t_i$  listed in a document, then the more critical it is to this document. It can be defined as: It can be defined as:

$$w_{x,y} = tf_{x,y} \times \log \left( \frac{N}{df_x} \right)$$

**TF-IDF**  
Term  $x$  within document  $y$


$tf_{x,y}$  = frequency of  $x$  in  $y$   
 $df_x$  = number of documents containing  $x$   
 $N$  = total number of documents

### 4.5.2 INVERSE DOCUMENT FREQUENCY

For the concept of inverse document frequency to be understood, we will first determine what document frequency is. Document frequency (DF) is actually an indication of how frequent term  $t_i$  appears in all documents  $C$  of total documents  $C$  By  $N(t_i, C)$ . The more the longer the term  $t_i$  appears in all documents  $C$ , the weaker the term  $t_i$  can be considered as the representative thing document  $d_j$ . Inverse document frequency suggests that the contribution of  $t_i$  to document  $d_j$  is measured, i.e., term  $t_i$  is counted or weighted. Its amount in all


documents  $N(t_i, C)$  is inverse proportion, which is represented by  $IDF(t_i)$ : its amount in all documents  $N(t_i, C)$  is inverse proportion, which is represented by  $IDF(t_i)$ :

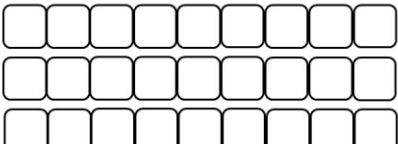
**TF-IDF**



$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$





### 4.5.3 WEAKNESS OF TF-IDF

The main drawback of TF-IDF is the inability of this algorithm to focus on semantics, which means text expression is just limited to frequency of word appearance in a document rather than the sense of words or phrase. same category of documents. The inverse document frequency principle is based on a straightforward premise -the more the term  $t_i$  repeats itself in a given document or collection of documents, the less often it appears, all things being equal. document, this one is more precise to the paper where it clearly communicates what the article is about. similarity or the difference in the matter of science where the terms may be in the same class in the situation. If the time  $t_i$  is increased instead of the number of temperature chances, this is the correct way to represent it. for instance, if a model is good enough to categorize a person correctly, it can be a great representative of the sort. Therefore, it is the combination of all the above mentioned language features which makes it possible to create a definition of the term. given higher weight. On the other hand, in equation 3.3 the IDF will decrease simultaneously with the increase in the number of power plant and the ECPs that we use to produce the electricity in the power electrical grid. but across the planet the temperature is gradually rising. So, we conclude: the amount of in a text which appear frequently. efforts should be jointly coordinated to make campaign in this group more effective. On the other hand, compared with  $t_2$ , it is evident that  $t_1$  exhibits a more normal outcome in recent websites. Thus,  $t_1$  is more likely to be selected. representative than  $t_2$ . If  $t_2$  occurs only in a small portion of the category and precisely in a couple documents of the category, then the probability of its relevance would be relatively small. the matter of relatively weak  $t_2$  connection to the category may appear, but it could be also accepted as  $t_2$  having a secondary role for it. This take into account that economic indicators are not very precisely representative, and they should be discarded as measures of world economic prosperity. TF-IDF cannot solve this situation neither

#### **4.5.4 IMPROVEMENT OF TF-IDF**

We reviewed legacy articles and found that this particular improvement supersedes only a few operations as reported by the literature. our analysis above. Paper adds an extra dimension to the category parameter with which a category is selected, so it places us into a unique position of an in-group member. views on the management of terms, the aspect which is mostly emphasized. The version that helps how to arrangement terms in category. canary in the coal mine means most of the documents have the same keywords, so the semantic distinction of documents which is ignored the fact that the term appears in the. 34 text samples which are a part of one group. Thus, in-category term frequency (TF-IDF) index which is the measure of information content is one of the key notions in the scientific retrieval method. important for the improvement. Here, 'term significance' means the number of times a term features in the documents, and 'term weight' is a measure of word relevance, which is positively related to the frequency of terms in the text. But it also emphasizes that the higher frequency vocabulary could be occur in more than just a part of the whole document. Such the terms are a misleading entity which is rather made for a whole document. We do not suggest any immediate notions but try to give readers the adequate view on the problem. solve the situation.

#### **4.6 VADER**

Sentiment analysis, the dominant tool, VADER (Valence Aware Dictionary and sentiment Reasoner) is so helpful that it contributes immensely, especially, to social media platforms' toolkit. As such, Reddit becomes a treasure-trove of different insights; all these because it receives feedback from an extremely, large audience every day. VADER was adapted for social media to detect the words. Based on a lexicon approach, it is exceedingly useful for organization of the individual words/phrases and shows the changes in sentiment using the pre-saved sentiment lexicons and rule. VADER, that speeds up performance with a precise sentiment analysis from Reddit posts discussions have made it worthwhile and effective. The feature that sets it apart is the fact that its vocabulary is a massive collection which consists of many words each of which has been tagged differently as highly negative, highly positive, and so on. This is what makes it able to see the diverse spectrum of sentiment expressed in social media messages. Considering that, VADER is rule-based and the system can manipulate the specific linguistic aspects like negator, intensity, etc. which further kicks the system. With that, it leads to the better performance when analyzing sentiments.

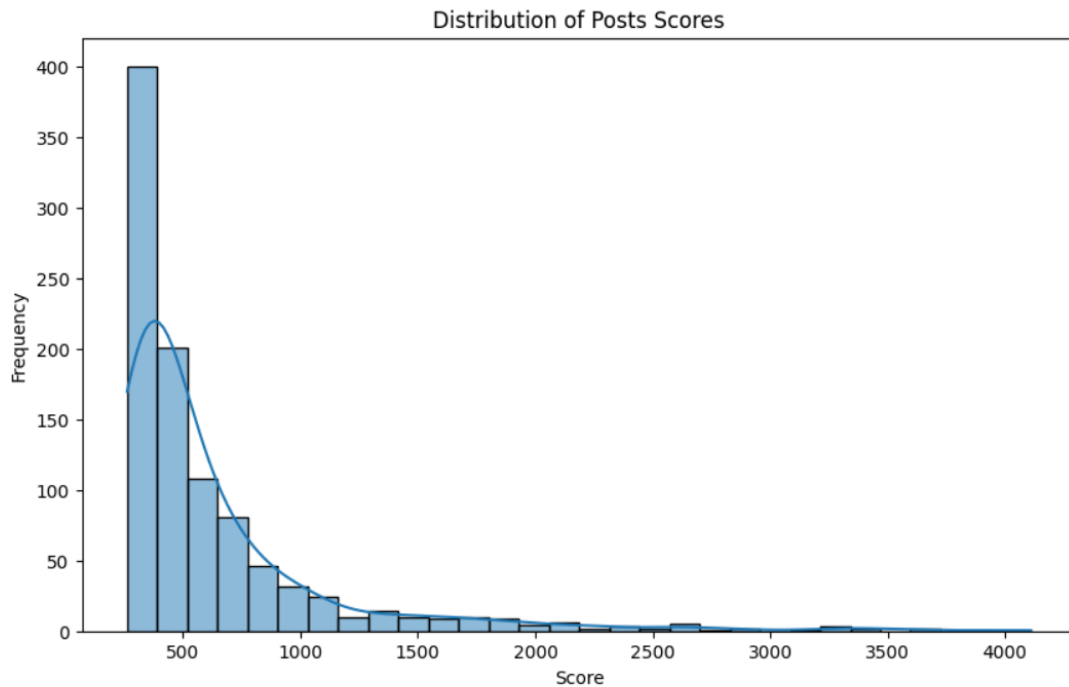


VADER is easy to adopt because of its simplicity of use is one of its advantages. Thanks to its easy installation and configuration, user can easily fit VADER in with their social media analysis pipeline and accessing insights shortly after deployment. In addition to that VADER is simple, fast and effective in the real time analysis of a big chunk of text data which could including but not limited to social media analytics and sentiment monitoring. Moreover, VADER is a reliable and accurate performance, proven across various social media platforms and domains after evaluating the high-performance. It also has a wide array of features to handle things like casual language, sarcasm and context specific expressions which is what makes it a powerful tool to grasp the intricate aspects of sentiment within the Reddit discussions.

