A Twitter Sentiment Analysis Application For Predicting The Success Of Upcoming Box Office Movies

To what extent can sentiment analysis accurately predict how well a film will do at the box office prior to its release?

Alex Stacey **SID - 6230867** **Date**

Dissertation submitted in completion for the degree of Bachelors of Science (Hons) in Computer Science.
Engineering, Environmental and Computing Department. Coventry University.

Project supervisor: Dr. Diana Hintea.

STATEMENT OF ORIGINALITY

300COM / 303COM DECLARATION OF ORIGINALITY

I Declare that This project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for the purposes of plagiarism prevention and detection.

STATEMENT OF COPYRIGHT

I acknowledge that the copyright of this project report, and any product developed as part of the project, belong to Coventry University. Support, including funding, is available to commercialise products and services developed by staff and students. Any revenue that is generated is split with the inventor/s of the product or service. For further information please see www.coventry.ac.uk/ipr or contact ipr@coventry.ac.uk.

STATEMENT OF ETHICAL ENGAGEMENT

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (https://ethics.coventry.ac.uk/) and that the application number is listed below (Note: Projects without an ethical application number will be rejected for marking)

Signed: AStacey Date: 30/04/2018

First Name:	Alex
Last Name:	Stacey
Student ID number	6230867
Ethics Application Number	P68008
1st Supervisor Name	Diana Hintea
2nd Supervisor Name	

1 Table of Contents

To what extent can sentiment analysis accurately predict how well a film will do at the box office prior to its release? 1

Statement of Originality 2

300COM / 303COM Declaration of originality 2

Statement of copyright 2

Statement of ethical engagement 2

2 Abstract 5

3 Technical terms and definitions 6

4 Introduction 7

4.1 Project background and Motivation 7

4.2 Aim and objectives 7

4.3 Research 7

4.4 Related projects 8

5 Literary Review 9

5.1 Methods for conducting sentiment analysis 10

5.2 Sentiment analysis projects in certain real-world areas 11

6 Primary Research 13

6.1 Gathering the Data 13

6.2 Tweepy Module and the Twitter API 13

6.3 SQL and Database structure 14

6.4 Limitations of the Twitter API 14

6.5 GetOldTweets Module 14

6.6 Receiving the tweets 15

6.7 Sentiment Analysis 17

6.8 Machine learning classifier 17

6.8.1 NLTK Library 17

6.8.2 Naive Bayes classifier 18

6.8.3 Training my classifier 18

7 Results and Analysis 19

7.1 Testing my classifier 19

7.2 Final Sentiment Classifier 20

7.3 Reviewing the Data 21

7.3.1 Polarity graphs 22

7.3.2 Subjectivity Graphs 23

7.4 Developing a Prediction System – Data mining 24

7.4.1 Weka and the J48 Decision Tree 24

7.4.2 The data used for the system 24

7.4.3 The Decision trees developed 25
7.4.4 Building my decision tree 26
8 My Final System 29
Screenshots Of The Final System 29
8.1 Database structure 31
9 Evaluation 32
9.1 Primary Research Evaluation 32
9.1.1 Data Gathering 32
9.1.2 Test Classifier 32
9.1.3 Data mining 32
9.1.4 Conclusion and other factors 33
10 Discussion 34
10.1 Achievements 34
10.1.1 Development of Text Classifier 34
10.1.2 Final System GUI and Prediction System 34
10.2 Deficiencies 34
10.2.1 Not enough data 34
10.2.2 Lack of Testing 35
10.2.3 Lack of Planning 35
10.3 How I would do it differently 35
11 Reflection 37
11.1 Ethics of my project 37
11.2 diligence 37
11.3 Problem encountered 37
11.3.1 Twitter API Limitation 37
11.3.2 Text classifier to naïve 37
11.4 Supervisor meetings and feedback response 37
11.5 Project timeline 38
12 Conclusion 39
13 Bibliography 40
14 Appendices 42
14.1 Appendix 1 – Project proposal 42
14.1.1 To what extent can sentiment analysis accurately predict how well a film will do at the box office prior to its release? 42
14.2 Appendix 2 – Github links 44
14.2.1 Primary Research Repository 44
14.2.2 Final System Repository 44

2 ABSTRACT

To what extent can sentiment analysis accurately predict how well a film will do at the box office prior to its release? This individual project aims to use Twitter sentiment analysis to find out the extent of its utility in predicting the future domestic gross of a film. The problem that my key aim will solve is the uncertainty that can surround certain films futures, the lack of knowledge about the general narrative surrounding a film can lead to poor advertisement and an ultimately financial loss for the production company. The solution I have created is a GUI system that acts as a frontend for the back-end system developed through research gained from gathering tweets about previously released films. Once the data about the films was gathered I conducted sentiment analysis and data mining on this data to develop a decision tree capable of using sentiment analysis results to output a prediction. The sentiment analysis was done using a combination of a machine learning classifier trained through a self-built training set and a simple text classifier that uses a bank of positive and negative words for comparison. Overall the project was a success to a degree as I built the final system and developed an accurate classifier and extensive back-end system, however the final prediction system is yet to be fully tested.

3 TECHNICAL TERMS AND DEFINITIONS

Twitter API - This online Application programming interface is what will be used as a middle man between the online twitter meta data I need to access to conduct the sentiment analysis. To access this, I had to create a developer account with Twitter and in return I receive access tokens that will be used with the code to allow my program to access the necessary data.

Sentiment analysis – the process of determining the sentiment (positive, negative or neutral) of a body of text in order to understand the general sentiment of the text so that this information can be used in data mining analytics.

Data Mining - the process of looking at a database of current data in order to generate new information such as predictions.

Tokenization – the process of taking a body of text and breaking it down into a list of its individual words

Stop word Removal – the process of going through a list of words produce through the tokenization process and removing the stop words (words that hold no sentiment such as 'and')

Regex – the process of removing special characters such as '@' from the tweet

Machine Learning Classifier - a text classifier created through training a classifier using training data sets that contain tweets with pre classified sentiments.

4 Introduction

4.1 Project background and Motivation

To what extent can sentiment analysis accurately predict how well a film will do at the box office prior to its release? The aim of this project is to investigate the utility of Twitter sentiment analysis in a specific subject area to see if it is possible to develop a program that is capable of an accurate prediction based on Twitter data (tweets) and the sentiment that this data has. The subject area I have chosen to investigate is movies given the fact that it is a very opinionated subject area, people often voice opinions about this subject area on Twitter prior to the films released due to production news and trailers being released. Social media carries a wealth of information that is continuously being generated and this data can prove invaluable in the deduction of the general narrative about a certain topic area. In a world where data is becoming more and more valuable companies and businesses have a need for this form of data analysis to prevent financial loss. The film industry is an ever-expanding market and a vast amount of films are released per year with many being great successes and others being large financial losses and I believe that the key to separating these lies in Twitter sentiment analysis as many people take to social media to voice opinions about movies. Using this data production companies could take certain measures to change the narrative about their film by potentially remarketing etc.

My motivation for this project is my interest in the subject matter and the concept of using existing data to develop a predictive system. Despite my existing interest in subject I believe that this project is important as in today's film industry given the large amount of films that are released which in turn has led to a large marketplace which many companies thrive and others suffer from financial losses. My system will aim to estimate the domestic financial outcome of a film so that production companies may be as successful as possible when releasing a new film as well as to provide useful information and data to help users understand the nature of the film industry through Twitter.

4.2 Aim and objectives

The key goal I hope to achieve within this project is to see as to what extent Twitter sentiment analysis can be used to make predictions about the potential future success of an upcoming box office film. To make such predictions I will have to use data mining to generate new information which will be the prediction. Within the backend of the system what I aim to develop will include these features:

1. A main module that uses access tokens to connect my application to the Twitter API so I can query it to receive tweets about a certain film
2. A text classification module that can perform the tokenization, stop word removal and regex special character removal. Once these stages have completed the text classifier will get the sentiment of the tweet by comparing the words within the tweet to a bank of positive and negative words so the sentiment can be determined.
3. A prediction system using Weka to create decision trees that use the results of the sentiment analysis to predict the potential future outcome of a films performance in the domestic market (prediction will be based on gross)

Once the backend modules have been developed in the primary research phase I aim to develop an application within Python that has a GUI for better user accessibility. The end goal of all this development is to investigate the effectiveness of the results gained from the sentiment analysis coupled with data mining analytic techniques in making an accurate prediction.

4.3 Research

To conduct my research, I will need a system that can produce results and collect the data that is required for my analysis. To do this I will be programming an alpha version of my application that will be used to gather and store the necessary data that I will used to preform data analysis. The method and explanation of which can be found in the primary research section of this report. Once all my data is gathered and stored it will then have to perform data mining analytics to draw out a prediction pattern in the form of a decision tree that can classify a films future based on the sentiment analysis.

4.4 Related projects

There have been many previous research projects around the area of sentiment analysis such as using it to gather information about political elections and the stock market, these studies have been outlined in the literature review in aid of the general understanding of the subject area.

5 Literary Review

Social media is a continuously growing area of technology that produces a constant stream of data and Twitter is no exception from this. Twitter the microblogging website that was started in 2006 and since that time millions of people are using the service to share opinions about certain topic areas. Many previous research projects have looked into key areas of my proposed project which include data mining and sentiment analysis to make predictions as to the events of the future. The sources I will be considering in this review are all in areas heavily related to my project such as how to build sentiment analysis systems, data mining projects and sentiment analysis systems that have be created to make predictions within certain topic areas such as the stock market. The aim of this review is to understand the techniques and methods used within the projects so that I may apply my understanding within my own project so that I may produce the best possible application.

The first research project I investigated was very similar to my proposed project in that it aims to predict the future success of a movie using sentiment analysis in which they deduce that movies make a great topic area for conducting this type of analysis due to the great interest people take in the subject matter and the real-world outcomes can be easily observed (Asur, S. and Huberman, B. 2010). To conduct this project, they started out by establishing rules before carrying out their research which included things such as picking films with unambiguous titles and ones that had a wide release (Asur, S. and Huberman, B. 2010) which is something that independent films may not have so it would cause outliers in the data. When collecting data, the method used was gathering a stream of data for 24 different movies over three months and the data that was taken for each movie was the information withdrawn over the *critical period* (Asur, S. and Huberman, B. 2010). The *critical period* is defined by the researchers as the period one week prior to a film's release and two weeks after (Asur, S. and Huberman, B. 2010) and this idea is one I will employ within my project given that it is a key period within the film industry as this is in theory is when the greatest volume of tweets will be available for the film. Prior to analysing sentiment this project looked a "tweet-rate" which is the number of tweets per hour (Asur, S. and Huberman, B. 2010), using this information they could deduce a positive correlation within a linear aggression model showing that the films that had a high tweet-rate in the first week of the critical period had a more profitable opening weekend (Asur, S. and Huberman, B. 2010). The idea of tweet volume and rate is one I will carry through into my project and my research as it can aid in the prediction progress.

This paper contained many ideas that could be applied to my project and one such idea was the concepts of polarity and subjectivity (Asur, S. and Huberman, B. 2010). From their research they concluded that sentiment is stronger during the second week of the critical period as people have seen the film and their opinions are based on personal experience rather than anticipation which means people are more likely to be influenced by the week two opinions, they define this measure as :

$$subjectivity = \frac{Total\ number\ of\ positive\ and\ negitive\ tweets}{Total\ number\ of\ neutral\ tweets}$$

Equation 1 – Subjectivity (Asur, S. and Huberman, B. 2010).

This idea once again is something I must consider for my project because I will have to use this knowledge to understand the patterns involved with the Twitter site so that I only extract necessary data. The final sentiment measurement this project employed was polarity which is a method of quantifying the sentiment of an individual movie so that it can translate this into a value to develop a prediction (Asur, S. and Huberman, B. 2010). Polarity is simply the ratio of positive tweets to negative and what they are concluding from the use of this measurement was that an increase or decrease in polarity it had an effect on the weekly revenue, furthermore adding this variable to the linear regression model from before caused improvements meaning sentiment analysis can affect the accuracy of the prediction (Asur, S. and Huberman, B. 2010). Polarity is a measure that will be heavily employed in my project given that I am mainly focusing on sentiment analysis as a method for prediction with the rate of tweets being an additional variable to improve predictions.

