

Sentiment Analysis Project Final Report



Date

19/04/2024

Supervisor

Greg Doyle

Student

Mantas Macionis (C00242178)

Academic Year

2023/2024

Contents

Date	1
19/04/2024.....	1
Supervisor	1
Greg Doyle	1
Student	1
Mantas Macionis (C00242178)	1
Academic Year	1
Introduction	3
Project Overview	3
Project Objectives	4
Project Structure	4
Landing	4
Analysis	6
History.....	8
User Interaction Diagram	10
Use Case Diagram.....	10
Project Outcomes	10
Challenges Encountered	11
Limitations / Not Achieved	12
Achievements.....	12
What I Would Change	13
Learning Outcomes.....	13
Conclusion	14
Acknowledgements	15

Introduction

This report will provide a comprehensive documentation of the product developed for my Final Year Project. It offers an overview of the significant challenges encountered, the limitations faced, the achievements realized, and concludes with a reflective analysis of the entire experience. Additionally, the report will outline the project's success and suggest areas for improvement.

Project Overview

The Reddit Sentiment Analysis Tool is a web application which allows users to choose a search term, it then retrieves reddit comments related to that search term and outputs a sentiment analysis. The sentiment analysis is conducted both with a machine learning model and with the ChatGPT api. Allowing the user to compare the results achieved by both methods. For the AI portion of the analysis, A user can choose if they would like their searchterm and related comments analysed by premade ai prompts which differ in their questioning style, or if they would like to attempt generating their own prompt which will be based off of the search term they chose, and optionally, extra emphasis terms which they specify, when a user gets their sentiment analysis result, they get displayed the following info:

- The terms they specified: Search Term, subreddit, sort order, time filter, comment sort order
- **In the Traditional (Machine learning) analysis:**
- The Overall sentiment label for the text analysed
- The percentage of positive comments
- The number of comments total analysed
- **In the ChatGPT api Analysis:**
- A quick sentiment summary label for the whole text analysed
- A detailed Analysis, the style of which will depend on which prompt the user chose
- **Visualisations:**
- A pie chart showing the percentages of positive and negative comments
- A word cloud showing the most common words encountered in the analysis

Project Objectives

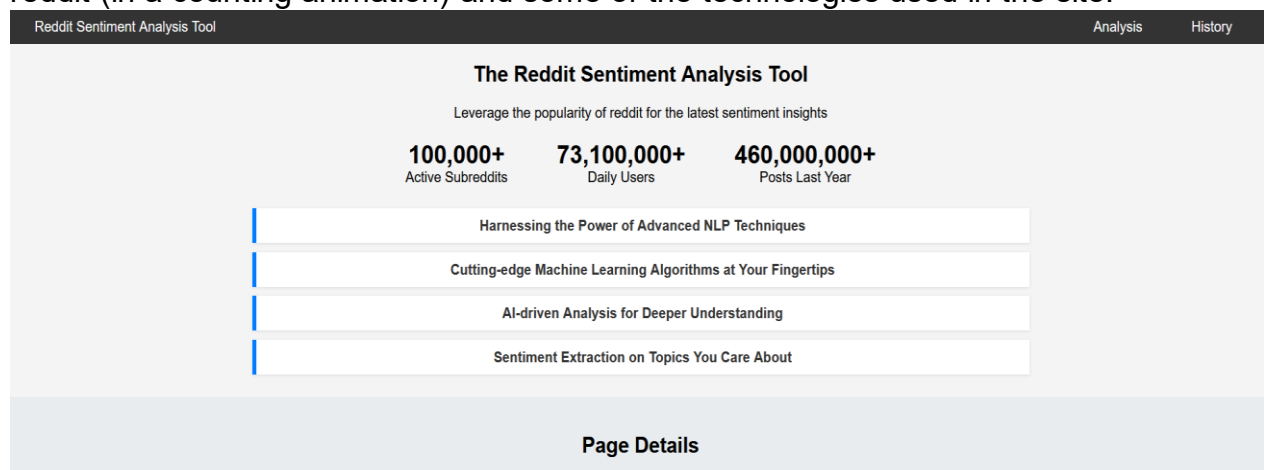
The objectives of the project included:

- Retrieving social media comments using an API.
- Preprocessing these comments to prepare them for analysis.
- Training a machine learning model to classify comments according to their sentiment.
- Creating tailored machine learning models for different types of information e.g. Financial or political commentary.
- Deploy the model to a site for use in the webapp implementation.
- Using an AI model to analyse the sentiment of retrieved comments, to see if the analysis provided is beneficial over a machine learning model.
- Prepare the analysis results in a visually appealing way.
- Generate visualisations based on the analysis results.
- Create a history page for users to view their previous searches conveniently.
- Create sentiment through time visualisations for users to visualise how sentiment has shifted on their topics of choice.