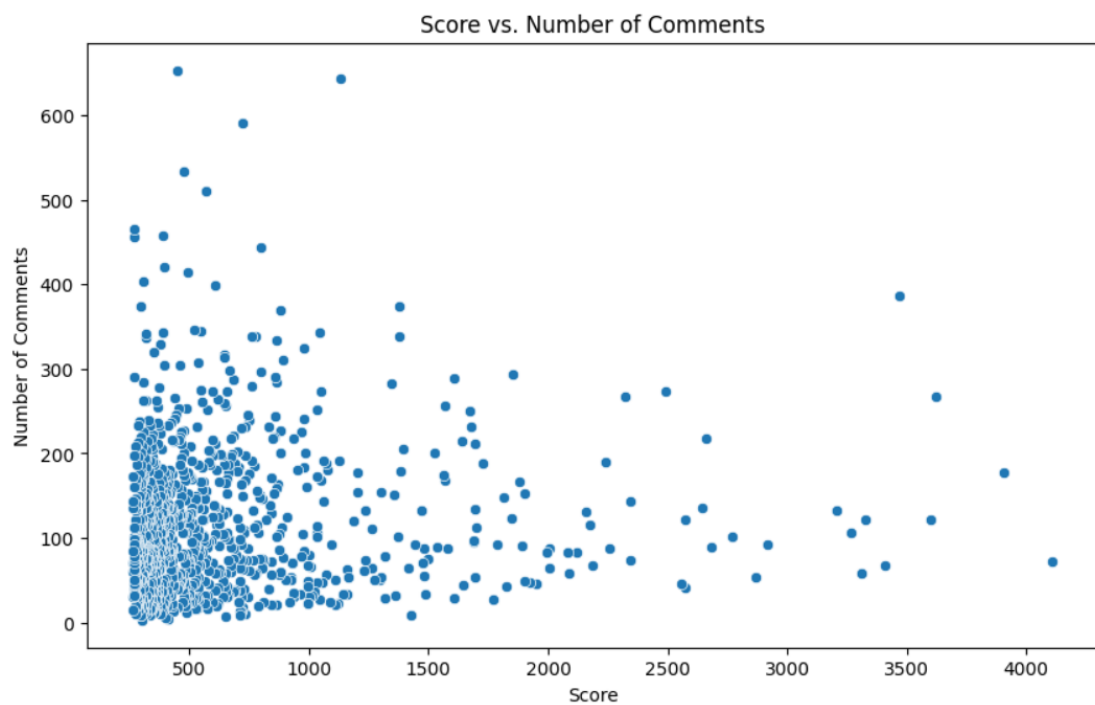
Overall, VADER is a mighty tool that can make sentiment analysis more manageable and easy to use in the case of the present study. The developers may realize potential information as the platform uses lexicon-based approach and rule-based heuristics at its core. Thanks to this, the sentiment analysis model gets high precision and accuracy making it possible to reveal the opinion expressed in the Reddit comments and the attitude of the overall public.

## 4.7 VISUALIZATION

1. Plot for Score Distribution for each Group vs Frequency: This graph presents the junction of post marks in the dataset. It displays the number of posts which gets different scores and in turn shows the user what is popular and who is engages on Reddit post.

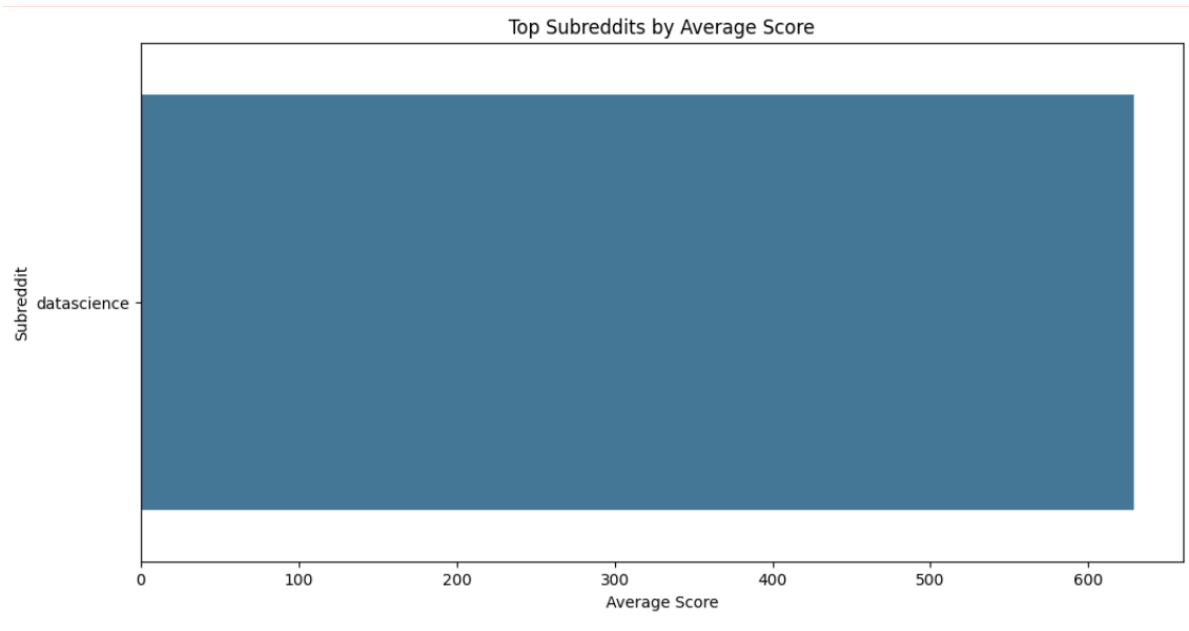


2. Scatter Type of Pictures of Comments and Times vs Score: The rom scatter plot which represents the correlating number of comments and the ratings of the Reddit posts is shown into the label. This helps us see a correlation between post popularity (score) and user reactions to the post (post comments). For example, we may see that a post with a high score has less comments or the opposite is true.