5.1 METHODS FOR CONDUCTING SENTIMENT ANALYSIS

The following projects consider sentiment analysis and the methods in which specifically the Twitter corpus is used for sentiment analysis. The first research paper I considered was concerned with creating and training a Twitter sentiment analysis system, to start this they firstly considered emoticons to gather sentiment given that when this was conducted Twitter posts were limited to 140 characters and therefore normally only one sentence, however they do remove these when filtering the tweet (Pak, A. and Paroubek, P. 2013). This concept is one I previously overlooked when visualising and developing this idea so it is one I will consider implementing within my system as it can provide a strong insight into the overall sentiment of the tweet. Another concept discussed was the use of n-grams, n-grams are defined as the set of co-occurring words within a string being processed by a computer (What Are N-Grams? 2014). N-grams are used within the research project, particularly higher-level ones such as trigrams as it is deemed better for capturing sentiment whereas unigrams are better for providing full coverage of the data (Pak, A. and Paroubek, P. 2013). The use of n-grams is one I considered at the start of this project as in the natural language it is necessary to consider n-grams in the following ways:

Tweet: “I do not like the movie Avatar”

1. Unigrams will be employed to consider each word within the tweet
2. Bigrams are trigrams are necessary to consider as “ do not like” and “not like” is a negative phrase where as “like” is positive

Finally, after testing the different n-grams within their system they determined that bigrams were the most effective as they combine the coverage of the unigrams with the sentiment analysis capabilities that come with the higher-level n-grams (Pak, A. and Paroubek, P. 2013) and that is why I will employ this technique within my individual project.

Another research project I considered that also had the goal of building a sentiment analyser for Twitter that used machine learning techniques on a training data set to develop their system. Similarly, to the previous project they mention the heavy use of emoticons within Twitter posts and how they can give insight to the sentiment but like in the afore mentioned project they removed them from the tweets within the training set as they can in fact lead to confusion and are defined as ‘noisy labels’ given that they may falsify a sentiment within a tweet (Go et al. 2009). Armed with this new information I will not rely heavily on emoticons as a method for determining sentiment however with proper implementation considering them may prove helpful and this is something to consider when conducting my primary research. Another idea raised within this paper is the use of a baseline for developing a sentiment for a tweet, a baseline is simply a bank of positive and negative words to gain an initial idea of the sentiment (Go et al. 2009). Of course, the idea of a baseline is a necessity for my project as the only way I can begin to get the sentiment of each individual tweet is to compare it to the words within to the English language. Once again as I read through this paper the use of n-grams was brought up however this paper provided more insight into the subject area, previously it was concluded that bigrams were the most effective but this was contradicted in this study. The use of bigrams within this study proved to lower the accuracy so in the end they concluded that the use of a combination of unigrams and bigrams was the most effective in increasing accuracy of the system (Go et al. 2009). The use of n-grams was also discussed previously but now it is clear that I must implement a combination of n-grams to provide coverage as well as establishing the sentiment of the tweet. Finally, the technique of POS-tagging was employed in this project with the conclusion that I wasn’t that useful (Go et al. 2009) which contracted the afore mentioned project given that they used it as a key feature alongside n-grams (Pak, A. and Paroubek, P. 2013). POS-tagging is something that I will not be considering for my project given how this research project claims that overall, they aren’t useful.

5.2 Sentiment analysis projects in certain real-world areas

Twitter sentiment analysis not only has been applied to the film industry but also other areas such as the stock market and politics. One of the first research papers I read involved using sentiment analysis in combination with other data analysis techniques to predict the movement of the stock market (Pagolu et al. 2016). Within this paper it is stated that prior to the development of sentient analysis using Twitter data the prediction of the stock market was rather inaccurate given the very unpredictable patterns the market follows in response to real world events (Pagolu et al. 2016). The method of pre-processing employed within this paper is one that I will implement into my system and works as follows:

1. Tokenization – the process of breaking a tweet down into its individual words so that they may be independently identified (Pagolu et al. 2016).
2. Stop word removal – after the tweet is tokenized a stop word removal process is conducted to remove the words that hold no sentiment (Pagolu et al. 2016).
3. Regex matching – the final step in the process is to remove the special characters and 'clean' the tweet (Pagolu et al. 2016).

The following stages of pre-processing are ones I am keen to implement into my final system as I believe that it is a logical process that is very effective in getting the data into a state that allows for sentiment analysis. Another interesting point raised by this paper is the fact that the analysis corpus must be made specific for the subject area meaning that for my project I must develop my own system from scratch rather than using pre-built modules and systems (Pagolu et al. 2016). Due to the fact this paper shares the goal of predicting the future using sentiment analysis as the backbone of the prediction the processes and methodologies used will cross over into my project.

Sentiment analysis has also been used in studies to help understand the general narrative of political elections and to even help predict the outcome of such elections. The first paper like this I read was concerned with the 2012 U.S presidential election which concluded the volume of tweets was driven by real world events (Wang et al. 2012). The idea of volume of tweets relating to real world events was one raised previously through the idea of the critical period which is a key concept that will be carried through to my system. Another key feature of this project was the use of a stream of data output that contained sentiment results such as volume of tweets with these results being continuously updated every five minutes (Wang et al. 2012). The idea of a continuous data stream is not one I will employ into my system however the concepts and ideas raised within this paper are interesting and will help in the understanding of the results my system will return. Another similar research paper I considered with the goal of monitoring election results and predicting their outcome found that the human language is complex when analysing sentiment overall (Bermingham and Smeaton 2011). What is meant by this deduction is the fact that the initial response to a real-time event may not reflect the overall sentiment and this is why the idea of the critical period is key so that analysis is spread over a period of time to get a better idea of the general sentiment (Bermingham and Smeaton 2011). Furthermore, this paper suggested that the volume of tweets can hold great insight as to the popularity of a certain topic or idea which is something that cannot be overlooked in my final system (Bermingham and Smeaton 2011). From these two projects I have concluded that maybe deciding whether tweet is simply positive or negative may not be enough to get an accurate idea about the overall narrative of a topic, rather volume of tweets can also provide great insight in combination with sentiment analysis.

In conclusion it is clear to see that there are many key practices and ideas that I can draw from these research projects to apply into my own. The ideas drawn from the movie sentiment analysis project such as the critical period, polarity and subjectivity will be used throughout my project as key concepts to be used in the results and analysis phase as well as the primary research phase. The methods for text classification drawn from the other projects such as n-grams will be employed within my project as it is clear that methods such as these are key for determining the sentiment of a tweet. The methods for pre-processing drawn from the stock market project will prove fundamental when I develop my text classifier as the data must be pre-processed before gauging the sentiment of the tweet as things such as emojis can act as "noisy labels" and lead to the incorrect classification of text. Finally, the ideas drawn from the political analysis projects will prove vital in the understanding of sentiment analysis as a whole, as it is from these projects I have understood that you must consider the volume of tweets as well as the sentiment over a period of time given that sentiment at one instance of time isn't enough to justify a predication.

6 PRIMARY RESEARCH

To conduct this project, I used the Python programming language to gather my data on films from 2017 – early 2018 that either performed well in terms of box office domestic gross or under performed. Upon gathering this data, I set out with the goal of creating the best possible text classifier through static text classification and later machine learning. Upon gathering the data, I will conduct data mining analytics to draw out patterns to help me develop a decision tree that will help generate a prediction based on the sentiment results that are gathered. Upon developing all this back-end to the final system I will code a GUI so that the back-end system can be easily manipulated to generate the required results (conclusion of API search and prediction).

6.1 GATHERING THE DATA

The data to be gathered will be ten films, some of which will be films that 'flopped' at the box office so I can understand what results of sentiment analysis correlate with a film that under performed. I'm defining under performed as a movie that only took less than 35% of its total gross from the domestic market and the reason for this is because I am looking at only English tweets and films that have a domestic release in America or the UK as often films can eventually claim their money back in the international market given its vast customer base.

Film Title	Gross	domestic %	Budget	Profit
Black Panther	$ 681,084,109.00	51.40%	$ 200,000,000	$ 481,084,109.00
Beauty and the Beast	$ 504,014,165.00	39.90%	$ 160,000,000	$ 344,014,165.00
Star Wars:The Last Jedi	$ 620,181,382.00	46.50%	$ 250,000,000	$ 370,181,382.00
Guardians of the Galaxy Vol. 2	$ 389,813,101.00	45.10%	$ 200,000,000	$ 189,813,101.00
Dunkirk	$ 190,068,280.00	36.00%	$ 100,000,000	$ 90,068,280.00
King Arthur: Legend of the Sword	$ 39,175,066.00	26.30%	$ 175,000,000	$ -135,824,934.00
Ghost in the shell	$ 40,563,557.00	23.90%	$ 110,000,000	$ -69,436,443.00
Valerian and the City of a Thousand Planets	$ 41,189,488.00	18.20%	$ 177,200,000	$ -136,010,512.00
A Cure for Wellness	$ 8,106,986.00	30.5%	$ 40,000,000	$ -31,893,014.00
Geostorm	$ 33,700,160.00	15.20%	$ 120,000,000	$ -86,299,840.00

Table 1 – Table showing a films domestic gross, the percentage of its overall earnings that the domestic gross contributed, the films production budget and the profit made in the domestic market (Box Office Mojo 2018)

For each of these films I will gather a certain number of tweets for each week in the critical period, the critical period which was defined in the literary review as the week before a film's release and the two weeks after the release date. The idea of the critical period is a very key part of my research method as it is an idea that will be carried through to the final system. Furthermore, the change in polarity and subjectivity will be key values to consider when generating a prediction or adjusting an existing prediction.

6.2 TWEEPY MODULE AND THE TWITTER API

To access the tweets, I had to access the Twitter API and to do this is had to sign up for a developer account and register my application. Upon registering my application, I was given a set of access tokens which will be used in a Twitter API class. The Tweepy module is a Python module that has built in methods to interact with the Twitter API. The first Python script I had to create was my Twitter API script that handles API interaction such as accessing the API and querying the API for tweets. The code for this was refined and developed from an example code on building a sentiment analysis system in Python (Kumar 2018). The tutorial that I based my API script on outlined how Tweepy works and how it can be used to gather tweets. My Twitter API script included a Twitter API class and methods to:

1. Query the Twitter API for the tweets about a given film
2. Create a database table to store the tweets gathered
3. Populate the database with the tweets gathered
4. Count the total number of tweets and the number of positive/neutral/negative tweets

So, when an object of the class was created the system would create a database table with the film title being the table name and for each tweet it would conduct sentiment analysis upon it and put all this data into a Python dictionary which then is used in a method that inserts the values into the database.

6.3 SQL and Database structure

To store all the data that was gathered I used SQLite with Python to store the data in a relational database. For the primary research stage of the project the database structure was very simple as it only contained tables for each film I would be researching to develop my prediction system. The design of the table is as follows:

<table>
<tr><th>Column Name</th><th>TweetID</th><th>Tweet</th><th>Clean Tweet</th><th>WeekNo</th><th>Sentiment</th></tr>
<tr><td><b>Data type</b></td><td>INT auto-increment</td><td>Text</td><td>Text (Python list)</td><td>INT (1,2 or 3)</td><td>TEXT</td></tr>
<tr><td colspan="6"><b>Examples Of actual tweets gathered</b></td></tr>
<tr><td><b>1</b></td><td colspan="2">Got my black panther tiks for next week, wasnt going to go but my sister (step) talked me into it i had to do reserve seating tho lol</td><td>['got', 'black', 'panther', 'tiks', 'next', 'week', 'wasnt', 'going', 'go', 'sister', 'step', 'talked', 'i', 'reserve', 'seating', 'tho', 'lol']</td><td>1</td><td>Positive</td></tr>
<tr><td><b>2</b></td><td colspan="2">The team behind “ Black Panther ” is facing allegations by a British-Liberian artist that her work was used without permission in Kendrick Lamar’s video for “All the Stars,” from the movie’s soundtrack http://nyti.ms/2Ek0rCk</td><td>['team', 'behind', 'black', 'panther', 'facing', 'allegations', 'british', 'liberian', 'artist', 'work', 'used', 'without', 'permission', 'kendrick', 'lamars', 'video', 'stars', 'movies', 'soundtrack', 'nyti', 'ms', '2ek0rck']</td><td>1</td><td>Negative</td></tr>
<tr><td><b>3</b></td><td colspan="2">I wonder why black panther is screening in the uk before America</td><td>['i', 'wonder', 'black', 'panther', 'screening', 'uk', 'america']</td><td>1</td><td>Neutral</td></tr>
</table>

The reason I have chosen this for my database is because I have a lot of experience with SQL and manipulating the data held within a relational database. In addition to this the data I needed was of a format that it could be properly stored in a relational database.

6.4 Limitations of the Twitter API

The Twitter API is a key part of my project as it provides me access to Twitter data, however it only provides tweets from a week prior to the current date in which you query the API. Given the fact that the films I have lined up to gather tweets on were mostly released in 2017 I couldn’t access tweets within the critical periods of any on the films I had planned to research. Upon learning this I had to look at alternate modules that would provide me access to tweets from as far back to the start of 2017 and during my research I found a module called get GetOldTweets.

6.5 GetOldTweets Module

The Twitter API as previously discussed has limitations as to how far back you can get tweets, however this module provides an alternative route for gathering tweets that allows you to go back to last year which is what was needed for my primary research. The way in which this works is that it goes through Twitter pages separately given the fact that when you go through a Twitter page and scroll through it a scroll loader starts which in turn provides access to more tweets which is made possible through calls to a JSON provider (Henrique 2018). For my primary research this module will be used to get tweets for each of the films critical period.

