**Using binary classifier for predicting success of technology release based upon sentiment analysis of social media data**

Ali Akbar Jilani

Department of Science and Computing,   
Waterford Institute of Technology, Waterford, Co. Waterford  
Ireland (Tel: 353-085-212-6940; e-mail: 20078735@mail.wit.ie).

**Abstract:** Textual Domain offers a unique proposition for sentiment analysis. After the popularization of online social networks, social media were naturally picked by the industry to evaluate the sentiments that related to their market segments. Twitter represents one of the most popular social media platform where people share their opinions and sentiments using within the 140 characters space made available to them. Keeping in view the volume of content growing on Twitter on daily basis, it was imperative to device some automated tool that performs this sentiment score calculation for a define-able volume of content. A number of techniques have been developed for different aspects of sentiment analysis i.e. document level, sentence level etc. This study is targeted at development of another sentiment analysis tool that uses these available techniques to find their mutual correlation. Something that it calls “axis of honesty” and connects these axes to develop the overall success score of a specific technology product from within the fed Twitter corpus.

*Keywords:* Social Media, Binary Classification, Text Mining, Sentiment Analysis, technology release.

1. **Introduction:**

Check if LNCS is required for format of dissertation template..

As online social networks (OSNs) have enabled worldwide consumers to openly communicate their experiences, it has created opportunities for technical communicators, marketing and public relations writers and pretty much any company or individual that want to monitor their reputation or get timely feedback about their products and actions. Social media platforms including Twitter, Facebook, Message Boards, Blogs and user forums offer an explosion of user-generated content (UGC) that can be tapped to ad-hoc corpus building processes thus creating word lists relevant to specific organizational interests. This way, technical communicators as well as marketers can listen to their external users and accurately identify area of need. Twitter is one of the most popular medium among social media that has contributed to reshaping the web from a mere static repository to a dynamic forum (microblogging service) where users can publish their thoughts and opinions along with other types of “user-generated content (UGC)” on any topic of interest. This content carries valuable information particularly for applications that require analysis of public opinion on a certain topic.

While sentiment analysis technology doesn’t stop us from employing artificial intelligence in a program to measure opinion scores on a specific subject, this study focuses on the current need of a much simpler approach of understanding public opinion patterns (sentiment analysis) about a certain technology while maintaining the socio-technical context from which these patterns emerge. It utilizes Amazon online public reviews about technology to train a binary classifier model on success of technology release and the same model is later used to predict success for upcoming technology products.

(Feldman, 2013) Defines ‘Sentiment Analysis’ or ‘Opinion Mining’ as the task of finding opinions of authors about specific entities. He explains how there is a huge explosion of ‘sentiments’ available from social media including Twitter, Facebook, message boards, blogs, and user forums. This opinionated information is a gold mine for companies and individuals that want to monitor their reputation or get timely feedback about their products and actions, may they be about product release. Sentiment analysis offers these organizations the ability to monitor the different social media sites in real time and act accordingly. Marketing managers, campaign managers, politicians, equity investors or even online shoppers can directly benefit from this sentiment analysis technology. (Feldman, 2013).

**1.1 Twitter:**

Twitter is one of the most popular microblogging platform (launched in 2006) that can also be an amazing mine for text and social web analyses. As a rough estimate taken in 2016 by (Giachanou and Crestani, 2016), Twitter had 284 million users who posted 500 million messages per day. Characterized by the ease of access and download of posts published through the system, it was considered one of the largest datasets of user-generated content. Twitter is considered an informal mode of social media content.

**1.1.1 Tweet:**

A single message posted on Twitter is called a “Tweet”. Its content may at maximum stretch over 140 characters that can vary from personal information or opinion about products or events to other content types such as photos, videos, news or even links.

**1.2**

It is relevant to mention here that this research is likely to have a universal application. The main outcome of this research is supposed to be a binary indicator representing prediction about the success of technology product release. This could either be ‘Success’ or ‘Failure’. The overall sentiment score across the Twitter corpus can finally be presented to the user using some visualization tool or mere as WordCloud.

