Twitter Data Stream Sentiment Analysis

By

Group 5

Minor project Software Design and Application 2017-2018

Tim Al 4439848  
Casper Spronk 4369475  
Michiel O’Herne 4387783

Course Professor: Dr. Georgios Gousios

The following report covers our minor project in big data analysis. Detailing the challenges faced and the progress made over the span of several weeks.

**Table of contents:**

1. Project start
   1. Objectives
   2. Motivation
   3. Tools used
2. In-depth project setup
   1. Obtaining the data
      1. Data Sources
      2. Converting obtained data
   2. Analysing the data
   3. Visualisation of analysis
3. Results
4. Addendum
   1. Similar projects
   2. Alternate possibilities

**Project start**

The main objective we set out to accomplish was performing sentiment analysis on the content of tweets, and to present this in several directly understandable manners.

This task can be split into three distinct sections: gathering and sorting the data, performing analysis on the data and lastly portraying this analysis in a useful medium.

**Objectives: (general description)**

* obtaining the data
* analysis
* Visualisation of the analysis

Our objectives for were to first be able to collect data from a twitter using a twitter spritzer. A spritzer is a kind of twitter stream that gives one percent of the tweets that meet a certain set of filters. Other types of twitter streams give more data and would be preferred but the spritzer is the only one that can be used free of charge. The only thing that is needed to be able to use this type of stream is a twitter account. The filters that can be selected include the language of the tweets, words used in tweets, hashtags in a tweet and many more. After the tweets are obtained they need to be analysed to judge whether or not the tweet is positive of negative about the subject that it’s talking about. This will be done through the usage of an open source sentiment analyser. This analyser will rate a tweets positive, negative and neutral sentiment. By combing the sentiment about a larger volume of tweets we can discover a trend in the sentiment. This trend can be anything, for instance the general sentiment about the weather is most likely better on a sunny day than on a rainy day. By plotting the general sentiment in a line graph, we can see this. Another way of visualizing these results is in a pie chart. By using this we can see whether the general sentiment on a subject was through the entirety of the time that the twitter stream was open. By using the location data that can be included in a tweet, this depends on whether or not the user has opted in to sharing his or her location data, you can also figure out what a countries sentiment is. An example of this would be that during the World Cup 2016 final, you can see that Germany would be more positive about than Argentina, especially after the 1-0 was scored. The problem with this is that not a lot of people opt in to sharing their location data. This means that this is very inaccurate and can potentially lead to no data being obtained.

**Motivation:**

The idea for this project came from our interest in major sporting events. Twitter can be used to quickly find out how a group of people think about a certain subject. For instance, when a major sports event is happening, lots of tweets will be sent out with hashtags relating to that event. By filtering these tweets on hashtags and then running a sentiment analysis on the tweets, one can figure out which of the teams is the preferred team or if the game is exciting. One of the biggest sporting events every year is the Super Bowl. This is the final to the NFL season and the winner of that game is crowned champion. Around the time leading up to this event lots of tweets are sent out containing hashtags with regards to the two teams competing and a general hashtag about the Super Bowl itself. By analysing the tweets containing these hashtags we can find out which of the two teams is favoured by the fans in general and which team is favoured to win. During the Super Bowl itself we can use the sentiment analysis to determine whether or not the game is seen as exciting. If it is, the general sentiment about the game will increase. When a play ends in one of the two teams scoring points we can most likely see that in the sentiment in the tweets that contain the scoring teams hashtag. One of the two will have a very positive sentiment while the other will be very negative. The general sentiment will also be likely to be more positive in high scoring but close games, neutral fans will be more likely to enjoy this type of game, while the sentiment will be a lot less positive if the game ends up being a blowout, the fans of the winning team will be very positive about the game while neutral fans and those of the losing team will send out more negative tweets about the game.

Description of the use of what we did  
Description of made choices? Vs include in in-depth description

**Tools used:.**

1. **Python:** Programming language with an aptitude for data analytics tasks.
   1. **NLTK**: Natural Language Toolkit, a suite of libraries focused on the analysis of language.
      1. **Vader Sentiment Analyser:** Part of NLTK focused solely on sentiment analysis.
   2. **Plotly:** a graphing library built to provide a multitude of graphing solutions.
   3. **Tkinter:** A graphic user interface package.
2. **Twitter API:** an api supplied by twitter to allow streaming of their data.
3. **Github:** Code development platform using a web-based git version control repository.
4. **PyCharm:** an IDE tailored to the Python language

**In-depth project setup:**

**Obtaining the data:**

We have chosen to concurrently use two methods of data gathering, i.e. already created data sets obtained from a database, along with live streamed data directly from twitter itself. Our focus is directed more towards the streaming of tweets, as this data opens possibilities to provide real time insight into trends and public opinions. The pre-made datasets however form a way for us to test our analysing tools in a consistent method.

**Data sources:**

Twitter has set up its own api that allows for live streaming of tweets in real-time, as such our data stream is built around this api. The output of Twitter their api is encoded in JSON, containing not only the tweets themselves, but also all possible related data such as number of like, retweets and reactions tweets have received.

**Converting obtained data:**

Tools used to obtain data + convert said data into a usable format:  
Used environment / packages  
from-to formats

**Analysing the data:**

* Divide the data into positive - negative tweets
* Give the overall opinion of the public (i.e. twitter users) of the topic
* Give the popularity of given subject

Tools used to sentiment analyse the data  
used environment / packages

**Visualisation of analysis:**

Description of used manners of visualisation

Detail on tools used:  
used environment / packages

**Results:**

Couple of graphs + part of tweet data set?

**Addendum:**

**Similar project / methods:**

Multiple project setup to do twitter analysis

Mainly business focused models renting their services for research and development purposes

**Alternate methods we could have used:**

Spark streaming s