Section D: Post Implementation Report

**Project Purpose**

Understanding the reception or sentiment towards a company’s products is getting increasingly difficult. Online publications and the proliferation of Artificial Intelligence generated articles has led user forums, the largest of which is Reddit, to be an increasingly common resource for people seeking and sharing opinions from other real people. When analyzing reviews from formerly reputable publications would once have been the benchmark to understanding the general sentiment towards a product or company, today, the crowdsourced discussions on Reddit are the richest source of data to understand sentiment.

**Datasets**

The data used for this application was scraped from various subreddits using the Reddit API. The top posts on a subreddit were retrieved. From those posts, the comments were retrieved and scored. The application was configured to retrieve comments from the top posts on /r/Excel, which is a subreddit dedicated to the spreadsheet software program from Microsoft, however this can be configured to run on any subreddit. The dataset includes the following:

* PostID
* CommentID
* Comment Text

This data was stored in a pandas DataFrame. The main limitation of this data was that that many comments were in reference to topics tangentially related to Excel. This means that while the sentiment of the comments in /r/Excel could be accurately calculated, said comments may not entirely represent a sentiment towards Excel.

**Process + Visualizations**

The application initially defines two global variables

* g\_subreddit: can be modified with the name of any subreddit the user wishes to analyze. In this instance the subreddit under analysis is /r/Excel
* g\_\_num\_posts: number of top posts to be analyzed (i.e. 10 = top 10, 100 = top 100). Can be modified with any integer under 2000

Using **praw** (Python Reddit API Wrapper), the reddit instance was created and connected to.

A function, given a subreddit, a timespan (day, week or month) and the number of posts desired, returned said number of posts from the aforementioned subreddit over the time period.

Another function would retrieve the comments from said posts and place them in a DataFrame.

A screenshot of a chat

Description automatically generated

The comment text was then run through the **VADER** (Valence Aware Dictionary and sEntiment Reasoner) Sentiment Intensity Analyzer. The body of the comment (comment\_text) in the DataFrame above was passed through the sentiment analysis engine.

The engine provided a compound score, which is the ‘normalized, weighted composite score’ of sentiment for the comments. The score is calculated by summing the valence scores of each word in the lexicon which is then normalized to be between -1 (most extreme negative) and +1 (most extreme positive. As per the standard practices in the documentation for VADER, the compound score can be classified using the following metric:

1. **positive sentiment**: compound score >= 0.5
2. **neutral sentiment**: (compound score > -0.5) and (compound score < 0.5)
3. **negative sentiment**: compound score <= -0.5

The compound score and the classified sentiment was then appended to the existing DataFrame

A screenshot of a computer

Description automatically generated

The resulting DataFrame was then sorted based on the value of the compound score to display the 10 most positive comments over a given timeframe.

A screenshot of a computer screen

Description automatically generated

The DataFrame contains the data needed to visualize and understand the sentiment in the subreddit. For example, the application was configured to generate a bar chart displaying the number of positive, negative and neutral sentiments over the past day/week/month.

A graph of positive and negative comments

Description automatically generated

A bar chart of the percentage of all comments over the various timeframes that were positive was generated to show how the general sentiment on /r/Excel has evolved.

A graph of positive and negative comments

Description automatically generated

**User Guide**

The application is run in a Jupyter Notebook. It can be published and run online, if hosted on a server, which would just require that the end user run the cells. To run the notebook locally Python needs to be installed.

Additionally, the following Python packages need to be installed for the application to run:

* praw
* pandas
* nltk
* vaderSentiment
* numpy
* matplotlib

A requirements.txt file will be provided in the repository.

Additionally connecting to Reddit requires the following credentials:

* client\_id
* client\_secret
* user\_agent

These credentials can be generated at the link below. <https://old.reddit.com/prefs/apps/>

**Summary:**

Completing this project was an extremely valuable experience. Until this project, my experience with programming had mostly been in completing and writing code snippets and solving problems. Completing an application start to finish, using all the different skills I learned was very helpful for me. I gained a better understanding of the complexities of tying together several skills and technologies.

Importantly, I now realize that even though my courses are completed, there will always be more to learn. There are always new technologies to familiarize myself with, and in order to be a successful programmer, it is essential to keep up to date and up to speed.

Learning comes in many forms, not just completing courses, reading documentation and running through tutorials. The assistance I received from those I asked for help taught me more than anything else over the project. The advice from my advisor Candice and my friend Prerak was invaluable. I would also like to thank my dad who helped me with a great deal of advice on designing and debugging the application.