Applying Sentiment Analysis to Social Media Posts to Gain an Insight into the Public's Opinion of Movies

BACKGROUND

Twitter is one of the most widely used social media platforms with 326 million active users (Twitter, 2018) and over 500 million tweets posted per day (Onmicore, 2018).
Tweets are limited to 240 characters and often contain opinions about a wide variety of topics ranging from politics to movies.

AIMS AND OBJECTIVES

The aim of this project is to create a system which provides an alternative rating system to the existing movie review aggregation websites by using real-time social media posts. This is achieved by collecting a number of tweets about a chosen movie from Twitter, calculating the polarity for each of these tweets, aggregating the results for each tweet and then returning the result to the user.
Similarly to IMDb's rating system, this will help potential movie watchers understand the general opinion of a movie.
Generally, sentiment analysis is a technique only used by businesses to gain an insight into customer's opinions. However, this project is a consumer-aimed system.
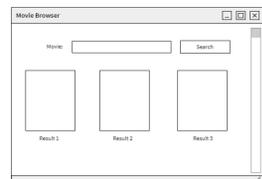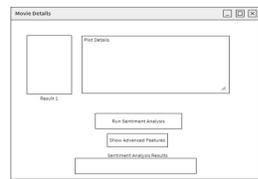

Figure 1: Movie Browser UI Wireframe

Figure 2: Movie Details UI Wireframe

SYSTEM

The system is comprised of three main components.
• Movie Browser - This allows the user can search for specific movies and retrieve information about them. The Movie Database (TMDb) API is used to access movie information. Figures 1 and 2 show plans for the two main forms.
• Tweet Retriever - This will retrieve tweets about the chosen movie using keyword matching using the Twitter API (free tier).
• Sentiment Analyser - This will calculate a polarity for each tweet that has been retrieved for the chosen movie.
A process flow chart is shown in Figure 3.

Movie Selection
The user would select a movie to be analysed
↓

Collect Tweets
Twitter API is used to collect tweets about the chosen movie
↓

Filter and Clean Data
This process eliminates noisy and irrelevant tweets
↓

Analyse Tweets
This will determine whether each tweet is positive/negative
↓

Result Aggregation
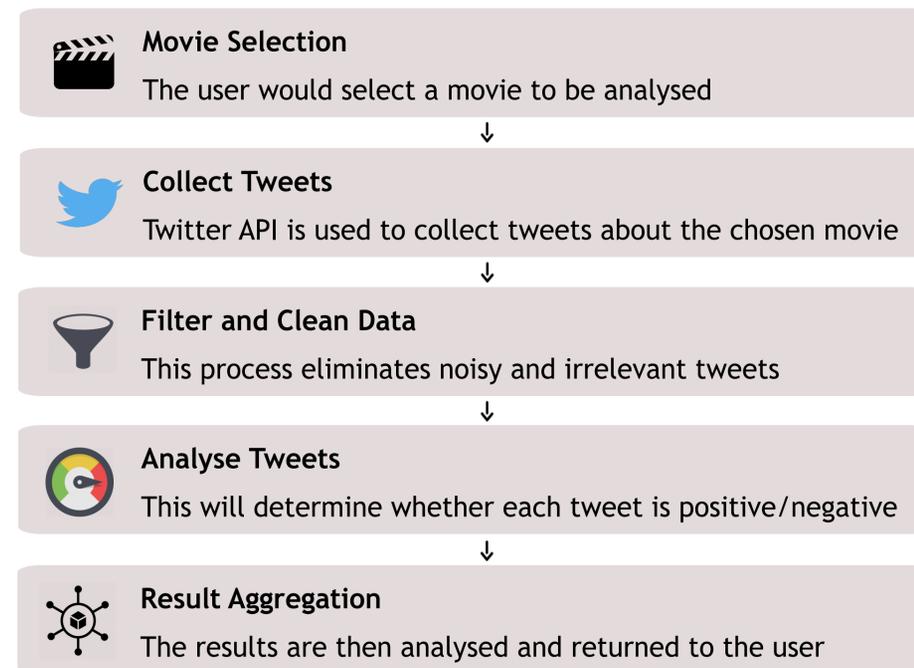The results are then analysed and returned to the user

Figure 3: Process Flow

MACHINE LEARNING

Sentiment analysis is achieved using supervised machine learning, so a suitable classification algorithm is required.
There are a number of suitable classifying methods but due to its simplicity and effectiveness compared to other methods, I will be using the Naive Bayes classifier. The equation is shown in Figure 4.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Figure 4: Naive Bayes Equation

The polarity of every word (B) is calculated using this equation to determine whether it is positive or negative (A). In order to calculate the polarity, the following datasets will be used to train the classifying algorithm.
• Stanford University Large Movie Dataset (Mass et al., 2011) - this is a set of 25,000 movie reviews retrieved from IMDb, split evenly between positive and negative reviews.
• Sentiment140 Dataset (Go, Bhayani & Huang, 2009) - this is a corpus of 1.6 million labelled tweets retrieved from Twitter.

ADDITIONAL FEATURES

• One additional feature that I would like to add is to give each aspect of the movie a polarity based on the tweet e.g. "Fantastic story but bad visual effects" would have a positive polarity for the story but a negative for the visual effects.
• Another feature I would like to add is to automatically display more recent/popular films to the user on a separate tab. This could help the user save time if they want to analyse a popular movie.

REFERENCES

Twitter (2018), Twitter Q3 2018 Earnings Press Release. Retrieved from https://investor.twitterinc.com/financial-information/quarterly-results .

Omnicore (2018), Twitter Statistics. Retrieved from https://www.omnicoreagency.com/twitter-statistics/.

The Movie Database (2018), API Overview. Retrieved from https://www.themoviedb.org/documentation/api?language=en-US

Twitter (2018), Twitter API Docs. Retrieved from https://developer.twitter/com/en/docs.html.

Maas, A, Daly, R.E, Pham, P.T., Huang, D., Ng, A.Y., & Potts, C. (2011), Learning Word Vectors for Sentiment Analysis. Retrieved from https://doi.org/978-1-932432-87-9

Go, A., Bhayani, R., & Huang, L. (2009), Twitter Sentiment Classification Using Distant Supervision. Retrieved from https://dl.acm.org/citation.cfm?id=2021114

Hamza Saeed | Supervisor: Simon Parkinson | University of Huddersfield School of Computing and Engineering