6.6 Receiving the tweets

Now that all the tools to conduct my primary research are in place it was time for me to gather all the tweets and to do this I used the following method within the Twitter API class:

```
def getCriticalPeriodTweets(self,query,releaseDate):
        """A function that returns tweets about a query within a certain time period date:yyyy-mm-dd"""
        try:
            rDate = str(releaseDate)
            q = str(query)
            #this method computes quite slow given that the module is work around to the twitter API week limit on tweets
            #given that user tweets are stored far back in time this seatches users timelines

            #first week of crtical period (the week prior to a films release)
            releaseDate1 = datetime.datetime.strptime(rDate, "%Y-%m-%d")
            week1StartDate = releaseDate1.date() - datetime.timedelta(days=7)
            releaseDate1Str = str(releaseDate1.date())
            week1StartDateStr = str(week1StartDate)
            print(week1StartDateStr)
            print(releaseDate1Str)
            tweetCriteria = got.manager.TweetCriteria().setQuerySearch(q).setSince(week1StartDateStr).setUntil(releaseDate1Str).setLang('en').setMaxTweets(50)
            #list to store tweet and sentiment
            tweets = []
            #list to just store tweet text to so we dont have duplicate tweets
            listOfTweets = []
            index = 0
            while index < 50:
                tweet = got.manager.TweetManager.getTweets(tweetCriteria)[index]
                if tweet.text in listOfTweets:
                    continue
                else:

                    #store tweet in dictionary
                    returnedTweet = {}
                    returnedTweet['tweet'] = tweet.text
                    returnedTweet['sentiment'] = self.getTweetSentiment(tweet.text)
                    returnedTweet['week'] = 1
                    tweets.append(returnedTweet)
                    index = index + 1

            #2nd week of the critical period
            week2StartDate = releaseDate1.date() + datetime.timedelta(days=1)
            week2EndDate = week2StartDate + datetime.timedelta(days=6)
            print(week2StartDate)
            print(week2EndDate)
            week2StartDateStr = str(week2StartDate)
            week2EndDateStr = str(week2EndDate)
            tweetCriteria = got.manager.TweetCriteria().setQuerySearch(q).setSince(week2StartDateStr).setUntil(week2EndDateStr).setLang('en').setMaxTweets(50)
            j = 0
            while j < 50:
                tweet = got.manager.TweetManager.getTweets(tweetCriteria)[j]
                if tweet.text in listOfTweets:
                    continue
                else:
                    # store tweet in dictionary
                    returnedTweet = {}
                    returnedTweet['tweet'] = tweet.text
                    returnedTweet['sentiment'] = self.getTweetSentiment(tweet.text)
                    returnedTweet['week'] = 2
                    tweets.append(returnedTweet)
                    j = j + 1

            #3rd week of the critical period

            week3StartDate = week2StartDate + datetime.timedelta(days=7)
            week3EndDate = week3StartDate + datetime.timedelta(days=6)
            print(week3StartDate)
            print(week3EndDate)
            week3StartDateStr = str(week3StartDate)
            week3EndDateStr = str(week3EndDate)
            tweetCriteria = got.manager.TweetCriteria().setQuerySearch(q).setSince(week3StartDateStr).setUntil(week3EndDateStr).setLang('en').setMaxTweets(50)
            k = 0
            while k < 50:
                tweet = got.manager.TweetManager.getTweets(tweetCriteria)[k]
                if tweet.text in listOfTweets:
                    continue
                else:
                    # store tweet in dictionary
                    returnedTweet = {}
                    returnedTweet['tweet'] = tweet.text
                    returnedTweet['sentiment'] = self.getTweetSentiment(tweet.text)
                    returnedTweet['week'] = 3
                    tweets.append(returnedTweet)
                    k = k + 1

            query1 = q.replace(" ", "")
            self.createDBTable(query1)
            cur.execute(''''SELECT COUNT (*) FROM sqlite_sequence WHERE name = ?''', (query1,))
            tableInDB = int(cur.fetchone()[0])
            #check if db table already exists
            if tableInDB > 0:
                pass
            else:
                for tweet in tweets:
                    tweetText = str(tweet["tweet"])
                    cleanTweet = str(TextClassifier.stopWordRemover(str(tweet["tweet"])))
                    sentimentOfTweet = str(tweet["sentiment"])
                    week = tweet["week"]
                    cur.execute(
                        """INSERT INTO """ + query1 + """ (tweet,cleanTweet,sentiment,weekNo) VALUES (?,?,?,?)""",(tweetText, cleanTweet, sentimentOfTweet, week,))

            con.commit()
        except tweepy.TweepError as error:
            #print error (if any)
                    print("Error : " + str(error))
```

How the code works:

1. The first stage of this code is to take the release date of the film and generate dates for the weeks within the critical period
2. Then for each week it gets tweets from that period of time using the GetOldTweets TweetCriteria() function
3. The tweets gathered are stored in a list of dictionaries, each dictionary contains the tweet, the sentiment given to the tweet and the week number

4. Then the list of dictionaries is iterated over and the values are inserted into the database table created (created when the function is called later in the code)

For each film I limited the tweets returned to 50 per week given the long computation time of the GetOldTweets module and I felt that 50 per week would be enough to draw out sufficient patterns and results. Once I had gathered all the tweets about each film (150 per film) it was time to review the data and perhaps improve upon my initial system as at this stage my sentiment classifier was very simplistic and not that effective at properly determining sentiment.

6.7 Sentiment Analysis

In the early stages of the development of the alpha version of my system I developed a very simple and naïve text classification approach that worked as follows:

1. Clean the tweet of links and any non-alphanumeric values
2. Tokenisation which involved splitting the clean tweet into a list of words
3. Stop word removal which is the process of removing stop words (and, you etc.). The stop words were identified via a text document I create that contained all the stop words in the English language. Each word in the list of words would be compared to each line in this text file (stop word) and If there was a match the word was removed.
4. The next stage was to get the sentiment from the final list that was returned from the stop word removal function. To get the sentiment of a tweet I used two text files that contained a bank of positive words in the English language and a bank of negative words. These text files were used to compare the words in the list with. When comparing the words in the list if a word matched a word in one of these text files then a positive or negative counter was increased depending on which word was matched. However, before this simple word check was conducted there was an n-gram analysis stage. N-gram analysis involved looking at tri-grams and bi-grams of the tweet e.g "not good" and "not very good" and if an n-gram contained a positive or negative word and also contained a word such as 'not' to imply the opposite meaning of the following word then the appropriate counter was increased.
5. Once all this analysis was conducted the counter with the highest value would determine the sentiment of the tweet and if they were equal then the sentiment would be returned as neutral.

This simple sentiment analysis system is the one I used when gathering the data and was the sentiment assigned to the tweet in the early stages of the system. Once I had gathered all the data it became clear that my text classifier was too naïve and incapable to producing accurate sentiment analysis so I looked into other methods for determining sentiment.

6.8 Machine learning classifier

Once I developed my initial text classifier it became clear that it was too naïve and needed to be enhanced if it was going to be able to accuracy determine sentiment as this would later affect the prediction generated. So, upon research into the topic area of sentiment analysis using Python I discovered that an effective method of sentiment analysis was to use a machine learning approach to train a text classifier with a given corpus.

6.8.1 NLTK Library

The Natural Language Tool Kit is a Python module that works with the human language that contained a wide range of corpuses such as a movie review corpus and a tweet corpus (Natural Language Toolkit — NLTK 3.2.5 Documentation 2018) . Furthermore, this module provides methods for text classification such as tokenization and POS tagging which will prove to be vital in developing an accurate text classifier (Natural Language Toolkit — NLTK 3.2.5 Documentation 2018). Finally, after more research into the NLTK library I discovered that you could in fact use a machine learning approach to develop a sentiment classifier, known as a Naive Bayes Classifier (Build A Sentiment Analysis App With Movie Reviews 2018)

6.8.2 Naive Bayes classifier

With the NLTK library within Python you can gain access to a huge range of text classifications tools and one such tool is a build-in function that allows you to create a Naïve Bayes classifier given a training data set such as a movie review corpus. The Naïve Bayes classifier is in fact a collection of algorithms that are based on Bayes theorem (Waldron 2015). The algorithms that are part of this collection share a common classification model which involves classifying each feature independently (Waldron 2015). In terms of my text classifier this means that each word is considered independently and then later as a whole may be considered positive, negative or neutral based on how the classifier is trained. However due to this independent evaluation technique it is regarded as a naïve approach given that the correlations between features may prove to provide more insight (Waldron 2015).

6.8.3 Training my classifier

When I started researching this topic area I soon discovered a very helpful tutorial sheet that showed you what form the training set needed to be in and explained how to train the text classifier. From this tutorial I found that the trainer in Python had to take the training data in this form:

[({‘Hello’:True,”World!”:True},sentiment)]

This is a list where each element is a tuple that contains a dictionary of the words within the tweet and the sentiment, this is known as getting the word features (Build A Sentiment Analysis App With Movie Reviews 2018). The function is as follows:

```python
def getWordFeatures(words):
    #stopword removal process
    usefulWords = [word for word in words if word not in stopwords.words("english")]
    dict1 = dict([(word, True) for word in usefulWords])
    return(dict1)
```

this word features dictionary is what allows the algorithm to classify each word independently. Once I had this code in place it was time to download and train classifiers using the movie review corpus and the tweet data corpus. After downloading these corpuses, it soon became clear that these corpuses only classify positive and negative so although they can provide an extensive training set they are only limited to providing positive and negative sentiment classifications. Given this limitation I decided that the best course of action was to generate my own training set with the Twitter data I gathered during the data gathering part of my method which linked back to an idea raised in my literary review. To do this I simply looped over all the tweets, read them and assign them a sentiment which I referred to as the ‘true sentiment’. Once I assigned each tweet in my database a ‘true sentiment’ I updated all the sentiment values within the database so I could create my training sets, for each film in my database I created a training set which was designed as shown above.

7 Results and Analysis

7.1 Testing my classifier

After creating all my training sets for each film in my database I had to test the accuracy of the classifier that these training sets could create so for each film in my database I create a classifier using all the training set except the one that applied to that particular film as this would lead to very high accuracies that wouldn't represent the true accuracy of the classifier. Here are the results:

To understand how good the classifier was I created a simple Python function that creates a classifier using the training sets (which are stored as pickle files) and had a counter for each sentiment value which kept track of the incorrectly classified values so that I could understand the shortcomings of the classifier.

Film	Accuracy	Number of Incorrectly classified Positive tweets	Number of Incorrectly classified Negative tweets	Number of Incorrectly classified Neutral tweets
Black Panther	55.3%	11	48	8
Beauty and the Beast	62.0%	5	48	4
Star Wars: The Last Jedi	46.0%	4	75	2
Guardians of the Galaxy Vol. 2	57.3%	14	43	7
Dunkirk	57.3%	12	50	7
King Arthur: Legend of the Sword	52.0%	22	49	1
Ghost in the shell	62.67%	5	47	4
Valerian and the City of a Thousand Planets	61.3%	25	28	5
A Cure for Wellness	53.3%	16	47	7
Geostorm	66.0%	15	30	6
Mean	57.32%	12.9	46.5	5.1

Table 2 – A table to show the accuracy of the Naïve Bayes sentiment classifier built using the training sets I created with the data I gathered in the early stages of my primary research

As you can see the naïve Bayes classifier that my system created is very inaccurate when it comes to classifying negative tweets given the fact that out of all three sentiment values this one appeared the least when I was classifying each tweet.

Film	Accuracy	Number of Incorrectly classified Positive tweets	Number of Incorrectly classified Negative tweets	Number of Incorrectly classified Neutral tweets
Black Panther	55.3%	18	11	38
Beauty and the Beast	60.6%	37	7	15
Star Wars: The Last Jedi	58.6%	29	20	13
Guardians of the Galaxy Vol. 2	65.3%	21	9	22

Dunkirk	71.3%	14	10	19
King Arthur: Legend of the Sword	66.6%	23	10	17
Ghost in the shell	66.0%	24	9	18
Valerian and the City of a Thousand Planets	67.3%	25	4	20
A Cure for Wellness	50.67%	36	16	22
Geostorm	56.0%	32	13	21
Mean	61.77%	25.9	10.9	20.5

Table 3 – A table to show the accuracy of my initial text classifier

As you can see from the from table 3 the initial text classifier I created although a naïve approach of sentiment classification has proven to work well when classifying negative tweets. In conclusion it became clear that the most accurate sentiment classifier I could build is in fact one combining my initial sentiment classifier with the Naïve Bayes classifier by allowing my initial classifier to classify the negative tweets and allowing the Naïve Bayes classifier to classify the positive and neutral tweets.

