College of Engineering and Informatics

CT413 Final Year Project Definition Document

[redacted]

17346123

[redacted]

Table of Contents:

Introduction

For my Final Year Project, I will create a Python application to analyse the topic, sentiment and political stance of tweets. This python application will have a simple Python Flask front-end, and will allow users to analyse the tweets within a specific hashtag, or by a specific user. The main data that will be analysed from the set of tweets will be the sentiment of the tweets, the most common words and phrases within the tweets and the political leaning of the tweets, left leaning to right leaning. This web app will also allow users to compare the activity within two separate accounts or hashtags to see similarities and differences.

Tools and software

- Python [0]
  - The majority of the code for this project will be written in the Python language. I chose to use Python as it is very widely used and well supported, it works with the twitter developer APIs and it allows for easy deployment as a web application.

- Python Flask [1]
  - Python Flask is a Python framework for building web applications with python. It requires some html to render web pages, but makes the process of deploying a Python web application relatively easy.

- Python unittest and pytest [2], [3]
  - Python unittest and pytest are two python testing tools which support easy building and deployment of python code tests. Each python file has its own set of unittest tests, which can be run together as one pytest test. If any one unit test fails, then the test set will fail, flagging the error. This makes it easier to keep code functional, and keep all the tests consistently up to date.

- Twitter Developer APIs [4]
  - The twitter developer APIs are written for Python, Ruby and Node.js. They provide an easy and well-maintained API system through which to fetch twitter data. This provides a very efficient way to analyse tweets, however the API has some limitations. When searching for tweets in a hashtag or user account, only tweets from the last 7 days can be accessed, which makes analysing past activity very difficult.

- Microsoft Azure Text Analytics Service [5]
  - Microsoft Azure provides a text analytics service, which can return a numerical (0 to 1) value representing the sentiment of unstructured text data passed to it. This Service is very efficient and incorporates well with the Twitter Developer APIs, however users with student accounts are limited to 5000 requests per month, which reduces the usability of this service within my project.

- NRC Emotion Lexicon [6]
    - The NRC emotion Lexicon is a list of English words, mapped along with their association with a range of basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). This dataset was collected through crowdsourcing by the National Research Council of Canada. The data is open source, and can be downloaded in csv format to allow for easy python processing. This emotion lexicon can provide an easy way to estimate emotions from a dataset of words used in tweets.

- Git, Github [7]
    - I used Github to manage my Project throughout its development. This allows for easy data backup and progress logging. Github workflows provide an easy way to automate unit testing on a github project. In my github repository, changes must pass these automated python pytest and unittest tests before they can be merged into the main branch. This helps ensure that all elements of the code continue to work with each change.

Deliverables

- Project Definition Document - 28th November 2020
- Final Project Report - 6th May 2021
- Project Demonstration and Viva Voce - 10th - 14th May 2021

Project Objectives

Core Objectives

Analyse Twitter Data within a hashtag

Build a system which takes a hashtag as input, and fetches the last n tweets within that hashtag. This set of tweets should then be analysed in a number of ways.
Some of the proposed data points to extract from a set of tweets include:

- The number of unique words within the set of tweets
- The most frequently used words within the set of tweets
- The most common emojis used within the set of tweets

Analyse twitter data for a user account

The system which handles hashtags can be extended to analyse data from a specific twitter account. The tweets from the user can be analysed to extract the same data points as stated above.

Estimate sentiment of set of tweets

Once a set of tweets has been fetched, whether through a hashtag or a user account, these tweets can be analysed to determine the overall sentiment of the set of tweets.
This can be approached in a number of ways

- Microsoft Azure Text Analytics Service
  - This service can easily analyse a set of tweets and return a sentiment score indicating how positive or negative the tweets are on a range from 0 to 1.
    0 being very negative, 1 being very positive.
    The limitations of this system are in the rate limits of the system. With a student account, only 5000 request per month are provided free
- The NRC Emotion Lexicon
  - This dataset requires more complex implementation, however it is open source and provides emotion mapping for over 14,000 english words.
    The limitations of this system are that most large sets of tweets will return similar results, so the applications of the data analysed with the NRC emotion lexicon is somewhat limited.
- Machine Learning
  - Another option for analysing the sentiment of a set of tweets would be through Machine Learning, by training a machine learning algorithm with a data set of tweets matched to sentiments. That algorithm could then be used to estimate the sentiment of a set of tweets with unknown sentiment.

