



DEPARTMENT OF COMPUTER SCIENCE

Are TV ratings possible with Twitter?

Andreas Georgiou

A dissertation submitted to the University of Bristol in accordance with the requirements
of the degree of Bachelor of Science in the Faculty of Engineering

Declaration

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Bachelor of Science in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Andreas Georgiou, May 2013

Abstract

This project explains the creation and the use of tweetTV, a system that collects and analyses tweets from the micro blogging site Twitter, with the aim of comparing them with TV ratings in order to identify meaningful relationships.

Using tweetTV, Twitter text messages were collected, processed and then classified, by applying some data mining techniques. By isolating tweets that referred to three specific popular TV shows, a significant number of tweets was analysed and compared with TV ratings available from the British Audience Research Board (BARB). Furthermore, sentiment analysis was implemented to investigate the temporal variability of positive, negative and informational tweets in conjunction with TV ratings and whether the awareness of tweet's attitude could enhance the accuracy of the system.

The running of tweetTV was successful in collecting sufficient amount of tweets to carry out the analyses required, which included qualitative and quantitative comparisons. A strong correlation was found between the number of tweets and the number of viewers, confirming the existence of a link between the two values that could be promising in accurately estimating TV ratings by only using Twitter. The small numbers of tweets collected for one show revealed a potential limitation of linking Twitter to some types of TV shows with relative explanations being discussed.

The use of sentiment analysis also proved to be useful in identifying trends related to TV shows such as a periodicity of positive tweets during the days prior to the show. This suggested that sentiment analysis could be used to improve the accuracy of tweetTV by weighting people's opinions before the shows and projecting estimations.

Overall, the performance of tweetTV was successful in collecting and filtering tweets and in conjunction with sentiment analysis; it has the potential to work as a real-time application that will provide TV ratings.

Table of Contents

Abstract	1
Table of Contents.....	2
1. Introduction	4
1.2. Twitter	4
1.1. Analysis Area	5
1.2. Statement of the Problem	5
1.3. Structure of Thesis.....	5
2. Background Chapter	6
2.1. Twitter	6
2.1.1. Hashtag.....	6
2.1.2. GeoLocated Tweets	7
2.1.3. Trending Terms.....	7
2.2. Approaches to Measurement	8
2.2.1. Recall Metering.....	8
2.2.2. Electronic Metering	8
2.2.3. Random Sampling.....	9
2.3. Sentiment Analysis	9
2.3.1. Text Polarity.....	9
2.3.2. Opinion Strength	9
2.3.3. Subjectivity/Objectivity Identification	10
2.3.4. Feature/Aspect-based Classification	10
2.4. Data Mining	10
2.5. Related Projects.....	10
2.5.1. “SentiStrength”	11
2.5.2. “TwitterPaul”	11
2.5.3. “Flu Detector”	11
2.5.4. “Predicting the Future with Social Media”	11
2.5.5. Evaluation.....	12
3. Requirements & Design	13
3.1. Requirements	13
3.2. Design	14
3.2.1. Data Mining	14
3.2.2. Dataset	15
3.2.3. Evaluation.....	15
3.2.4. Baseline	15
3.2.5. Tweet Rate	15
3.2.6. Tweet Polarity.....	16
3.2.7. Bag of Words	16
3.2.8. Hash Maps.....	16
4. Implementation	18
4.1. Twitter APIs	18
4.3. Data Gathering & Selection:	19
4.4. “tweetFilter”	20
4.5. “tweetSent”	20
4.5.1. Pre-Processing	21

4.5.2.	Tokenization	21
4.5.3.	Data Cleaning.....	21
4.5.4.	Classification.....	21
4.5.5.	Dictionaries.....	21
4.5.6.	Slang Dictionaries	22
4.5.7.	Emoticons	22
4.5.8.	Exclamation Mark	22
4.5.9.	Capitalization	22
4.5.10.	Word Classification.....	23
5.	Analysis & Results.....	24
5.1.	Qualitative relationship between TV ratings and Twitter	24
5.2.	Correlation between TV ratings and Tweets	26
5.3.	Sentiment analysis Testing	27
5.4.	Relationship between ‘mood’ of tweets and TV ratings	28
6.	Discussion & Conclusions	30
6.1.	Conclusions.....	30
6.2.	Correlation between Twitter traffic and viewers.....	30
6.3.	Semantic Analysis	30
6.4.	Trending Phenomena	31
6.5.	Failure of detection	32
6.6.	Technical Issues	33
6.6.1.	Data Management and Storage.....	33
6.6.2.	Twitter Demographic.....	33
6.6.3.	Language limitation of sentiment analysis	33
6.6.4.	Other Limitations.....	34
6.7.	Further Work	34
6.7.1.	Algorithm Improvements	34
6.7.2.	Web Application	35
7.	References	36
8.	Appendices	39
8.1.	Appendix A – Generating API Keys	39
8.2.	Appendix B – Implementation code	39
8.3.	Appendix C – “tweetFilter” results – “Analysis.txt”	42

1. Introduction

Television has probably been one of the most influential media until nowadays, along with the radio and more recently the internet. The wide availability of TV channels is translated into a very strong competition that requires successful marketing and good strategies in order to thrive in the challenging environment of the media. The recording of people's watching preferences has been one of the essential tools to extract precious information related to the commercial and social aspects of TV. The methods used to record TV-viewing preferences have been changing along with the progress of technology; however, there is a constant effort to increase the accuracy of these measurements and reduce the time taken to collect them. Knowledge derived by these records can be the key to business success and planning of effective strategies.

1.1. General aim

The general aim of the project is to discover any relationships between the text messages that users post in Twitter with the real TV ratings generated by the Broadcasters Audience Research Board (BARB) and identify opportunities for further understanding of this new concept. In order to achieve this, natural text language is collected and processed using a program that has been developed for this purpose. An essential part of the project is also the development of an opinion miner which is able to classify tweets according to their polarity. This classification of tweets is required to examine whether people's attitude is a factor to be considered when comparing tweets with TV ratings.

1.2. Twitter

During the last decade, there has been a boost in the usage of social media, where anyone can have access and publish anything for everyone to see. Twitter is the third largest growing network, recently surpassed by Google+, where its users post their opinions almost about everything, including their views on television series [1]. This has been regarded by this study as a major opportunity to use the data originating from Twitter, as it is an extremely popular social medium with millions of users. Material published by a huge number of people can be particularly useful in having access to people's minds, without having to go through costly and lengthy surveys, whilst the size of sample taken can be substantially larger. Furthermore, data collection from Twitter can be done in real-time, thus saving time lost between collection and analysis of tweets. Twitter's upper limit in characters is another advantage which makes analysis easier, since analysis of long texts is avoided.

Twitter has also several functions that facilitate the identification and grouping of information. The use of hashtags is an example of such a function, since it enables picking tweets relevant to a specific subject by avoiding irrelevant tweets to some extent. Furthermore, geo-location tags can also give an insight into patterns related to location that may be revealed. Finally, the easy access to internet in most places

around the UK contributes to the phenomenon of trending, which can be particularly revealing about people's preference for a channel or show, especially in cases of a recent or forthcoming event

1.1. Analysis Area

The project focuses on two main areas, with the first being the collection of useful data and its subsequent comparison with TV ratings, as an attempt to investigate whether there is a correlation between television ratings and Twitter traffic referring to specific TV shows.

Secondly, this project covers into a certain depth some possible ways that semantic analysis of the tweets messages can be conducted. A classifier determines whether the text messages collected are positive, negative or neutral. Considering these results, indications of the popularity of a show in the cyberspace and its connection with TV ratings are sought.

1.2. Statement of the Problem

The main idea behind this project is to identify whether there is a capability of substituting the traditional ways of monitoring television audience with modern and more accurate methods by using the data provided by social networks in real time. Harvesting a sufficient quantity of data with the appropriate content is another part of the problem where the importance of successful filtering is highlighted. The influence of a TV programme is also questioned, depending on the time that people tend to post on Twitter their short text messages, i.e. whether it is during or after watching a specific TV program.

Concluding, finding a formula to be able to generate accurate metrics is extremely difficult; however by taking into account the number and the context of the tweets, the primary goal is to investigate this relationship and discuss the conclusions generated throughout the project.

1.3. Structure of Thesis

Initially, this study goes through a literature review by exploring current and past television audience measuring methods, along with sentiment analysis techniques used by other projects, looking for lessons learned and ways to benefit from them. Other related projects involving the handling of Twitter data were also used to contribute to this research. All the requirements for the project to be considered as successful and complete are also identified, including descriptions and explanations of the algorithms and the design concept behind the constructed system. A detailed description of the implementation stages in all three phases of the project is then outlined, referring to the specific techniques used to collect, process and classify tweets, explaining how each part of the tweetTV system operates. Collected tweets are compared with TV ratings qualitatively and quantitatively, employing charts to better demonstrate the findings. Finally, all results are discussed with the conclusions about tweetTV's performance being outlined; simultaneously capturing potential for improvements and future work.

2. Background Chapter

In the audience measurement sector a significant amount of money is invested in the development of accurate methodologies for TV ratings. Producers and shareholders rely on the metrics to plan their marketing strategies and increase their profits. In many cases it is the only tool for broadcasting channels to decide on the future of television programs. The literature review that follows gives an insight into Twitter and how it works, along with past and modern approaches to measuring TV audiences. Furthermore, a review of techniques and examples from relevant studies is also provided.

2.1. Twitter

Twitter is a relatively new real-time social network that was founded in 2006. The users of the micro-blogging site are only allowed to post short text messages similar to “SMS” commonly known as “tweets”. Tweets have a limit of 140 characters and may also include the geographical location of the device used to post them. A random tweet is shown in Figure 2.1.1.



Figure 2.1.1- “Example of Tweet from Twitter.com”

Twitter has become extremely famous counting more than half a billion users and it is ranked as the third largest social network worldwide [1]. At least 10 million UK users are actively engaging with the site and around 90% of online conversations consist of discussions about TV series and programs [2].

Tweets are by default publicly posted online, therefore, anyone who has access to Twitter is able to observe and interact with any of the tweets or follow any of the Twitter’s users. On the other hand, the site gives the option to a user to lock their account and let only approved followers to see their tweets. Nevertheless, most of Twitter’s users choose to widely publish their tweets, therefore a creation of a huge database may offer valuable information to researchers and analysts about the behaviour and nature of people’s feelings.

2.1.1. Hashtag

Among the features of Twitter is the hash-tag (#), a prefix in front of a word or a sequence of words that describes the content and sometimes the emoticon of the author. It also enables keyword searching and facilitates content categorisation. This is one of the main advantages of Twitter over other social networks, since it allows users to tune in easily on the same subject and follow on the flow of information about an event in real time. For example, if a Twitter user posts a message about the “Top

Gear” TV series, then it is most likely that one of the hash-tags will be “#topgear”.

Hashtags are sometimes promoted as official by broadcasting channels or brand managers to boost discussion for their products or television show. Following the example of many companies, BBC TV series “Top Gear” has “#BBC_TopGear” as the official hash-tag for their broadcasted show.

2.1.2. GeoLocated Tweets

Another important feature of Twitter is the recording of the place or location where the user posted the tweet, by producing tweets with real latitude and longitude coordinates. Locations can be then mapped on to Google Maps and show the approximate location of a user, thus giving an indication which regions of the UK have special interests (see Figure 2.1.2.1). This feature is optional; therefore most of the users tend to avoid posting their location online for privacy reasons. However, it is still quite interesting for data mining purposes to be able to acquire information about location of users.



Figure 2.1.2.1 – “An Example of Geolocated tweet”

2.1.3. Trending Terms

Trending is the term to use in Twitter to call something that has been around and becomes an object of temporary interest. Twitter uses a complex algorithm to generate a list with trending terms in real-time. The algorithm takes into account the velocity of the trending term, the volume of the tweets and the location of the network. This is why it is harder for a term once joined the list of popular trends to stay on top. Figure 2.1.3.1 shows the trending terms on Twitter as captured on the 29th of April.