7.2 Final Sentiment Classifier

My final sentiment classifier that would be used in the final version of my system would be a combination of the two classifiers I created. The way in which the code would classify the tweets is as follows:

1. Use my text classifier module's stop word removal function to clean, tokenize and remove stop words from the tweet.
2. After I have got the clean tweet use my text classifier to get the sentiment of the tweet, if it is negative then return negative however if classifies it as positive or neutral we use the Naïve Bayes classier (which was stored as a pickle and now is loaded in) that was create using all the training sets for each film in my primary research database.
3. If the second classifier returns a negative then we reclassify it using the initial text classifier so that the Naïve Bayes classifier never classifies a tweet as negative given its inaccuracy with classifying negatives.

Here is the code:

```
def getFinalSentiment(tweet):
    #from my tests i found that my text clasification process was very good a classifiing negatives
    #correctly whereas the classifier i trained classifies to many tweets as negative
    #so to produce the most accurate sentiment analysis system I have combined both classifiers
    tweetList = stopWordRemover(tweet)
    #sentiment form the text classifier above
    textSentiment = getSentiment(tweetList)
    if textSentiment == "negative":
        return("negative")
    else:
        #now I will
        classifierSentiment = getClassifierSentiment(tweet)
        if classifierSentiment == "positive":
            return("positive")
        elif classifierSentiment == "neutral":
            return("neutral")
        #if the classifier returns negative then we will reasign the sentiment based on the text classifer
        else:
            tweetList = stopWordRemover(tweet)
            # sentiment form the text classifier above
            textSentiment2 = getSentiment(tweetList)
            if textSentiment2 == "positive":
                return("positive")
            elif textSentiment2 == "neutral":
                return("neutral")
            else:
                return("negative")
```

After developing this new classifier system, it was time to test it to and get the accuracy of the classifier and to so this I tested it in a very similar way as I tested the original Naïve Bayes classifier. I tested the final classifier by testing it out on all the films in my primary research database however when a film was tested for accuracy I excluded it's training set from the training set that was used on the classifier as this would lead to higher accuracies that were not accurate. For the above code all the training sets for each film has been used to train the classifier meaning that the accuracy could slightly vary from the mean accuracy found but given that more data of a similar format has been used to train it we can assume the accuracy has increased.

Film	Accuracy of combined classifier
Black Panther	64.0%
Beauty and the Beast	76.0%
Star Wars: The Last Jedi	65.3%
Guardians of the Galaxy Vol. 2	71.3%
Dunkirk	72.6%
King Arthur: Legend of the Sword	65.3%
Ghost in the shell	67.3%
Valerian and the City of a Thousand Planets	66.0%
A Cure for Wellness	59.3%
Geostorm	66.0%
Mean Accuracy	67.31%

Table 4 – a table to show the accuracy of the combined text classifier

As you can see from table 4 the accuracy has increased due to the combining of the two text classifiers. Because of this, this combined classifier will be used in my final system as is it the most accurate classifier I built, however because this isn't 100% accurate it will influence the predictions my final system will ultimately make.

7.3 Reviewing the Data

Part of my primary research plan was to develop a predictive system that could use the sentiment analysis results and make a prediction about the films future success. However, before I could build this predictive model I had to take all the data that I had gathered and visualise it so that I would begin to understand how the sentiment analysis results related to the actual success of the film. In the end I used the matplotlib library in Python to produce graphs that showed the polarity and subjectivity for each week in the critical period.

Polarity – The ratio of positive to negative tweets (the greater the polarity the more positively the film is being received)

Subjectivity – The number of positive and negative tweets over the number of neutral tweets (Higher the subjectivity the more opinions are being shared about a certain film)

7.3.1 Polarity graphs

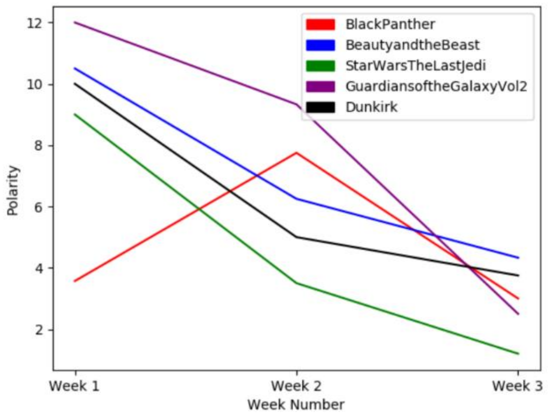


Figure 1 – a graph to show the change in polarity over the critical period for the five films that performed well at the box office

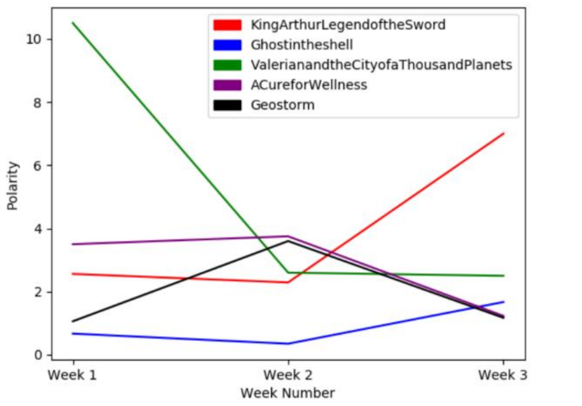


Figure 2 – a graph to show the change in polarity over the critical period for the five films that underperformed at the box office

7.3.1.1 Conclusion from polarity graphs

From my research that went into my literary review it was suggest that an increase in polarity was sign that a movie was being well received and therefore likely to do well at the box office, with this in mind I expect to see with the successful films was a slow increase in polarity or a consistently high polarity. As you can see every film that performed well in the box office started with a very high polarity except for black panther which was the most successful film of all the ones I chose to research. Other than black panther all these films in figure 1 are steadily decreasing in polarity, three of which drop below 4 in the final week. Black panther was the most successful in the domestic market out of all these films and the rapid increase in polarity during the second we may be the reason for this. Star Wars: the last Jedi is an interesting film as the polarity rapid decreased but the film had the second highest profit of all the films, I believe the reason for this is that it is a franchise film meaning it has a dedicate fanbase who supported it, however it was a film that divided many people and I believe this is the reason for the decrease in polarity.

Figure 2 much like figure 1 shows some interesting results because the polarity remains slightly more consist between weeks in the critical period. The big conclusion I drew from figure 2 is the fact most of the polarity values per week are under 4 which isn't the case in figure 1, this could be a key factor when making a prediction about a films future success. The film Ghost in the Shell is an interesting example as the polarity remained very low for its entire critical period and I believe this was due to the fact there was controversy surrounding this film due to the fact it was an anime adaptation that was accused of 'white washing' the cast. Valerian however started with a very high polarity that rapidly dropped during the second week and I believe this is due to good marketing of the film which interested people but after the film was released and people saw it they shared more negative views due to the quality of the film which in turn lead to a decrease in polarity. Overall it is clear that a film that under performs will likely never have a polarity rating greater than 4.

7.3.2 Subjectivity Graphs

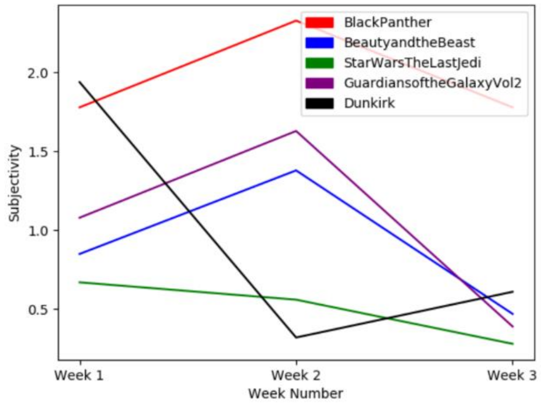


Figure 3 – a graph to show the change in subjectivity over the critical period for the five films that performed well at the box office

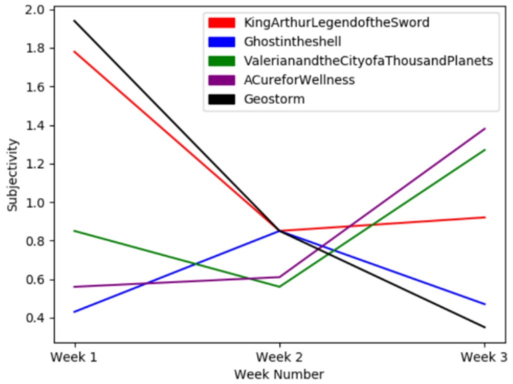


Figure 4 – a graph to show the change in subjectivity over the critical period for the five films that under performed at the box office

7.3.2.1 Conclusions from subjectivity graphs

From the research conducted during my literary review it was suggested that the subjectivity should increase during the second week of the critical period as people would have seen the film and have real opinions about it to share rather than preconceptions. From figure 3 you can see that for the majority of the films this is the case as the subjectivity does increase with Dunkirk and Star Wars being exceptions. Figure 4 is more inconclusive given the wide spread of results. Although subjectivity isn't directly related to understanding how positively a film is being received I believe that used in combination with the polarity results it can provide vital information that could help determine the future success of the film. For example, Ghost in the Shell had a consistently low polarity and during its second week it's subjectivity increased whereas it's polarity decreased which means people were being overall more negatively opinionated about this movie which is very bad for a films future success in the box office.

Now that I had all these results it was time to use them to create a decision tree so that I could generate a prediction about a film's future success based on the sentiment analysis results and to do this I used the data mining tool Weka.

7.4 Developing a Prediction System – Data mining

Data mining is the process of using existing data to draw out predictive patterns (An Introduction To Data Mining 2018). Given this information it was clear that data mining was the key in developing a prediction pattern to use in the final version of my system. To start I looked at each of the films I gathered data on from my primary research and created a table of the film's title, its domestic gross and how that related to the profit when compared to the production budget, remember that I am only looking at the domestic gross given that a film can likely recoup it budget in the international market. This was the table I created:

Film Title	Gross	domestic %	Budget	Profit
Black Panther	$ 681,084,109.00	51.40%	$ 200,000,000	$ 481,084,109.00
Beauty and the Beast	$ 504,014,165.00	39.90%	$ 160,000,000	$ 344,014,165.00
Star Wars:The Last Jedi	$ 620,181,382.00	46.50%	$ 250,000,000	$ 370,181,382.00
Guardians of the Galaxy Vol. 2	$ 389,813,101.00	45.10%	$ 200,000,000	$ 189,813,101.00
Dunkirk	$ 190,068,280.00	36.00%	$ 100,000,000	$ 90,068,280.00
King Arthur: Legend of the Sword	$ 39,175,066.00	26.30%	$ 175,000,000	$ -135,824,934.00
Ghost in the shell	$ 40,563,557.00	23.90%	$ 110,000,000	$ -69,436,443.00
Valerian and the City of a Thousand Planets	$ 41,189,488.00	18.20%	$ 177,200,000	$ -136,010,512.00
A Cure for Wellness	$ 8,106,986.00	30.5%	$ 40,000,000	$ -31,893,014.00
Geostorm	$ 33,700,160.00	15.20%	$ 120,000,000	$ -86,299,840.00

Table 1– Table showing a films domestic gross, the percentage of its overall earnings that the domestic gross contributed, the films production budget and the profit made in the domestic market
(Box Office Mojo 2018)

After creating this film, I created five categories that a film could fall into based on the domestic profit and it is as follows:

Category	Criteria to fit into category (profit based)
Very Good	Domestic Profit > $200,000,000
Good	$200,000,000 >= Domestic Profit > $100,000,000
Okay	$100,000,000 > = Domestic Profit >= $0
Bad	$0 > Domestic Loss >= -$100,000,000
Very Bad	Domestic Loss < -$100,000,000

Each film in Table 1 will be assigned one of these values and this data will then be used as the foundation for my data mining.

7.4.1 Weka and the J48 Decision Tree

Weka is software program that is used for data mining, in short is it simply a collection of machine learning algorithms that are applied to data mining tasks such as creating decision trees (University Of Waikato 2018). Within Weka the key data mining tool I would be using was creating J48 decision trees as I felt that a decision tree would be appropriate for my system given the five prediction values a film can be assigned. The j48 decision tree is developed from the C4.5 classification algorithm, which constructs decision tree from a dataset that has already classified the data (Li 2018).