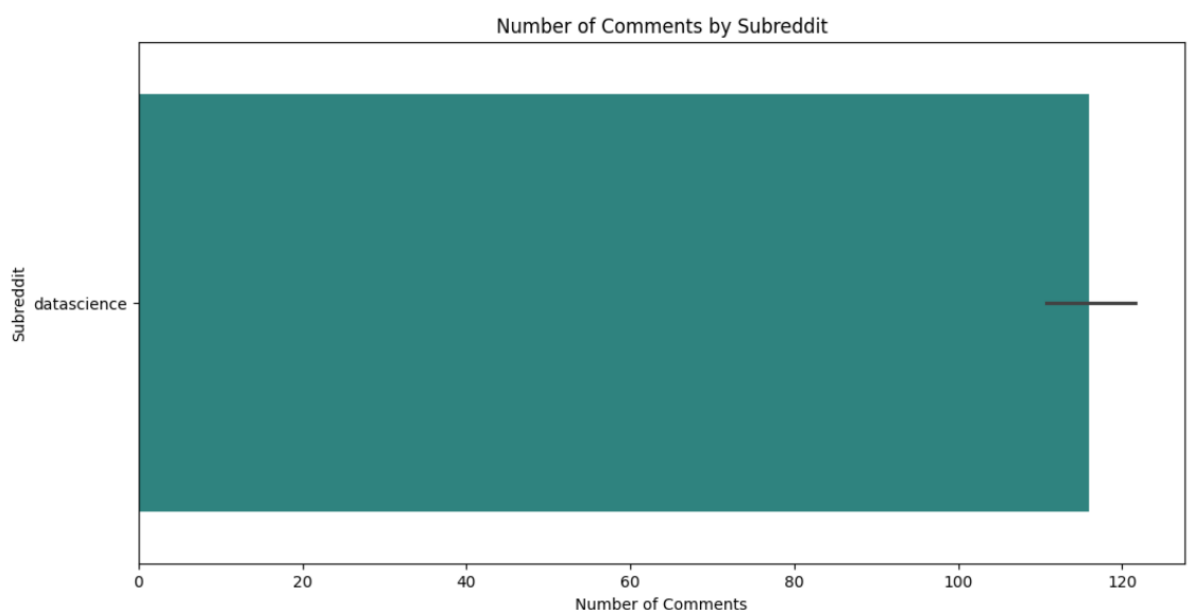




3. Subreddits as a column and their corresponding GDP as a row have been used for the creation of the bar graph vs Score: This chart illustrates the distribution of posted across different subs on the given Reddit thread. Through this, it allows one to be able to identify subreddits posting content that attract relatively low or high score engagement over the other communities.

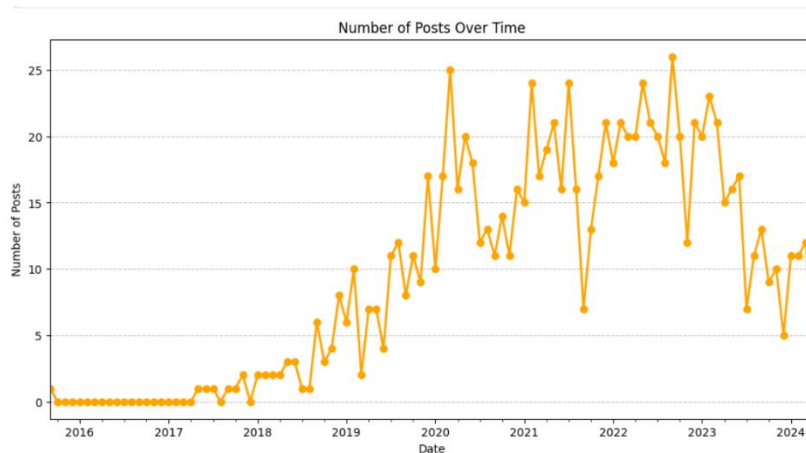


4. The graph above also shows the subreddits engagement of teenagers ages, bar plot- subreddits-vs-ages. Number of Comments: Just like the yarn above, this plot cut across the reddit boards in relation to the number of comments generated. It gives the net relation of level of distress in subreddit communities that could be used to decide which subreddits are the sites where the most discussions happen.



5. From the graph of time below, we can see that the number of posts clearly increases

over time. Date: This time graph illustrates a gradual tendency of post frequency when authors publish various posts during different dates or time intervals. It helps in exploring tender motions, and discovering the regular patterns and temporal trends of those discussions, for example the peaks of activity or seasonal patterns.



6. Word Cloud Graph: A word cloud, depicting the most occurring words from the Reddit dataset, becomes the size of its frequency in a webpage. It provides a thematic and qualitative view of the most dominant topics and keywords within that discussion. In addition, that helps in the identification of common trends in twitter.



## 4.8 EMOJI'S

VADER (Valence Aware Dictionary and sentiment Reasoner), however, recognizes emoji's that are dominant source of communication rendering on social media, and thus the emotional context. Emoticons have a crucial role in sentiment analysis as they make the manner in which text sentiment is analyzed more comprehensible. In VADER emoji's are treated as

special characters which sentiment scores derived from the semantic meaning and context of the emoji's. Specifically, some emoji's have been linked with `positive emotion`, while some others are seen as `negative or neutral sentiments`. For example, emoji's such as



generally, indicate positive things and are regarded as high positive lexicon by VADER. Alternatively, emoji's like angry, sad, and scared are expressive of the negative words and obtain the low negative sentiment scores. Furthermore, the simplest emoji's, such as 😊, 😐, and 😞 demonstrate medium simplicity and carry almost no sentiment weight at all. For sentiment analysis based on social media texts VADER's aptitude in the translation of emoji ultimately improves its respective performance in association with sentiment finding in text data. Through its novel lexicon-based approach that incorporates emoji's, VADER becomes able to explore the issue of social media sentiment in depth, giving us an in-depth view of public opinions and attitudes as related to different subject matter conversions. Putting forward the addition of emoji's into VADERs analysis work process help to create a more powerful and apt instrument in tracking down the various factors of the dull emotions and indirect text messages in socializing media.

## CHAPTER 5

### DEPLOYMENT

The deployment process on Streamlit was a long & complex one and so I had to make sure that my system configurations were done right to enable smooth integration and excellent performance. The code was also decorated with additional description so that it can be communicated properly to the Streamlit platform. Among the tasks which were completed was, create environment variables, define port configuration, and ensure integration with Streamlit runtime environment. As a further adjustment, deployment switches were configured to improve resource utilization and much more better system stability. Identifying the existence of dependencies was the first step in ensuring successful deployment, and this must have been done by assessing the correct packages and versioning compatibility. Pretty much handling the dependency management means provide trying to come up with an comprehensive list of required packages so as to hold their versions consistent across different programs. Resolving the dependencies issue used to take place

automatically and libraries or frameworks useful during deployment would be automatically installed or updated as required. As a result, we introduced a cache that defined patterns of dependency to accelerate the deployment and prevent situations where two or more versions of libraries could trigger errors during runtime.