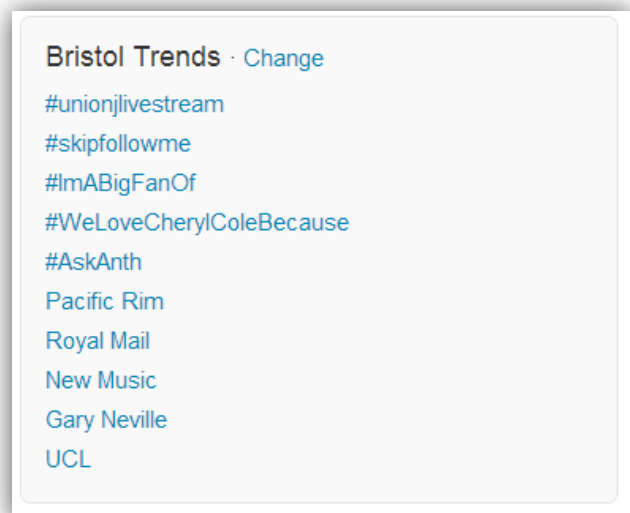


Figure 2.1.3.1 – “Trending Terms on 29/04/2013”

2.2. Approaches to Measurement

In the last few decades of television's life, companies that specialise in this sector invested money in various technologies starting from Diaries to Electronic devices and historically this process was carried out with surveys and forms filled out by the public. There are three ways that measurement of audience is done, namely recall, electronic and predictive methods.

2.2.1. Recall Metering

People were given a diary ("sweeps") and were instructed to write about which television programs they used to watch and the corresponding duration [3]. The diaries were then given back to the measuring company and the data were then analysed producing the television ratings. This process was not only unpractical and time consuming for the users but it was proven to be a very inaccurate method [4]. This way of tracking the viewers' preferences was deemed unreliable, since people tended to forget or unintentionally made mistakes. As the human factor was playing the leading role in surveys and diaries, they were always suffering a constant error percentage [4]. Therefore, even if online surveys were available, this would not guarantee an error-free approach.

2.2.2. Electronic Metering

Television viewership measurement has been significantly improved by the installation of a mechanical device called "audience meter". "Audience meter" is essentially a recording device usually referred to as the "black box"; it is embedded in a television device and records the duration and the channels the viewer has watched by recording the frequency of each channel. Despite its accuracy, it has often been suffering from various problems since satellite channels have been introduced to the market, due to its inability to provide recordings of satellite broadcast. The downside of each approach is the inability to provide the demographic view of the viewers to the marketing industry [5].

In the 1980's a new technology was introduced called the "local people meter" (LPM), when a dispute started about the quality of the ratings [6]. LPMs were installed on devices and enabled users to login and record minute-to-minute whatever each user watched. Users had the control to start and stop recording of their viewing habits at any time. LPM and diaries can provide a demographic view of the ratings, since they provided specific information on the personal details of each user (sex, age, ethnic origin) [6]. Broadcasters' Audience Research Board (BARB) installed this type of device in 5,100 households, monitoring 5,000 homes across different regions of the United Kingdom to generate estimations [7].

Nielsen recently launched a cross-platform measuring applications that are able to log user's behaviour from television, mobile and computer devices [8]. Not only do they effectively monitor the conventional television viewing time but they also record the time that a particular user spends on the internet either

watching online TV or surfing the internet [8]. As incrementally more users spend time in the cyberspace, it is essential to create a complete profile to watch what websites a user enters either during the TV show or after.

All the methods mentioned above describe different ways of collecting viewing preferences. Nevertheless, in all cases the data analysis that occurs after collection is based on a statistical approach called “Statistical Sampling”, which is explained next.

2.2.3. Random Sampling

The companies randomly choose a sample that comprises approximately 1% of the total audience in order to identify the percentage of viewers’ preference to a particular channel [9]. Statistical sampling is generally accepted to be the most accurate method to reflect trends and preferences of wholes and can be parallelized to an exit poll on the elections day. Even though it has been proven that increasing the sample size will definitely decrease the error percentage, this is only true until a threshold is reached [4]. Unfortunately, it seems to be impossible to use this kind of statistical analysis in this project as the only object of analysis is the text messages that users send through the micro-blogging site “Twitter”, whereas any further information about the users is not available.

2.3. Sentiment Analysis

Sentiment Analysis, or ‘opinion mining’, is a combination of machine learning methods used to determine the opinion of the author on a specific subject [10]. The process of analysing natural unstructured language has now more than ever troubled the scientific world. Different applications of this technique were implemented using machine learning algorithms. Lately, the sentiment analysis is used commercially to rate customer reviews or to discover the popularity of a specific brands and businesses on the internet [11].

2.3.1. Text Polarity

There are various methods with sentiment analysis which are all based on the probability of a word or a sentence to be positive, negative or neutral. By using traditional machine learning algorithms such as the Naive Bayes classifier, the features from each text extracted are independently evaluated to calculate the probability and classify them correctly based on the trained data set supplied [12].

2.3.2. Opinion Strength

Opinion strength, or also known as ‘scaling’, is a more advanced method that measures the strength of the emoticons captured by the author [13]. The scoring system uses a scale that usually spans from –10 to +10, with negative to positive values being determined according to the power of emoticons.

2.3.3. Subjectivity/Objectivity Identification

A different approach to sentiment analysis is the classification of the corpus of data based on a subjectivity/objectivity score. The algorithm examines whether the context can be classified as subjective or objective and then determines the sentiment strength of the text [14].

2.3.4. Feature/Aspect-based Classification

A complete and more complex method is the Feature/Aspect-based analysis. The algorithm of this method tracks down the desired feature in a document and identifies the opinion expressed for that feature. For example, a feature can be “smartphone” or “PC”; therefore, this method distinguishes opinions in objective/subjective and determines its strength for a specific aspect of the feature [15].

2.4. Data Mining

The huge amount of information generated today from the social media makes data mining a particularly stimulating in the research community. Mining is the automated processing of large amount of information and extract meaningful value. Common method in Data mining is the 5 stages Knowledge Discovery in Database (KDD) processing along with other variations of the same concept. A six stage KDD process is mainly used in businesses is Cross Industry Standard Process for Data Mining (CRISP-DM). The six stages are, “Business Understanding”, “Data Understanding”, “Data Preparation”, “Modeling”, “Evaluation” and “Deployment” [16].

There are two main techniques for data mining; regression and classification. Regression is based on the analysis of data with an aim to produce a mathematical formula. Its purpose is to find the best formula that describes the numerical data as accurate as possible. Then the formula can be used to predict the behaviour of new sets of data. The main drawback of this method, however, is the fact that it performs well only with continuous data.

The second data mining technique is by classification, which tries to classify data according to a set of features. Instead of a formula, a decision tree is used to classify the data to classes. Classification results using this technique are much easier to be interpreted and explained. The form of the data set in this project makes the use of the second technique more suitable than regression.

2.5. Related Projects

In October 2012, Nielsen company, a major player in information measurement cooperation, announced the launch of a project that will try to produce television metrics based on the social activity from Twitter [17]. Despite that this announcement was made recently, no previews attempts have been made to achieve a solution using social networks before. On the other hand, many research papers and journals have produced interesting predictive models by analysing users’ text messages which are referred below.

2.5.1. “SentiStrength”

“SentiStrength” is a powerful tool developed by a group of scientists and professors. The classifier uses a variety of different ways to successfully classify tweets. For each corpus of tweets, it generates a negative and a positive score. Among the features used, a list of negative and positive words, a list of emoticons and a detection mechanism of abbreviations grammar correction were included.

In addition, the capitalisation and punctuation marks were taken into account to boost the. SentiStrength also includes a dictionary used to autocorrect grammar mistakes in the text processed. The algorithm was trained with text messages from different social networks such as MySpace, Facebook and Twitter. The datasets were then rated by scientists and were used to improve the performance of the algorithm. Comparing it with other algorithms including Naive Baise, SentiStrength has achieved significantly higher accuracy. [13]

2.5.2. “TwitterPaul”

“TwitterPaul” is a great example of handling Twitter messages as the main data input and implementing data mining techniques. It was designed to predict the 2010 FIFA World Cup tournament results by extracting predictions from users’ tweets. The system used approximately half of a million text messages and among the techniques used, two important conclusions were found to be relevant to this project. First, the fact that historical features derived from users were employed in order to boost the system’s accuracy had no effect, subsequently leading to abandoning them in the final version of “TwitterPaul”. Secondly, it was outlined that high precision methods on smaller datasets give much more accurate results than low precision methods on huge datasets. Therefore, for the purposes of this project, the text analysis precision should be maximized by trying multiple algorithms. [18]

2.5.3. “Flu Detector”

“Flu Detector” is another similar project that was able to simulate the flu around the United Kingdom with a high precision accuracy of Influenza-like Illness (ILI) rates. It was built to monitor flu by collecting and aggregating tweets based on some selective features in 3 different regions (Central England, South England and Wales). From this project, the methodology used was also quite valuable to collect tweets from three different urban areas. By using the “geolocation” tags on the tweets, “Flu Detector” collected tweets only from 49 urban areas in a distance of 10km. The system collected around 50 million tweets. [19]

2.5.4. “Predicting the Future with Social Media”

This projected success to predict the movie box office sales by analysing the sentiments from tweets. The classifying algorithm processed 2.89 million tweets from 1.2 different million users to examine the correlation with movies sales in cinemas. As Asur’s and Huberman’s research clearly suggests, tweets

reflect the willingness of people, if they are going to see the movie or not. This project used the method of creating a predictive model based on the tweet traffic (tweets, retweets and replies) posted on Twitter, where a high linear correlation of 0.90 between tweet rate and box office sales was obtained. Furthermore, the text classification of the tweets proved that sentiment analysis of the tweets is a good indicator of what people think about a movie. It should be noted that Asur's and Huberman's research also underlines the value of sentiment analysis in social media including Twitter and that the success of this model could be followed in other examples as well. [20]

2.5.5. Evaluation

Evaluation of sentiment analysis is usually done by comparing what humans think with algorithm-generated values. According to this research, people only agree up to 79% of facts/arguments; as opposed to machines which cannot understand ironies, sarcasm or humour in text [21]. Therefore, a classifier that is able to achieve an approximate accuracy of 70% may be considered as accurate as a human judgement would be. In other words, a perfect algorithm classifying everything correctly would still be 20% off. For evaluation purposes, correlation measurement is more accurate than precision because the first can show how close the expected results are from the actual ones.

3. Requirements & Design

3.1. Requirements

The purpose of all processes and applications is first identified and is based on the requirements that arise from the problem; therefore these requirements will act as the specification guide for designing processes and components of the proposed system. The data to be collected and analysed focus on three different areas. These areas comprise a) the total number of tweets discussing about a particular subject, b) the rate of frequency that a tweet related to a subject is posted and c) the mood/nature of the tweet by distinguishing them in positive/negative. These are taken as the main parameters of the project that determine the stages of Design and Implementation.

In addition to the nature of the data required, the principal processes that constitute the proposed system are accurately specified in this section. The initial identification of all stages is important, as this enables to build the structure of the desired system in a logical way. These stages involve collection, filtering and analysis of data/tweets. These system components are captured in three stages and for convenience the programs developed should carry appropriate names. There should be a program that will collect tweets off the internet (tweetCollector), one that will filter and clean the data (tweetFilter) and one that will classify the data (tweetSent). The proposed system is therefore named tweetTV.

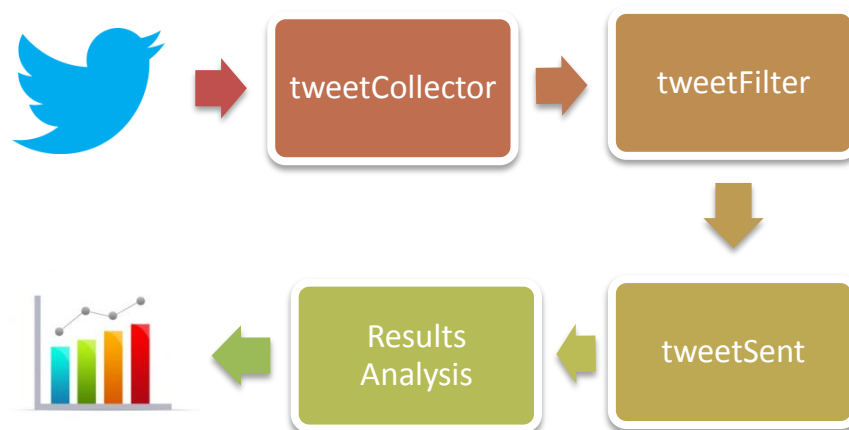


Figure 3.1.1 – “An illustration of the System”

The operation of the tweetTV is defined through the list of specifications below considering all the phases, from collection to delivery of results. It is expected that by fulfilling this set of requirements tweetTV should reach an optimum performance.