7.4.2 The data used for the system

The data used to create my decision tree had to represented in a particular way and saved as an .arff file. The data sets I create to build decision are as follows:

Trees to build:

- Polarity/subjectivity values for each film per week
- Mean of all three polarity/subjectivity values for each film
- The two changes in polarity/subjectivity between the three weeks

The data to build these trees is in this form (polarity per week data set):

```
@relation sentiment


@attribute performance {veryGood,good,okay,bad,veryBad}

@attribute week1Polarity real

@attribute week2Polarity real

@attribute week3Polarity real


@data

veryGood,3.57,7.75,3.00

veryGood,1.05,6.25,4.33

veryGood,9.00,3.50,1.20

good,1.20,9.33,0.25

okay,1.00,5.00,3.75

veryBad,2.56,2.29,7.00

bad,0.67,0.35,1.67

veryBad,1.05,2.60,2.50

bad,3.50,3.75,1.23

bad,1.06,3.60,1.17
```

7.4.3 The Decision trees developed

For each of the trees I said I would create there were three that had the highest accuracy these where:

7.4.3.1 Change in polarity tree

This decision tree returned with a 70% accuracy when classifying the 10 films within my dataset.

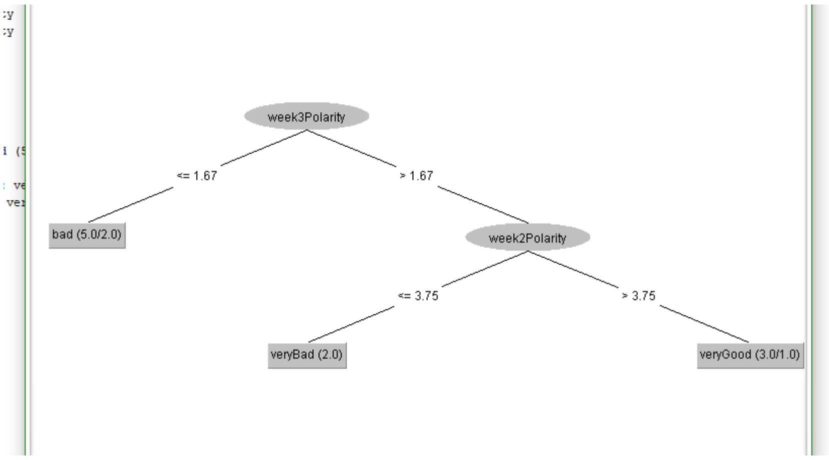


Figure 5 – decision tree to classify a film's future performance based on polarity

The reason I will be using this tree is because of its accuracy and the fact polarity is main measure of sentiment and the most important factor within my project.

7.4.3.2 Change in polarity tree

This decision tree returned with an accuracy of 60% when classifying the 10 films within my primary research system.

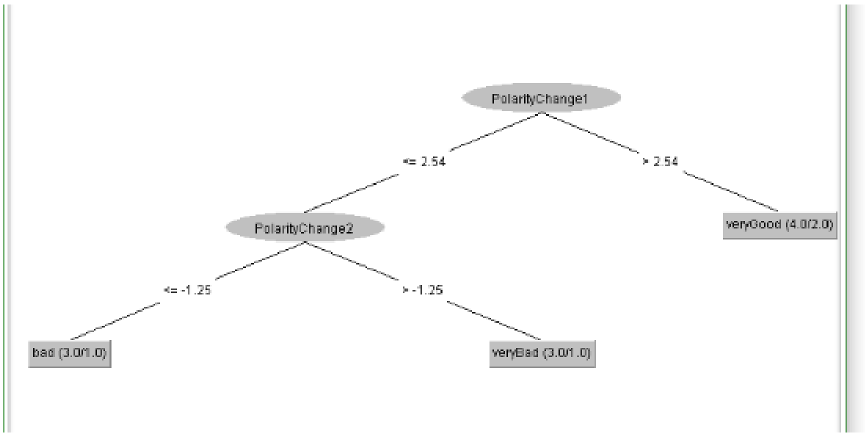


Figure 6 – decision tree to classify a film's future performance based on polarity change

The reason this tree will be used when developing my system is much the same as the reason why I will be using the polarity per week decision tree and that is because of that accuracy as well as the fact that polarity change in another key measure concerned with sentiment analysis.

7.4.3.3 Change in subjectivity tree

This decision tree returned with a 70% accuracy so it was necessary to consider it when developing my final system.

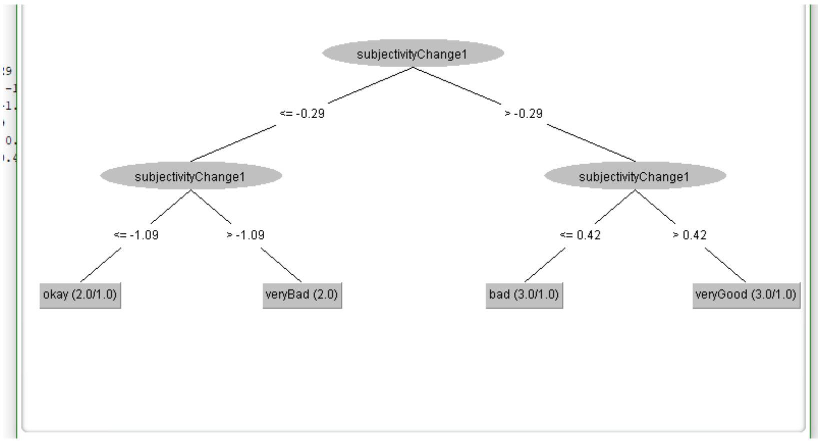
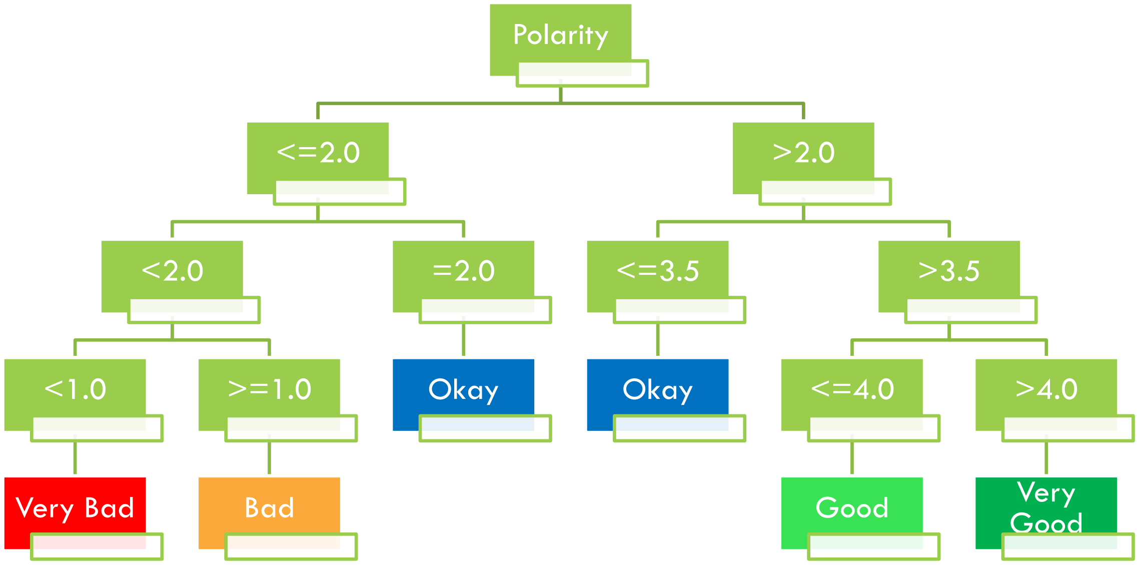


Figure 7 – decision tree to classify a film's future performance based on subjectivity change

Subjectivity is more of a secondary measure within sentiment analysis as what it tells you is how many options (positive or negative) are being shared about a certain film however I believe in combination with polarity it can provide insight as to the future of the film's success and that is why I aim to include this tree in my final decision tree. Furthermore, it provided the widest spread of classifications and can aid in my understand of how the results can related to a prediction classification.

7.4.4 Building my decision tree

Now that I have extracted the most accurate decision trees from my data mining analysis I had to combine them so that I could implement them into my final system. From the trees above you'll notice that the ratings 'Good' and 'Okay' aren't classified so I had to develop the given trees into one capable of classifying films in all five categories. The first tree I developed simply looked at the polarity value and will be used to develop a prediction when there is no change in polarity or subjectivity data available e.g in the first week of the critical period.



Prediction	Initial polarity
Very Good	>4.0
Good	3.5< polarity <=4.0
Oka	2.0 <= polarity <= 3.5
Bad	1.0<=polarity<2.0
Very Bad	<1.0

The above hierarchy and table shows how I will initially determine the prediction for the film, each prediction relates to a domestic profit estimate for that film e.g Very good means that the film is estimated to reach $200 mil or greater. From the polarity per week decision tree we saw that to be considered “bad” a film had to have a polarity less than or equal to 1.67 which is between 2 and 1. Furthermore to be considered “Very good” a film had to have a polarity greater than 3.75 which my decision tree adheres to however it was necessary to push the limit up because I had to classify films as ‘Good’ so I rounded the value to the nearest whole number and from there I filled in the blanks of the tree with the remaining values to get a full tree.

Now I had to develop a tree that dealt with change in polarity and subjectivity, these trees were far less complex than the initial decision tree as they simply would adjust the initial prediction. The rules for these trees were as follows:

1. If the polarity increases by two then move the prediction up a tier unless it was already classified as “Very Good” e.g if it was ‘Bad’ it would be pushed to ‘Okay’
2. If the polarity decreased by two then the prediction would be moved down a tier unless it was initially classified as “Very Bad” e.g if it was ‘Good’ it would be moved down to ‘Okay’.
3. If the polarity increased and the subjectivity increased by one than the prediction would be moved up a tier, however if the polarity decreases and the subjectivity increases than the prediction tier is pushed down.

7.4.4.1 Justification for these rules

Rule 1 – the reason that this rule is in place is because the in the polarity change decision tree if the change was greater than 2.54 it received a “Very Good” rating so because of this I adjusted the figure slightly and felt that if the polarity increased by two this was reason enough to move the prediction up a tier.

Rule 2 – From figure 6 you can see that if a film experienced a polarity change of less than -0.29 than it would either lead to a 'Okay' or 'Very Bad' classification, given an increase of two would lead to an increase in tier (based on my system) I felt it was appropriate to apply the reverse if the polarity decreased by two. This decrease is far greater than the on the Weka decision tree suggested but I believe to decrease the prediction tier is only to be done if there is a great change in the results.

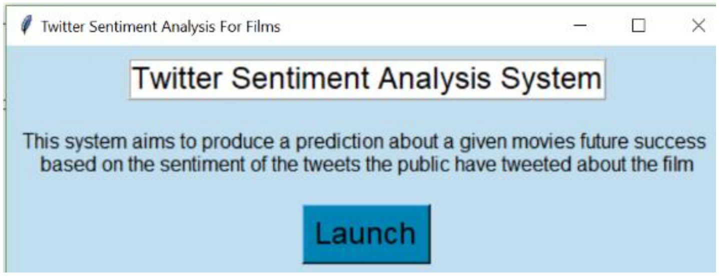
Rule 3 – From figure 7 you can see that if there is a small increase in subjectivity it can lead to a "Very Good" classification and this is the similar if there is a decrease in subjectivity as this can lead to a "Very Bad" classification. So, because of the accuracy of this tree I felt it was necessary to incorporate this information, however just using subjectivity change isn't a strong enough decider to change the prediction tier so that is why I incorporated polarity with this rule. If the subjectivity increases by one and the polarity also increases than this implies people are sharing more positive options about the film given equation 1 in the literary review and if the subjectivity increases with a decrease in polarity than this implies people are sharing more negative options about a film. Provided with this information I felt it was appropriate to adjust the prediction tier if the rule was satisfied.

8 My Final System

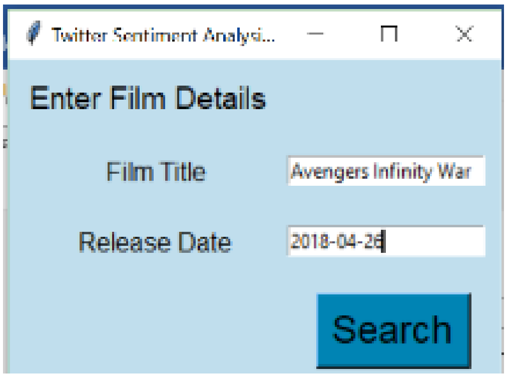
After developing my primary system and analysing the results to develop a prediction system it was time to implement this into a final version of the system that was concise and easy to use for a user who didn't have the understanding of the system that I had.

Screenshots Of The Final System

This first screen is a simple explanation of what the system aims to achieve.