The deployment details over Streamlit encompassed a plethora of areas including the setting of the environments, the specification of runtime, designing of the deployment pipelines, and many others. Parameters like user interface layout, theme personalization and interface responses were among those that could be adjusted in application settings. Runtime environment settings entail setting up runtime packages, environment variables, and execution parameters to deliver better efficiency and cross-compatibility with the Streamlit platform. Deployment pipelines were implemented to automate deployment, ensuring smooth fulfillment of continuous integration and delivery (CI/CD) workflows, enabling accelerated deployment of updates and extras to the application. Compiling multiple libraries and framework being necessary for a deployment is a tricky issue; it means constant testing to make sure that all of them are available and compatible. The usage of automatic dependency resolution mechanisms help in the auto installation and updating of dependencies as the modules are required, thus making the deployment process more straightforward and free of version conflict and runtime errors. What was employed in the process was a checkpoint caching mechanism for dependency management so as to make the process efficient and reduce network loads, the network load occurring as a result of redundant dependencies

## **5.1 DEPLOYMENT 1: DEPLOYMENT CONFIGURATION**

The development of uploading and analysis application on Reddit's API utilizing Streamlit API involves the following critical features to be put in place for flawless operation and easy to use terms. Defining consequently, the settings of the application is crucial such as the interface settings, data processing and how the application behaves. The task in this stage encompasses multi-steps, which include the arrangement of the application such as layout of input fields, buttons and result displays to ensure a seamless, intuitive, and user-friendly interface for uploading and analyzing text data

Also, debugging runtime environments is crucial for configuring and monitoring the application's performance and compatibility with given environment. This can be summarized as the process of specifying the run time dependencies for environment such as Python libraries for data processing and analysis and subsequently any environment variables that would be necessary for authenticating with the Reddit API or accessing other external services or resources. Besides, the execution parameters setting, for instance, memory and CPU limits will improve application efficiency, and therefore, it is possible to meet processing requirements of text data social media monitoring.

The creation of deployment pipelines which will be based on automation and will serve to bridge information exchange between the various CI/CD workflows will enhance the workflow process. This means that no matter what kind of launch we have (whether it is a huge one or a small one), updates and enhancements to the app will still be released seamlessly and will be able to be rolled out to users quickly. More so, putting log and monitoring approach enables us to have the ability to see if there are any performance issues or a problem that we had thought about during the course of operation so that they can be fixed timely.

After that, apply security measures which will accomplish the purpose of secure customers' data along with guaranteeing the accuracy and confidentiality of the social media text analysis. These comprise the use of authentication to validate user identity and access to privileges to what encryption provides for data exchange and storage security. Moreover, the approach that is input validation and sanitization approaches helps in common security issues such as injections attacks and cross-site scripting (XSS). These helps in making the site more secure.

## **5.2 MANAGING DETAILS**

Ensuring that libraries/frameworks used in the app Streamlit for handling text analytics based on Reddit API are accessible and workable is a typical task when dealing with dependencies. To do this let's provide the requirements for the Python modules that will be used for the data pre-processing, text analysis, and interface with the Reddit API. These tools can resolve interdependencies among libraries and modules, provide unified packages for specific operating systems, and support multiple versions of programs.

We should compile a list of all the necessities and include exact versions so to guarantee a complete consistency across all the different environments of deployment. This line usually comprise libraries for data processing and analytics (such as pandas and numPy), NLP (nlTK and spaCy), sentiment analysis (e.g. VADER), and the Reddit API interaction (e.g. praw, etc).

Minimizing dependency version conflicts and runtime errors through the use of dependency resolution mechanisms is a way to smooth the deployment process and put it on the right track. This entails the indication of unique version specifications for each dependency and automatic resolving every issues during installation. Also the caching of dependencies can be taken into consideration which will store and reuse resolved dependencies based on previously shipped and installed packages thus avoid bothersome of downloading and installing the same things again.

Having the option to do dependency pinning also poses the setting of flexibility in different

production environments. Via the definition of exact version numbers of the components in the requirements.txt or environment. Producers of software can design such environments with consistency which is crucial for the stable work and prevents wearing effects or unclear behavior. In addition, permanent and dynamic checks of the dependencies are the only way to address the related security issues and guarantee the stability and reliability of the application. To fill the gap, frequently checking for updates and patches which are security related for their dependencies could help to reduce the risks and keep the general system of the application secure. Automation is done with the use of dependency management tools and services, which shall efficiently give you a quick update and change in functionalities with less disruption to the application.

## **CHAPTER 6**

### **CHALLENGES AND SOLUTIONS**

#### **1 Technical Challenges:**

a. Data Acquisition and Management: One of the meaning technical obstacles that is needed to deal with and handle large amount of data promptly. Acquiring fast processed data in real time from platforms such as Reddit which is high throughput but low-latency is this fact that raw data can be received in time nevertheless, it should be restructured to fit the desired output format such as json.

b. Natural Language Processing (NLP) Complexity: The design and introduction of NLP innovations for emotion analysis, auto-correction, and text production was not an easy task. Through the challenge to grasp language intricacies like subtleties, slang, and context-dependent interpretation, the mini-projects were able to accurately and extensively achieve this goal.

c. Integration and Compatibility: Successfully putting together all of the technical pieces together and ensuring smooth integration of various APIs like PRAW for Reddit data fetching and Gemini API for multilingual text generation was a really restless task. To enhanced the overall performance there arose difficulties of interfacing the diverse systems and part of the results were that the stability of the application suffered.

#### **2. Solutions Implemented:**

a. Advanced Machine Learning Models: We put the NLP intricacies into our jobs using the best ML models and tools available to us. Taking advantage of such libraries as NLTK and spaCy, we come up with custom-trained models on a sizable dataset. This capability ultimately leads to high accuracy in text analysis tasks.

b. Robust Integration Practices: We were able to deal with integration issues by an introduction of the containerization architecture and the microservice. Docker containers and orchestrated deployments via Kubernetes are two approaches we are using in order to keep the environment of the different entities consistent and also for the easy deployment and integration process.

### 3. Lessons Learned:

a. Importance of Modular Design: The module-based design method of architecture was among the techniques that were pointed out as what was learnt. Such design approach contributed greatly to the development team that with this philosophy they were able to step by step develop, see the development, and extend the portal. Parts of the system could be faulty but when refurbished or replaced would not need to the operations being carried out stopped.

b. Continuous Testing and Monitoring: Continuous running and monitoring of this application became the cornerstone of the substitutional scenario where this application was run with elimination of the bugs. We adopted the Test and Monitoring automation technology that provided us with quick detection of issues thus reducing the downtime hence improving our customers' satisfaction.

c. User-Centric Development: Making conversation with customers right after they use our product, and before they share their opinion with others, was a crucial part of our strategy. Both real life users and technical testing's were key elements of the process all the way through the building process since they could reveal things that technical tests could not identify. This technique had a dual benefit: It made the process for filling the user requirements and expectations easier, which in turn resulted in a product that was more user-oriented in nature.

## CHAPTER 7

### RESULTS AND DISCUSSION

#### 1. Analysis of Sentiment Trends:1. Analysis of Sentiment Trends:

a. Sentiment Tracking Over Time: We obtained insights into potential public mood and attitude changes as per the evolved sentiment tendencies in relation to different subjects discussed on Reddit. We can capture the sentiment of data with time-series analysis and those discussions or events that were linked to unusual changes in sentiment these events can then be isolated. This gave some intrigues about the pattern and time of public reactions related to what truly happened in the world surrounding.

b. Impact of Sentiment on User Engagement: The searched results indicated an association between sentiment trends and user engagement factors like upvotes, comments, and shares. The posts which mainly had positive emotions did trigger growing user activity as studies showed. Therefore, the providers could increase the number of people taking an active part by understanding and reacting promptly to the trend in sentiment.

c. Sentiment Distribution Across Topics: What is obvious is that different topics and conversation about themes presented different sentiment distribution. One way to draw this conclusion is that there was an obvious difference in ratings by when firstly entertainment related topic was sentimental and when later on political discussions were posted. These arrays of results generate novel ideas both for creating message and marketing that fit into the emotional architecture of the preferred customer segment.