- Continuous streaming/harvesting of tweet 24hours a day, for one week, in order to ensure capturing a variety of trends.
- A fast rate of filtering of tweets should be ensured either by keyword or by location.
- The algorithm should classify tweets successfully and as closely as possible to objective parameters, enabling the delivery of meaningful results.

In principle, the algorithm should be able to give provide collected data for analysis and presentation of graphs of numbers of tweets versus time. The tweetTV system should be able to produce the amount of tweets collected for each program during the period of time of seven days. In addition it will be necessary to identify other useful data about users posting the tweets such as location. Such a function of the system will enable the tracking and harvesting of tweets that are only posted from users inside the United Kingdom.

Tweets should be filtered by the tweetFilter program and passed to the tweetSent that will determine the polarity of the text. Concluding, these series of processing should produce a log which will be represented graphically to facilitate comparisons and identification of trends; thus comprising the main object of analysis.

tweetSent should be evaluated and the values generated should indicate some correlation with the 'human' evaluated tweets, in order to accept its suitability for classifying tweets. As a minimum acceptable threshold, given the complex content of tweets, the sentiment analysis tool should be able to produce an accuracy rate higher than 50%, which is the random chance for flipping a coin.

Indications of any relationship should be sought between TV ratings and tweets as long as with the nature of the tweets (positive, negative and neutral).

3.2. Design

The main approach adopted in the project will be the Knowledge Discovery in Databases (KDD) process. KDD is considered as the fundamental technique for data mining which usually consists of the following stages: Data Gathering & Selection, Pre-Processing, Data Mining and Evaluation (Figure 3.2.1).

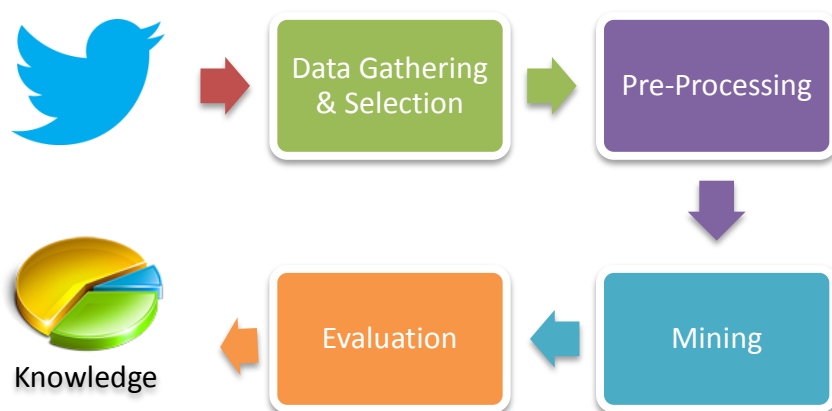


Figure 3.2.1 – “The 4-stage KDD process”

3.2.1. Data Mining

The next step is to “mine” the data. By analyzing the context of the tweets, the system classifies them into three categories “Like”, “Dislike” and “Neutral / Informational”. In reality, it will be attempted to construct a system, where tweets will act like votes and votes will weight differently according to the

category they have been assigned. Therefore, it will be possible to rank the TV programs from the tweets extracted by the system and by taking into account the “tweets”, “retweets” and hash tags associated with them.

3.2.2. Dataset

Since the subject of the project is focusing on tweets, Twitter will be the source of data/tweets. Three TV shows with high popularity in the UK were chosen, namely ‘*MasterChef*’, ‘*Broadchurch*’ and ‘*One Born Every Minute*’. Table 3.2.2.1 gives the details of broadcast of each show. Two weekly periods in April were chosen to harvest tweets about these TV shows, anticipating large logs of tweets due to the forthcoming ending of the TV season. The first weekly dataset spans from the 2nd to the 9th of April and the second week covers the period from the 17th to the 24th of April. No upper limit in collecting tweets is set, as the larger the sample the better the expected accuracy of results.

Channel	TV Show	Broadcasted
BBC1	Masterchef	20:00 Wednesday
BBC2	Masterchef	20:00 Thursday
Channel 4	One born every minute	21:00 Wednesday
ITV	Broadchurch	21:00 Monday
BBC1	Master Chef	20:30 Friday
BBC1	Top Gear	Non-Broadcasted

Table 3.2.2.1 - “TV Programme and shows of interest”

3.2.3. Evaluation

The last stage of the process will be the creation of an abstract form of the data in a way that an easy visualization of the findings will be enabled. At this stage, the main objective will be to discover relationships with “real” data provided by the Broadcasters' Audience Research Board (BARB). In all cases the evaluation will focus on qualitatively analysing the observed similarities and differences, attempting to detect sources of errors and opportunities of improvement.

3.2.4. Baseline

This will be the total number of tweets that will help to construct a baseline and generally will be the primary detection, assuming that if something is a hot topic on the social media, it will have more views on TV as well. Recently TV broadcasting channels promote official hash tags such as (#BBC_TopGear) to promote online discussions.

3.2.5. Tweet Rate

The tweet rate will record the amount of traffic during the times which the show is on air. In other

words, this is another way to examine whether users tend to post relevant messages while watching the show. Later in Chapter 5, the results are displayed allowing a detailed discussion about the correlation of the tweet flows and the real users of watching the show.

3.2.6. Tweet Polarity

By mining the opinions of the viewers through their messages on tweets, meaningful information is extracted about the popularity of the TV production. It is an indirect way to learn about the influence of a show has on the viewer and it could not only help to improve the quality of a show but even potentially enhance the planning of commercial strategies. This is why a significant part of the project is dedicated on Semantic Analysis of natural language text.

3.2.7. Bag of Words

'Bag of words' is a common algorithm used for natural language analysis and it has recently been applied in the vision recognition field as well. This is one of the main methods of semantic analysis that is used for this project. The concept of using the Bag of Words model is to ignore grammar and logical sequence of words, and to consider every word individually. Therefore, each word is taken as a distinguished token and each token is compared with a database of words.

An emoticon strength indication is given to each word and a tweet is then classified either as positive or negative and a sentence is then similarly labelled as positive or negative (see Table 3.2.7.1). In this way it is ensured that all words are weighed and measured so a score can be generated that will define the tweet's nature. Very often, a sum of negative words does not necessarily mean that a sentence is positive and vice versa. However, due to the huge amount of tweets expected, a manual correction of all tweets is impossible and therefore one of the assumptions of this model that need to be considered is that a concentration of negative/positive words equal a negative/positive tweet.

Category	Keywords in Tweets
Positive	"Think I am watching the best tv programme ever...", "I love top gear !!!"
Negative	"I HATE Top Gear, so boring . . ."
Neutral / Informational	"#masterchef is tonight at 20:00",

Table 3.2.7.1 – "Classification Example"

3.2.8. Hash Maps

Hash maps over other data structures are faster and return results, Collisions in hashmaps are avoided since every key is unique. Considering this project, there is only one word in either dictionary, positive or negative, therefore hashmap was the ideal choice to proceed for our project.

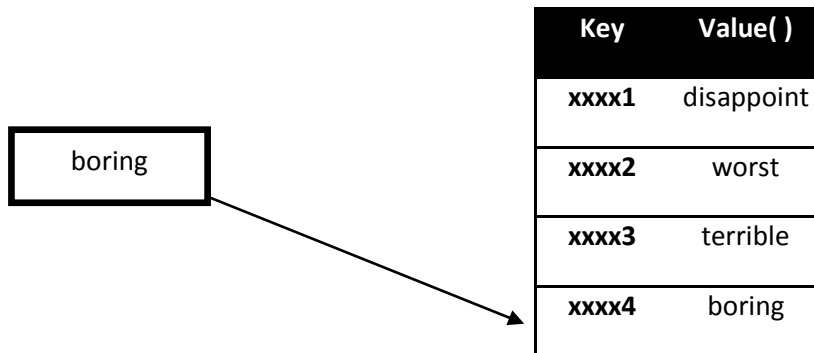


Figure 3.2.8.1 – “Hash Map Example”

Hasmaps by default have a load factor of 0.75. Load factor is the relationship between the initial capacity and how full the map is, which make them a good choice for space and time.

4. Implementation

At this stage, the use of various tools and applications to build the tweetTV is explained in detail, giving information about the procedures followed. The implementation phase technically described and is supported by abstracts of the code used that are included in the Appendices.

4.1. Twitter APIs

API is the abbreviation of “Application Programming Interface” and it allows developers to build their own application and extract data through the Twitter network. There are currently three Twitter platforms, namely REST, Stream and Search API. REST API is used to access timelines of users. Search API helps to search for specific keywords and returns recent results. Finally, the Stream API allows connecting to Twitter database as long as a user desires to stream public statuses in real time. For the purposes of this project the Stream API v1.0 is used.

4.2. Using JAVA with Twitter

Among many programming languages available, JAVA was chosen as the main way to proceed with the implementation of this project. JAVA language is platform independent and closely related with JAVA script, with these offering the potential to develop a web application of the program in the future.

Therefore, in order to proceed with this realising tweetTV, the Twitter4J was used in conjunction with several libraries. “Twitter4J” is a library which enables the easy use of “Twitter API” and “Twitter Stream API” through Java. Twitter4J is built in OAuth support and is fully compatible with the new version of Twitter API v1.1. [22]

Requests are done in the form of URLs and the process is described graphically in Figure 4.2.1 below.

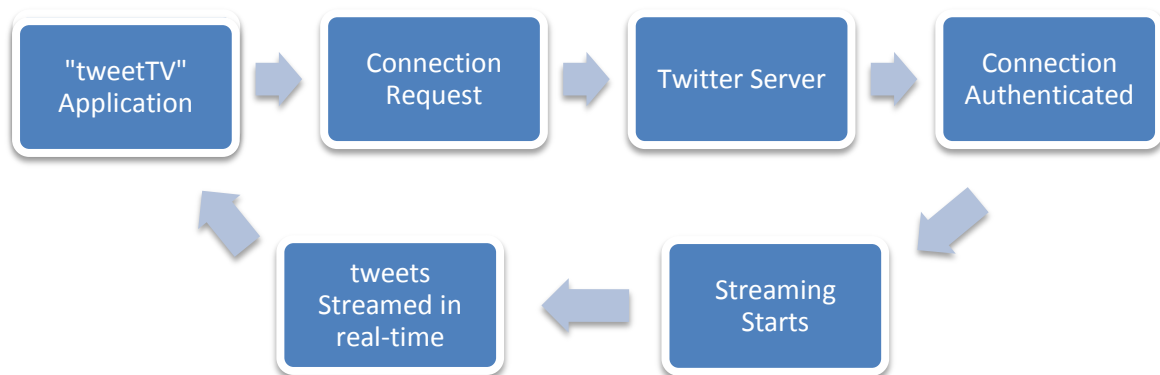


Figure 4.2.1 – “Stream API calls to the Twitter Server”

4.2.1. Authenticating with Twitter

In order to be eligible to collect any data, requests need to be authenticated with the Twitter servers. Creating a Twitter API application it is required to give details about the reason of using Twitter data and the exact purpose of the application. Since it is a third party application, an open standard for

authorization called OAuth v1.0 needs to be used [23]. Generating a key for OAuth enables to make API calls (Application-Only authentication) which in the case of Twitter have an upper bound. In Appendix B Figure B1 shows a screenshot showing part of the code where the API keys were used.

4.3. Data Gathering & Selection:

A stream is created with the help of Twitter4J and using the keys obtained for the application, the stream establishes a connection with the Twitter server. Using the “StatusListener()” class a thread is set up retrieving tweets. The stream is scheduled to stay up and running harvesting tweets 24hours every day for one week. Two versions of the “tweetCollector” were developed, one version of streaming was done by keyword and the other was done by Location. The 1st week’s tweets were collected only by keyword and the second only by location. A more detailed explanation is given in the following sub-chapters.