According to (Feldman, 2013), More than 7,000 articles have been written about sentiment analysis and various startups are developing tools and strategies to extract sentiments from text. The scope of this study is therefore being drifted more towards contribution to quality (as compared to variety) of results and to make it more manageable, it is also being confined only to *subjective sentences* (that contain opinions, beliefs and views) as opposed to *objective sentences* (that contain factual information). Subjective sentences carry the essence of sentimental information (opinions, beliefs and views) while objective sentences contain factual information that is more suitable to areas like stock picking. A Tweet may vary in the number of sentences it contains. These sentences may carry different opinions about the same entity. In order to develop an accurate and fine-grained view of different opinions, the proposed tool is required to attach sentiment annotations to individual sentences within a tweet. However, In order to limit the scope of research, following assumptions are being made.

* That only English tweets will be considered in this study. Re-tweets are excluded from the analysis.
* Since a tweet comprises up of more than a sentence, it may be assumed that the entire tweet contains an opinion on one main object expressed by the “Twitterati” (more reasonable in the context of document-level sentiment analysis).
* That we know the identity of the entity discussed in the sentence.
* It is assumed that each phrase in the sentence also contains just one opinion.
* To further relax the situation, it is assumed that there is a single opinion in each sentence.

**Ethical considerations:**

Privacy in Twitter is not an issue since Twitter allows users to post messages on its platform after a registration phase. During registration, the user is asked to select a unique pseudonym (username) that further serves as the user’s identity. Users may choose to use their original identity instead. All “Mentions” in a tweet indicate the username the tweet is directed at and in order to refer to other users, it uses ‘@’ followed by the username to which it is directed (@username). Across all interactions (replies, follows, retweets), user keeps control over the choice to disclose his/ her original identity or to use a pseudonym. Twitter even gives a user the option to decide if his/ her tweets will be visible to everyone or only to his/ her approved followers.

The study is designed around sentiment analysis of a particular subject that will limit the scope at group level, not an individual user. The topic of interest is also related to “Technology Release” that lies in the public domain and does not pose any privacy challenges. Furthermore, the scope of this study at every level will be defined after a detailed consideration of all possible privacy aspects. The possibility of a misuse or breach of privacy will be minimized.

* 1. **Research Questions:**

The purpose of this research is to investigate public sentiments from their experiences of technology and to tally them as they occur on social media as well as in technology reviews found in sites such as Amazon. This research also involves the correlation between these sentiments as recorded through both the above sources. The underlying research question can therefore be defined as the following.

**RQ1:** Does aggregate social media sentiment agree with online user reviews of technology?

**2. Preliminary Literature Review:**

**2.1 Text Mining:**

According to (Giachanou and Crestani, 2016), Twitter APIs provide an easy access to large amounts of tweets that have specific characteristics. This allow the creation of a data filtering system that can define scope for data collection e.g. having certain terms or emoticons.

1. **Social Media:**

(Carol et al., 2013) defines social media or Web 2.0 technologies as “innovative online tools designed to enhance communication and collaboration”.

Sentiment Analysis is possible across the broad range of social media microblogging platforms such as Tumblr, FourSquare, Google+, and LinkedIn etc.

**(literature review)**

***Why Twitter:*** Sentiment Analysis is possible across the broad range of social media platforms available today. Below are some of the unique characteristics of twitter that distinguish it from other microblogging platforms such as Tumblr, FourSquare, Google+, and LinkedIn for sentiment analysis.

1. *Standard length:* Tweets have a standard length limitation of 140 characters which gives enough room to the Twitterati to explain his/ her opinion while remaining relevant to the topic.
2. *Informal type of medium:* Twitter seems to be the most suitable out of all other social media platforms as it offers an informal medium of expression (more suitable for subjective content) to its registered users while limiting them to 140 characters which helps control content relevance. Other microblogging platforms are either formal (LinkedIn) or are less popular than Twitter (Tumblr, Google+).