Project Structure

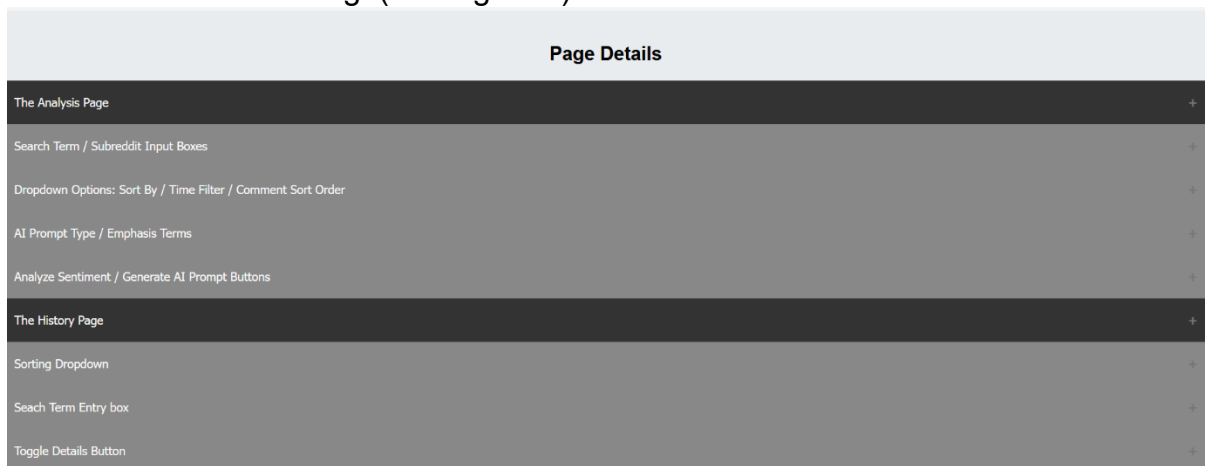
Landing

The user begins in the landing page(landing.html), they are greeted with info about reddit (in a counting animation) and some of the technologies used in the site.

