A Reddit Bot for Analysis of Political Commentary

-- Detailed Proposal --

Albin Vincent
(Student ID: 201317521)

Supervisors:
Primary Supervisor: Stuart Thomason
Secondary Supervisor: Terry Payne

Table of Contents

Project Description

Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify and study affective states and subjective information *(Sentiment analysis, 2020)*.
Some users on the social media platform Reddit, often make posts or comments in regards to politics. These posts are usually written to support or oppose certain politicians or political policies. From the user's language and choice of words, we can try and deduce if they are left-wing, right-wing or neither.

The main aim of this project will be to create a program that can trawl through specified subreddits (a forum dedicated to a specific topic on Reddit) to extract public comments, analyse these comments, and do political sentiment analysis on these comments. One of the areas we are interested in is categorising the users in these subreddits in terms of their political opinion, i.e. if a user is left-wing or right-wing. Another thing to be done will be to look for connections between a specified user's posts in other subreddits. Examples of this can be, does a user who has been identified as left-wing, post right-wing views on other subreddits or if multiple users who are left-wing have similar views in other subreddits such as liking pictures of cats. All of this can be done using data mining, natural language processing and machine learning.

After the program has been trained successfully to identify user's political sentiment, the next stage will be to represent this data in a visually interesting way. The visual representation should allow the user of the program to navigate through the data with ease.

Aims and Objectives

Aims:

- The program should be able to trawl through a specified subreddit to extract all comments made.
- Given a Reddit username, the program should be able to extract all comments made by the user
- The program should be able to categorise the comments in terms of the political sentiment using machine learning
- The program should output the results as a user-friendly visualisation

Objectives

- The program can retrieve all comments made on a specified subreddit
- The program can retrieve all comments made by a specified user
- Accurate machine learning model made using a training data set
- The program can place users on the political spectrum (left-wing to the right-wing) based on comments they previously made accurately
- The program should be able to identify similarities within a group of people who are on the same point on the political spectrum (e.g. left-wing people holding right-wing views on other subreddits)
- Interface as a website
- Interface easy to navigate
- Interface outputs the results in an aesthetically pleasing way
- Make a website to get input parameters and output results
- The program should adhere to data privacy rules and have ethical consideration when processing data such as hiding usernames when outputting results

Key Literature & Background Reading

The first piece of reading done was on the Reddit API *(reddit.com: api documentation, 2020).* This website has the documentation for the API and contains all the information regarding its syntax and usage. Using this API, posts and comments from subreddits can be retrieved for processing. When using this API, one must be sure to follow the rules for the API (api - reddit.com, 2020) and only use it as intended with the correct OAuth 2 authentication in place.
For processing the data retrieved using the Reddit API, Spark and Hadoop will be used. To learn how to use this technology, lectures from the module COMP336 (Bakhtiar, 2020) will be very useful.
The main programming language that will be used is Python. The reason for this is that it is one of the most common programming languages amongst AI developers (therefore a lot of support with common issues will be available) and since it has a huge collection of machine learning libraries, this language will be ideal for this project. Online tutorial websites such as W3schools (Introduction to Python, 2020) will be used to find out more about Python and specific tutorials on machine learning with Python such as the one provided by W3schools (Python Machine Learning, 2020) and another found on the datacamp website (Simplifying Sentiment Analysis in Python, 2020) will be useful for this project. As well as this, multiple Python libraries will be useful for this project. Some key libraries are NumPy (Overview — NumPy, 2020), TensorFlow (Tutorials | TensorFlow Core, 2020), Matplotlib (Matplotlib documentation, 2020) and PRAW (PRAW: The Python Reddit API Wrapper documentation, 2020).

To select the best AI model to use for this project, an article written on the Dzone website was used (Top 10 Most Popular AI Models - DZone AI, 2020). From this article, the Naïve Bayes algorithm was selected as it best matches the project description. In the use case of this project, the algorithm would calculate the probability to see where each comment lies on the political spectrum. To see more information in terms of linking multiple comments together to see similarities between users who have the same political opinion, the K-Means algorithm will be used. This is an unsupervised machine learning algorithm and an explanation of this algorithm was published in an article written on the Stanford University website (K-Means, 2020).

Development process and methods

This project will be taking on a variation of the tradition waterfall model. The only difference from the original model being that there will be several iterations of the design and implementation stages. The reason for this is because the best machine learning model for this task is not yet known and therefore a lot of experimentation will be needed to figure out what the best model is and also the design may need to be changed during the implementation stage to make the program as optimal as possible.