*Volume of content:* Over the years, Twitter's interface has remained simple, which is why a lot of tweets take place through third-party sites and applications that make the experience more useful. There could be other sources considered but volume and content relevance become important questions when you consider analyzing sentiments in products that are yet to be announced. In the context of technology release, there is a better chance of finding pre-release product centered content on twitter than any other social media platform also because of its popularity.

1. **Sentiment Analysis Methods:**

(Peng , 2014) Invented a method for Social Media Sentiment Analysis. Accordingly, embodiments generally relate to systems and methods for generating a sentiment dictionary and calculating sentiment scores of adjectives within the sentiment dictionary. A set of seed words can be identified and expanded using synonyms and antonyms of the set of seed words. Social media data can be parsed to identify adjectives that link to the set of seed words with the words "and" or "but." Matrices representing the attraction and repulsion among the linked adjectives can be generated. A factorization algorithm can be minimized to determine an output matrix that comprises positive and negative sentiment scores for each of the adjectives. In embodiments, a sentiment score for part of all of the social media data can be calculated using the output matrix, and one or more parts of the social media data can be classified as a positive or negative sentiment.

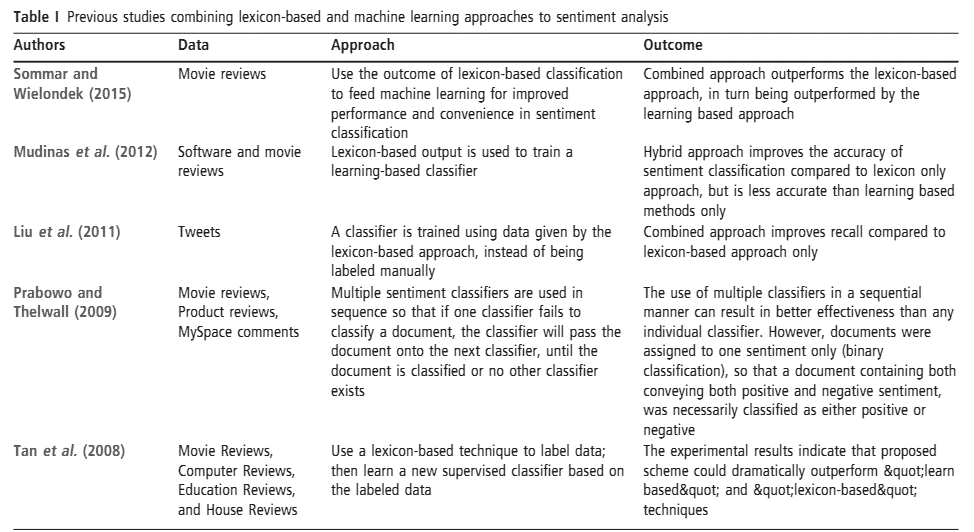


Table : Previous studies combining lexicon-based and machine learning approaches to sentiment analysis.

(Jeong et al.) Explains different approaches to identification of author’s sentiment along with the degree of sentiment

1. Lexicon-based:

This approach uses predefined dictionaries that define sentiment words and their corresponding sentiment values. E.g. SentiwordNet. A number of sentiment lexicon are available to be used for sentiment analysis. Given below are a few references.

* General Inquirer Lexicon: http://www.wjh.harvard.edu/~inquirer/ spread-sheet\_guide.htm.
* Emotion Lexicon: http://www.purl.org/net/emolex
* Financial Sentiment Lexicons: http://nd.edu/~mcdonald/Word\_Lists.html.
* MPQA Subjectivity Lexicon: http://www.cs.pitt.edu/mpqa/subj\_lexicon.html
* SentiWordNet: http://sentiword-net.isti.cnr.it/
* Sentiment Lexicon: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

(Ribeiro et al., 2016) have conducted a benchmark comparison (apple-to-apple) of twenty-four popular sentiment analysis methods (called the state-of-the-practice methods) across eighteen labeled datasets, originating from messages posted on social networks, movie and product reviews, as well as opinions and comments in news articles was conducted as they are used in practice, across multiple datasets.

**2. Sentiment Analysis Algorithms:**

There are multiple methods for measuring sentiments, including Classification method (supervised machine learning) and lexical-based method.