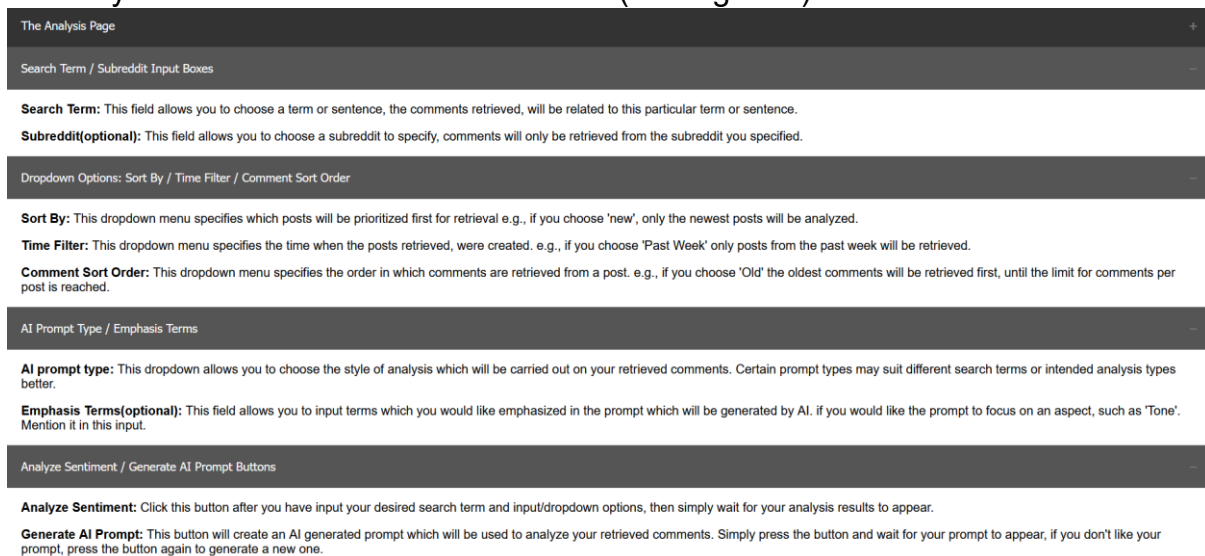


Reddit Sentiment Analysis Final Report

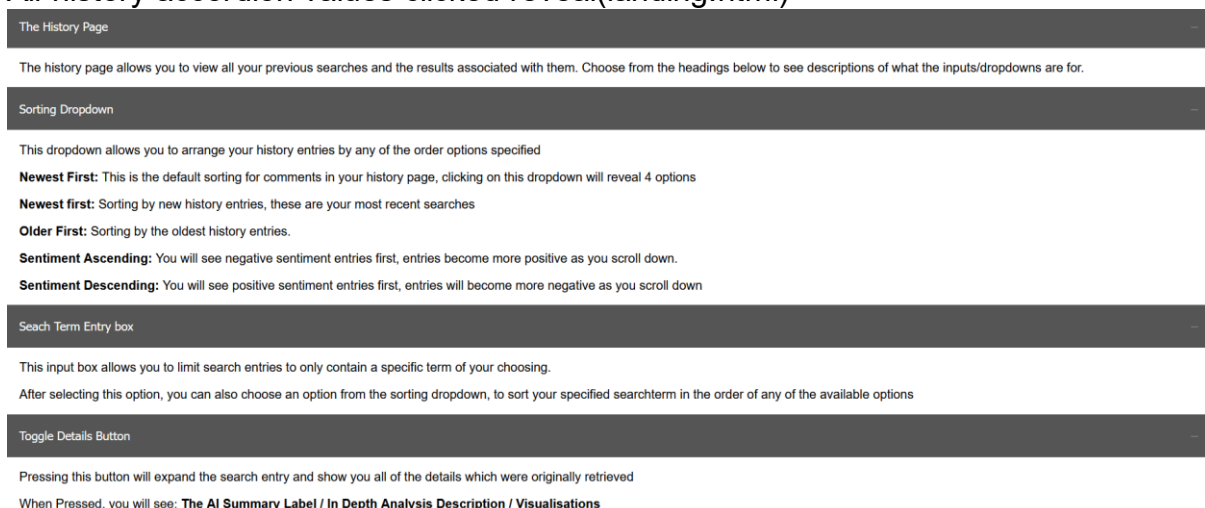
The user can scroll down and look at descriptions of elements in the site, in a clickable accordion design(landing.html)



All analysis accordion values clicked reveal(landing.html)



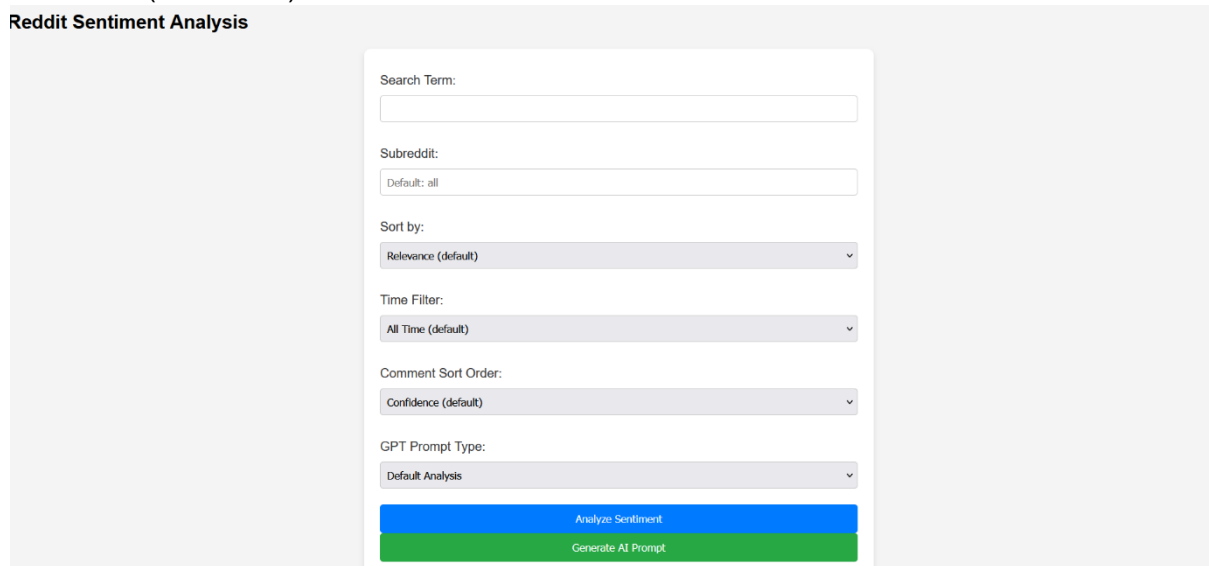
All history accordion values clicked reveal(landing.html)