4.3.1. Collection by Keywords

Tweets are filtered through the JAVA environment only if they fulfil a series of parameters. Among the crucial parameters is choosing the appropriate keywords and hashtags (#). An example of keywords that were used is shown in Table 4.3.1.1.

In order to check the strength of each hashtag (#), an online service called HashTags.org was used, that identifies the trending hashtags over the whole Twitter network. This constituted the first means to identify which keywords should be adopted in order to filter the initial volume of tweets and track other related hashtags on the same topic.

Channel	Programme	Broadcasted	Top Hashtag	Other
BBC2	Top Gear	BBC Two, 20:00 Sunday	#TopGear	#BBC_TopGear
Channel 4	One born every minute	Channel4, 21:00 Wednesday	#OneBorn	#oneborneveryminute
ITV	Broadchurch	ITV, 21:00 Monday	#BroadChurch	#DavidTennant
BBC1	Master Chef	BBC1. 20:30 Friday	#MasterChef	-

Table 4.3.1.1 – “Table of Hash Tags monitored”

Keywords were passed in the form of an array and only when keywords were they found in the tweet content, the tweet was recorded in a (.txt) format file. Each unique tweet is stored in a different text file along with other information recorded including username, content of tweet, date and time sent and the tweet ID.

4.3.2. Collection by Location

This project is interested in investigating only the domestic Television broadcasting metrics. Therefore it was quite crucial for this research to collect data within the United Kingdom only. Nine different urban city centres were chosen under the assumption of them being the places where the highest Twitter activity is likely to occur, due to their significant numbers of population. Therefore, tweets originating from these hotspot-urban centres were only collected.

City Locations were entered as a “long” type double array, where the geographical coordinates are given in two pairs. Each coordinate represents the North-West and South-East corner of the bounding box that surrounds those centres. Figure B2 in Appendix B illustrates the list of cities with their coordinates, given in the form of NW{Latitude,Longitude},SE{Latitude,Longitude}.

4.4. “tweetFilter”

The next step involves the screening of data using tweetFilter. This is a small JAVA program designed to record and select the data containing any of the desirable keywords. The process of filtering tweets; i.e. matching the tweet’s content to keywords is partly demonstrated in Appendix B, Figure B4. In the end, it generates a file containing all the selected tweets, the file locations and the date and the time tweets were broadcasted online. This process was done as a separate class to allow the independence or elimination in any future versions.

4.5. “tweetSent”

tweetSent is the third part of the project and is responsible for opinion mining on the tweets. tweetSent was run through all the tweets collected and results are displayed in Chapter 5. Below, Figure 4.5.1 demonstrates the main processes and how they are interrelated in a simple diagram.

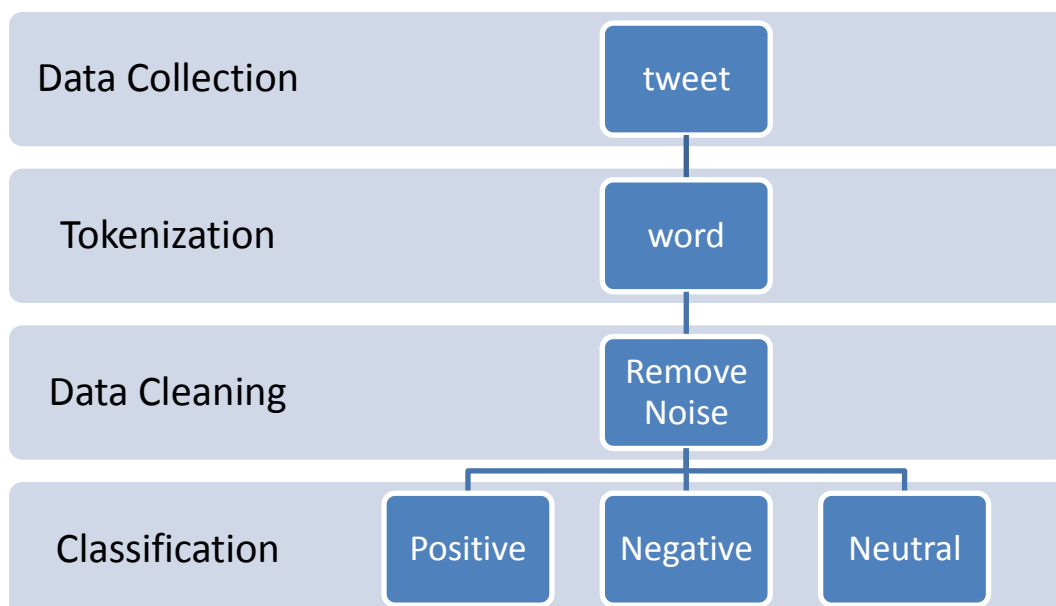


Figure 4.5.1 – “Flowchart of the processes that goes through a tweet”

4.5.1. Pre-Processing

A cleaning process of data follows by removing as much noise as possible. Comprising an additional screening, the pre-processing of data includes a series of different methods in order to increase the number of accurate matches. For this purpose it was attempted to manually exclude fake accounts or Twitter accounts that belong to news broadcasting websites.

4.5.2. Tokenization

Every block of text usually has two or three sentences at most. Each sentence is split into a number of tokens and every token is a separate word and is divided by space. Tokenization is part of the final algorithm for classifying the tweets.

4.5.3. Data Cleaning

Before data are actually processed, tweets go through a cleaning process among steps taken to increase accuracy. Simple stemming algorithm was used to removing the suffixes “ing” and “ed”. Following, the “Cleaning_Data” function includes removing any non-alphanumeric characters from the word. In addition, letters repeated more than once at the ends of a word were removed.

4.5.4. Classification

Data, as described before in the design chapter, will be classified in three categories; Positive, Negative and Neutral, describing the opinion or emoticon the writer has about the TV broadcasting show. Among the different algorithms that were tried, it was concluded that using the “Bag of Words” algorithm is the most suitable one for the objectives of this project.

Each feature is described below and forms a sum of points, essentially creating a sum of negative and positive points. At the end of the processing of tweets, each tweet has a score which is classified as positive when it is above zero or classified as negative if it is below zero.

4.5.5. Dictionaries

Two dictionaries were constructed using one list of positive and one list of negative words. The lists were manually constructed by merging adjective lists showing emoticons. Three categories of pleasant feelings and four categories with adjectives describing unpleasant feelings such as “Sadness, Anger, Depression, and Confusion” were employed as ‘emotion’ dictionaries. Training of the algorithm aided to track down anomalies whenever words were removed or added to increase accuracy.

Each dictionary of words is loaded in the program as a HashMap data structure. The lists of words are stored in a text file and weighted with scores ranging -3 and 3 respectively. Text files (.txt) can be easily read and modified by almost all languages and software. Therefore it was considered as a primary solution. Figure B3 in Appendix B demonstrates the code used to load the dictionaries into HashMaps.

4.5.6. Slang Dictionaries

Among the features used to increase accuracy were urban dictionary InternetSlang.com, the current top 50 trending terms and other slang dictionaries such as NetLingo, all found available online to transform abbreviations into part of sentences. It was very important to use multiple sources of words and phrases that constantly develop or are uncommon in official writing, since posting on Twitter accommodates all types of backgrounds of people. For example in the formal English language, “brb” is not a proper word but in the cyber space is translated to and is widely recognizable as “be right back”. A part of the list is shown in Appendix C, Figure C2.

4.5.7. Emoticons

Emoticons are often seen as strong indicators about the content of the tweet. Text containing emoticons that may reflect both negative and positive emoticons is classified as neutral. Based on various sources, including the sentiment analysis and various sources found on the internet, a table with all the possible emoticons is constructed and a weight value associate with them. Table 4.5.7.1 shows part of this list of emoticons.

No.	Emoticon	Weight
1.	%-(-1
2.	%-)	1
3.	(-:	1
4.	(:	1
5.	(^ ^)	1
6.	(^~^)	1
7.	(^.^)	1
8.	(^_ ^)	1
9.	(o:	1
10.):-:	-1

Table 4.5.7.1 – “Part of Emoticon List”

4.5.8. Exclamation Mark

Exclamation mark in a natural text is a clear sign of emphasis. At this point, detection and evaluation marks were also adopted in the algorithm. One exclamation gives one point to the word before the mark. More than one exclamation mark attributes two points to the pre-successor word.

4.5.9. Capitalization

Detection of words which only consisted of capital letters was interpreted in assigning that word with extra points. Assuming that in Cyberspace capital letters are equal to “shouting”, it was decided to

include a function that checks each word individually for capital letters. If a word only consisted of capitals, then 1 point extra is given to the word.

4.5.10. Word Classification

Finally the classification of tweets is done by comparing each word as a key in the HashMaps generated with all the dictionaries mentioned above. If no match is found it gets zero score. Otherwise a negative or positive score will associate with the word and taking into account any of the features, extra weight will be assigned as well.

5. Analysis & Results

After carrying out the processes of gathering and filtering tweets, as explained in the previous chapters, the next step is to analyse the data in order to draw meaningful results that may shed light on the relationships between the tweets and TV viewings. The analysis is split in to investigating the qualitative relationship between tweet flow and TV shows, and the ratio between positive/negative tweets.

5.1. Qualitative relationship between TV ratings and Twitter

A first impression of the variability of the harvested tweets is demonstrated in Figure 5.1.1. All tweets collected during the week 02/04/2013-/9/04/2013 are plotted against a 24-hour time axis, as an attempt to compare the time and volume of tweet flow with the time of broadcast of the TV shows. Even though initially there is no usage of any filters to isolate tweets related with TV, Figure 5.1.1 demonstrates that there is a significant increase in Twitter traffic during the evening TV prime time period, i.e. 19:00-23:00, which justifies the investigation of the relationship in place.

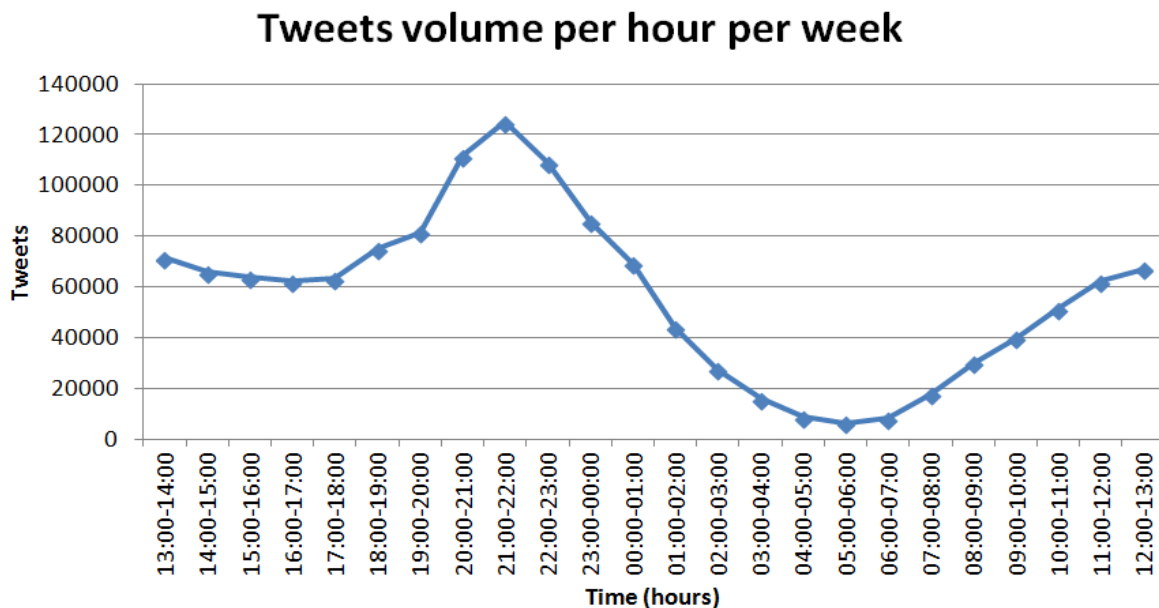


Figure 5.1.1 – “Total number of tweets against time over week 1.”