Estimate Political Leaning of set of tweets

Once a set of tweets has been fetched, whether through a hashtag or a user account, these tweets can be analysed to estimate the political leaning of the set of tweets.

The most sensible way to implement this would be to find an existing dataset of tweets with known political leanings, and analysing the most used words, hashtags and emojis within the manually categorised dataset.
This can allow for the collection of ‘political flags’ which can be then compared against any set of tweets to estimate the likelihood that a user or set of tweets are politically aligned with a category for which ‘political flags’ are known.

Develop a nice user interface for the web application

A nice user interface will be built to allow users to easily use the system, and to easily understand the results that the system provides.

The data should be displayed in a way that avoids overwhelming the user with data, but allows the user to access more complex data if they wish.

If full tweets or twitter accounts are shown as the output to any query, they should look similar to how they look within twitter's native application.

A significant importance will be placed on sensible data visualization within the User Interface. There will be a lot of data returned from any single query, and it is important to display this data in a sensible and understandable way.

Extra Objectives

If I have sufficient time upon completing the primary objectives of this project, I hope to implement some, if not all, of these secondary objectives.

Estimate Probability of twitter account being fake

Fake twitter accounts are a prevalent problem within the politics of many countries. Online radical communities have been known to create fake accounts masquerading as members of minority groups, or people of different nationalities, to manipulate political discussions and political sentiments for specific issues.

There are some tools already available which can return a probability of a twitter account being authentic, such as Botometer [8] or BotSight [9].

Botometer seems like the best tool to use in my use case since it is well supported and has a Python API.

If none of the bot detectors available work for my specific use case, I will build my own simple Bot detection system. This system would work by analysing the user data that can be fetched with the twitter API such as account creation date, account location and follower numbers, along with other key determinants of a fake account, to give a probability of an account's authenticity.

Since there is no way to know for sure whether any twitter account is a fake account or not, this probability, generated from an established system or my own system, will only be a guess, however it can still provide some useful information. This would allow the system to analyse the prevalence of fake twitter accounts within a hashtag.

Estimate Political Polarisation within a Hashtag

Many discussions and hashtags on twitter are very political in nature however lots of these discussions happen within a political group who share the same ideas. This can lead to people overestimating the support for their ideas, and underestimating the strength of the opposition to those views.

Once the political sentiment of a set of tweets can be found, that system can be extended to analyse the range and strength of political sentiments within a hashtag, and thus the polarization of said hashtag.

Implement Machine Learning system to train one or more of these algorithms

One way to potentially gather some very interesting data would be through machine learning algorithms such as tensorflow.

To implement a machine learning algorithm to analyse tweets and come to some conclusion such as the political leaning of the tweets, one would need a significant dataset of tweets with known political

sentiment, to see if the machine learning algorithm can make any connections that I missed within the data.
The amount of work required to implement machine learning in this project is quite significant, with no guarantee of a useful output, so this feature will only be implemented if time allows for it.

Allow web-app results to be easily shared to Twitter or other platforms

A nice feature to implement into the web-application would be an easy integrated way to share the analysis returned by the webapp on twitter.
This feature would probably not be too difficult to implement, and it would be a nice addition to the User Interface if implemented well

Background Research

Before beginning my Project, I wanted to take a look at what research had been done in this field in the past.

The most relevant research to my project was done by S. Stieglitz and L. Dang-Xuan of the University of Duisburg-Essen, who published a research paper entitled “Political Communication and Influence through Microblogging - An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior”. [8]
In this research paper, written for the 45th Hawaii International Conference on System Sciences in 2012, Stieglitz and Dang-Xuan analysed a dataset of 64,431 political tweets with respect to how often tweets were ‘retweeted’, and the correlation between tweet sentiment and retweet rate. This paper was very interesting to read and provided much insight into how certain tweets and ideas can spread quickly across the platform of twitter.