Analysis

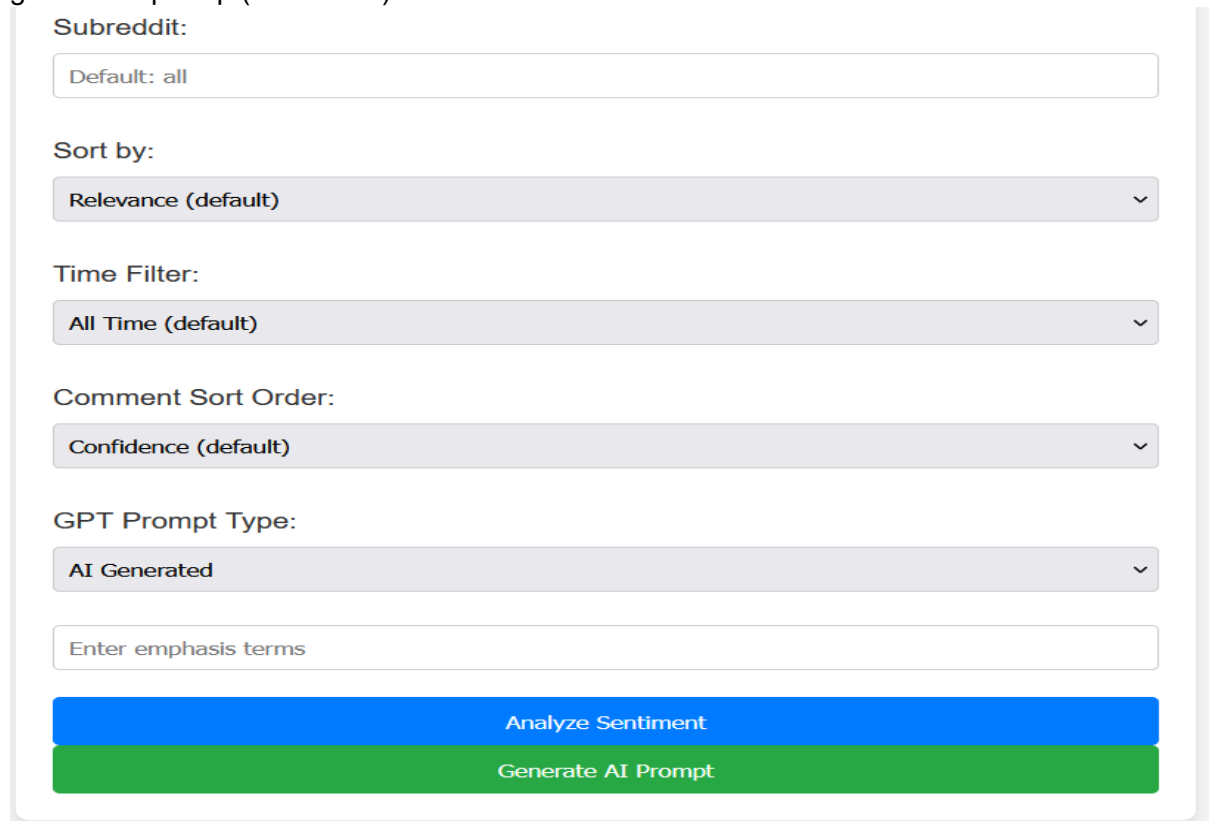
The user is greeted with an option list where they specify a searchterm and any dropdown options, **Within the sort by menu are options:** Relevance, New, Hot, Top. **Within the Time Filter menu are options:** All Time, Past Hour, Past Day, Past Week, Past Month, Past Year. **Within the comment sort order are options:** Confidence, Top, New, Controversial, Old, Random, Q&A. **Within the GPT prompt type are options:** Default analysis, Contextual Analysis, Emotional Analysis, Comparative Analysis, Impact Analysis, AI Generated(index.html).