The main programming language for this project will be in Python as it is both a very popular language and one with a lot of support for machine learning. In addition to this, Python has a lot of libraries that will be useful for this project such as NumPy (supports large multi-dimensional arrays and allows for high-level mathematical functions to be applied to it), PRAW (a Reddit API wrapper), Matplotlib (to plot the graphs in python). Initially signing up for the Reddit API is required where a username and password can be made. This information is needed to use the PRAW library in python. Apache Spark may be used for storing the data sets and applying machine learning algorithms on it. Finally, the front end for this application will be done using HTML, CSS, JavaScript and Bootstrap.

For version control, Git will be used and all code will be made available on a private GitHub repository.

Data Sources

The main data source that will be used will be the data retrieved from Reddit. This will be accessed through the Reddit API which requires you to sign up to the service using a username and password. This provides you with legal access to retrieve data from Reddit using the API. This API will be used to get posts/comments made by a user and also to get all posts/comments made in a specified subreddit. When using this source, great care must be taken to ensure that personal information such as real names, usernames or any identifying information for a user does not get outputted by the application. To ensure that the user's anonymity is protected, instead of using the usernames as the identifier for the user, a non-identifiable hash will be used. This means that the username isn't passed around the program and therefore there will be no chance of accidentally revealing identifying information.

In addition to this, evaluation questionnaires will be used towards the end of the project. This will be to evaluate this application and measure both the ease of navigation on the website and also how easy it is to understand the information outputted by the application. When taking in these forms, the names and any other private information will be redacted for privacy.

Testing & Evaluation

A very brief way in which this application will work will be: retrieving the data-set from Reddit, cleaning the data, training/building the AI model, process the data-set, outputting the results on a website.
To test that we can get data from Reddit successfully, unit tests can be written to retrieve data from Reddit and compare it to the data found on the website. This can be useful if the format of the reply from Reddit changes in the future. If this were to happen, then the unit tests for this would break and you would know where the error is.
Component Testing will be done in the stage where the AI is trained. This is to test it the AI can sort a user according to their political sentiment accurately. To test this, users can be identified randomly and once their political sentiment has been found, the user/comment can be passed into the AI to check the results. If the results do not match, then the AI can be retrained with the correct output. Once the AI has been trained and if it was to incorrectly identify the political sentiment for a specific user/comment at a later stage then the AI would be retrained with the correct values. This will be an ongoing stage throughout the development.
To test that the results are displayed on the website correctly, manually looking at the visualisations can be useful to check if the representation of the results is accurate.
As well as this, UAT (User Acceptance Tests) will be done by giving the volunteer testers evaluation surveys and this will be used to check if the system is easy to use and that the data is being represented in a way that is easy to understand. Finally, smoke testing will be done to check that the system works as a whole and that all the objectives are met.

Ethical Considerations:

One of the ways in which I must act ethically is when collecting data from Reddit, it must not be scraped from the website directly and has to be retrieved using the Reddit API. The reason for this is because scraping directly from the Reddit website is illegal and one has to accept the terms and conditions of the API to be able to use it.
Another way in which I must act ethically is when collecting evaluating surveys, I should make sure to follow all the university guidelines (such as hiding the identity of the person who filled out the survey).

I confirm that I have read the university ethical guidelines *(Ethical Conduct, 2020)* and will follow it.

BCS Project Criteria

One of the BCS requirements for the honour's year projects is to be able to apply practical and analytical skills gained during the degree programme. I will do this by applying as many of the skills I have learnt during my studies at university for this project. Some of the topics that I have studied that directly applies to this project are processing big data, machine learning/AI, software engineering, using APIs etc...
Another requirement is to have innovation and creativity in my solution. Big Data and Machine Learning/AI are two areas of study that are relatively new and very innovative. I would have to use these technologies to get results and I would need to come up with a creative approach to display these charts in a way that is easy to understand.
A requirement is for you to combine information, ideas and practices to provide a quality solution with an evaluation of that solution. I will fulfil this requirement by using all the information and ideas I have learnt in lectures so far as well as some of the good practices I have learnt during my software development placement (such as commenting code, regularly backing the project up, a reflection of the project, etc…) for my project.
Another one of the requirements is for the project to meet a real need in the wider context. My project satisfies these criteria as the solution can be changed slightly to get different types of sentiment i.e. views on abortion (anti-abortion vs pro-abortion) or secularity. The AI needs to be retained for the specific category however the rest of the workings of the solution can remain the same. In addition to this, since the machine learning model has already been made, this can then be reused in a different context without having to change the model using ideas of transfer learning.
Another one of the requirements is to self-manage a significant piece of work. I will be fulfilling this requirement as I will be completing this task by myself and this is a significant piece of work with a lot of planning and technical skills required.
Finally, the last requirement is the critical self-evaluation of the process. I will do this by carrying out a high-level review of the work that I have done regularly to check if anything needs to be changed to improve the solution.