The peaks of the total number of tweets do not prove any correlation between Twitter and TV viewings; they are merely indicative of increased posting on Twitter during the time of broadcast of highly viewed TV programmes. Therefore, a more specific analysis is carried out that involves the full utilisation of tweetTV. Once isolating tweets that refer to the three TV shows during two different weeks (dataset is specified in Chapter 3), several comparisons are made between the TV ratings taken from BARB and tweets.

Figure 5.1.2 shows the number of tweets posted on the day of broadcast and the number of viewers plotted for every programme during week 1. There are some similarities and several inconsistencies. There is an overall similarity between the two sets of columns, i.e. a decrease/increase of the Tweets’

column height coincides with a similar change in the height of the Viewers' columns. In other words, fewer tweets are interpreted by a smaller number of viewers. This observation is valid, however, only on a qualitative point of view, since by looking at the absolute numbers, there are several quantitative differences.

Weekly Tweets vs. TV ratings

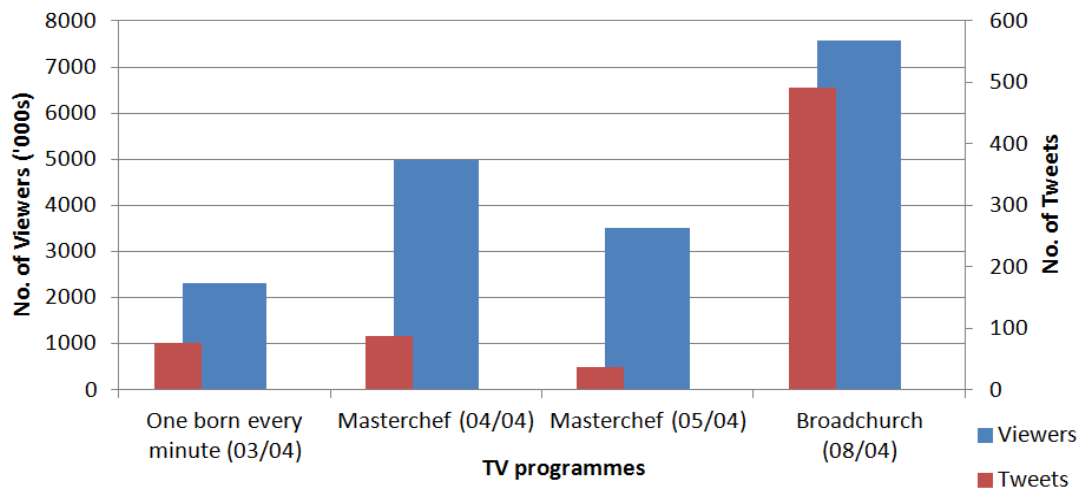


Figure 5.1.2 – “Total No. of tweets per show on the day of broadcast during week 1.”

‘One born every minute’ has the lowest TV rating but more tweets were posted about it when compared to *Masterchef* broadcast on the 5th of April. Furthermore the relative differences between tweets and TV ratings for every show are not proportional.

Weekly Tweets vs. TV ratings

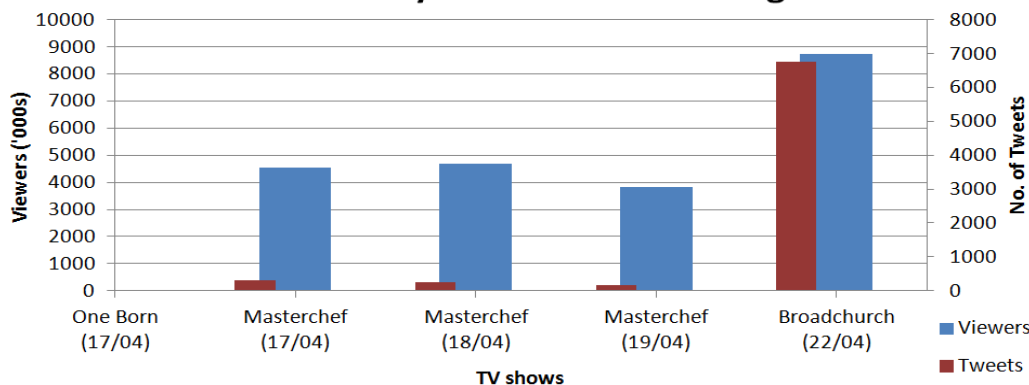


Figure 5.1.3 – “Total No. of tweets and viewers per show on the day of broadcast during week 2.”

Figure 5.1.3 demonstrates tweets posted on the day of broadcast and corresponding TV ratings for week 2. It is notable that during week 2, a particularly larger number of TV-related tweets was collected, especially for the shows *Broadchurch* and *Masterchef*. Another feature of Figure 5.1.3 is the extremely low number of Tweets identified with content related to *One born every minute*. As far as the qualitative comparison between the numbers of tweets and TV ratings, again there are certain similarities and differences. *Broadchurch* gathers the highest number of tweets and the highest number of viewers, showing a consistency between week 1 and week 2. Another feature observed on both weeks is the lower number of *Masterchef*'s tweets posted on the Friday show, compared to the Thursday show,

which is also consistent with fewer viewers watching the Friday show than the Thursday show.

Nevertheless, the most important characteristic that is also identified in week 2 is the similar pattern between TV ratings and tweets. Again, when comparing two shows, a smaller number of viewers is also reflected by fewer tweets, thus indicating that there is a link between the number of viewers and the number of tweets posted on Twitter. In Figure 5.1.4, an isolation of *Masterchef* tweets posted during both weeks compared with the number of viewers seems to confirm the validity of a relationship.

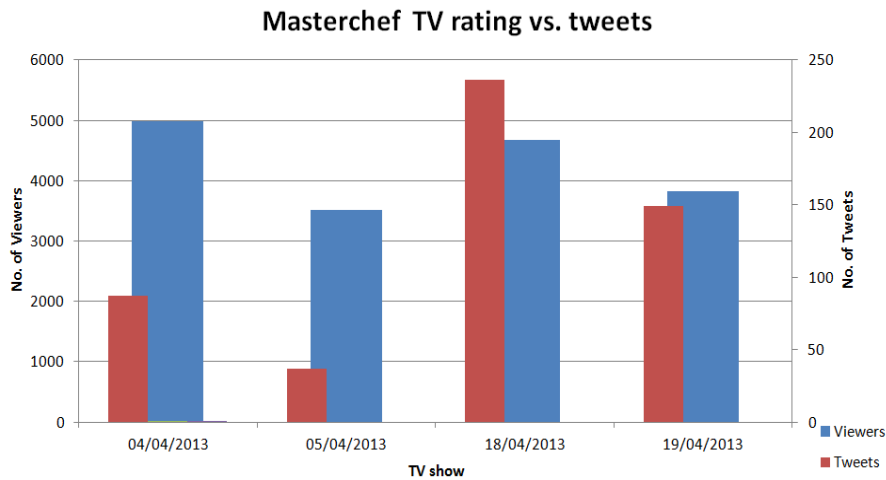


Figure 5.1.4 – “Total No. of tweets and viewers for Masterchef on the day of broadcast on weeks 1 & 2.”

5.2. Correlation between TV ratings and Tweets

The various graphical qualitative comparisons between the number of viewers and tweets in section 5.1 give extensive evidence of a potential correlation. The Pearson Correlation Coefficient was chosen to investigate how strongly these quantities are related. In order to enable calculating the Correlation Coefficient, the total number of tweets posted during the day of broadcast of every show is plotted against the corresponding TV rating, due to the absence of continuous daily TV ratings.

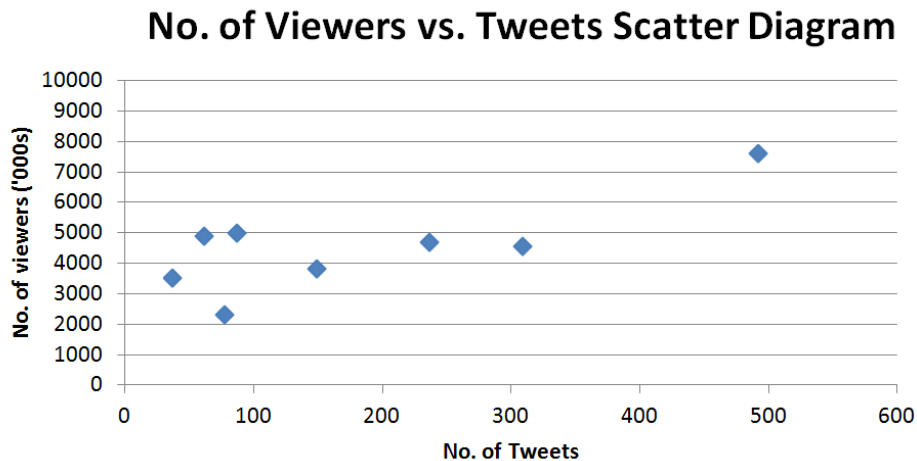


Figure 5.2.1 – “Scatter Diagram and positive correlation between Viewers and tweets”

Figure 5.2.1 illustrates the scatter diagram between TV ratings (Y-axis) and tweets (X-axis) which

correspond to all shows during both weeks. The distribution pattern suggests a high positive correlation which is verified by a Pearson Correlation Coefficient of $r=0.74$. This result quantitatively confirms the initial speculation of a strong relationship between Twitter and TV viewings; however the nature of Twitter implies several considerations to be highlighted. Having in mind that Twitter is based on trends and current news, it is very probable that extreme values become more difficult to relate to an established model. For example, the first or the last episode of a TV show is expected to gather a larger number of viewers - and subsequently tweets - than an episode in the middle of the season. However, *One Born Every Minute* is a good example of a popular TV show that failed to reach the numbers of *Masterchef's* and *Broadchurch's* tweets; therefore an estimation of its TV rating would be dramatically inaccurate. Nevertheless, shows such as *Masterchef* and *Broadchurch* that are widely viewed by a large spectrum of people and classes appear to generate more tweets which better correlate with TV ratings.

5.3. Sentiment analysis Testing

The second part of the analysis consists of the recognition of “mood” by “tweetSent” and the comparison of the ratio of positive/negative tweets against the TV ratings. However, the classification of tweets as positive, negative or neutral is first checked through a testing process which measures the accuracy of “tweetSent” against human and thus realistic judgement.

The evaluation of “tweetSent” was carried out by comparing the classification generated by tweetSent to a Human classification. Two different random samples of tweets were scored by three different people and then averaged to one value, in order to reduce human subjectivity. A value was assigned to each tweet using a scale from -5 to 5, with zero representing neutral/informational tweets. In order to have an idea of the accuracy of a sentiment analysis tool, the score obtained by the application “SentiStrength” is also provided. The first sample of 120 tweets was randomly collected from the week 1 pool, with tweets not being filtered by “tweetFilter”. This decision ensured that the collected tweets did not have any common patterns in content or time of post. The comparison between the calculated average of the Human classification, “tweetSent” and SentiStrength is summarised in Table 5.3.1.

Table 5.3.1 shows the variability between the two sentiment analysis tools and the Human classification. tweetSent achieves a good accuracy in capturing neutral/informational tweets; however, there is an underestimation of both positive and negative tweets.

classifier	Positive	Negative	Neutral
tweetSent	25	13	40
Human Classification	46	30	44
SentiStrength	30	14	31

Table 5.3.1 – “Initial evaluation of tweetSent” algorithm

The second sample consisted of 324 tweets which were randomly selected from the filtered week 2 pool. In other words, all tweets contained content relevant to the keywords detected by the algorithm. The second testing sample was chosen to assess the efficiency of tweetSent on the TV-related tweets. Again, all tweets were rated by three different people and then compared to tweetSent's generated classification (see Table 5.3.2).

classifier	Positive	Negative	Neutral
tweetSent	120	17	187
Human Classification	166	42	116

Figure 5.3.2 – “Second evaluation of tweetSent algorithm”

The percentages given in Table 5.3.3 reflect the tweetSent's accuracy for each category of tweets. Having an understanding of the accuracy achieved by tweetSent is particularly important when interpreting the moods that dominate during the broadcast of the TV shows in interest. The percentages obtained indicate some significant differences between the initial and the second evaluation. The second evaluation that was carried out on TV-specific tweets gives an accuracy of 72.73% which is substantially higher than the 54.33% indicated by the initial evaluation. The relatively small difference between the accuracies acquired for negative tweets implies that there is a failure of identifying a large proportion of them. As far as the neutral/informational tweets are concerned, the high 90.91% of the initial evaluation shows a successful capture of these tweets whereas the overestimation of the second evaluation confirms a general weakness of such semantic analysis tools to distinguish idioms and phrases that may imply negative thoughts or feelings.

classifier	Positive	Negative	Neutral
Initial evaluation	54.35%	43.33%	90.91%
Second evaluation	72.73%	39.53%	161.21%

Figure 5.3.3 – “Accuracy of tweetSent”

An example of sentiment analysis failure is given below. The system limits as natural processing is extremely hard to predict.