*Classification Method Algorithms:*

(Feldman) Classification algorithms used Given the training data, the system learns a classification model by using one of the common classification algorithms such as Support Vector Machine (SVM), WSVM, C4.5 tree, AdaBoost, Linear kernel SVM, Naïve Bayes (NB), MaxEnt, Multi Naïve Bayes (MNB), CRF, Perceptron with Best Learning Rate, Voted Perceptron, Ensemble Method, Logistic Regression, or kNN.

*Lexicon Based Algorithms:*

SentiStrength, SentiCircles, Clustering-based word sense disambiguation (WSD), Lexicon-based classifier, Rule-based and ESSA.

This classification is then used to tag new documents into their various sentiment classes.

As per (Giachanou and Crestani) these characteristics themselves pose challenge to sentiment analyst.

(Bermingham and Smeaton) have concluded that classifying tweets is easier than classifying longer documents such as blogs.

**3. Challenges in sentiment analysis:**

(Giachanou and Crestani) have explained below characteristics of twitter as the main challenges faced by sentiment analysts.

1. *Text Length:* have explained tweet length limitation (140 characters) and informality of medium as challenges.
2. *Topic Relevance:* many researchers of twitter sentiment analysis have been considering presence of a word in a tweet as an evidence of topic relevance while others studies consider the hashtag symbol as a strong indicator or topic relevance. To a certain degree, these approaches may be correct as commonly the sentiment does target the topic.
3. *Incorrect English:*  Length Limitation and informality of communication make the language used in tweets is very different than the one used in other geners (web, blog, news etc.)
4. *Data Sparsity:* Owing the large volume of incorrect English and misspelled words, tweets contain an extensive amount of noise called “Data Sparsity” that negatively impacts sentiment analysis. Another reason for this noise is the use of non-standard textual artefacts such as emoticons and informal language. (Jeong et al., 2017) have also mentioned emoticons (‘^^’, ‘:-D’) and onomatopoeic words (‘haha’, ‘blah’) as a type of noice.
5. *Compositional Sentiments:*  (Feldman) has expressed the need for better modeling of Compositional Sentiments. At sentence level, this means more accuracy is required in overall sentence sentiment calculation from sentiment-bearing words, the sentiment shifters and the sentence structure.
6. *Anaphora and Auto-Entity Resolution:* Typically in an informal mode of communication, a product may be referred to by multiple names within a context. Anaphora resolution refers to aspect extraction e.g. “battery life” and “power usage” both mean the same thing(Feldman).
7. **Technology Release**

(Jeong et al.) See social media as an emerging source of customer voice since it assumed the form of a channel for exchanging and storing consumer-generated, large-scale, and unregulated voices about products. The authors have proposed a 4 step opportunity mining (identification of product opportunities) approach based upon topic modeling and sentiment analysis of large-scale customer generated social media data using open APIs. Below are the different steps discussed in the approach.

1. Use topic modeling to identify latent product topics used by product customers in social media
2. Quantify the importance of each product topic.
3. Use sentiment analysis to evaluate satisfaction level of each product using sentiment analysis.
4. Use the opportunity algorithm that uses product topic importance and satisfaction to determine opportunity value and improvement direction of each product topic from a customer centered view.

As a case study, opportunity mining of Samsung Galaxy Note 5 has been described as performed through the use of AIChemyAPI included in IBM’s Watson platform.

**3. Theory/ Hypothesis**

**4. Methodology and Research Design:**

* 1. Introduction.. purpose of this section: outlines the methodology adopted and r.design used in order to test the hypo stated/ presented in section 3…

As per section.. lit. review..… a tweet is .. 140 char.. There are multiple methods for measuring sentiments, including lexical-based and supervised machine learning methods. This study is designed to be conducted through development of a software tool (R Script) that will gather relevant data through a supervised process using “twitter” package. The supervised approach assumes that there is a finite set of classes into which data should be classified and training data is available for each class.