Software and Hardware Resources

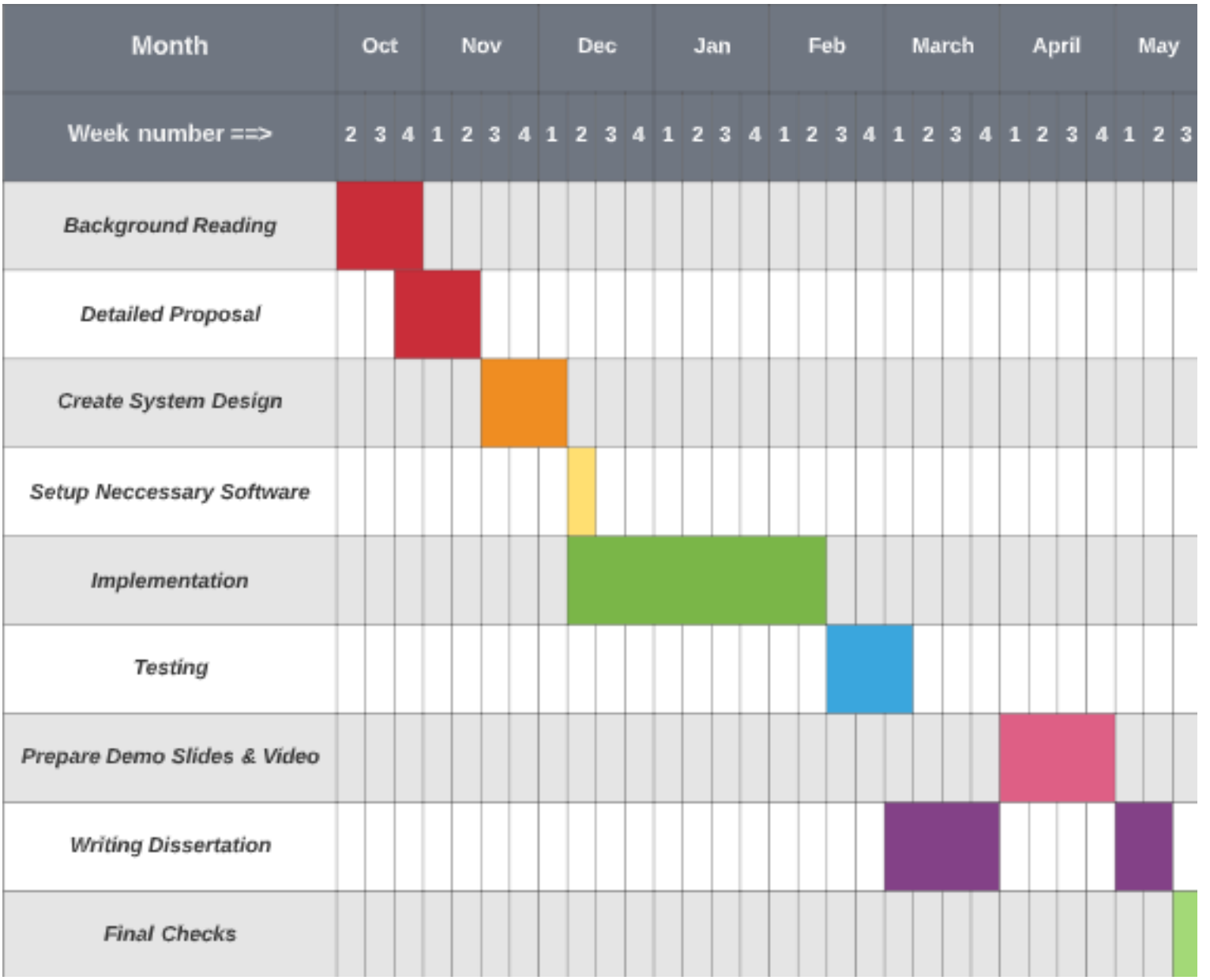
Software:

- Python 3.9 - Open source programming language
- Python Libraries (e.g. NumPy, Matplotlib)
- Apache Spark 2.4.7- Free to use
- GitHub – free to use – student account
- Reddit API – free to use for personal use – Need to sign up on the website
- Editors – Atom and IntelliJ

Hardware:

- My PC:
  - CPU - AMD Ryzen 5 3600 3.6 GHz 6-Core
  - RAM - 16 GB DDR4-3200
  - GPU - MSI Radeon RX 570 8 GB
  - SSD - 480 GB

Project Plan



Risks & Contingency Plans

Risks	Contingency	Likelihood	Impact
Hardware Failure	Backup code and all files regularly using Git.	Low	Medium – I will have all the software needed to create the program on my Laptop so if my PC fails, I can keep working on the laptop, pulling all the up-to-date files down using git.
Lack of programming ability	I will be learning topics related to Big Data in semester 1 so by Semester 2 I should be able to process Big Data. Machine Learning is something new for me – I will try and watch as many tutorials online to understand it better.	Low	High – If I do not know how to code certain bits of the project, then a lot of the project may end up not working. Very unlikely as I am confident in my programming abilities
Running out of time	I have planned out how long each stage of this project will take on the Gantt Chart on the previous page. If the project takes a lot longer than expected, then certain functionality will have to be prioritised i.e. completing the machine learning algorithms will be prioritised over having a fancy front end webpage	Medium	High – Unable to submit a project that is fully completed
The project does not meet all requirements	The requirements set for this task are all related to the core functionality of this project. If I cannot meet certain requirements, I will speak with my project supervisor to find out the best possible remedy for this.	Low	High – When the project does not meet all requirements this usually means that the project is not complete and therefore will not work fully as expected.
Breaking Reddit Regulations for API usage	I have thoroughly read the terms and conditions and will try my best to adhere to the terms of usage.	Low	Very High – Legal action may be taken

References

Reddit.com. 2020. *Reddit.Com: Api Documentation*. [online] Available at: <https://www.reddit.com/dev/api/> [Accessed 8 November 2020].

Reddit.com. 2020. *Api - Reddit.Com*. [online] Available at: <https://www.reddit.com/wiki/api> [Accessed 8 November 2020].

En.wikipedia.org. 2020. *Sentiment Analysis*. [online] Available at: <https://en.wikipedia.org/wiki/Sentiment_analysis> [Accessed 8 November 2020].

Bakhtiar, A., 2020. *COMP336*. [Lectures] University of Liverpool, Liverpool, UK.

W3schools.com. 2020. *Introduction To Python*. [online] Available at: <https://www.w3schools.com/python/python_intro.asp> [Accessed 8 November 2020].

W3schools.com. 2020. Python Machine Learning. [online] Available at: <https://www.w3schools.com/python/python_ml_getting_started.asp> [Accessed 8 November 2020].

DataCamp Community. 2020. *Simplifying Sentiment Analysis In Python*. [online] Available at: <https://www.datacamp.com/community/tutorials/simplifying-sentiment-analysis-python> [Accessed 8 November 2020].

Numpy.org. 2020. *Overview — Numpy*. [online] Available at: <https://numpy.org/doc/stable/> [Accessed 8 November 2020].

TensorFlow. 2020. *Tutorials | Tensorflow Core*. [online] Available at: <https://www.tensorflow.org/tutorials> [Accessed 8 November 2020].

Matplotlib.org. 2020. *Matplotlib Documentation*. [online] Available at: <https://matplotlib.org/contents.html> [Accessed 8 November 2020].

Praw.readthedocs.io. 2020. *PRAW: The Python Reddit API Wrapper Documentation*. [online] Available at: <https://praw.readthedocs.io/en/latest/> [Accessed 8 November 2020].

dzone.com. 2020. *Top 10 Most Popular AI Models - Dzone AI*. [online] Available at: <https://dzone.com/articles/top-10-most-popular-ai-models> [Accessed 8 November 2020].

Stanford.edu. 2020. *K-Means*. [online] Available at: <https://stanford.edu/~cpiech/cs221/handouts/kmeans.html> [Accessed 8 November 2020].

University of Liverpool. 2020. *Ethical Conduct*. [online] Available at: <https://student.csc.liv.ac.uk/internal/modules/comp390/2020-21/ethics.php> [Accessed 8 November 2020].