## 2. Usefulness of Text Generation:

a. Enhanced Content Creation: The application of the Gemini API for the generation of the auto text proved to be exceptional in producing diverse and multi-lingual content. The system could give people the power to create texts in multiple languages, expanding the reach of the podcast. Thus, it not only did away with man hours but also encouraged creativity, letting out ideas and drafts automatically.

b. Cross-Language Barriers Overcome: Text generating techniques became particularly important to solving the language barrier problem. The feature would enable users to search posts in one language and generation in another, thus encouraging inclusiveness among the users of different linguistic background. This functionality was increasing the platform's variety and making it not only good for domestic, but for international market.

c. Practical Applications in Marketing and Engagement: In the real life applications, marketing, and customer engagement areas, auto-generated texts were taken into consideration. Automating social media responses and generating creative ideas, the businesses could keep their social media platforms alive, interactive, and attractive, thus increasing their visibility as well as the user engagement.

## CHAPTER 8

### CONCLUSION

#### 1. Summary of Achievements:

a. Innovative Integration of Technologies: The project successfully incorporated some cutting-edge technologies among them Reddit API, Gemini text-automation API and natural language tools to deal with and interpret social media data. Thus, the integration resulted in a detail adaptation, hence facilitating the attainment of the goals.



b. Development of a User-Centric Web Application: Thanks to our development of Streamlit platform we could create the web-application that is really easy to use and provides a fast real-time data analyzing and visualizing. The facility to analyze and visualize in an easy to understand form highly complex data was a prominent breakthrough that directly improved the user experience repeatedly.

c. Advancement in Text Analysis Capabilities: Our project set an important milestone for the text analyzation by implementation of real-time auto-correction and sentiment sensing. They feature multi-faceted aspects that help the users to discover the intentions from a large quantity of text data thus initiating a revolution in text analytics.

## 2. Results

Enhanced Access to Information: Customers can now parse through Reddit data through our application to allow them to obtain knowledge through the means of simple user interface and input, which leads to more correct and wider perception of the world.

Insightful Analysis: The production of information which represents trending topics and the trend of reddit discussions becomes possible by making use of data visualization. it is then a subject of study that helps researchers, marketers, and policymakers to understand trend public opinion and trends.

Interactive Engagement: Finally, by combining Gemini API and UI so that the user gets all the insights without leaving the network, ZOD will provide the desired know-yourself experience filled with entertainment and coziness.

Quality Assurance: Promoting data accuracy and credibility is guaranteed with automatic sentence correction feature, because it becomes useful and will be effective in decision-making after needed in-depth investigation and explanation.

Social Impact: The project we are carrying will be useful to the extent of helping to control over a certain sentiment and consequently cultivate an internet expression that is free of toxicity. This could be the possibility to solve the problems like cyberbullying and fake news.

Therefore, the system that we have devised is a tremendous breakthrough in the science of social media text analytics that will enable researchers, analysts, and community leaders to magnificently dig deep into the data, and instantly form valuable perspectives about the virtual communities across the globe.

## REFERENCES:

### i. API Documentation:

The API and documentation disseminated by Reddit contains endpoints (or entry points) as well as the methods via which these endpoints can be used to perform various activities on Reddit (post or get data). You can find it at [https://www.reddit.com/dev/api/\(stat\)](https://www.reddit.com/dev/api/(stat)).

### ii. Libraries and Frameworks:

PRAW (Python Reddit API Wrapper): Python Library that has the feature easy to find, so accessing to the Reddit API can be made. It would cover all the stuffs that I have mentioned such as posting, commenting and also these bits of user information. You can find it at <https://praw.readthedocs.io/en/latest/> does not offer specific instructions, but it is a comprehensive guide to programming.

### iii. SnooWrap:

Internet technology has made things easier by allowing people to communicate faster like a skincare app developer who uses Python scripts. Programmable Implementation is similar to writing a file with Python codes, thus simplifying the process and making it uniform.

### iv. Articles and Online Resources:

Inside the text, an overview of the Python text processing basics will be provided as well as sentiment analysis, text clustering, and other extras.

#### Analyzing Reddit Data:

The Best Practices: Elimination and Analysis of Reddit Data: An Example of Python Tools at Work.

Here this tutorial will tell how data is taken from reddit and then can be analyzed using python libraries such as pandas and PRAW. You can find it at [https:-\(https://www.datacamp.com/community/tutorials\) /scraping-reddit-with-python-and-scrapy"](https://www.datacamp.com/community/tutorials/scraping-reddit-with-python-and-scrapy).

Sentiment Analysis on Reddit News Headlines with Python's Natural Language Toolkit (NLTK): Shown as the next sentence is a guide that helps to sentimental analysis of the Reddit headlines using the NLTK in Python. You can find it at [https:https://datacamp.com/community/tutorials/sentiment-analysis-python](https://datacamp.com/community/tutorials/sentiment-analysis-python)

# PROJECT CODE

GITHUB LINK:

<https://github.com/annepally/reddit-sentiment-analysis/tree/main>

```
In [1]: pip install praw
```

```
Collecting praw
  Downloading praw-7.7.1-py3-none-any.whl (191 kB)
      191.0/191.0 kB 2.9 MB/s eta 0:00:0
Collecting prawcore<3,>=2.1 (from praw)
  Downloading prawcore-2.4.0-py3-none-any.whl (17 kB)
```

```
pip install vaderSentiment
```

```
Collecting vaderSentiment
  Downloading vaderSentiment-3.3.2-py2.py3-none-any.whl (125 kB)
      0.0/126.0 kB ? eta -:--:--
```

```
import praw
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation
from wordcloud import WordCloud
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
import warnings
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import SVC
from sklearn.metrics import classification_report
```

```
reddit = praw.Reddit(client_id='fpqm-JgqdYpiAmZodSh8Pw',
                      client_secret='LCrn_D_tPMfn1_uj3pxVSXz_gWjjZw',
                      redirect_uri='http://localhost:8080',
                      user_agent='gojoinfinity1')
```

```

subreddit_input = input("Enter the subreddit(s) you want to fetch data from (comma-separated): ")

# Convert user input to a list of subreddits
subreddits = [sub.strip() for sub in subreddit_input.split(',')]

# Initialize post dataframe
posts_df = []

# Fetch data for each subreddit
for subreddit_name in subreddits:
    subreddit = reddit.subreddit(subreddit_name)
    posts = subreddit.top(time_filter="all", limit=10000)

    for post in posts:
        posts_df.append({
            'post_id': post.id,
            'subreddit': post.subreddit.display_name,
            'created_utc': post.created_utc,
            'selftext': post.selftext,
            'post_url': post.url,
            'post_title': post.title,
            'link_flair_text': post.link_flair_text,
            'score': post.score,
            'num_comments': post.num_comments,
            'upvote_ratio': post.upvote_ratio
        })

# Create a DataFrame from the collected data
df = pd.DataFrame(posts_df)
print(df)