Other interesting articles I read in researching the area of twitter analysis were “Finding interesting posts in twitter based on retweet graph analysis” by M.-C. Yang, J.-T. Lee, S.-W. Lee, and H.-C. Rim, published in Proceedings of SIGIR in 2012 [9], and “Style Matters!: Investigating Linguistic Style in Online Communities”, written by Padmini Srinivasan and Osama Khalid and published in ICWSM in 2020 [10].

My work on this project began in October of 2020 and on the third of November in 2020 there was a presidential election in the United States of America. This provided a very interesting dataset to analyse throughout October and November of 2020 as the US operates a two party system and thus most voters identify strongly with one of the major political parties. Most US news networks cater predominantly to one side of the political spectrum, and this leads to a very polarised voting demographic, and thus very polarised twitter activity.
I did a lot of research throughout the election on the beliefs within each political party, and the main reason why voters were voting for either one of the parties.
This made it easier to determine what I wanted to analyse within the data. For example, I wanted to see which presidential candidate used a wider range of words within their tweets, to see what education level they were targeting their tweets at, and how often they repeat the same information.

In developing the first prototype of my application, I used the Microsoft Azure text analytics service to determine the sentiment of a set of tweets. This gave me valuable insight into how sentiment tends to vary across different hashtags, and in different communities across twitter. Although the Microsoft Azure text analytics service is too limited for use in the final version of my application, it was very useful as a research and prototyping tool.

Planning and timeline

Schedule is based on: 4hrs FYP work per week in semester 1
8hrs FYP work per week in semester 2

Task ID	Tasks	Estimated Time (Work Hours)	Dependencies	Planned Completion Date	Actual Completion Date
t0	Receive FYP allocation	-	-	-	13/10/2020
t1	Research Twitter Developer API	4h	-	20/10/2020	20/10/2020
t2	Investigate previous research in this area	4h	-	27/10/2020	21/10/2020
t3	Working command line system - analyse hashtag and get most used words	8h	t1	10/11/2020	22/10/2020
t4	Working system - analyse user or hashtag, get sentiment (Azure)	8h	t3	24/11/2020	29/10/2020
t5	Create flask application to run implemented system	4h	t4	1/12/2020	05/11/2020
t6	Setup flask application to work on college linux server	4h	t5	8/12/2020	12/11/2020
t7	Apply weighting to sentiment analysis	4h	t4	15/12/2020	12/11/2020
t8	Create unit tests and setup github repo to properly use them	8h	t2	05/01/2021	19/11/2020
t9	Estimate political leaning of tweets	8h	t4	05/01/2021	-------
t10	Analyse more user account data - location	8h	t8	19/01/2021	-------
t11	Research, rescope and re-estimate	16h	-	02/02/2021	-------
t12	Design and Implement nice looking GUI for project	16h	t5, t6	16/02/2021	-------
t13	implement account authenticity estimator	16h	t11	02/03/2021	-------
t14	implement political polarisation estimator	16h	t11	16/03/2021	-------
t15	allow web-app data to be tweeted from within web app	4h	t11	23/03/2021	-------
t16	Write Final Year Project Report	32h	all	20/04/2021	-------
	Report Due Date			06/05/2021	
	Project Demonstration & Viva Voce			10/05/2021	

