Initial plan - Sentiment Analysis

"Using NLP and machine learning to analyse a dataset of tweets"

Cardiff University School of Computer Science and Informatics

CM3203 – One Semester Individual Project - 40 credits

Author – Callum Haine

Supervisor – Jose Camacho

Project description

The social media platform 'Twitter' has been steadily growing in popularity since being founded over a decade ago in 2006. With over 330 million active monthly users, and 500 million tweets being sent every day [1], it provides a constant stream of user generated data which is openly available through the twitter API. Twitter is a platform on which users freely share their opinions on everything, from newly released products to famous public figures. For this project, I intend to analyse this dataset as a means of gauging user's sentiments on different companies, and tracking how they change over time. I have decided to pick a set of 10 to analyse, based on the size of their following on twitter. Currently, the 10 most followed brands on twitter [2] are:

1) Playstation (@playstation)
2) Xbox (@Xbox)
3) Chanel (@Chanel)
4) Samsung Mobile (@SamsungMobile)
5) Starbucks Coffee (@Starbucks)
6) Victoria's Secret(@VictoriaSecret)
7) Android (@Android)
8) Nintendo of America (@NintendoAmerica)
9) Rockstar Games (@RockstarGames)
10) SpaceX (@SpaceX)

Although I have picked these brands based on the size of their followings, I do not intend to solely collect tweets which are directed at the accounts themselves. Using Twitter's streaming API, I will be able to filter incoming data based on a set of criteria which includes usernames, locations, and most importantly keywords. By formulating a keyword list, I will be able to filter the stream for tweets corresponding to each of the brands respectively, even if the accounts themselves are not mentioned explicitly.

The project itself can be divided into three main areas; collecting the dataset, analysing sentiment, and tracking how sentiment changes over time. I will now discuss the first of these three in more detail.

In order to gain access to a stream of twitter data, I intend to use the python library 'Tweepy', which will allow me to access the twitter stream API in real time. Using this, I can then filter the stream by hashtags or keywords, and save them to a file. I will store the dataset for each company in .TXT format. As I intend the system to track how user's sentiment changes over time, I will need to ensure that the data frame for every tweet contains the date it was created alongside its textual contents.

Assuming I now have a dataset of tweets for each company to work with, I will next consider how best to analyse the sentiment of a given tweet. For this project, I am only concerned with whether the sentiment of the tweet appears to be positive or negative, or the sentiment polarity of the tweets. I intend to use machine learning techniques to determine the sentiment polarity.

Broadly, if I intend to use machine learning to analyse the dataset, I will first need to annotate the sentiment of a set of tweets myself, in order to train the system. Then, I can use a test dataset to supervise the algorithm, in order to form the predictive model which will be used on the real data. I will implement this using the python library 'sklearn'. When annotating, I will consider not solely relying on my own judgement, but also ensuring a subset of the tweets are annotated by another party. This is to ensure that my own biases do not affect the predictive model formed.

After achieving a fully working model for calculating the sentiment of the tweets, I will then be in a position where I will be able to calculate the mean sentiment for each company based on their respective datasets.

Finally, I intend to visualise this data for each company, so that user's opinions can be observed over a timeframe of several weeks, from when I begin collecting data to the end of the project. By including the date attribute in the data frame when collecting tweets, I can easily divide the dataset by day, week, or month and use machine learning to model how sentiment changes. This can be plotted to make it presentable. Doing this will provide an interesting insight into how positively the userbase views each company respectively, as well as how their opinions fluctuate over time.

Ethical considerations

Because this project involves the use of data taken from a social media platform, it is important that I gain ethical approval before beginning. I must complete an ethical approval form before I collect the dataset used later in the project.

Another ethical consideration is the fact that if I ask other individuals to annotate tweets when building the predictive model, I must ensure they have given consent prior.

Aims and objectives

The aim of this project is to use machine learning to analyse the overall sentiment held by Twitter's userbase, and visualise this over time. I will aim for the capability of the algorithm used to determine sentiment to have an acceptable level of accuracy, precision, sensitivity, and specificity. I hope that once I have a solid foundation, I can continue to train the algorithm in order to improve its performance.

In order to achieve this aim, I will first have several research objectives to complete:

- Research differences between twitter APIs, and different methods of utilising the API (e.g 'Tweepy')
- Research different methods for pre-processing raw text in order for it to be optimised for machine learning
- Research machine learning, how it works fundamentally, and how it can be implemented (e.g. with 'Sklearn')

After completing these research objectives, I will then have several objectives which I will need to complete:

- Use twitter API to build a dataset of tweets for the specific brands mentioned previously.
- Take this dataset and pre-process it to optimise for analysis, using the method decided on during supervisor meetings and from research.
- Successfully use machine learning to build a predictive model which can take the processed data and use it to calculate sentiment.

Work plan

I have outlined below the coding and research I intend to complete for every week of the project. As this is an initial plan, there may be unforeseen circumstances which require me to deviate from this, for example if a certain aspect is more time consuming than I initially believed it to be. There is also the possibility that I decide on different methodologies for analysing the sentiment after.

I will also need to consider the fact that, for this project to be successful, I will have to conduct research into topics which I may be less familiar with. I have tried to take this into account when creating my work plan by providing ample time at the start to build on my pre-existing knowledge.

At the start of the project, it will be of great importance that I start collecting data to form my dataset as soon as possible. This is because all subsequent work in the following weeks relies on the use of the datasets.

Week 1 (27/01/20)

- I will conduct Initial research into Twitter API and machine learning techniques.
- Discuss different machine learning implementations with supervisor.
- Apply for a twitter developer account so that I can access Twitter API in subsequent weeks.

Week 2 (03/02/20)

- Continue research into different machine learning implementations.
- Implement method for collecting and storing tweets using Twitter API and 'Tweepy'
- Discuss with supervisor most suitable format to store tweets in dataset.

Week 3 (10/02/20)

- More research into machine learning. By this stage I should have a good idea as to exactly how it will be implanted, possibly using 'sklearn'.
- Begin collecting the datasets. Collect a specific number of tweets periodically. I will assume to use the time period and number of tweets discussed in project description, although this is open to change.

Week 4 – 5 (17/02/20)

- Continue to build dataset.
- Take subsets of tweets to be used as training, validation, and test sets
- Begin work on pre-processing the training and validation sets, considering how features will be chosen.

Week 6 (02/03/20)

- Finalise method for pre-processing the data.
- Begin annotating the training set. When doing this, consider having a 3rd party check over some of my annotations in order to ensure that my own subjective views do not affect the predictive model.

Weeks 7- 8 (09/03/20)

- Finish annotation of training set.

- Work on using ‘sklearn’ alongside training and test sets to build predictive model.

Week 9 (23/03/20)

- By this stage, I plan to have a working model for analysing the sentiment of tweets in the dataset. I can now tweak this model in order to improve its effectiveness.

Week 10 (30/03/20)

- Use model on larger dataset, gain information on how the sentiment for each company changes over the time period I have been collecting tweets for.
- Collate this data into a presentable format.

Weeks 11- 12 (13/04/20)

- Finalise collected data.
- Write the final report.

References

[1] "Twitter Statistics," [Online]. Available: https://www.oberlo.co.uk/blog/twitter-statistics.

[2] "Statistics," [Online]. Available: https://www.socialbakers.com/statistics/twitter/profiles/brands.