```

```
In [6]: df['created_utc'] = pd.to_datetime(df['created_utc'], unit='s')
```

```
In [7]: df.head()
```

```
Out[7]:
```

	post_id	subreddit	created_utc	selftext	post_url	post_title	link_flair_text	score	num_comr
0	k8nyf8	datascience	2020-12-07 19:49:55		https://d5Intlv9vhjr4.cloudfront.net/posts_ima...	data siens	Fun/Trivia	4111	
1	oeg6nl	datascience	2021-07-05 20:57:20		https://i.redd.it/yqnunwryjg971.jpg	The pain and excitement	Fun/Trivia	3909	
2	hohvgq	datascience	2020-07-10 03:45:31	I've been lurking on this sub for a while now ...	https://www.reddit.com/r/datascience/comments/...	Shout Out to All the Mediocre Data Scientists ...	Discussion	3622	
3	xdv6nz	datascience	2022-09-14 07:11:15		https://i.redd.it/k102dyo0yrn91.jpg	Let's keep this on...	Fun/Trivia	3600	
4	tj3kek	datascience	2022-03-21 04:34:37		https://i.imgur.com/TAex5zG.jpg	Guys, we've been doing it wrong this whole time	Meta	3468	

In [10]: `df.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 989 entries, 0 to 988
Data columns (total 10 columns):
#   Column              Non-Null Count  Dtype
---  -
0   post_id              989 non-null    object
1   subreddit            989 non-null    object
2   created_utc          989 non-null    datetime64[ns]
3   selftext             989 non-null    object
4   post_url             989 non-null    object
5   post_title           989 non-null    object
6   link_flair_text      887 non-null    object
7   score               989 non-null    int64
8   num_comments         989 non-null    int64
9   upvote_ratio         989 non-null    float64
dtypes: datetime64[ns](1), float64(1), int64(2), object(6)
memory usage: 77.4+ KB
```

In [11]: `df.describe()`

```
Out[11]:
```

	created_utc	score	num_comments	upvote_ratio
count	989	989.000000	989.000000	989.000000
mean	2021-08-18 08:34:19.971688704	628.576340	115.917088	0.949232
min	2015-09-24 14:09:16	262.000000	3.000000	0.660000
25%	2020-06-16 15:51:02	335.000000	55.000000	0.940000
50%	2021-10-13 06:37:30	441.000000	96.000000	0.960000
75%	2022-10-22 18:37:04	688.000000	154.000000	0.980000

In [13]: `df.isnull()`

```
Out[13]:
```

	post_id	subreddit	created_utc	selftext	post_url	post_title	link_flair_text	score	num_comments	upvote_ratio
0	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...
984	False	False	False	False	False	False	True	False	False	False
985	False	False	False	False	False	False	False	False	False	False
986	False	False	False	False	False	False	False	False	False	False
987	False	False	False	False	False	False	False	False	False	False
988	False	False	False	False	False	False	False	False	False	False

989 rows × 10 columns

```

In [14]: df.isnull().sum()

Out[14]: post_id      0
subreddit    0
created_utc  0
selftext     0
post_url     0
post_title   0
link_flair_text  102
score        0
num_comments 0
upvote_ratio 0
dtype: int64

In [15]: df.isnull().sum().sum()

Out[15]: 102

In [16]: data = df.fillna("data not available")
data

Out[16]:
```

	post_id	subreddit	created_utc	selftext	post_url	post_title	link_flair_text	score
0	k8nyf8	datascience	2020-12-07 19:49:55		https://dslntlv9vhjr4.cloudfront.net/posts_ima...	data siens	Fun/Trivia	4111
1	oeg6nl	datascience	2021-07-05 20:57:20		https://i.redd.it/yqnunwryjg971.jpg	The pain and excitement	Fun/Trivia	3909
2	hohvgq	datascience	2020-07-10 03:45:31	I've been lurking on this sub for a while	https://www.reddit.com/r/datascience/comments/...	Shout Out to All the Mediocre Data Scientists ...	Discussion	3622

```

]: df.isnull().sum().sum()

]: 102

]: data = df.fillna("data not available")
data

]:
```

	post_id	subreddit	created_utc	selftext	post_url	post_title	link_flair_text	score
0	k8nyf8	datascience	2020-12-07 19:49:55		https://dslntlv9vhjr4.cloudfront.net/posts_ima...	data siens	Fun/Trivia	4111
1	oeg6nl	datascience	2021-07-05 20:57:20		https://i.redd.it/yqnunwryjg971.jpg	The pain and excitement	Fun/Trivia	3909
2	hohvgq	datascience	2020-07-10 03:45:31	I've been lurking on this sub for a while now ...	https://www.reddit.com/r/datascience/comments/...	Shout Out to All the Mediocre Data Scientists ...	Discussion	3622
3	xdv6nz	datascience	2022-09-14 07:11:15		https://i.redd.it/k102dyo0yrn91.jpg	Let's keep this on...	Fun/Trivia	3600
4	tj3kek	datascience	2022-03-21 04:34:37		https://i.imgur.com/TAex5zG.jpg	Guys, we've been doing it wrong this whole time	Meta	3468
...	...	...	...	...	...	...	...	...

989 rows × 10 columns

```
4]: X = data['text']
    y = data['link_flair_text']

5]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

6]: vectorizer = TfidfVectorizer(max_features=5000) # You can adjust max_features as needed
    X_train_vectors = vectorizer.fit_transform(X_train)
    X_test_vectors = vectorizer.transform(X_test)

8]: svm_classifier = SVC(kernel='linear')
    svm_classifier.fit(X_train_vectors, y_train)

8]: SVC(kernel='linear')
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

]: y_pred = svm_classifier.predict(X_test_vectors)

]: print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
Career	0.38	0.17	0.23	30
Career Discussion	0.00	0.00	0.00	3
Challenges	0.00	0.00	0.00	1
Discussion	0.42	0.92	0.58	64
Education	0.50	0.07	0.12	14
Fun/Trivia	0.42	0.61	0.50	28
Job Search	0.50	0.08	0.14	12
Meta	0.00	0.00	0.00	7
Networking	0.00	0.00	0.00	1
Projects	0.00	0.00	0.00	11
Tooling	0.00	0.00	0.00	6
data not available	0.00	0.00	0.00	21
accuracy			0.42	198
macro avg	0.19	0.15	0.13	198
weighted avg	0.32	0.42	0.31	198

```

In [42]: joblib.dump(svm_classifier, 'svm_model.joblib')

Out[42]: ['svm_model.joblib']

In [43]: joblib.dump(vectorizer, 'vectorizer.joblib')

Out[43]: ['vectorizer.joblib']

In [44]: import pandas as pd
from sklearn.model_selection import train_test_split
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
from sklearn.metrics import classification_report

In [45]: analyzer = SentimentIntensityAnalyzer()

In [46]: def get_sentiment(text):
    scores = analyzer.polarity_scores(text)
    if scores['compound'] >= 0.05:
        return 'positive'
    elif scores['compound'] <= -0.05:
        return 'negative'
    else:
        return 'neutral'

In [47]: # Apply the get_sentiment function to each text in the DataFrame
data['sentiment'] = data['text'].apply(get_sentiment)

In [48]: print(classification_report(data['link_flair_text'], data['sentiment']))

In [48]: print(classification_report(data['link_flair_text'], data['sentiment']))