Reddit Sentiment Analysis



The screenshot shows a web form titled "Reddit Sentiment Analysis". It contains several input fields and dropdown menus: "Search Term:" with a text input; "Subreddit:" with a text input containing "Default: all"; "Sort by:" with a dropdown menu showing "Relevance (default)"; "Time Filter:" with a dropdown menu showing "All Time (default)"; "Comment Sort Order:" with a dropdown menu showing "Confidence (default)"; and "GPT Prompt Type:" with a dropdown menu showing "Default Analysis". At the bottom of the form are two buttons: a blue "Analyze Sentiment" button and a green "Generate AI Prompt" button.

If the user specifies GPT Prompt Type 'AI Generated' They are greeted with an Emphasis terms input box, which allows them to input terms which will be emphasized in their generated prompt(index.html)



This screenshot shows the same form as above, but with the "GPT Prompt Type:" dropdown menu set to "AI Generated". Below this dropdown, a new text input field has appeared with the placeholder text "Enter emphasis terms". The "Analyze Sentiment" and "Generate AI Prompt" buttons remain at the bottom.

Reddit Sentiment Analysis Final Report

The user chooses to have an AI prompt generated while their search term is 'Joe Biden Speech'(index.html)

Message (optional)

GPT Prompt Type:
AI Generated

Enter emphasis terms

Analyze Sentiment

Generate AI Prompt

Generated Prompt: Prompt for sentiment analysis of comments related to Joe Biden's speech:
Please analyze a collection of comments and opinions regarding Joe Biden's recent speech. Focus on sentiments expressed towards the clarity of his message, effectiveness of his delivery, perceived authenticity, and overall impact on the audience. Provide insights into the varying emotions and attitudes conveyed in the comments, highlighting any recurring themes or contrasting viewpoints. Additionally, assess the level of positivity, negativity, or neutrality in the overall sentiment towards Joe Biden's speech.

The user searched for the term 'Joe Biden Speech' with the default GPT prompt style and has obtained their analysis. They are first greeted with info of their specified search(index.html)

Results For:

Search Term: **Joe Biden Speech**
Subreddit: **all** (if specified)
Sort Order: **Relevance**
Time Filter: **All**
Comment Sort Order: **Confidence**

The user scrolls down to look at the analysis results(index.html)

Traditional Sentiment Analysis:
Overall Sentiment: Neutral
Positive Percentage: 47.9%
Total Comments Analyzed: 1000

GPT-3.5 Turbo Sentiment Analysis:
Quick Sentiment Summary: **Highly Positive**
Analysis Details:

Positive Points:

- President Biden's passionate speech against Donald Trump resonated with many listeners.
- The speech highlighted Biden's commitment to upholding democracy and the rule of law.
- Biden's assertive stance and strong message were well-received, showing his leadership and resolve.
- Many individuals expressed appreciation for Biden's empathy, understanding, and genuine care for others.
- Biden's speech was commended for addressing important issues and challenging Trump's actions and behaviors.

Negative Points:

- Some comments referenced negative reactions from conservative individuals or Trump supporters to Biden's speech.
- There were mentions of complaints or criticisms despite the positive reception of Biden's speech.

The user scrolls down to look at their visualisations(index.html)

Analysis Visualisations:

Sentiment Distribution:

Sentiment	Percentage
Positive	47.9%
Negative	52.1%

Word Cloud of Comments:

History

The user has just arrived at the history page (history.html)

Reddit Sentiment Analysis tool Analysis Home

Search History

Newest First Search View All

Joe Biden Speech
Subreddit: all
Overall Sentiment: Neutral Date/Time: 19/04/2024 21:43
GPT Prompt Type: default
Toggle Details

Joe Biden Speech
Subreddit: all
Overall Sentiment: Neutral Date/Time: 19/04/2024 19:13
GPT Prompt Type: default
Toggle Details

The user chooses the 'Toggle Details' option for one of their entries(history.html)(pie chart partially not pictured)