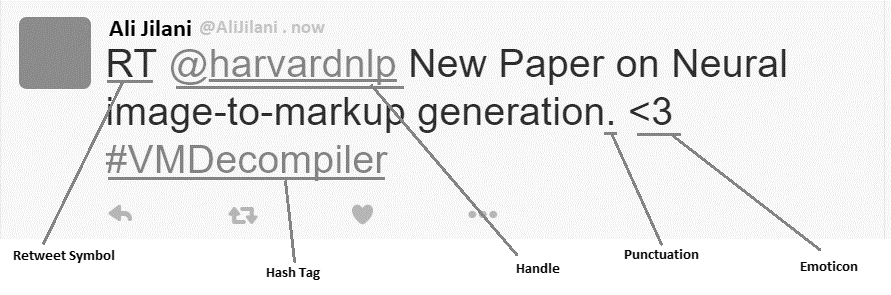


Figure : Components of a typical (re)-tweet.

The script pulls 1000 English tweets matching the specified keywords. E.g. iPhone 4. These tweets form the corpus that is to be analyzed using text mining algorithms. It is therefore converted into a data frame being more suitable for text analysis. The harvested tweets are extracted from this collection to form a corpus comprising Text vectors. A series of data cleaning steps is performed using the text mining package TM. Thus the corpus is processed to remove stopwords, Retweet RT and @Handle labels, #Hashtags, different types of URLs (tiny and normal), all special characters other than English letters and spaces, Numbers and Punctuations. These vectors are then homogenized into lower case before stripping off whitespaces. Other techniques like removal of stem-words etc. are also applied to prepare the corpus for analysis.

Frequencies of different words are then calculated per tweet in the form of Document Term Matrix (DTM). A DTM is a matrix that arranges documents along rows while individual terms/ words in the tweets are arranged into columns. These terms comprising both positive as well as negative terms are compared with a lexicon of positive and negative opinion/ sentiment words.

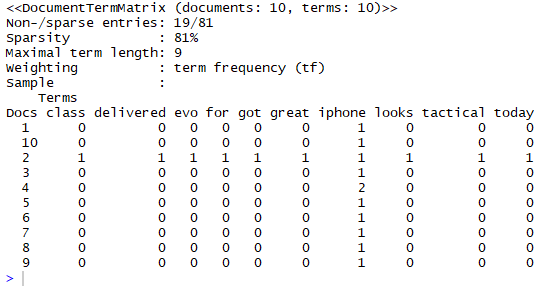


Figure : A sample Document Term Matrix (DTM).

The negative sentiment words lexicon comprises 4782 frequently occurring words while the positive sentiment words lexicon comprises 2006 frequently occurring words.

The entire DTM of 1000 rows is combined and their term frequencies summed up to form a single data row.