```

	precision	recall	f1-score	support
AI	0.00	0.00	0.00	1.0
Analysis	0.00	0.00	0.00	3.0
Can we impute it?	0.00	0.00	0.00	1.0
Career	0.00	0.00	0.00	148.0
Career Discussion	0.00	0.00	0.00	20.0
Challenges	0.00	0.00	0.00	1.0
Coding	0.00	0.00	0.00	1.0
Discussion	0.00	0.00	0.00	324.0
Education	0.00	0.00	0.00	66.0
Fun/Trivia	0.00	0.00	0.00	137.0
Job Search	0.00	0.00	0.00	72.0
Let's Discuss This	0.00	0.00	0.00	1.0
ML	0.00	0.00	0.00	1.0
Meta	0.00	0.00	0.00	27.0
Monday Meme	0.00	0.00	0.00	5.0
Networking	0.00	0.00	0.00	6.0
Projects	0.00	0.00	0.00	42.0
Tooling	0.00	0.00	0.00	30.0
Tools	0.00	0.00	0.00	1.0
data not available	0.00	0.00	0.00	102.0
negative	0.00	0.00	0.00	0.0
neutral	0.00	0.00	0.00	0.0
positive	0.00	0.00	0.00	0.0
accuracy			0.00	989.0
macro avg	0.00	0.00	0.00	989.0
weighted avg	0.00	0.00	0.00	989.0



```

from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

# Initialize the VADER SentimentIntensityAnalyzer
analyzer = SentimentIntensityAnalyzer()

# Sample text to test sentiment
sample_text = "i dont want to study"

# Get sentiment scores using VADER
scores = analyzer.polarity_scores(sample_text)

# Determine sentiment label based on compound score
if scores['compound'] >= 0.05:
    sentiment = 'positive'
elif scores['compound'] <= -0.05:
    sentiment = 'negative'
else:
    sentiment = 'neutral'

print("Sample text:", sample_text)
print("Sentiment label:", sentiment)
print("Sentiment scores:", scores)

```

Sample text: i dont want to study  
Sentiment label: negative  
Sentiment scores: {'neg': 0.234, 'neu': 0.766, 'pos': 0.0, 'compound': -0.0572}

```

joblib.dump(analyzer, 'analyzer.joblib')

```

```
['analyzer.joblib']
```

```

pip install -q -U google-generativeai

```

```

import pathlib
import textwrap

import google.generativeai as genai

from IPython.display import display
from IPython.display import Markdown

def to_markdown(text):
    text = text.replace('•', ' *')
    return Markdown(textwrap.indent(text, '> ', predicate=lambda _: True))
# Used to securely store your API key
from google.colab import userdata

```

```

genai.configure(api_key='AIzaSyCaZ6FjKaz6nc3dnjaNgEYQJdYjBOISDuI')

```

```

for m in genai.list_models():
    if 'generateContent' in m.supported_generation_methods:
        print(m.name)

```

```

models/gemini-1.0-pro
models/gemini-1.0-pro-001
models/gemini-1.0-pro-latest
models/gemini-1.0-pro-vision-latest
models/gemini-1.5-pro-latest
models/gemini-pro
models/gemini-pro-vision

```

```

model = genai.GenerativeModel('gemini-pro')

```

```
%%time
response = model.generate_content("Machine Learning")
response.text
```

CPU times: user 158 ms, sys: 14.5 ms, total: 173 ms

Wall time: 11.5 s

```
'**Introduction**\n\nMachine learning (ML) is a subfield of artificial intelligence (AI) that enables computers to learn from data without explicit programming. It allows computers to identify patterns, make predictions, and perform tasks that would normally require human intervention.\n\n**Types of Machine Learning**\n\n* **Supervised Learning:** Algorithms learn from labeled data (input data has known outputs) and make predictions on new data. Common algorithms include:\n    * Linear regression\n    * Logistic regression\n    * Decision trees\n    * Support vector machines\n* **Unsupervised Learning:** Algorithms learn from unlabeled data (input data has no known outputs) and identify patterns or structures. Common algorithms include:\n    * Clustering\n    * Dimensionality reduction\n    * Anomaly detection\n* **Reinforcement Learning:** Algorithms learn through interactions with an environment and receive rewards or penalties, adjusting their behavior based on feedback. Common algorithms include:\n    * Q-learning\n    * Policy gradients\n    * Deep reinforcement learning\n\n**Applications of Machine Learning**\n\nML has a wide range of applications in various industries, including:\n\n* **Healthcare:** Disease diagnosis, drug discovery, personalized medicine\n* **Finance:** Fraud detection, credit scoring, investment analysis\n* **Manufacturing:** Predictive maintenance, quality control, supply chain optimization\n* **Transportation:** Autonomous vehicles, traffic management, route optimization\n* **Customer Service:** Chatbots, recommendation systems, sentiment analysis\n* **Media and Entertainment:** Content recommendation, image processing, video analysis\n\n**Techniques**\n\n* **Deep Learning:** A class of ML algorithms that use neural networks with multiple hidden layers.\n* **Natural Language Processing (NLP):** ML techniques for understanding, generating, and manipulating human language.\n* **Computer Vision:** ML techniques for processing and interpreting visual information.\n* **Time Series Analysis:** ML techniques for analyzing and forecasting data over time.\n\n**Benefits of Machine Learning**\n\n* **Automation:** Automates tasks that would otherwise require manual labor.\n* **Efficiency:** Improves efficiency by optimizing processes and reducing errors.\n* **Predictive Analytics:** Provides insights and predictions to help make data-driven decisions.\n* **Customization:** Allows for personalized experiences and tailored services.\n\n**Challenges**\n\n* **Data Quality:** The quality of data used for training ML algorithms is crucial for accurate results.\n* **Bias:** ML algorithms can inherit biases from the data they are trained on.\n* **Interpretability:** Understanding how ML algorithms make predictions can be challenging, which can limit their acceptance and adoption.'
```

```
response.prompt_feedback
```

```

1  import streamlit as st
2  import pandas as pd
3  import praw
4  import os
5  from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
6  from autocorrect import Speller
7  import google.generativeai as genai
8  import textwrap
9  from dotenv import load_dotenv
10
11  # Load environment variables
12  load_dotenv()
13
14  # Initialize Reddit API client
15  reddit = praw.Reddit(client_id='fpqm-JgqdYpiAmZodSh8Pw',
16                      client_secret='LCrn_D_tPMfnl_uj3pxVSXz_gWjjZw',
17                      redirect_uri="http://localhost:8080",
18                      user_agent='gojoinfinity1')
19
20  # Load VADER Sentiment Analyzer
21  analyzer = SentimentIntensityAnalyzer()
22
23  # Configure genai
24  genai.configure(api_key=os.getenv("GOOGLE_API_KEY"))
25
26  # Helper function to format text as Markdown
27  def to_markdown(text):
28      text = text.replace('•', ' *')
29      return textwrap.indent(text, '> ', predicate=lambda _: True)
30