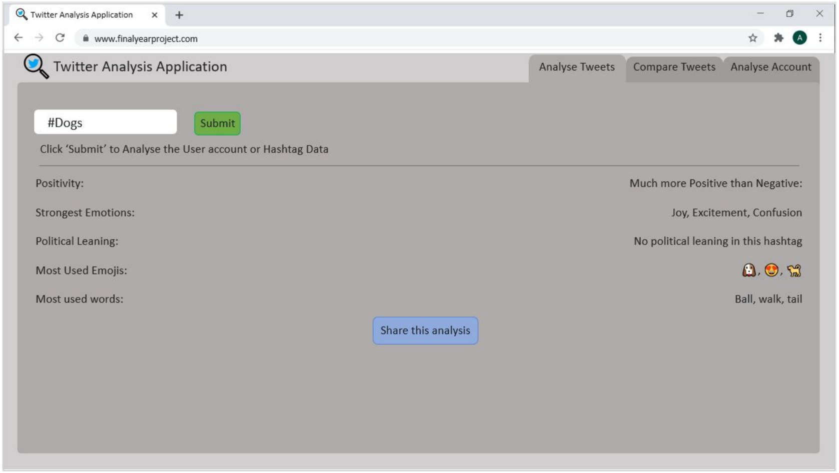
Risk assessment

Risk	Consequence	Impact	Risk response strategy
Hardware Issues with laptop	Unable to work on project, unable to meet deadlines	High	Stay ahead of schedule
Loss of already completed work	Work needs to be redone from scratch	High	Maintain versioning System (Github)
Unable to work on Final Year Project due to demands of other modules	Unable to meet deadlines	Medium	Start work early, and stay on top of deadlines for all modules
Previously working feature is now broken	Work needs to be fixed or redone	Medium	Implement consistent unit tests on github repository
Limitations of API make it too difficult to gather large dataset	Results of data analysis are less interesting and less useful	Medium	Gather and store datasets to better train and understand data
Rate limit of 500,000 tweets per month from twitter developer API used up	Unable to analyse any new or current data	High	Store dataset of 10,000 tweets from various hashtags/accounts to allow for offline development
Insufficient time to complete goals	Lower quality project	High	Avoid by staying on schedule, if it happens, handle by re-planning and re-estimating

Most of the risks investigated can be mitigated, or avoided, by staying ahead of schedule and maintaining a consistent, functional version of the project in a Github repository.

GUI Mockup

This is a mocked up image of what the final user interface of the user interface of the application might look like.

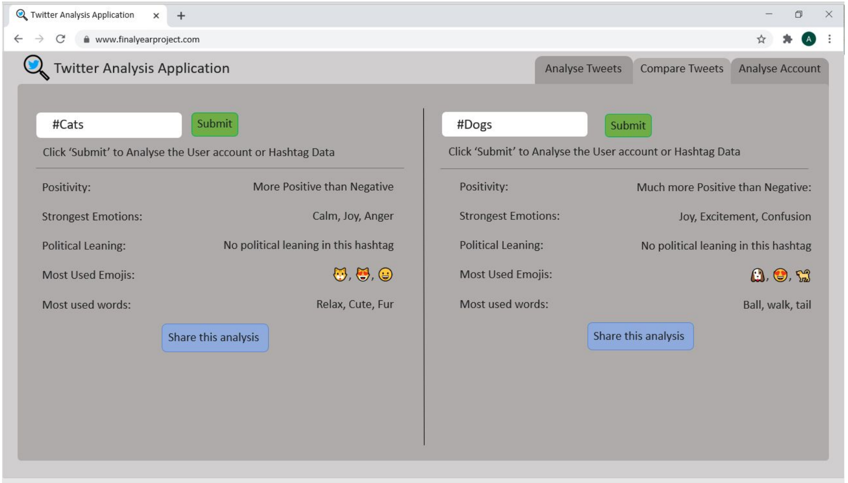


Twitter Analysis Application

#Dogs

Click 'Submit' to Analyse the User account or Hashtag Data

Positivity:	Much more Positive than Negative:
Strongest Emotions:	Joy, Excitement, Confusion
Political Leaning:	No political leaning in this hashtag
Most Used Emojis:	🐶, 😍, 🐕
Most used words:	Ball, walk, tail