Toggle Details

ChatGPT 3.5 Summary: Highly Positive

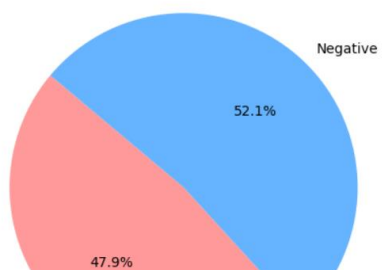
Analysis Description:

Positive Points:

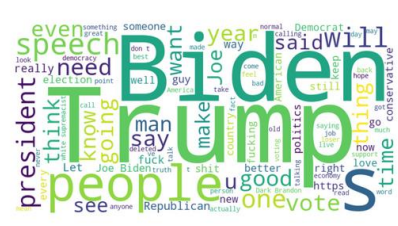
- President Biden's passionate speech against Donald Trump resonated with many listeners.
- The speech highlighted Biden's commitment to upholding democracy and the rule of law.
- Biden's assertive stance and strong message were well-received, showing his leadership and resolve.
- Many individuals expressed appreciation for Biden's empathy, understanding, and genuine care for others.
- Biden's speech was commended for addressing important issues and challenging Trump's actions and behaviors.

Negative Points:

- Some comments referenced negative reactions from conservative individuals or Trump supporters to Biden's speech.
- There were mentions of complaints or criticisms despite the positive reception of Biden's speech.



Sentiment	Percentage
Negative	52.1%
Positive	47.9%



Reddit Sentiment Analysis Final Report

The user inputs a searchterm 'Leo Varadkar' in the input box, they are shown relevant history entries(history.html)

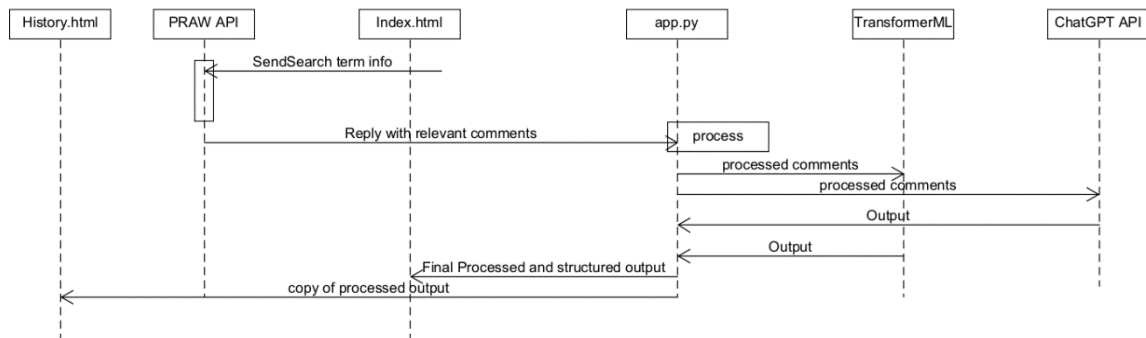
The screenshot shows the 'Search History' section of the 'Reddit Sentiment Analysis tool'. At the top, there is a navigation bar with 'Reddit Sentiment Analysis tool' on the left and 'Analysis' and 'Home' on the right. Below the navigation bar, the 'Search History' title is followed by a dropdown menu set to 'Newest First' and a search input field containing 'Leo Varadkar'. There are two buttons: 'Search' (blue) and 'View All' (grey). The search results are displayed in two white boxes. Each box contains the search term 'Leo Varadkar', the subreddit 'all', the overall sentiment 'Mostly Negative', and the GPT prompt type 'ai_generated'. The date and time for each entry is '19/04/2024 17:53' and '19/04/2024 17:47' respectively. Each entry has a 'Toggle Details' button.

