Initial Plan

Cardiff University School of Computer Science and Informatics

ANALYSING PREDICATIVE FACTORS OF COVID-19 SENTIMENT ON TWITTER

CM3203 - One Semester Individual Project

Author: Akansha Garg

Supervisor: Steven Schockaert

Project Description

The Covid 19 Pandemic has been described as the most challenging crisis humanity has had to face since World War II [1]. It has affected everyone and everything; the impacts have been long lasting and will change people will live their lives in the future. With new information being released almost every day in the United Kingdom regarding the rules and regulations, people's sentiment towards the pandemic is constantly fluctuating. Twitter is a social networking site that was released in 2006 and has grown in popularity as one of the most used social media platforms with over 145 million daily users [2]. It allows its users to post a short message of up to 280 characters in the form of a 'tweet'. Around 6,000 tweets are sent out every second [2]. It is a free public site where people can openly express their opinions from government restrictions to favourite lockdown activities.

This project focuses on identifying and analysing predictive factors that effect Covid 19 sentiment on Twitter. I have decided to select python as the programming language for my project, as it is a language I am familiar and comfortable with. It also has a lot of different libraries and tools that will be useful for this project. The first stage in this project is to create a dataset. By connecting to the Twitter API, I will be able to collect tweets over a short period of time that have included, or hash tagged one of the keywords related to the Covid 19. Many new phrases and words have become increasingly popular in the last year and are now nationally associated with the pandemic e.g., Lockdown, Isolation, Social Distancing etc. I will also ensure that the tweets are posted by users in the United Kingdom, which can also be achieved using the API. The tweets will then need to be stored in a format that can be used for analysis later.

After collecting the tweets and creating a dataset, I will need to focus on analysing the sentiment of the tweets. There are many different natural language processing tools that work with python i.e., NTLK. I will have to research the tools and find which is best suited for my project.

While collecting the tweets, I will need to identify relevant data sources that can provide information on key activities that have taken place during the tweet collection period. For example, what different regions went into a local lockdown, if there is there is a travelling ban, if the number of new cases were rising etc. These factors will be key in analysing the sentiment and can help to find predictive factors.

The final part of the project is the analysis, there are again many business intelligences tools that allow for visualizations to be created from a dataset such as Microsoft Power Bi. I will want to analyse how the sentiment varies around the country and if there are in key factors that impact the sentiment. For example, if the country is put back into a 'tier' system for restrictions do different levels of restrictions effect sentiment? A heatmap visualisation could be a useful analysis tool that could be used for this. It will also be useful to compare sentiment at the start of the project when I initially collect tweets to towards the end of the project where I can re stream new tweets with the same filters.

Ethical considerations

I will be collecting data and information from an online platform and therefore need to seek ethical approval. The university offers a course on Research Integrity Training which I must pass before I can apply to get ethical consent. Once completing the course, I will submit an ethical consent form and not begin collecting any data from twitter until I have received approval.

I only plan on storing the text contents of the 'tweets' and the location of the 'tweeter'. However, when using the Twitter API all information is returned and I will have to ensure anonymity of the 'tweeter'.

Project Aims and Objectives

The project can be divided into 4 main aims:

1. Creating a tweet dataset
2. Defining the sentiment of each of the tweets
3. Identifying the key factors at the time of collection
4. Analysing how those factors impact the sentiment of the tweets.

With these in mind there are key objectives that must be met to ensure the project is successful. I will have to conduct a relatively in-depth amount of research into the different tools available for achieving the aims which will take significant time that has been explained with the objectives and has been acknowledged in the workplan below as well.

Key Objectives:

- A set of hashtags or key words that will be the search terms for the tweets.
  - This will require both research and personal knowledge to find terms and words that will produce relevant tweet streams.
- A dataset of tweets collected via the Twitter API that fall into the terms defined.
  - Will require a twitter developer account and a python library to stream the tweets. Research will be required to decide on which library to choose.
- A number of data sources that can identifying the factors and help predict the sentiment.
  - This will require research online statistics, government guidance etc.
- A sentiment analysis model that will output the sentiment scores for a positive, negative, or neutral emotion.
  - There are many different sentiment models that work with python and a research will be required to be able to choose which model to use.
- A dashboard with visualisations showing the impacts different factors make on sentiment.
  - Will require different visualisations that help display the sentiment analysed and will be used to compare the change over the time of the project. research will be required to find the business intelligence tool best suited for the analysis.

Work Plan

I aim to follow each week as closely as possible. However, there is always a chance of unforeseen circumstances or incorrect time estimates which may lead to adaptations of the plan below. I have also not planned any work for easter which allows for any overflow of work to be completed if needed.

Key:

🏆 Milestone!

- Task to complete.

Week 1 - 01 February 2021

- Complete Ethical training course and submit Ethical consent form.
- Research:
  - What machine learning models are currently available
  - What Twitter API to use and what details can be extracted from tweets
  - What data visualisation and business insight tools are available
- Apply for Twitter Developer account
- Write initial plan.

Week 2 - 08 February 2021

- 🏆 Submit initial plan.
- Continue research and conclude what tools to use after discussing possible options with supervisor.
- Ensure business intelligence tool can analyse the factors to the intended level, discuss with supervisor.
- Decide the best method to store and analyse the tweets.

Week 3 - 15 February 2021

- Have a clear idea on how to implement the sentiment classification model.
- Define the key words and the predictive factors to analyse the data on

If Twitter Developer access is given:

- Begin writing Python programme to stream tweets by using the Twitter API.

Week 4 - 22 February 2021

- 🏆 Ethical consent form should have been approved at this stage.
- Start saving tweets with the correct factors being met. E.g., uses the word #Lockdown and was tweeted from the United Kingdom.
- Begin looking at ways the dataset can be analysed.
- Test creating visualisations from test data if insufficient number of tweets are collected.

Week 5 - 01 March 2021

- Continue saving tweets for dataset.
- Begin writing sentiment analysis model.

Week 6 - 08 March 2021

- Continue streaming and saving tweets.
- Continue writing sentiment analysis model.
- Continue creating a template for the analysis so once real data is extracted can easily be implemented into the dashboard.

Week 7 - 15 March 2021

- 🏆 Streaming of the tweets should be completed.
- 🏆 Sentiment model should be finished.
- Run the dataset of tweets through the sentiment model and create a dataset including the tweet contents, the location, and the predicted sentiment.

Week 8 - 22 March 2021

- Using the dataset create an analysis of the tweets.
- 🏆 Last working week before Easter break, the tweets should be analysed and stored with their relevant sentiment. A good dashboard analysing the tweets and showing the sentiment around the Country.

Week 9 - 19 April 2021

- Improve the analysis of the dashboard and try and find correlations.
- Try and improve the sentiment analysis of the tweets.
- Collect new tweets to compare sentiment at the start of the project with the sentiments towards the end of the project.

Week 10 - 26 April 2021

- Finish any tweak for the model and analysis.
- Begin writing final report.

Week 11 - 03 May 2021

- Continue writing up final report.

Week 12 - 10 May 2021

- Finish final report.
- 🏆 Submit project.

References

[1] https://www.aa.com.tr/en/health/covid-19-most-challenging-crisis-since-wwii-un-chief/1787121

[2] https://www.oberlo.co.uk/blog/twitter-statistics#:~:text=There%20are%20330%20million%20monthly,are%20between%2035%20and%2065.

[3] https://www.dsayce.com/social-media/tweets-day/