“What’s this ‘fifth gear’ ??? Fake version of top gear”

5.4. Relationship between ‘mood’ of tweets and TV ratings

Making the most of the tweetSent's ability to classify tweets as positive, negative and neutral ones, the following analysis is attempting to reveal how the content of tweets may be used to reflect the TV ratings.

All tweets that refer to Masterchef were filtered through tweetSent and were plotted against their time of post (Figure 5.4.1). The shows that were broadcasted on the 18th and 19th of April were chosen to compare the temporal variability of positive and negative tweets during the time interval 18:00-00:00. The reason for choosing several hours before and after the time of broadcast (20:00) was to observe the pattern of changes in number of volumes of positive and negative tweets and whether that could be used to predict TV ratings.

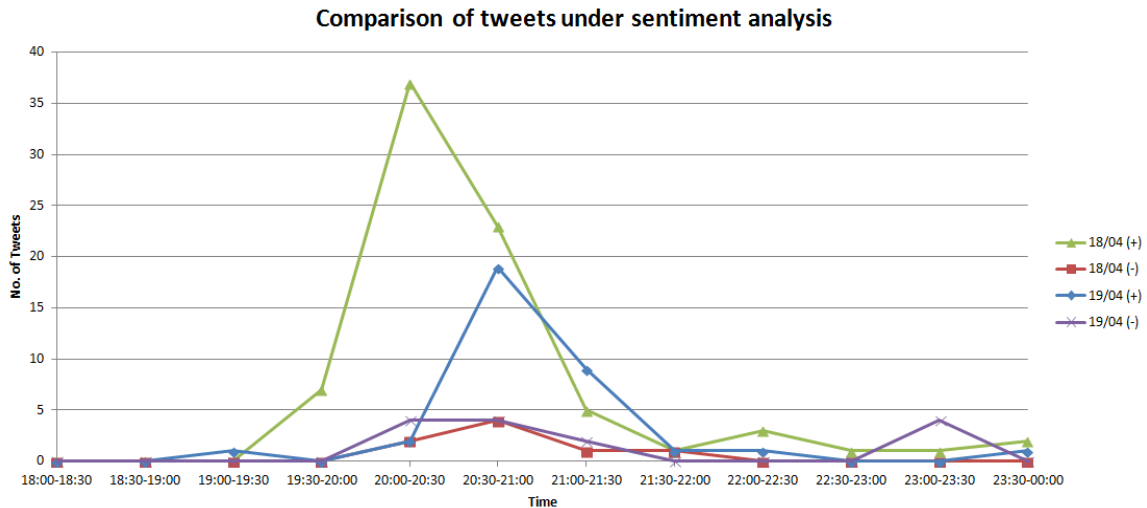


Figure 5.4.1 – Masterchef: Total number of positive and negative tweets on the 18th and 19th of April from 18:00-00:00.

Figure 5.4.1 shows a significantly increased number of positive tweets on the 18th of April which concentrated a total of 4,680,000 viewers, compared to 4,540,000 viewers that watched the show on the 19th (that started at 20:30). The lower number of viewers on the 19th also coincides with a lower value of positive tweets at the time of broadcast, which again confirms the existence of a relationship between TV ratings and Twitter. Another important feature of Figure 5.4.1 is the flow of both positive and negative tweets before and after the broadcast of the show. There is a distinct increase in positive tweets during the hour before and the hour after Masterchef is broadcasted on both days, indicating that perhaps those times could be potentially considered in a projection of the actual TV ratings. Figure 5.4.2 also indicates the existence of a periodicity that is observed in both weeks, during the days before the show, thus suggesting that the period used to project TV ratings could be extended to several days.

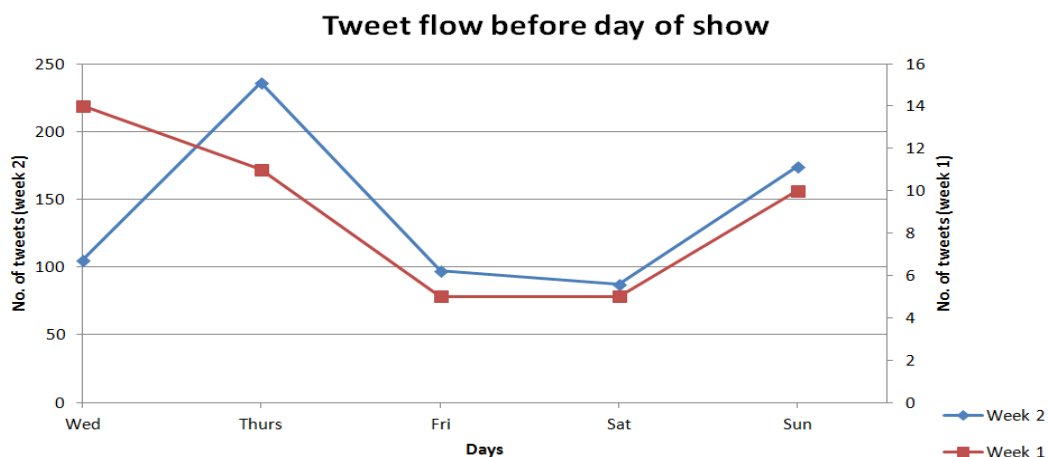


Figure 5.4.2 – Broadchurch: Tweet posting during a period of 5 days before the show

6. Discussion & Conclusions

6.1. Conclusions

Data mining of Twitter users' messages is something relatively new in the field of Computer Science; nevertheless it appears to be very promising. As explained in the Background Chapter, many researchers managed to use Twitter in order to predict events or estimate accurately the outcome of others. This project's main objective was to investigate an area of Twitter that has been remaining unexplored and ultimately discover potential areas of development by taking research one step forward. By investigating the relationship between Twitter users and TV viewers, some conclusions were drawn based on the qualitative trends that were identified through the analysis of harvested tweets and the observation of the corresponding TV viewing rates.

By implementing the suggested system, namely tweetTV, several difficulties and limitations were also identified, the overcoming of which is regarded as crucial in order for such a principle to be adopted and applied in the commercial industry in the future.

6.2. Correlation between Twitter traffic and viewers

It was expected that from the very beginning of this project that while monitoring the tweet traffic, a distinct peak should occur while the show was played on TV. This was verified by all plots of the number of tweets against time, meaning that the assumption that people are tweeting using their smartphones, PCs or other devices while watching the show is valid. The most important conclusion, however, is that the temporal variability of tweets significantly coincides with some of the highly viewed TV shows and this clearly indicates that a correlation between Twitter and TV viewing is there.

On the other hand, no results were expected to be as precise as the conventional ways of obtaining TV ratings are. Evaluation showed that quantitative methods of analysis that produce more accurate results require a larger training set. Furthermore, the unavailability of a continuous TV rating for every programme made impossible to quantify how closely related are tweets and TV ratings. In addition, different analysis method showed that more specific analysis on the content of the text messages has minor importance and the actual existence of the tweet shows the interest of the user in the specific subject. In this way the assumption that a user commenting about the show is watching the show was actually enforced and verified. On this statement the whole project was based and justifies a small gap for error.

6.3. Semantic Analysis

Semantic analysis is a challenging domain of natural language processing. One of the main aims of the project was to prove a correlation between the human behaviour – in terms of watching a particular TV show - and the text messages posted on the social network Twitter. Again, the analysis of tweets' mood

as classified by tweetSent showed that a positive flow of tweets was translated into a higher viewing of a specific show. However, the most important conclusion is the existence of tweets that refer to a show that are posted on different hours or even days than the broadcast day. The attitude of those tweets can give an indication of current viewing but also it can pave the way to predicting viewings on a future dates by measuring current posts through a semantic analysis approach. Despite the difficulties in language, partly, the analysis proved a relationship between the emoticons and famous TV shows broadcasted on television.

A large amount of the tweet messages collected, approximately 45%, was neither positive nor negative; instead they were purely informational, usually discussions on the plot of the television show. Therefore the cyber chatting on Twitter tends to outline the opinion of the users about the show indicating that they are already watching the program but without actually emotionally expressing their feelings.

6.4. Trending Phenomena

It was noticeable that whenever a term was trending throughout the day in Twitter, it was very likely that the number of mentions referring to that issue would dramatically boost. This type of increase in tweets would cause a bias during the process of harvesting data, since only a percentage of the available tweets is collected. Rebroadcasting or replying to Twitter messages on the same topic increases the chance of trending. The graph shown in Figure 6.4.1 represents how usually a topic trends on Twitter [24], following a specific pattern of people talking why a topic is so famous triggering a chain of events.

Twitter applies a threshold to all streaming applications including tweetTV by limiting the number of tweets to 1% of all Twitter traffic. Full access to the Twitter API, also known as ‘firehorse’ service, is restricted only to a few elite companies.

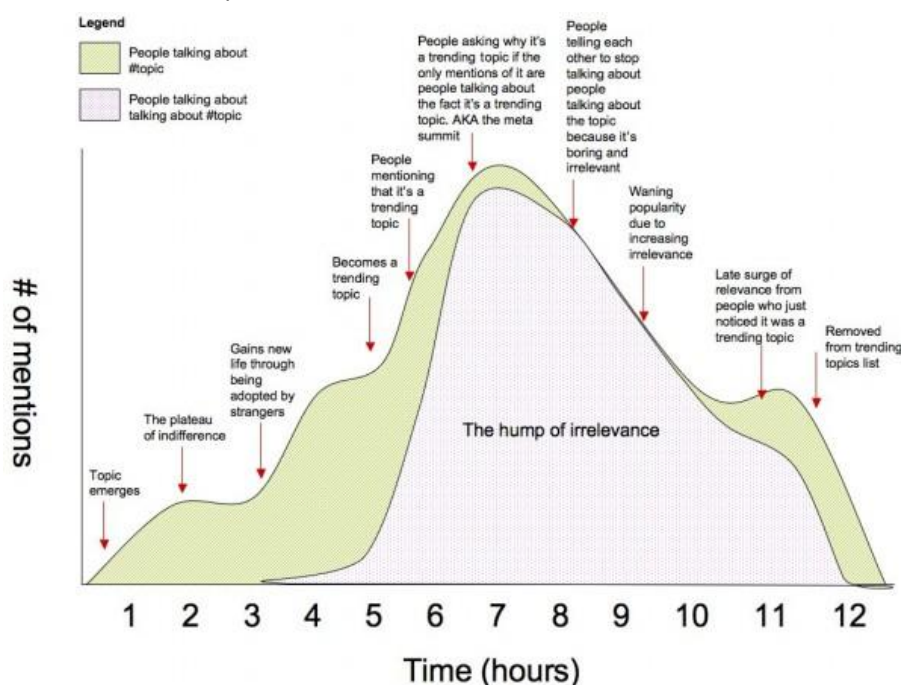


Figure 6.4.1 – “How a topic becomes popular in social networks” [24]

Figure 6.4.2 shows the curves of the positive, negative and neutral flows of tweets on the 8th of April when Broadchurch was played on TV. In all three curves there are signs of trending, which starts early in the afternoon as the topic emerges and eventually reaches its peak when the show is broadcasted at 21:00pm. The gradual increase and decrease of all curves and thus number of tweets that express all sorts of opinions match with a trend of people to initially become interested in the show and then ultimately watch it. This is a very critical conclusion, as controlled trends could potentially increase the ratings of a show, in a case where Twitter unconsciously pushes the user to TV due to a temporarily popular matter discussed in the Cyberspace.