Twitter Sentiment Analysis For Films
Twitter Sentiment Analysis System
Launch

This system aims to produce a prediction about a given movies future success based on the sentiment of the tweets the public have tweeted about the film



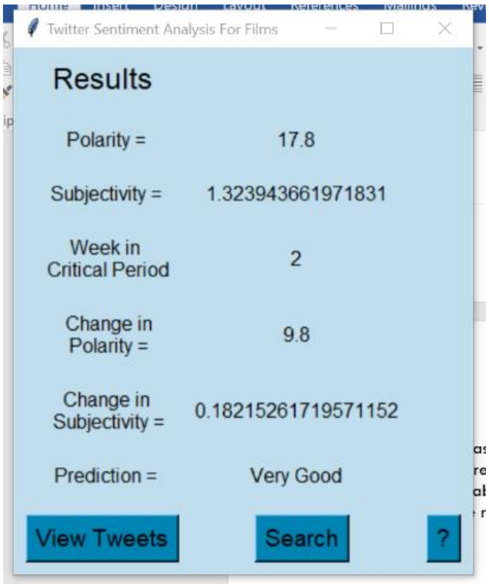
Film Title: Avengers Infinity War

Release Date: 2018-04-26

The Next screen asks the user to enter a film title and the release date so that it can get the week number for the film (week in critical period). The film I have searched for is Avengers: Infinity War as when testing the final system it was in its first week of its critical period, as of taking these screenshots the film is in the second week so it was good for testing the full extent of my prediction system.

Once this information is entered the user can search for the tweets, if the tweet is already in the database then it isn't added. If the search for tweets for this film is yet to be executed than it will take a minute or two to generate the results as it uses the Tweepy cursor method to search for tweets and my sentiment classifier to classify each tweet.

Once the search has been executed fully the user is taken to the results page. The results page shows a range of results as well as a button to view all the tweets, a search button to search for tweets about another film and finally a help button that opens a text file with information about what all the results mean.



Polarity = 17.8

Subjectivity = 1.323943661971831

Week in Critical Period 2

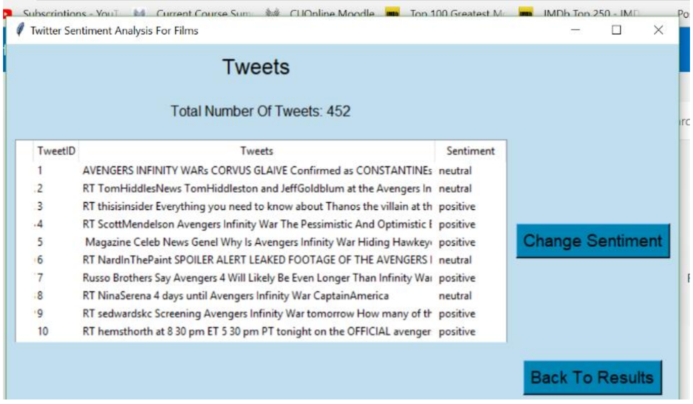
Change in Polarity = 9.8

Change in Subjectivity = 0.18215261719571152

Prediction = Very Good

As you can see from the results page the film is in the second week of its critical period so the change in polarity and subjectivity information is available (N/A is shown if the change in polarity and subjectivity can't be calculated). The Initial prediction for this film was "Very Good" and as you can see this hasn't change because it can only be increased given the change in polarity and subjectivity results, however because it is at the highest tier no change is made.

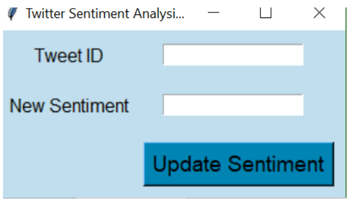
If a user clicks the view tweets button they are taken to the following screen:



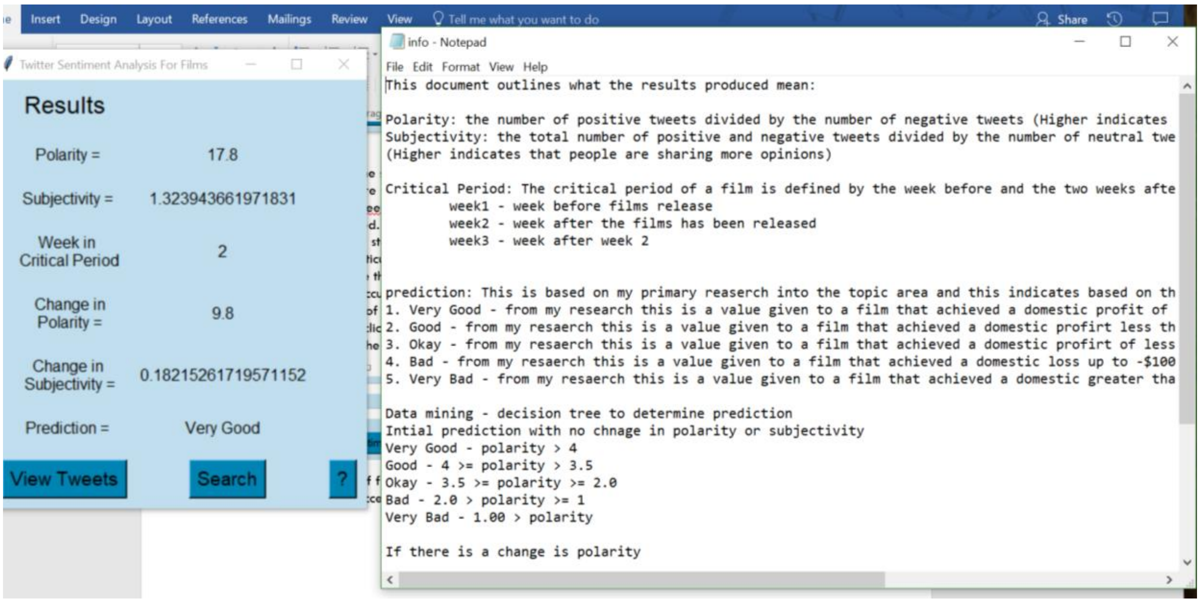
Total Number Of Tweets: 452

TweetID	Tweets	Sentiment
1	AVENGERS INFINITY WARs CORVUS GLAIVE Confirmed as CONSTANTINEs	neutral
2	RT TomHiddlesNews TomHiddleston and JeffGoldblum at the Avengers In	neutral
3	RT thisisinsider Everything you need to know about Thanos the villain at th	positive
4	RT ScottMendelson Avengers Infinity War The Pessimistic And Optimistic I	positive
5	Magazine Celeb News Genel Why Is Avengers Infinity War Hiding Hawkey	positive
6	RT NardInThePaint SPOILER ALERT LEAKED FOOTAGE OF THE AVENGERS I	neutral
7	Russo Brothers Say Avengers 4 Will Likely Be Even Longer Than Infinity Wa	positive
8	RT NinaSerena 4 days until Avengers Infinity War CaptainAmerica	neutral
9	RT sedwardskc Screening Avengers Infinity War tomorrow How many of th	positive
10	RT hemsthorth at 8 30 pm ET 5 30 pm PT tonight on the OFFICIAL avenger	positive

As you can see from the screenshot to the right all the tweets are put into a treeview where the tweetID, Tweet and sentiment are displayed. Furthermore, the total number of tweets stored in the database for that particular film as well as an option to change the sentiment of a tweet is present. I have included the option to change the sentiment of a tweet given that fact that my text classifier isn't 100% accurate and if a user was inclined they could update the sentiment of a tweet to increase the accuracy of the results produced. The screenshot below shows the simple GUI window that is opened when a user clicks the change sentiment button and this works through a simple SQL update function that changes the sentiment of a tweet given the ID which can be seen in the treeview.



The final small piece of functionality that I included in my final system is a small information text file that can be accessed by pressing the "?" button. This text file tells the user what polarity and subjectivity means, as well as what the critical period is and what the prediction means and relates to.


Insert
Design
Layout
References
Mailings
Review
View
Tell me what you want to do
Share
Twitter Sentiment Analysis For Films
info - Notepad
File Edit Format View Help
View Tweets
Search
?

Results

Polarity = 17.8

Subjectivity = 1.323943661971831

Week in Critical Period 2

Change in Polarity = 9.8

Change in Subjectivity = 0.18215261719571152

Prediction = Very Good

```
This document outlines what the results produced mean:

Polarity: the number of positive tweets divided by the number of negative tweets (Higher indicates
Subjectivity: the total number of positive and negative tweets divided by the number of neutral twe
(Higher indicates that people are sharing more opinions)

Critical Period: The critical period of a film is defined by the week before and the two weeks afte
        week1 - week before films release
        week2 - week after the films has been released
        week3 - week after week 2


prediction: This is based on my primary reaserch into the topic area and this indicates based on th
1. Very Good - from my research this is a value given to a film that achieved a domestic profit of
2. Good - from my resaerch this is a value given to a film that achieved a domestic profirt less th
3. Okay - from my resaerch this is a value given to a film that achieved a domestic profirt of less
4. Bad - from my resaerch this is a value given to a film that achieved a domestic loss up to -$100
5. Very Bad - from my resaerch this is a value given to a film that achieved a domestic greater tha

Data mining - decision tree to determine prediction
Intial prediction with no chnage in polarity or subjectivity
Very Good - polarity > 4
Good - 4 >= polarity > 3.5
Okay - 3.5 >= polarity >= 2.0
Bad - 2.0 > polarity >= 1
Very Bad - 1.00 > polarity

If there is a change is polarity
```

8.1 DATABASE STRUCTURE

Each film's tweets are stored in the exact same way as my primary system, however there is an additional table that stores the analysis results so that they can be used to populate the results page. This table is designed as follows:

Column	FilmID	FilmTitle	Polarity	Subjectivity	Polarity2	Subjectivity2	Prediction
Data Type	INT Auto increment	Test	INT	INT	INT	INT	Text

The reason there is two polarity and subjectivity columns is because it is necessary to store two values so the system can get the change in these values. During week 1 of the critical period the results are stored in the first polarity and subjectivity column, the week 2 results are stored in the second column and finally the week 3 results will be stored in the 2nd columns with the week 2 results being pushed to the first set of columns and so on if it is past the critical period.

9 Evaluation

Overall, I believe that my project was successful in answering my proposed research question given the extent of my research into the topic area. Sentiment analysis is a very effective tool for making predictions in this topic area as well as others such as the stock market which is a concept brought up in the literary review. In terms of to what extent sentiment analysis can be used to make an accurate prediction I believe that to a significant enough extent it can provide real insight into the future performance at the box office. What I mean by significant extent is the fact that provided with the information that my final system can produce you can form a rough idea about how well a film will do at the box office. Although the true accuracy of my final system is yet to be tested from the example shown in the screenshots I can confirm that with further development this can be a very accurate system. The main reason I say my final system looks promising is because the example I provided has certainly been assigned a very accurate prediction as the film is a major release franchise film, however these films are easy to predict given their existing quality assurance. In the long run I feel my system may suffer when trying to classify independent films as these may provide a very high polarity ratings that won't necessarily related to a high domestic profit given the fact people can't guarantee the quality of the film.

9.1 Primary Research Evaluation

9.1.1 Data Gathering

Gathering my own data to use in my primary research was a key part of this project as with no data I would have no project. Given the importance of this I felt that if I had more time and greater resources I could have gathered far more data and in turn created a better overall final system. More data would have led to a greater accuracy in the long run when it comes to the machine learning and data mining as both these areas of computer science work better within big data projects given the vast amount data they can use. Despite lacking an extensive amount of data, I felt that what I was able to accomplish with the data I gathered was extensive enough to develop a final system that satisfied my goal of answering my proposed research question.

9.1.2 Test Classifier

The final text classifier that I developed and deployed in my final system was one of a high accuracy (67%+ from my analysis). I believe that the reason for this high accuracy was the use of my own training data so my classifier was tailor made to my individual system and didn't use a general solution. The existing corpuses provided by the NLTK library in Python were extensive however they lacked the option of a neutral selection and were very general, for example the tweet corpus just contains a large bank of tweets about anything and everything which wasn't ideal for my system as the tweets I gathered pertained to movies specifically. Furthermore, during analysis I determined what the key weakness was within the machine learning classifier and how I could strengthen this by utilising the initial text classifier I developed which at first, I thought was naïve and not effect for text classification. Due to the extent and depth of research coupled with analysis of my text classifiers I developed an accurate text classifier that in theory could adapt and get smarter overtime if more training data is used to create the classifier.

9.1.3 Data mining

When conducting the data mining analysis to draw out a prediction system for my final system I felt that with amount data I gathered I develop quite an accurate prediction system due to the fact I tested many sets of data and developed my decision tree from the most accurate ones produced by Weka. In an ideal world I would have had far more data to use when data mining as much like machine learning, data mining is improved when it is provided with more data. The lack of data is perhaps the biggest weakness my project suffered from for example on the decision trees not all prediction results could be attained due to a lack of data so this meant I had to build my own decision trees using the small trees that I produced through Weka which could very well lead to a low accuracy within my final system.