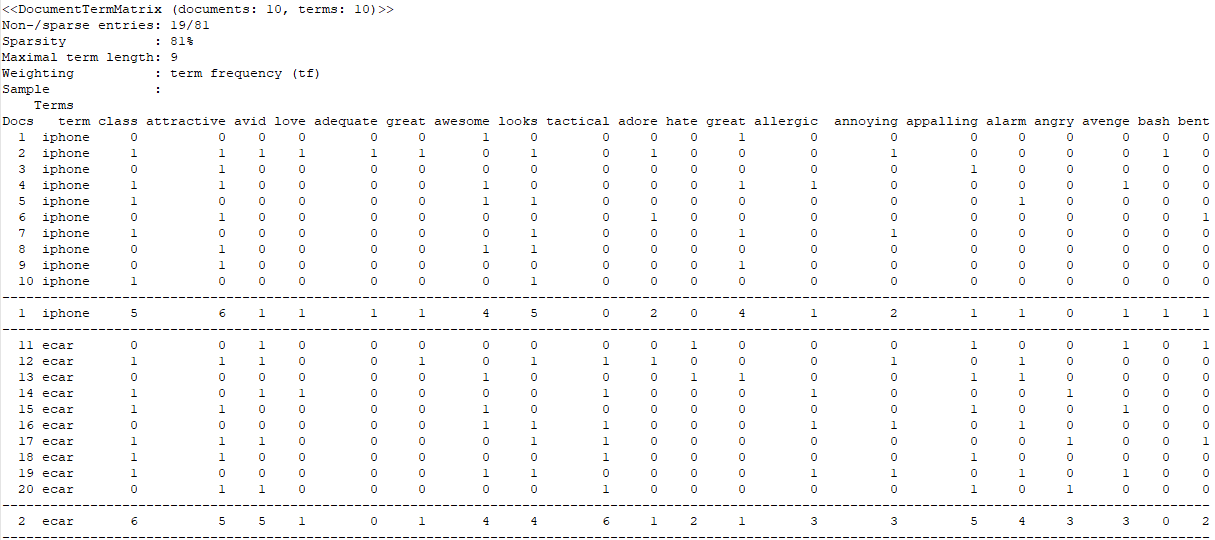


Figure : Term frequencies summed up to form a single data row.

The same process is repeated per technology keyword to form a matrix having rows equal to the number of technologies being investigated in the experiment and columns equal to the union of columns from all the respective DTMs. This would certainly mean that a term not found in one DTM will assume all zeroes as the respective frequencies against columns of words not found in that particular tweet.

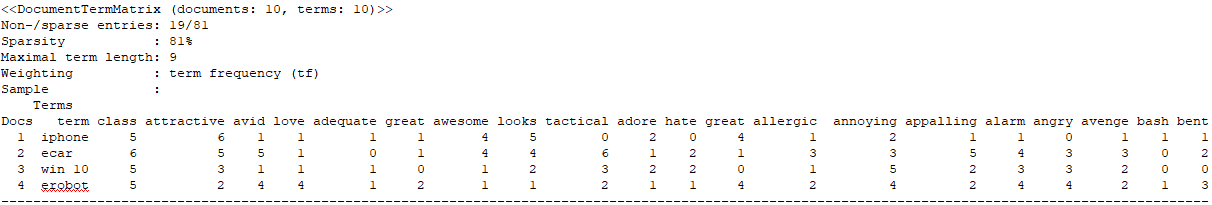


Figure : A matrix of all technologies summarized per data row.

.

A Sentiment Analysis Algorithm will further be employed to calculate the aggregate sentiment score per tweet (say on the scale of 1 to 10).

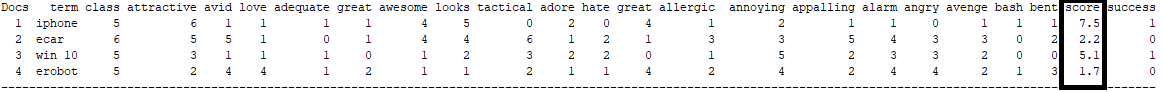


Figure 8: A matrix of twitter data and the calculated aggregate sentiment score.

.

On the other hand, product reviews are inspected over reputable websites to gauge an average public liking for the same technology product over a unified scale. Amazon is one such reputable website that covers the largest number of technology products.