The user chooses to sort their search term by 'Oldest First'(history.html)(note the date changes)

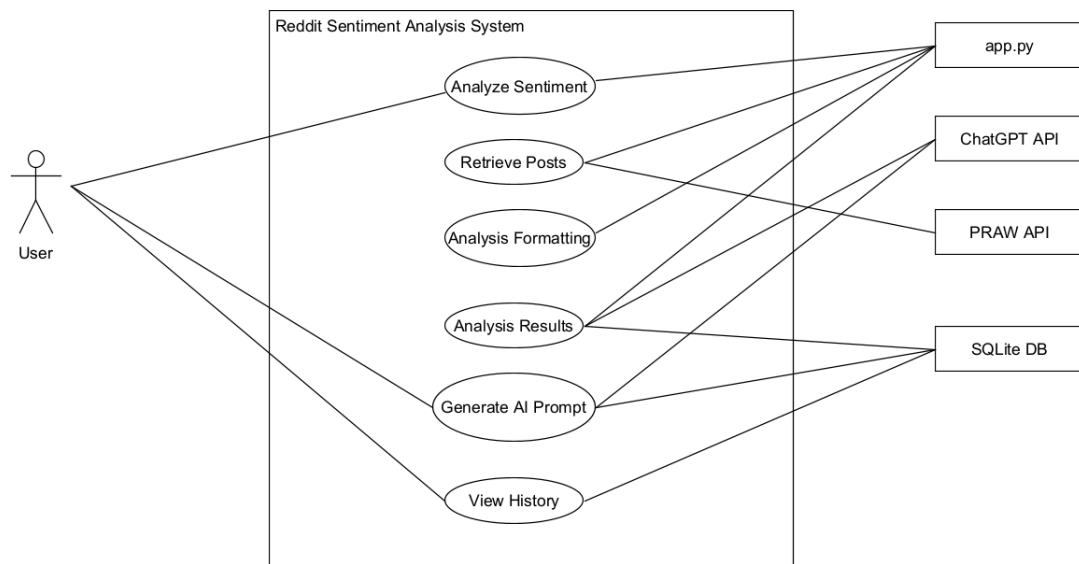
The screenshot shows the 'Search History' section of the 'Reddit Sentiment Analysis tool'. At the top, there is a navigation bar with 'Reddit Sentiment Analysis tool' on the left and 'Analysis' and 'Home' on the right. Below the navigation bar, the 'Search History' title is followed by a dropdown menu set to 'Oldest First' and a search input field containing 'Leo Varadkar'. There are two buttons: 'Search' (blue) and 'View All' (grey). The search results are displayed in two white boxes. Each box contains the search term 'Leo Varadkar', the subreddit 'Ireland', the overall sentiment 'Mostly Negative', and the GPT prompt type 'default'. The date and time for each entry is '18/04/2024 22:20' and '18/04/2024 22:21' respectively. Each entry has a 'Toggle Details' button.

User Interaction Diagram

The user chooses their search term and presses the 'Analyse Sentiment Button'



Use Case Diagram



Project Outcomes

The final product achieved the following:

- Successfully and efficiently retrieved reddit comments using the PRAW API.
- Pre-processed comments, using techniques such as stop word removal, lemmatization, unnecessary character removal.
- Successfully trained two machine learning models intended for use in the webapp, an LSTM and a transformer, both trained on a social media dataset of over 1.5 million comments.
- Deployed the transformer model to a flask website.

- Successfully implemented the OpenAI API into the webapp, specifically using gpt3.5 turbo for ai related functionality.
- Created two visualisations from the sentiment output – percentage of positive comments visualised, wordcloud of most common words visualised.
- Created a history page using SQLite for storage, allowing users to conveniently view their previous results in a visually appealing way, with sorting functionality.

Challenges Encountered

This project has been a major learning experience, before taking on this project, I had no experience with any of the main technologies which would be used for the process of sentiment analysis. These include key libraries such as Tensorflow, Keras and NLTK, but also both the traditional machine learning models and the Open AI api functionality.

I had a large learning curve to overcome when I first began this project, originally the project was intended to be based on twitter comments, but due to prohibitive costs of the API. I started researching for alternative suitable sites immediately, many options were considered but the reddit website was chosen due to the availability of the API and also, having social media style posts, while covering a large range of topics being discussed in a moderated style.

The largest challenge initially was the machine learning models, I had no previous experience with the process of training any models, my supervisor guided me in the right direction in terms essential research which needed to be conducted before beginning coding, and later, the order in which I should begin attempting to train models.

Due to having no previous experience, it was necessary for me to spend time experimenting with more relatively basic models which I knew would likely not be the final choices for my webapp.

Another challenge encountered was the hardware requirements which machine learning models needed, machine learning models ideally require a GPU for training, due to the resource intensive nature of the process. I overcame this issue by using google colab which allows users to utilise a free GPU for a limited period.

These challenges contributed to the project progressing quite slowly in its initial stages.