```

```

# Function to fetch response from Gemini API
def get_gemini_response(question):
    model = genai.GenerativeModel('gemini-pro')
    response = model.generate_content(question)
    return response.text

# Function to perform Auto Text Correction
def auto_text_correction():
    st.title("Auto Text Correction")
    spell = Speller(lang='en')
    user_input = st.text_input("Enter your text:")
    if st.button("Correct Text"):
        corrected_text = spell(user_input)
        st.write("Corrected Text:", corrected_text)

# Function to perform Reddit Sentiment Analysis
def reddit_sentiment_analysis():
    st.title('Reddit Sentiment Analysis')
    emojis = {'positive': '😊', 'negative': '😞', 'neutral': '😐'}
    user_input = st.text_area("Enter your text here:")
    if st.button('Predict Sentiment'):
        sentiment = predict_sentiment_vader(user_input)
        emoji = emojis[sentiment]
        st.write("Predicted sentiment:", sentiment, emoji)

# Function to predict sentiment using VADER
def predict_sentiment_vader(text):
    # Preprocess the text
    processed_text = text
    # Analyze sentiment using VADER
    # Analyze sentiment using VADER
    scores = analyzer.polarity_scores(processed_text)
    # Determine sentiment label based on compound score
    if scores['compound'] >= 0.05:
        return 'positive'
    elif scores['compound'] <= -0.05:
        return 'negative'
    else:
        return 'neutral'

# Streamlit interface for Auto Text Generation
def auto_text_generation():
    st.title("Auto Text Generation")
    input_text = st.text_input("Input: ", key="input")
    submit = st.button("Ask the question")
    if submit:
        response = get_gemini_response(input_text)
        st.subheader("The Response is")
        st.write(response)

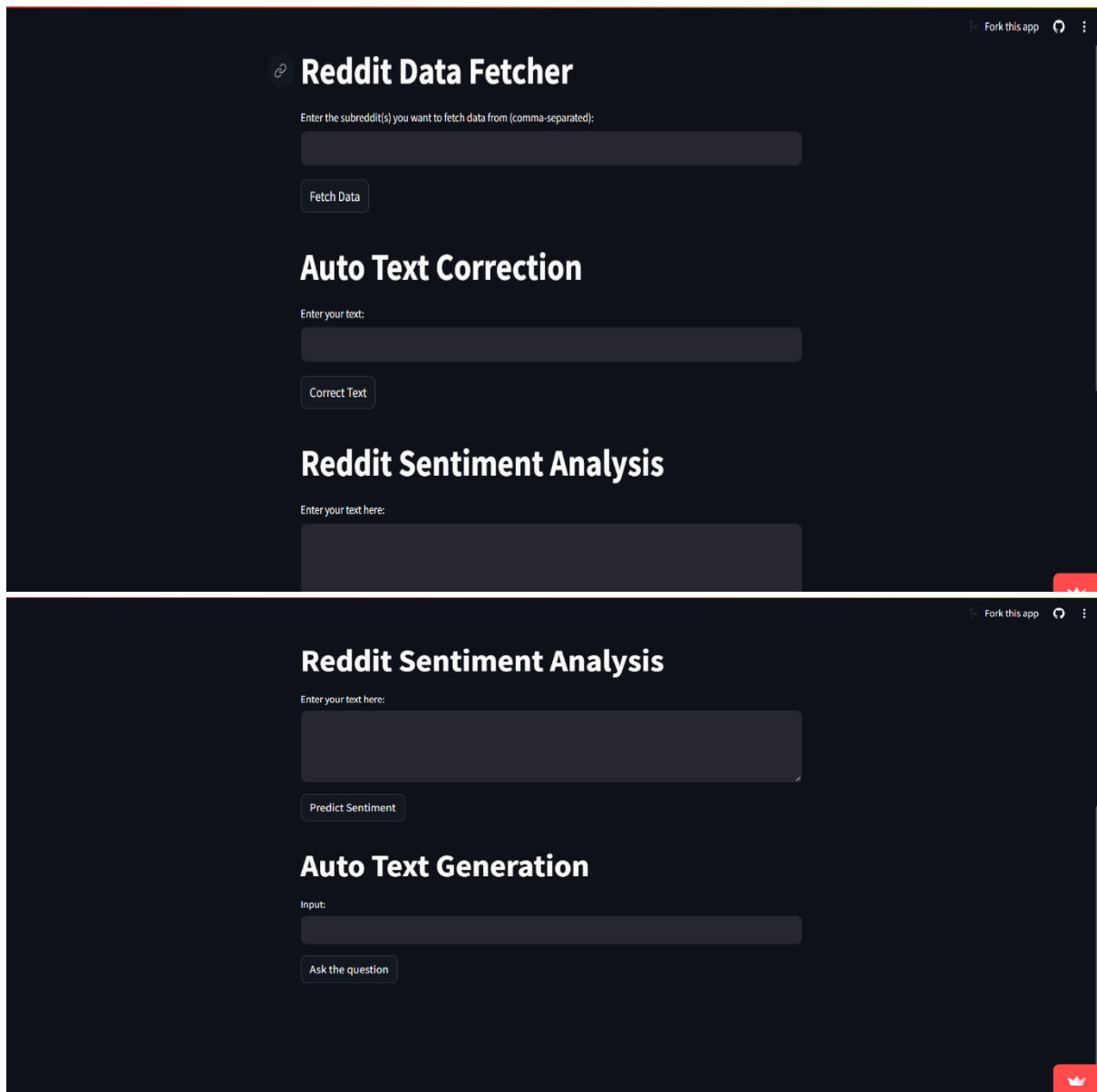
# Streamlit interface for Reddit Data Fetcher
def reddit_data_fetcher():
    st.title('Reddit Data Fetcher')
    subreddit_input = st.text_input("Enter the subreddit(s) you want to fetch data from (comma-separated): ")
    if st.button('Fetch Data'):
        subreddits = [sub.strip() for sub in subreddit_input.split(',')]
        if subreddits:
            posts = []
            for subreddit_name in subreddits:

```

```

89         subreddit = reddit.subreddit(subreddit_name)
90         for post in subreddit.top(time_filter="all", limit=100):
91             posts.append({
92                 'post_id': post.id,
93                 'subreddit': post.subreddit.display_name,
94                 'created_utc': post.created_utc,
95                 'selftext': post.selftext,
96                 'post_url': post.url,
97                 'post_title': post.title,
98                 'link_flair_text': post.link_flair_text,
99                 'score': post.score,
100                'num_comments': post.num_comments,
101                'upvote_ratio': post.upvote_ratio
102            })
103        df = pd.DataFrame(posts)
104        st.write("Data Fetched Successfully!")
105        st.dataframe(df) # Display the dataframe in the app
106    else:
107        st.write("Please enter at least one subreddit.")
108
109    # Main function to run the Streamlit app
110    def main():
111        st.set_page_config(page_title="Social Media Text Analysis")
112
113        reddit_data_fetcher() # Reddit Data Fetcher at the top
114
115        auto_text_correction()
116        reddit_sentiment_analysis()
117        auto_text_generation()
118
119    # Run the app
120    if __name__ == "__main__":
121        main()
122

```



THANK YOU

## Checklist for Dissertation-III Supervisor

Name: \_\_\_\_\_ UID: \_\_\_\_\_ Domain: \_\_\_\_\_

Registration No: 12017831

Name of student: Annepally Sanjay Kumar

Title of Dissertation:

SOCIAL MEDIA SENTIMENT ANALYSIS ON REDDIT

- 
- ☐ Front pages are as per the format.
  - ☐ Topic on the PAC form and title page are same.
  - ☐ Front page numbers are in roman and for report, it is like 1, 2, 3.....
  - ☐ TOC, List of Figures, etc. are matching with the actual page numbers in the report.
  - ☐ Font, Font Size, Margins, line Spacing, Alignment, etc. are as per the guidelines.
  - ☐ Color prints are used for images and implementation snapshots.
  - ☐ Captions and citations are provided for all the figures, tables etc. and are numbered and center aligned.
  - ☐ All the equations used in the report are numbered.
  - ☐ Citations are provided for all the references.
  - ☐ **Objectives are clearly defined.**
  - ☐ Minimum total number of pages of report is 50.
  - ☐ Minimum references in report are 30.

Here by, I declare that I had verified the above-mentioned points in the final dissertation report.

Signature of Supervisor with UID