9.1.4 Conclusion and other factors

In conclusion the answer to my question is that to a certain extent sentiment analysis can in fact accurately predict how well a film will do at the box office however I am of the belief that it is not the only factor that can be considered when making these types of predictions. Other factors that I believe can affect a movies performance at the box office are things such as genre because in today's world the genre of superhero movies have excelled when it comes to turning a profit even if the film isn't received very well, for example the film Batman v Superman: Dawn of Justice took a domestic profit of $80,360,194 (Batman V Superman: Dawn Of Justice (2016) - Box Office Mojo 2018) which should provide an 'Okay' prediction however it wasn't well received only achieving a 27% rating on the movie review site Rotten Tomatoes which is a very low score to receive from this website and would have resulted in a low polarity that perhaps would have led to a lower tier prediction (Batman V Superman: Dawn Of Justice 2018). Furthermore, I'd like to point out that two of the films in my primary research data set that performed well were in fact superhero movies (Guardians of the galaxy vol. 2 and Black Panther). Another key factor I believe that should be considered when making a prediction is the production company behind the film and whether it belongs to a franchise, for example Black Panther and Star Wars: The Last Jedi are two of the highest grossing films I gathered tweets about and they both belong to massive cinema franchises, in addition to this Beauty and the Beast is a Disney film and Disney is a company that nearly always granites quality and for this reason produces very successful films. The final factor that could provide insight into a films future is the people who are or directing or are in the film as many people have a favourite actor or director, for example Dunkirk is not a superhero film and doesn't belong to a franchise but was directed by Christopher Nolan who has a large following as a director due to the high quality of his previous work. If we look at these factors when considering the films that underperformed we can see that they do help a film perform well at the box office as none of these films are part of a franchisee or are superhero movies. In summation you can see my system can to an extent provide insight into a films future but will not guarantee the highest degree of accuracy as I believe that an increase in accuracy lies within gathering more data about more films as well as the other contributing factors discussed which could be implement into my final system in combination with the sentiment analysis to provide greater insight and accuracy.

10 Discussion

Overall, I believed that I achieved a lot in this project, enough to where I could comfortable answer my proposed research question with a good degree of accuracy and justification given my research and analysis. The two biggest achievements I got from this project are the development of my final system and the development of a high accuracy text classifier. Despite these achievements I also had many deficiencies within my project such as a lack of primary data and a lack of testing within my final system. This discussion will address all these achievements and deficiencies.

10.1 Achievements

Throughout the course of this project I achieved several things and in this section, I will be discussing them in depth.

10.1.1 Development of Text Classifier

During the primary research phase of my project it was essential for me to develop a text classifier to be used for sentiment analysis. Existing text classifiers were available within Python however upon review they were not very accurate in my opinion so I felt that developing my own was essential if I wanted my system to have a high accuracy and work in the way I wanted as I had a specific vision for my final system. Initially I developed a very naïve text classifier based on my pre- determined idea about how a text classifier works that was enhanced by knowledge gained from my literary review. After I realised that this wasn't going to be an effective method for classifying text it was time to use a more advanced approach to text classification which lead me to Machine learning within Python using the NLTK library. Machine learning was a concept I wanted to employ in my system but I was unsure at the start of the project if I would be capable of the implementation. Once I was able to implement this I was very proud of what I was able to achieve given the vast amount of training and analysis I put into developing the most accurate text classifier I could, given the amount of training data I had. Furthermore, the fact I used my own training data set that I created from the data I gathered meant my text classifier was tailor made and this was a great achievement within my project.

10.1.2 Final System GUI and Prediction System

The final GUI system I developed was one that I was proud of given the time scale I was provided with. The reason I felt this is an achievement is the fact it performs the range of functions I had envisioned at the very start of my project and in my opinion despite the Tkinter module not being the best GUI development system when it comes to aesthetics the GUI looks good with an appropriate colour scheme to match the theme of Twitter. The final system I developed not only had all the features I wanted to implement, it also featured a prediction system that I felt was effective enough considering the lack of data that went into its development. Data mining was something that I had intended to implement at the start of the project given that it is the analysis technique that is used to get new information (predictions) from existing data and I believe that I accomplished this goal within my final system. Although untested I believe that given the data my decision trees where built upon I developed a prediction system that may prove sufficient enough to gauge a general idea about a film's future with further testing which was the gaol of my project.

10.2 Deficiencies

Like many research projects mine was meet with deficiencies that if I could I would avoid so that overall my project would be improved.

10.2.1 Not enough data

The key deficiency within my project is the fact that I felt I didn't get enough data as I only considered 10 films (5 that performed well and 5 that didn't) and only gathered 1500 tweets in total which may seem a lot but is nowhere near the big data project scale that would have been more appropriate for a big data project which mine could be considered as. Machine learning and data mining are two areas of computer science that are fuelled by data and can't be employed effectively with a lack of data. Given the amount of data I had I was able to effectively create a text classifier using the Naïve Bayes machine learning approach however in not unwise to think that with more training data I could have created a better text classifier especially given the fact it had a major weakness with classifying negative tweets. The reason my text classifier suffered from this weakness was due to the lack of negative tweet data which was a very key weakness in my system and why I had to combine text classifiers, which isn't an ideal solution in the long term. Finally, the lack of data had a massive effect on the quality of the decision trees I was able to develop, the lack of data lead to a lack of branches in my trees as the majority of the time they only lead to 3/5 prediction results given how the predictions such as 'Okay' were only assigned to one film I gathered data about. The lack of data when creating a decision tree lead to me creating my own using the tress I developed as a base, however due to a lack of testing it isn't possible to provide a precise accuracy.

10.2.2 Lack of Testing

One area I felt that I servilely lacked was the testing of the final predictive system I developed for my final system. Due to time constraints and limitations with the Twitter data I couldn't test my final system extensively to get an accuracy percentage which is quite a weakness as it would have helped me to justify and cement a firm answer to my research question. The limitation with the Twitter API which I got around within my primary research could have been applied to my final system however its takes a long time to gather the tweets and this was time I didn't have near the end of the project as a deadline had to met. For these reasons I wasn't able to fully test my system which is perhaps the largest deficiency in my project.

10.2.3 Lack of Planning

In my opinion other than my proposal and primary research plan I did have a lack of planning in my project. What I mean by this is that up until the development of my final system I didn't have a plan in place to develop a prediction system which is something I should have done when I decided what films I would use in my primary research as I could have classified them and got more films with a wider range of prediction classifications which would have helped my decision tress be more extensive.

10.3 How I would do it differently

In hindsight this project could have be conducted in an overall more organised way so that I would have been more thorough and get a definitive and conclusive answer to my research question. The first thing I would have done differently is to do more research into the processes involved with sentiment analysis within Python and how a text classifier could be built within Python, in addition to this I would have researched data mining with Weka so I would know the data required to create better decision trees. After conducting all this research, I would have created a plan that followed the process which I used to create this project. The plan that I would have created would have helped me manage my time more effectively and could have allowed for me to build on what I produced within this project such as gathering more data and providing sufficient time for testing the final system. As previously mentioned in the evaluation stage if I could go back I would have gathered far more data about more films and not necessary view films as under performer/good performers, rather consider them more as a scale like I did when developing a prediction system and if I had done this than I wouldn't have been as selective with the films I chose to use in my primary research opening me up to the option of gathering far more data. Gathering more data would have strengthened my project as it would have improved the machine learning and data mining analysis stages of the project. The final thing I would have done differently if I was to do the project again is more testing of the final system as this was a key part of my project I had to overlook due to limitations in time and the way in which I would have had to gather tweets which was discussed previously. Furthermore, it is clear that testing the final system would have helped when answering my research question as it would have provided justification for my final answer.

In conclusion it is clear that throughout this project I achieved a great deal but my project is not without flaws and these flaws did hold my project back from achieving its full potential. In spite of these deficiencies I feel that I was able to answer my research question with a certain degree of accuracy and justification.

11 Reflection

11.1 Ethics of my project

Overall the ethics we quite simple in my project given the fact that I used public Twitter data from the Twitter API that I was provided access to through a Twitter developer account. Furthermore, I never stored data about the user just the tweet itself which are public and in addition this data is simply being used to make predictions about films using the sentiment of the tweet in general not specifically what the tweet is saying. Given my ethics was approved and the fact I didn't breach what I outlined in my ethics application it is clear that there were no ethical violations within my project.

11.2 Diligence

Throughout the project I attempted to maintain a level diligence particularly during the primary research phase given the importance of this phase in the implementation process. Personally, I think at the start of the project I lacked the work ethic shown during the implementation phase. Finally, during the analysis stage I believe I worked very diligently as I effectively used the data I gathered to perform data mining analysis such as prepping the data sets to be used by Weka.

11.3 Problem encountered

11.3.1 Twitter API Limitation

The first key problem I encountered was the limitation with the Twitter API whereby the only tweets available were ones from a week ago which was a major setback given the fact I needed tweets from the critical period of films from 2017. After much research I was able to overcome this issue through the use of the GetOldTweets module. Overall this problem did set me back on the timeline I had for the project.

11.3.2 Text classifier to naïve

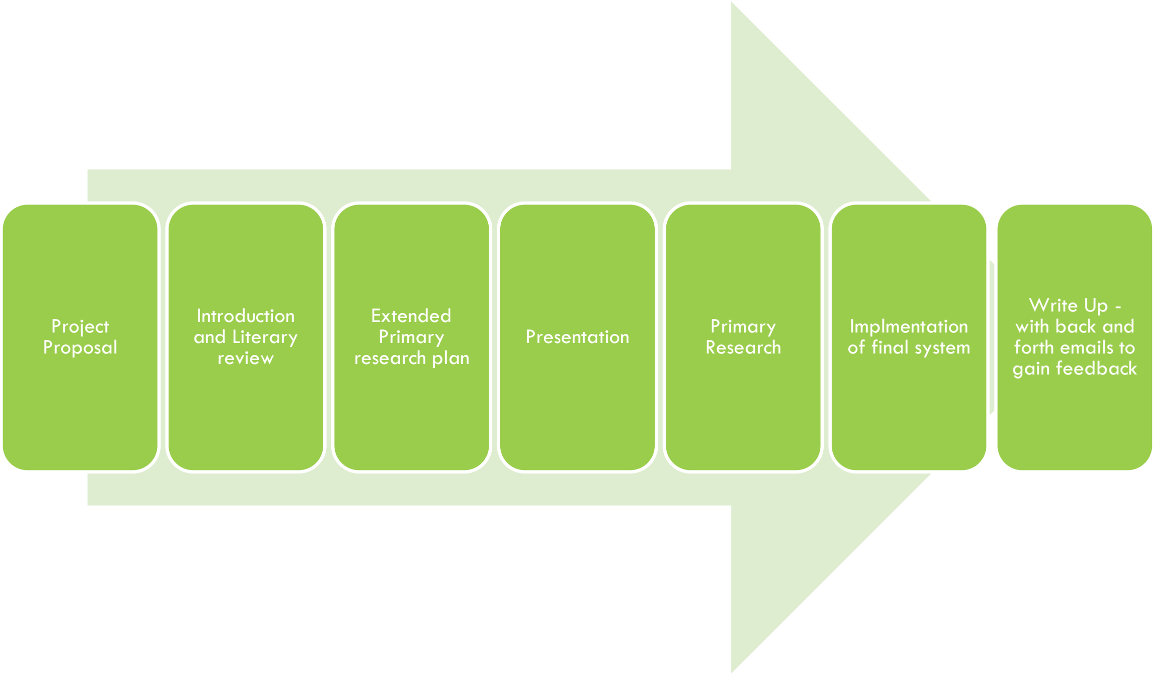
The second setback I faced was the fact that it quickly became clear that my initial text classifier wasn't an effective text classifier and I would have to develop a more advanced text classifier through machine learning. This was a setback as I had to research and develop my text classifier as well as create a training set that could classify neutral sentiments as well as positive and negative sentiments.

11.4 Supervisor meetings and feedback response

Throughout the course of the project I kept in touch with my supervisor for advice and feedback on my work. The first meeting I had with supervisor was a group meeting to discuss how the project would work and what was expected of us which took place before Christmas and before I started my project. After this meeting I researched into computer science projects and came up with two key ideas I wanted to develop and emailed my supervisor for feedback, after receiving this I selected this Twitter sentiment analysis project as I had great interest in the subject matter and I was confident I could complete the project.

The next key meeting I had with my supervisor was my presentation as I felt that emails and online resources were enough for me to conduct my project successfully. During my presentation I presented my idea and the work I had up to that point which was the introduction, literary review and extended primary research plan. The key feedback I received was the I had to start the implementation phase of the project so that I could answer my research question. Upon receiving this feedback, I set out to implement my research plan and develop my final system while continuously emailing my progress to my supervisor for advice and feedback.