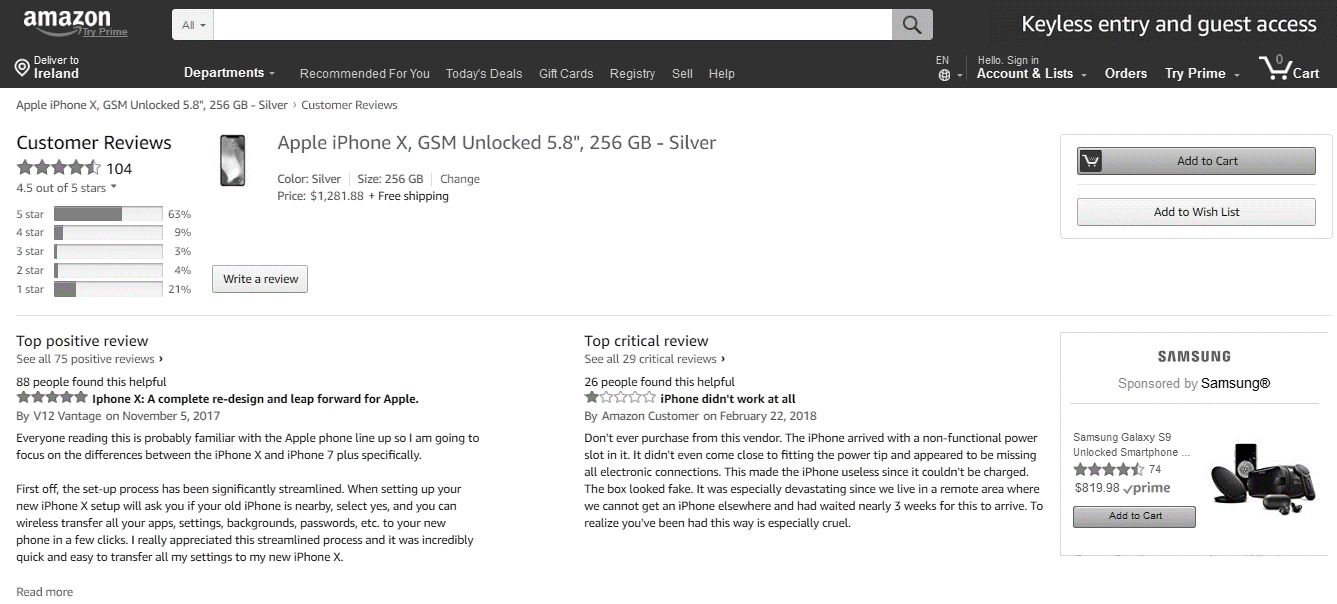
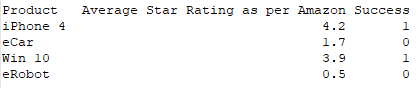


Figure : Amazon reviews on iPhone X.

.

The same set of technology products are inspected to identify their average review over Amazon. A typical Amazon review is marked over a scale of 1 to 5 where 1 represents the least rated product while 5 marks a highly rated product. For the sake of clarity, a threshold value of 3 is decided to mark the boundary value between a successful and non-successful product. i.e. a product given rating <=3 will be considered non-successful while any product rated above 3 (three) will be considered successful.



Pearson’s R will be employed to find a correlation between the overall Unified Sentiment Scores and

The employed machine learning model will learn from the success determined using Amazon average star ratings and will relate the twitter data rows in the document term matrix to their respective success indicators. Below figure represents the arrangement of information from this view.

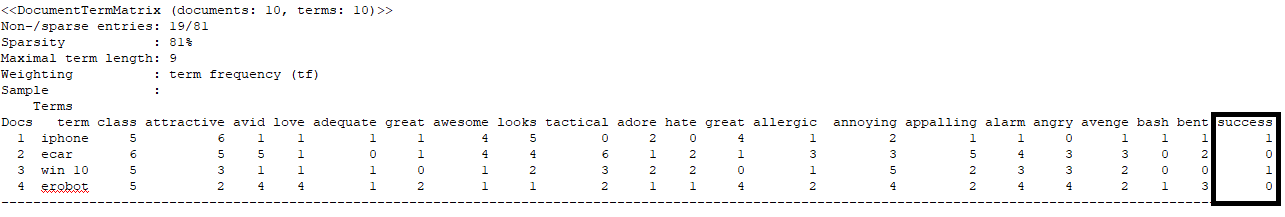


Figure : Success of a technology product is internally related to the DTM data row within the sentiment analysis algorithm.

.

The Machine Learning model using Binary Classification will be trained on these values representing success. Once trained with sufficient data, the model will be utilized in predicting success.

Goes to machine learning for sentiment analysis in Literature review..

Add a column for success and failure.

Above 3 Stars is success below is failure. Below 3 star should have negative Sentiment.

Column for product ID, column for success/ failure (+1/-1).

We will need sentiment scores as outcomes in case of correlations..

Co-relational measure will be easier to implement in R

Algorithms...

Usually used.. Piersons R - Correlation Coefficient to identify correlation of sentiments on twitter with reviews..

P value and R values

P will be below 0.05 (less than .05 percent of error possible)

R will be between 0 and 1

Can you get to correltate SVM with reviews.