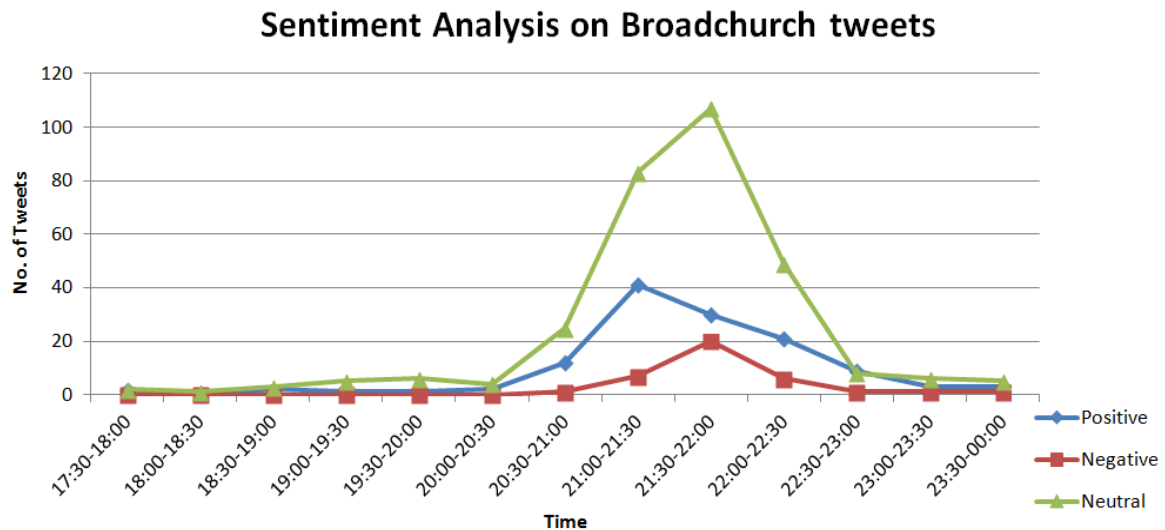


Figure 6.4.2 – “Positive, negative and neutral tweets prior, during and after the broadcast of BroadChurch”

6.5. Failure of detection

An unexpected feature of the performance of tweetTV was the failure to collect a sufficient amount of tweets for certain shows, such as “One born every minute”. A show which according to the official stats from BARB has a viewing rating of more than 2 million viewers, almost failed to be traced from the system at all. An explanation for this phenomenon may be because the age of viewers of the specific show bounces off the Twitter demographic, since older people or people below or above the range of 20-35 do not have access to Twitter. Therefore, potentially not all TV shows can they be measured using Twitter, since there are programs that specifically target these ages.

On the other hand, shows such as Top Gear was always among discussions and tweets as one of the favourites, although no new episodes were broadcasted on British channels. TV programs that are more appealing to a younger audience can be easily traced on the social networks.

6.6. Technical Issues

6.6.1. Data Management and Storage

Week 1 data was gathered by only being based on the geolocation tag attached to the tweets, having around 1.8 million tweets being harvested off the network. During the process of collecting tweets, it was discovered that this amount of data required an efficient strategy of handling them. Directory based storage was not the most efficient way to store and retrieve data. Although it was the best way to enable any read-and-write actions by statistical software to generate the figures displayed in the analysis chapter.

The amount of tweets would commence an SQL database to store and retrieve, but setting and maintaining a database server online during the project would be extremely time-consuming and some more complex IT difficulties could occur. Therefore, compared to other possible solutions it was considered to generate one text file for each file harvested, always making the most of the advantage of the cross-platform ability of text files to be read and written relatively easily

6.6.2. Twitter Demographic

As social media is an invention of the last few years, not everyone in the United Kingdom has created a Twitter account. Furthermore, it should not be assumed that those who are subscribed to Twitter always post their personal opinions about their favourite television shows [25]. Therefore Twitter users may not fully represent the TV audience. In contrast users of the site vary from company accounts, celebrities and politicians [26]. In terms of the system itself it should also be mentioned that Twitter allows only the open accounts to be harvested; tweets from locked accounts cannot be collected. So a large proportion of tweets belonging to inaccessible accounts was inevitably left out of the harvested pool.

Fake accounts and bots are gradually increasing in the Twitter world. Tweets were blacklisted in an effort to limit invalid tweets or accounts that were not thought to be eligible, due to re-tweeting meaningless or identical tweets. This type of account creates fake traffic by re-tweeting or posting automated messages through multiple tweets in order to increase the tweet traffic of specific accounts. It is estimated that there are more than 20 million fake accounts. [27]

6.6.3. Language limitation of sentiment analysis

One of the most critical limitations encountered during the sentiment analysis was the inability of TwitterSent to recognise complex expressions or oddly structured sentences. People usually do not use a standard language or language that strictly follows grammatical rules. Slung language and the metaphoric meaning of the words create complex sentences, as new terms show up every day in the cyberspace, where a machine learning algorithm could be a false negative or false positive [13]. More

importantly, slang language, incorrect grammar and idioms are the type of language that is often used, reducing the accuracy of a sentiment analysis tool. Sentiment analysis is hard to be applied on text messages due to the default inability of computers to understand feelings in particular [28].

6.6.4. Other Limitations

The constructed system will not be able to generate “overnight” results for all the TV channels due to the volume of the data and parameters needed to be under consideration. Instead, the system will produce ratings for specific TV programs at the end of each week. This is mainly due to the fact that the official metrics given by the Broadcasters' Audience Research Board (BARB) authority do not allow daily measurements. Data from BARB provide information such as Channel, Average Daily Reach, Average Weekly Viewing, Share and are given for each week of the month [29].

6.7. Further Work

With Twitter being significantly unexplored, it was shown through this study that there are a lot of opportunities to examine its aspects and establish innovative ways to measure TV ratings. The task is difficult due to the involvement of many complex and unpredictable factors. Nevertheless, the idea of combining sentiment analysis with tweet harvesting seems to be promising in terms of understanding the behaviour of trends and eventually become able to predict them. In terms of the algorithm used, the accuracy was rather high compared to the academic project “sentiStrength” which used the same techniques.

6.7.1. Algorithm Improvements

In the future different methods could be attempted to improve the accuracy of tweetSent by altering the basic algorithm, bag of words. Although the main purpose of this project was not to focus only on the development of opinion mining algorithm, the tweetSent has managed to hit high levels of accuracy.

A certain way to improve results could be to program the algorithm to be self-trained through in real time. As the amount of data in internet is gradually increasing, training new terms and words will increase the probability for new words to be added on the dictionary and will most definitely increase the amount of correct classifying. Self training algorithm with some manual verification should produce better results.

On a more general point of view, collecting more data for a longer period of time will produce more interesting results. The 3 weeks period may not actually be enough to make definite conclusions; however it is sufficient to capture the indicative similarities and relationships between tweets and TV ratings. A greater volume of data is required which subsequently implies more processing time and a more sustainable infrastructure which is not available at the time due to limited resources.

6.7.2. Web Application

Taking tweetTV forward, the next step should be the creation of a web application that will track and generate plots of the Twitter traffic related to the television programs. The web application could provide sophisticated statistics and the ability to give emoticon strength of the total tweet collected for the UK television shows in real-time. Streaming Twitter statues from the internet and producing the popularity comparison on the webpage.

Additional feature should be considered the option for the site visitors to help training the algorithm by stating the emoticon strength of the given tweet messages. Therefore trained data will automatically generated from the website visitors will improve the accuracy of the system.

7. References

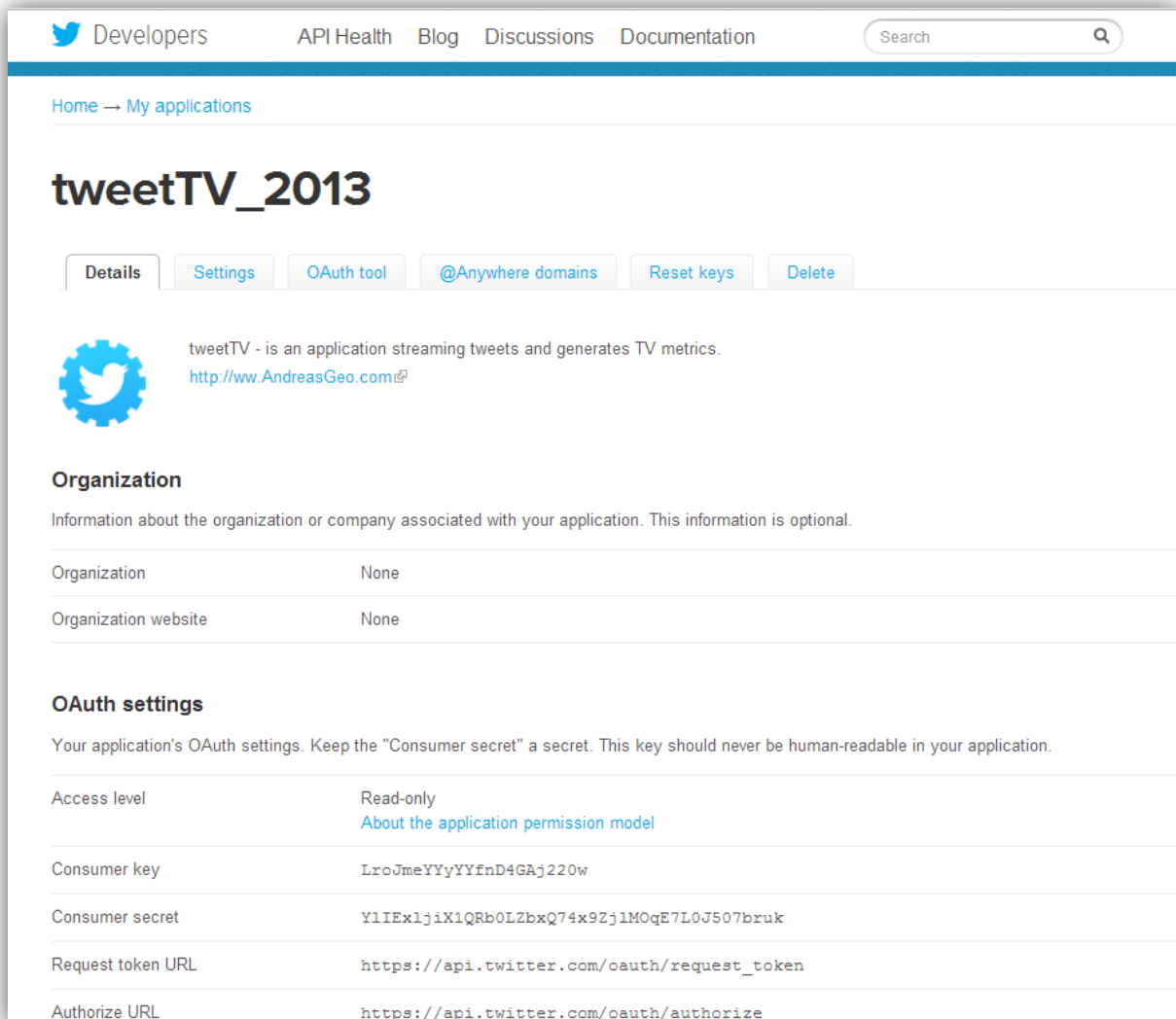
- [1] Stinson, M., 2013. *Search Engine Journal*. [Online]
Available at: <http://www.searchenginejournal.com/google-plus-surpasses-twitter-to-become-second-largest-social-network/57740/>
[Accessed 5 February 2013].
- [2] Arthur, C., 2013. *The Guardian UK*. [Online]
Available at: Boot up: Chromebook impressions, Anonymous and Swartz, Twitter and TV, and more
[Accessed 17 January 2013].
- [3] Hong, Jongpil; Leckenby, John D., 1996. Audience Measurement and Media Reach/Frequency Issues in Internet Advertising. Texas, Department of Advertising College of Communication, University of Texas.
- [4] Starkey, G., 2004. Estimating audiences: sampling in television and radio audience research. *Cultural Trends*, 1(13), pp. 3-25.
- [5] Jongpil, H. & Leckenby, J. D., 1996. *Audience measurement and media reach/frequency issues in Internet advertising*. Vancouver B.C, Proceedings of the 1996 American Academy of Advertising.
- [6] Napoli, Philip M., 2008. *Audience Economics, the Diversity Principle, and the Local People Meter, Communication Law and Policy*, Bronx, NY: Donald McGannon Communication Research Center.
- [7] BARB, 2013. *Broadcasters Audience Research Board*. [Online]
Available at: http://www.barb.co.uk/resources/reference-documents/how-we-do-what-we-do?_s=4
[Accessed 17 January 2013].
- [8] UK, Nielsen, 2013. *Measurement*. [Online]
Available at: <http://www.nielsen.com/uk/en/measurement.html>
[Accessed 18 January 2013].
- [9] RTE media sales, 2012. *Raidió Teilifís Éireann*. [Online]
Available
at: http://www.rte.ie/mediasales/television/Content/PDF's/TV%20Audience%20Measurement%20-%20Guide_2.pdf
[Accessed 20 January 2013].
- [10] Nisbet, R., Elder, J. & Miner, G., 2009. *Handbook of Statistical Analysis*. 1st ed. San Diego: Academic Press.
- [11] Gräbner, D., Zanker, M., Fliedl, G. & Fuchs, M., 2012. *Classification of Customer Reviews based on Sentiment Analysis*. Helsingborg, Springer.
- [12] Zambonini, D., 2010. *Self-Improving Bayesian Sentiment Analysis for Twitter*. [Online]
Available at: <http://danzambonini.com/self-improving-bayesian-sentiment-analysis-for-twitter/>
[Accessed 25 January 2013].
- [13] Thelwall, M., Buckley, K., Paltoglou, G. C. & Kappas, A., 2010. Sentiment Strength detection in short informal text. *Journal of the American Society for information Science and Technology*, 12(61), pp. 2544-2558.
- [14] Su, F. & Markert, K., 2008. *From Words to Senses: A Case Study of Subjectivity Recognition*. Manchester, Proceedings of the 22nd International Conference on Computational Linguistics.