11.5 Project timeline


Project Proposal
Introduction and Literary review
Extended Primary research plan
Presentation
Primary Research
Implmentation of final system
Write Up - with back and forth emails to gain feedback

12 CONCLUSION

In conclusion the final system I developed does go some way to help solve the problem discussed in my introduction, the problem of films not truly knowing how well they will perform which can sometimes lead to a financial loss in the domestic market which can be recouped in the international market however this isn't guaranteed. I believe my project can be used as a tool in combination with the consideration of other factors I suggested in my evaluation that can help gain a general idea about a films future. My final system not only provides a prediction that can provide a rough estimate to a films future success, it also provides vital information in a concise format. What I mean by my previous statement is the fact that this data by itself can prove to provide insight into the general narrative surrounding a film which without the prediction can be effective in understanding how well a film will perform at the box office. Finally, the results of my research suggest that my system provides accurate sentiment analysis that could be improved upon given the back bone of the classifier is created with machine learning and despite the fact the accuracy of the prediction system is yet to be tested I believe that given my research it works well as a general solution to a complex problem.

13 BIBLIOGRAPHY

Asur, S. and Huberman, B. (2010) Predicting The Future With Social Media [online] available from <https://arxiv.org/pdf/1003.5699.pdf> [19 March 2018]

An Introduction To Data Mining (2018) available from <http://www.thearling.com/text/dmwhite/dmwhite.htm> [25 April 2018]

Batman V Superman: Dawn Of Justice (2016) - Box Office Mojo (2018) available from <http://www.boxofficemojo.com/movies/?id=superman2015.htm> [28 April 2018]

Batman V Superman: Dawn Of Justice (2018) available from <https://www.rottentomatoes.com/m/batman_v_superman_dawn_of_justice/> [28 April 2018]

Bermingham, A. and Smeaton, A. (2011) On Using Twitter To Monitor Political Sentiment And Predict Election Results, pp 9 [online] available from <http://www.aclweb.org/anthology/W11-3702> [25 January 2018]

Box Office Mojo (2018) available from <http://www.boxofficemojo.com/> [25 April 2018]

Build A Sentiment Analysis App With Movie Reviews (2018) available from <http://pythonforengineers.com/build-a-sentiment-analysis-app-with-movie-reviews/> [25 April 2018]

Go, A., Bhayani, R. and Huang, L. (2009) Twitter Sentiment Classification Using Distant Supervision [online] available from <http://www.yuefly.com/Public/Files/2017-03-07/58beb0822faef.pdf> [20 March 2018]

Henrique, J. (2018) Jefferson-Henrique/Getoldtweets-Python [online] available from <https://github.com/Jefferson-Henrique/GetOldTweets-python> [24 April 2018]

Kumar, N. (2018) Twitter Sentiment Analysis Using Python - Geeksforgeeks [online] available from <https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/> [24 April 2018]

Li, R. (2018) C4.5 Data Mining Algorithm In Plain English - Hacker Bits [online] available from <https://hackerbits.com/data/c4-5-data-mining-algorithm/> [28 April 2018]

Natural Language Toolkit — NLTK 3.2.5 Documentation (2018) available from <https://www.nltk.org/> [25 April 2018]

Pagolu, V.,Challa, K., Panda, G., Majhi, B. (2016) Sentiment Analysis Of Twitter Data For Predicting Stock Market Movements [online] available from <https://arxiv.org/pdf/1610.09225.pdf> [25 January 2018]

Pak, A. and Paroubek, P. (2013) Twitter As A Corpus For Sentiment Analysis And Opinion Mining [online] Universit´e de Paris-Sud. available from <http://crowdsourcing-class.org/assignments/downloads/pak-paroubek.pdf> [19 March 2018]

University Of Waikato (2018) Weka 3 - Data Mining With Open Source Machine Learning Software In Java [online] available from <https://www.cs.waikato.ac.nz/ml/weka/> [25 April 2018]

Waldron, M. (2015) Naive Bayes For Dummies; A Simple Explanation - AYLIEN [online] available from <http://blog.aylien.com/naive-bayes-for-dummies-a-simple-explanation/> [25 April 2018]

Wang, H., Can, D., Kazemzadeh, A., Bar, F. and Narayanan, S. (2012) A System For Real-Time Twitter Sentiment Analysis Of 2012 U.S. Presidential Election Cycle pp 115 - 120 [online] available from< https://dl.acm.org/citation.cfm?id=2390490 > [25 January 2018]

What Are N-Grams? (2014) available from <http://text-analytics101.rxnlp.com/2014/11/what-are-n-grams.html> [20 March 2018]

14 Appendices

14.1 Appendix 1 – Project proposal

14.1.1 To what extent can sentiment analysis accurately predict how well a film will do at the box office prior to its release?

To what extent can sentiment analysis be used to accurately predict how well a film will do at the box office? What I hope to answer with this question is if analysis of what people are saying prior to a film's release can be used as an accurate method of prediction. The goal of this project is to see if a prediction is accurate enough based on twitter sentiment analysis to aid production companies in the hope that they can avoid financial loss or at least reduce it as much as possible. In today's world many films are being released per year with many becoming financial losses due to many people taking to social media to voice negative opinions about the films based on their own conclusions drawn from trailers and news about the film in production. What I hope to achieve is with this project is to create a twitter sentiment analysis application that can show the user how positively or negativity the film is being talked about on twitter as well as being able to visualise this data and draw out a prediction on the films future in the box office. I aim to achieve this by initially collecting data using the twitter API about a few films that have already been released to gauge an idea of how the sentiment analysis results relate to real world situations, then take data about an upcoming film to see if I can make an accurate prediction based on the pattern/relationship I have drawn out from my primary research. To complete this task, I will use python to build my analysis software, python has a module that can perform sentiment analysis however I will build up my own module that can perform sentiment analysis to fit the needs of my project. To store the data, I will employ the use of an sqlite database to store the tweets as python can interact very well with sqlite. Finally, I believe building a GUI within python would help to allow easy user interaction with the system I intend build. Of course, there are existing twitter sentiment analysis applications but because I intend focus in of the film industry and I want to use my analysis alongside techniques such as data mining to spot patterns that then allow for predictions.

14.1.1.1 Keywords

Twitter API; Hashtag; SQlite Database; Sentiment analysis; Word classification; Python Libraries

14.1.1.2 Project title

A sentiment analysis application for predicting the success of upcoming box office movies.

14.1.1.3 Client, Audience and Motivation

I believe my project is important in today's film industry as we are at a time where we have loads of data from the customers about a film prior to its actual release such as opinions based on trailers. To allow such data to go to waste without analysis is simple negligent so that is why this is important to consider. My project is important to film production companies such as universal because no company wants to take a loss and every company wants to profit or if that is seeming impossible reduce losses to a bare minimum. My project will generate new knowledge about upcoming films as it will provide predictions that I hope will in turn provide aid to production companies so that they may be as successful as possible and perhaps in the long-term help companies produce great films people wish to see armed with the knowledge my app will provide.

14.1.1.4 Primary Research Plan

Within my project primary research will be key as before I can create a prediction system I will have to gather data to develop my system and draw out a relationship/pattern between the results of the analysis and the real-world success of the film. To access the data, I will use the twitter API which provides access to a portion of twitter's 'tweet' data. To implement my system, I will be using python and some external modules such as tweepy which will allow me to access the twitter API and gather the data. Upon setting up a connection between python and the twitter API I will query the data and store the results in an sqlite database, for example the query will be a film title and I will store all the data about that film in a sqlite table. Once I have the data stored I can then move to stage two which is the implementation of the system and as mentioned previously I will use python to do this. For the primary research stage of my project I will implement an alpha version of my final system that will can produce a set of results (such as percentage of positive tweets). Armed with these results I will then look at the success of these films to determine a pattern with the results I get and the real-world success of the film, the pattern/relationship developed will be used to generate a predictive algorithm of sorts. Finally, I will need to test my system and to do this I will perform two types of test the first being a unit tests of the initial system I develop to see if it contains any syntax or runtime error when in operation. The final test I will perform will be functional testing to see if the system produces the correct results for me to develop my predictive application.

14.1.1.5 Literarily review

In today's world social media is a wealth of data such as customer reviews and opinions, twitter is a power house in the social media world as it is a fast way to share opinions/ideas that have the potential to be viewed by the people and companies to whom you have an opinion about. Within this project I intend to use this data from twitter to perform sentiment analysis to determine whether a film is viewed positively, neutral or negatively. Upon making this determination I will use techniques such as data mining and predictive algorithms to make predictions based on the data to determine how successful a film will do at the box office. This project will involve using the twitter API to gather tweets which will act as my data that will be stored in an SQLite database. My application will be built in python and my predictive algorithm will be prototyped and adjusted based on data that I will gather from previously released films.

14.1.1.6 Initial/Mini Literature Review (500 words – 750 words)

Twitter sentiment analysis is a field of computing that has become increasingly popular to use and develop given the fact that it is very beneficial in generating a general narrative on a certain topic area. Furthermore, upon greater research into this topic area it is clear that sentiment analysis in combination with data analytics techniques can indeed perform predictions and this has been utilised in the stock market area (Pagolu et al. 2016). Prior to the development of this analytical technique historical data was used as a method of prediction for the stock market, however it lacked accuracy given the fact that there are large fluctuations in the market due to real world events (Pagolu et al. 2016). In a study published in 2016 with the goal predicting stock market movements they used a sentiment analysis approach which involved a process of data pre-processing prior to the sentiment analysis. Firstly, to filter the tweets so they can be analysed they used a process called tokenization which breaks down each tweet into its individual word, then a stopword removal process is used to remove words that express no sentiment and finally a regex matching is deployed to remove special characters (Pagolu et al. 2016). The method of pre-processing of the data is one I intend to employ into my system as I believe it is effective in getting the data into a useable state for sentiment analysis. Within this study it states that the analysis corpus must be specific for each subject area so what this means is that I will have to develop a unique analysis machine in my project (Pagolu et al. 2016). This research project relates heavily to mine as it to shares a goal of predicting future events of a certain topic area so the methodologies and processes used will be employed within my research project, particularly the process of data preparation for analysis.

Another sentiment analysis research paper looked at sentiment analysis as a way to observe the general narrative of the 2012 U.S presidential election. Due to the very opinionated nature of the election there were great fluctuations within the sentiment surrounding this topic area. Furthermore, this paper deduced that the volume of tweets was largely driven by real time events (Wang et al. 2012). To achieve this real-time sentiment analysis, they employed the use of a tokenizer which handled all data pre-processing. Furthermore, they used aggregation to output continuous data such as tweet volume and sentiment figures such as the number of positive tweets that are produced every five minutes given the fact tweets are continuously received (Wang et al. 2012). Although not directly related to my project the ideas presented by this research paper are interesting and will prove useful for me and help me understand more about the area of sentiment analysis because as I predicted real time events effect the volume tweets which is key in sentiment analysis as stated in another research paper which also had the goal of monitoring election results and even predicting them (Bermingham and Smeaton 2011). Within another research project they found that the human language can be complex and when analysing sentiment as an initial response to a single event may not reflect the overall inner perspective of the person, so it is suggested that the volume of the tweets could potentially hold a greater insight as it could indicate popularity of a certain topic (Bermingham and Smeaton 2011). From these two projects I can deduce the perhaps simply deciding whether a tweet is positive or negative is not enough, rather consider other factors such as volume of tweets as this has proven to help provide further insight into the overall narrative of a topic area.

In conclusion I intend to draw ideas from these previous projects with the goal of applying them to the subject area of film. The data pre-processing of the stock market project is one I am keen to adapt and implement into my own system into order to get the data into a usable state so I am able to perform sentiment analysis. In terms of producing results, I intend to draw from the U.S presidential election project as the results they produced such a volume of tweet and number of positive and negative tweets are results I wish to produce to gauge an idea of the general sentiment surrounding the film. Furthermore, following real-time events is something I will have to do as the volume of tweets will be influenced by real-time events such as trailers being released which is an idea that is key in the afore mentioned project. Finally, Within the real-time election analysis they claim that their research could be expanded upon to gauge sentiment on “films and actors surrounding Oscar nomination and selection” which of course relates heavily to my project and suggests that what I aim to achieve is possible given this past research (Wang et al. 2012).

14.1.1.7 Bibliography

Bermingham, A. and Smeaton, A. (2011) On Using Twitter To Monitor Political Sentiment And Predict Election Results, pp 9 [online] available from <http://www.aclweb.org/anthology/W11-3702> [25 January 2018]

Pagolu, V.,Challa, K., Panda, G., Majhi, B. (2016) Sentiment Analysis Of Twitter Data For Predicting Stock Market Movements [online] available from <https://arxiv.org/pdf/1610.09225.pdf> [25 January 2018]

Wang, H., Can, D., Kazemzadeh, A., Bar, F. and Narayanan, S. (2012) A System For Real-Time Twitter Sentiment Analysis Of 2012 U.S. Presidential Election Cycle pp 115 - 120 [online] available from< https://dl.acm.org/citation.cfm?id=2390490 > [25 January 2018]

14.2 Appendix 2 – Github links

14.2.1 Primary Research Repository

https://github.coventry.ac.uk/staceya4/Dissertation-Primary-Research

14.2.2 Final System Repository

https://github.coventry.ac.uk/staceya4/Dissertation-Final-System