Twitter Analysis Application

#Cats

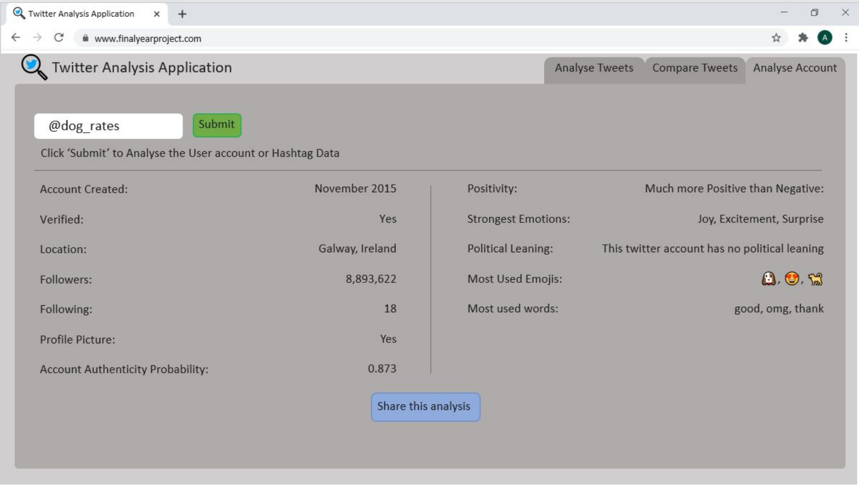
Click 'Submit' to Analyse the User account or Hashtag Data

Positivity:	More Positive than Negative
Strongest Emotions:	Calm, Joy, Anger
Political Leaning:	No political leaning in this hashtag
Most Used Emojis:	🐱, 😻, 😀
Most used words:	Relax, Cute, Fur

#Dogs

Click 'Submit' to Analyse the User account or Hashtag Data

Positivity:	Much more Positive than Negative:
Strongest Emotions:	Joy, Excitement, Confusion
Political Leaning:	No political leaning in this hashtag
Most Used Emojis:	🐶, 😍, 🐕
Most used words:	Ball, walk, tail



Twitter Analysis Application

@dog_rates

Click 'Submit' to Analyse the User account or Hashtag Data

Account Created:	November 2015
Verified:	Yes
Location:	Galway, Ireland
Followers:	8,893,622
Following:	18
Profile Picture:	Yes
Account Authenticity Probability:	0.873

Positivity:	Much more Positive than Negative:
Strongest Emotions:	Joy, Excitement, Surprise
Political Leaning:	This twitter account has no political leaning
Most Used Emojis:	🐶, 😍, 🐕
Most used words:	good, omg, thank

Conclusion

I hope to develop a useful tool for twitter analysis, that will provide deeper insight into the use of twitter as a platform for political discourse, and the implications of twitter on political discussion and political rhetoric. I will be testing the tool throughout the process of building on known hashtags of various political leanings, and I will be determining how accurate the tool is based on how well it can categorize and analyse these known hashtags. I plan to stick to my self-imposed deadlines and through this, and regular meetings and discussion with my Final Year Project Supervisor, I expect to deliver a high quality, well tested and well implemented tool.

References

[0] - Python 3.8 documentation
https://developer.twitter.com/en

[1] - Python Flask Documentation
https://flask.palletsprojects.com/en/1.1.x/

[2] - Python unittest Documentation
https://docs.python.org/3/library/unittest.html

[3] - Python Pytest Documentation
https://docs.pytest.org/en/stable/contents.html

[4] - Twitter Developer API Documentation
https://developer.twitter.com/en

[5] - Microsoft Azure Text Analytics Service Documentation
https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/

[6] - NRC Emotion Lexicon Documentation
https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[7] - Github Documentation
https://docs.github.com/en

[8] - Botometer https://botometer.osome.iu.edu/

[9] - BotSight https://download.botsight.nlok-research.me/

[8] - Political communication and influence through microblogging
S. Stieglitz and L. Dang-Xuan.
an empirical analysis of sentiment in twitter messages and retweet behavior.
In 2012 45th Hawaii International Conference on System Sciences.

[9] - Finding interesting posts in twitter based on retweet graph analysis
M.-C. Yang, J.-T. Lee, S.-W. Lee, and H.-C. Rim
In Proceedings of SIGIR, 2012.

[10] - Style Matters!: Investigating Linguistic Style in Online Communities
Padmini Srinivasan, Osama Khalid
ICWSM, 2020