Limitations / Not Achieved

Initially, I had planned for users to be able to select a category which their search term relates to, and depending on this category, a model which has been trained on that particular category would be chosen. The theory being that the model will be more accurate in a certain domain as it is trained on data which relates to that domain.

When I first planned this idea, I did not realise how large of a challenge it would be to find enough datasets which would be tailored to specific categories only. I stopped developing on this idea quite early in the projects development, as even if I did get tailored datasets to a specific topic, such as politics or finance, there would likely not be enough sheer data, which would lead to issues such as overtraining of the machine learning models, as the data is not complex enough.

Another limitation is the sentiment by time visualisation I had planned, this visualisation was intended to take a sentiment label, and along with the time it was generated, map it to a visualisation, allowing users to get a comparison of two of the same label types , One created by a machine learning model, and one by the ChatGPT API, graphed through time, showing how the sentiment has shifted through time on a particular search term. I was not able to implement this feature in a refined manner due to time constraints, I created a working example of this feature, but I felt in terms of design, it did not suit the aesthetic of the history page in which it resided, leading me to cut it from the final webapp.

Achievements

In terms of functionality for the project, I feel like the main objectives I set out to achieve were reached, the web application created is quite refined, featuring highly tested error handling for a range of scenarios. A fast retrieval and analysis system (taking approximately 55 seconds on average to retrieve a full sentiment analysis)

In terms of personal growth. This project has been major for developing my skills in all technologies used throughout the project development. I am now very comfortable with using complex libraries such as TensorFlow, Keras, NLTK. Throughout the project I also completed many examples of fully functioning trained machine learning models which I used for text classification, I now understand the process involved with choosing an effective dataset, preprocessing the data in a way which suits the intended goal, and choosing hyperparameters which will lead to increase in the models effectiveness and metric scores. In terms of hyperparameter tuning I experimented with factors such as the addition of LSTM layers, dropout layer addition, learning rate customization and epoch customization. This process of experimentation which I conducted during the course of my project, has made me a much more knowledgeable and competent programmer.

What I Would Change

If I was to restart the project from scratch, I would begin by first looking at examples of similar projects online and competitor web applications. When I first began my project, I spent a lot of time researching algorithms and methods a lot of the theory behind sentiment analysis, in my final project, I never used much of this information for the development of my project. I feel like it would have been much more time efficient to start straight working straight away on technologies which are commonly used for text classification, So I could built upon examples.

Another thing I would change is my mindset towards finding an appropriate dataset, when I initially began coding for this project, I only ever considered datasets which were large and already pre-processed to a high extent, not many datasets of this nature exist, So I was quite limited in the choices I could make, as I became more experienced through the training of machine learning models, I started realising its possible to manually create an effective dataset. By finding small datasets and combining them, a more effective final dataset can be reached. Leading to a more accurate model which is trained on more diverse data examples.

Learning Outcomes

This project has been my first experience creating a relatively large code base which features the information exchange of two API's, a machine learning model and a database, while visualizing information in a user friendly web application.

Working on a project for such an extended period of time, working through all the issues encountered during the different elements of site design: Backend database, backend functionality, front end user orientated design, API interactions with a webapp, model deployment on a live site. The skills which I have acquired during the course of my project, will have a much higher carry over to the professional software development environment then anything I have previously done.

Aside from technical skills, communicating with my supervisor about the trajectory of the project, issues encountered and future plans, have given me insight more realistic communication styles which will be expected in a working environment, where developers are expected to be problem solvers.

Conclusion

To conclude, I am satisfied with the amount of work achieved during the course of the project and the learning outcomes I experienced.

All the intended core functionalities were achieved and tested to a high degree, I feel like the shortcomings in the final product, such as some of the web design elements, could have been more refined if I didn't face some of the challenges I discussed earlier, which limited my productivity in the early stages of development.

The technical skills gained, and valuable personal skills such as time management, problem solving, organisational skills, planning skills and communication skills will stay with me in the future.

Acknowledgements

I would like to thank my project supervisor, Dr. Greg Doyle, for assisting me throughout the development of my final year project. Weekly meetings with Greg provided me with consistent valuable information and a guideline for the direction the project should be heading in at all stages. The assistance provided by Greg was insightful even until the very last week of development, where I was guided in adding extra features which turned out to make a considerable difference to the quality of the final product.