- [15] Liu, B., 2012. *Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies)*. 1st ed. California: Morgan & Claypool.
- [16] Chapman, P., 1999. *The CRISP-DM User Guide*. [Online]
Available at: <http://lyle.smu.edu/~mhd/8331f03/crisp.pdf>
[Accessed 29 January 2013].
- [17] Panzarino, M., 2012. *The Nielsen Twitter TV Rating just made Twitter the first purely social ranking of US TV program popularity*. [Online]
Available at: <http://thenextweb.com/twitter/2012/12/17/twitter-becomes-tv-ratings-system-nielsen/>
[Accessed 27 January 2013].
- [18] UzZaman, N., Blanco, R. & Matthews, M., 2012. *TwitterPaul: Extracting and Aggregating Twitter Predictions*. [Online]
Available at: <http://arxiv.org/pdf/1211.6496v2.pdf>
[Accessed 23 January 2013].
- [19] Lamos, V., De Bie, T. & Cristianini, N., 2010. Flu Detector - Tracking Epidemics on Twitter. *Lecture Notes in Computer Science*, Volume 6323, pp. 599-602.
- [20] Zubiaga, A., Spina, D., Amigó, E. & Gonzalo, J., 2012. *Towards Real-Time Summarization of Scheduled Events from Twitter Streams*. New York, Proceedings of the 23rd ACM conference on Hypertext and Social Media.
- [21] Ogneva, M., 2010. *How Companies Can Use Sentiment Analysis to Improve Their Business*. [Online]
Available at: <http://mashable.com/2010/04/19/sentiment-analysis/>
[Accessed 29 January 2013].
- [22] Twitter4J, 2010. *Twitter4J*. [Online]
Available at: <http://twitter4j.org/en/>
[Accessed 5 February 2013].
- [23] hueniverse, 2012. *OAuth 1.0*. [Online]
Available at: <http://hueniverse.com/oauth/>
[Accessed 6 February 2013].
- [24] Doctorow, C., 2009. *Graph of how #topics get played out on Twitter*. [Online]
Available at: <http://boingboing.net/2009/05/16/graph-of-how-topics.html>
[Accessed 3 March 2013].
- [25] Java, A., Song, X., Finin, T. & Tseng, B. 2007. Why We Twitter: Understanding Microblogging. *Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007*, 12 10, pp. 56-6
- [26] Pak, A. & Paroubek, P. 2010. *Mining, Twitter as a Corpus for Sentiment Analysis and Opinion*. Paris, European Language Resources Association (ELRA).
- [27] Pelroth, N., 2013. *Fake Twitter Followers Become Multimillion-Dollar Business*. [Online]
Available at: <http://bits.blogs.nytimes.com/2013/04/05/fake-twitter-followers-becomes-multimillion-dollar-business/?ref=technology>
[Accessed 10 April 2013].
- [28] Wright, A., 2009. *Mining the Web for Feelings, Not Facts*. [Online] Available
at: <http://www.nytimes.com/2009/08/24/technology/internet/24emotion.html> [Accessed 15 January 2015].

[29] BARB., 2013. Top 10 Programmes - BARB. [Online] Available at: <http://www.barb.co.uk/viewing/weekly-top-10>? [Accessed 15 January 2013].

8. Appendices

8.1. Appendix A – Generating API Keys



The screenshot shows the Twitter Developers 'My applications' page for an application named 'tweetTV_2013'. The page has a navigation bar with links to 'Developers', 'API Health', 'Blog', 'Discussions', and 'Documentation', along with a search bar. Below the navigation bar, there's a breadcrumb 'Home → My applications'. The application name 'tweetTV_2013' is prominently displayed. Below the name are several tabs: 'Details' (selected), 'Settings', 'OAuth tool', '@Anywhere domains', 'Reset keys', and 'Delete'. The 'Details' tab shows a Twitter logo icon, a description 'tweetTV - is an application streaming tweets and generates TV metrics.', and a website link 'http://www.AndreasGeo.com'. Below this is the 'Organization' section, which states that information is optional and shows 'Organization' and 'Organization website' both set to 'None'. The 'OAuth settings' section follows, explaining that the 'Consumer secret' should be kept secret. It lists several settings: 'Access level' (Read-only), 'Consumer key' (LroJmeYYyYYfnD4GAj220w), 'Consumer secret' (Y1IExljiX1QRb0LZbxQ74x9Zj1MOqE7L0J507bruk), 'Request token URL' (https://api.twitter.com/oauth/request_token), and 'Authorize URL' (https://api.twitter.com/oauth/authorize).

8.2. Appendix B – Implementation code

```
public class tweetCollector2 {

    public static void main(String[] args) {

        ConfigurationBuilder config = new ConfigurationBuilder();
        config.setDebugEnabled(true);
        config.setOAuthConsumerKey("osUTHffmCm8cspVZc5ihg");
        config.setOAuthConsumerSecret("exKuPFTG7IJnVMcdyTPjjIAiDrzjV22abgbUGMKA");
        config.setOAuthAccessToken("221385138-wm1UCOKcncLdbCkkWHNgJov9wFUM9X0P0PKu8Ia7");
        config.setOAuthAccessTokenSecret("QevoD6ovtV0JLiPA5MpM9lZRJ2ElAyPZKd3jkTZow");

        TwitterStream twitterStream = new TwitterStreamFactory(config.build()).getInstance();
```

Figure B1 – “tweetCollector – API keys”

```

169     double[][] City_Locations = {
170         // London City
171         {-0.6069, 51.2459}, {0.461, 51.737},
172         // Birmingham City
173         {-2.017434, 52.385999}, {-1.709829, 52.568876},
174         // Manchester City
175         {-2.300097, 53.399903}, {-2.147087, 53.544588},
176         // Liverpool City
177         {-3.008748, 53.326744}, {-2.811504, 53.503907},
178         // Newcastle City
179         {-1.781082, 54.95944}, {-1.532605, 55.045304},
180         // Nottingham City
181         {-1.24829, 52.889}, {-1.091834, 53.019045},
182         // Sheffield City
183         {-1.663959, 53.30455}, {-1.334953, 53.486883},
184         // Glasgow City
185         {-4.393201, 55.781279}, {-4.071717, 55.929641},
186         // Cardiff City
187         {-3.282382, 51.443575}, {-3.116437, 51.560906}
188     };

```

Figure B2 – “tweetFilter - Keyword matching”

```

20 public static void main(String[] args) throws IOException {
21
22
23     // List of Positive Words
24     HashMap<String, Integer> pos_list = new HashMap<String, Integer>();
25     pos_list = load_list("pos_words.txt");
26
27     // List of Negative Words
28     HashMap<String, Integer> neg_list = new HashMap<String, Integer>();
29     neg_list = load_list("neg_words.txt");
30
31     // Load List of Emoticons;
32     HashMap<String, Integer> emoticon_list = new HashMap<String, Integer>();
33     emoticon_list = load_list("emo_list.txt");
34
35     // Load List of Slung Dictionary
36     HashMap<String, String> slang_list = new HashMap<String, String>();
37     slang_list = load_abbreviations("slang_list.txt");
38
39     // Current Word

```

Figure B3 – “HahMaps Loading Dictionaries”

```

193 // Exclamation Marks - Extra Points for Emphasis
194 public static int exclamation_marks( String line){
195
196     int points = 0 ,count=0;
197     char chr;
198
199     for( int i = 0; i < line.length( ); i++ )
200     {
201         chr = line.charAt( i );
202
203         if( chr=='!' )
204             count++;
205     }
206     switch(count){
207     case 1: points = 1;
208             break;
209     case 2: points = 2;
210             break;
211     default:
212         if (count>2) points=2;
213         break;
214     }
215     return points ;
216 }
217

```

Figure B4 –“Exclamation marks function – Code Snippet”

```

{
    if (wordExists = pos_list.containsKey(currentWord.cleaned))
    {
        total = total + pos_list.get(currentWord.cleaned) + currentWord.extraWeight;
    }else if (wordExists = neg_list.containsKey(currentWord.cleaned))
    {
        total = total + neg_list.get(currentWord.cleaned) - currentWord.extraWeight;
    }else if (wordExists = emoticon_list.containsKey(currentWord.cleaned))
    {
        total = total + emoticon_list.get(currentWord.cleaned);
    }
}
}

```

Figure B5 – “Word classification and Points’ calculation”

8.3. Appendix C – “tweetFilter” results – “Analysis.txt”

```
// [ keywords_analysis.txt generate by tweet-Tv ]
// List of tweets containing the keywords
// topgear, top gear, bbc_topgear, top_gear

1. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1013933.txt 08/04/13 00:26:54
2. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1014988.txt 08/04/13 00:32:36
3. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1037171.txt 08/04/13 05:24:51
4. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1039491.txt 08/04/13 07:17:57
5. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1044537.txt 08/04/13 08:36:48
6. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1046245.txt 08/04/13 08:58:19
7. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1060268.txt 08/04/13 11:11:29
8. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1062053.txt 08/04/13 11:25:24
9. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1062639.txt 08/04/13 11:29:55
10. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1064016.txt 08/04/13 11:40:01
11. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1064091.txt 08/04/13 11:40:34
12. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1066623.txt 08/04/13 11:58:52
13. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1067471.txt 08/04/13 12:04:38
14. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1069048.txt 08/04/13 12:14:56
15. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1069419.txt 08/04/13 12:17:17
16. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1069496.txt 08/04/13 12:17:45
17. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1070275.txt 08/04/13 12:22:43
18. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1071269.txt 08/04/13 12:29:08
19. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1071370.txt 08/04/13 12:29:49
20. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1072280.txt 08/04/13 12:35:44
21. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1074900.txt 08/04/13 12:52:18
22. C://Users//Andreas//Desktop//tweetsExtr//zip1//tweet_1082953.txt 08/04/13 13:27:02
```

Figure C1 – “tweetFilter output”

```
lml "love my life"
sml ""story of my life"
ikr "i know, right"
lol "laughing out loud"
rofl "rolling on floor laughing"
imao "in my arrogant opinion"
hwp "height weight proportional"
mfw "my face when"
pfa "please find attached"
rachet "crazy"
eta "estimated time of arrival"
ditto "the same"
ty "thank you"
icymi "in case you missed it"
fao "for attention of"
gg "good game"
lms "like my status"
kmsl "killing myself laughing"
bff "best friends forever"
```

Figure C2 – “Part of the slang dictionary in slang_list.txt”