Sentiment scores correspond to stars on reviews. Are they the same as social media buzz.

does the star review go up when sentiment score goes up.. and vise versa..

Success Criteria

1. how many items were sold until the next model was launched.. could be a criteria of success.

Instead of combining features, we can add their totals at the end to come up with one grand total sentiment.

train your data with test data.. success again.. again.. split into two.. by identifying the patterns of success clear enouph pattern .. able to converge into ..

words or phrasess or sentiments on success sides that can lead us to judge success based upon them..

**Sentiment Analysis Tool:**

Twitter Sentiment Analysis has been performed by several researchers using a variety of tools developed using different languages such as Java. As part of this study, the tool however is planned to be developed using the programing language “R”. The language R is considered a gold standard for development in the numerical analysis and machine learning space. R’s biggest advantage worth a mention here is its package ecosystem (Krill, 2015).

On the other hand “R” is known to have issues in security and memory management. Since the Sentiment Analysis Tool to be developed in this study, is a stand-alone application with no other modules, security should not be an issue. Memory management however is something that requires serious attention especially when talking about large datasets. This risk in application performance should not be serious as a tweet is typically a “bag” of 140 characters in its anatomy. If however the issues arises, it can be addressed by some trade off in terms of simplicity of sentiment analysis algorithm.

**Sentiment Lexicon:**

Phrase (Mohammad Saif M, 2014)

**Data Collection:**

As per (Giachanou and Crestani, 2016), Twitter provides an easy way for developers and researchers to access and gather twitter data using either of two APIs namely, “REST” and “Streaming”. Twitter APIs provide an easy access to large amounts of tweets that have specific characteristics. This allow the creation of a data filtering system that can define scope for data collection e.g. having certain terms or emoticons. Usually, lists of entities, hashtags or emoticons are used to crawl Tweets that are returned in JSON format. JSON is widely used for storing and exchanging data, and can be easily parsed by many programming languages. The metadata returned by the Twitter APIs include information such as publication date, author’s username, location, hashtags, retweets, followers, and many other data.

(Giachanou and Crestani, 2016) pointed out at Streaming API as a preferred data gathering approached for most researchers as it provides unlimited and real time access to tweets that meet a specific requirement. Same approach was followed for this study. The R specific flavor of twitter APIs come under the “twitteR” package that offers neat options of pulling live tweeter data. The study is focused around tweets with only language constraints (English language but no specific time constraint). This makes the fresh tweets data equally relevant without having to apply any specific filters. Yet, in favor of homogeneity, data gathering was (and will be) spread over a period of one week to minimize any effects of different (non)/working days. Time of the day may also have some effect but that is not considered interesting for this study. It is also relevant to note here that the study tries to limit the impact of product (technology) release dates on the results of the study by keeping the focus on technologies already launched do not have a significant impact on the study.

The number of tweets gathered as a whole is decided to be 1000. Below are the technology categories and keywords used in the search criteria.

|  |
| --- |
| 1. Mobile Phones: (13 models) |
| #iPhone X, iPhone 8 Plus, iPhone 8, iPhone 7, iPhone 7 Plus, iPhone 6, iPhone 6 Plus, iPhone 6s,  #iPhone 5, iPhone 5c, iPhone 5s, iPhone 4, iPhone 4s. |
| 2. Autonomous Cars: (10 models) |
| #Tesla Model S, Audi A7, Genesis G90, Cadillac CT6, Mercedes-Benz CLS-Class, BMW 6-Series,  #Cadillac XTS, Kia K900, Lincoln Continental, Acura RLX. |
| 3. Air Drones (14 Quadcopter models) |
| #DJI Phantom 3, DJI Phantom 3 Standard, DJI Phantom 3 Advanced, DJI Phantom 3 Professional, DJI PHANTOM 4 PRO, DJI Mavic Pro, DJI Spark #Yuneec Typhoon Q500, Typhoon H 4k, Yuneec Breeze. #Parrot AR.Drone 2.0, Bebop 2,  Parrot Disco FPV, Parrot Mambo |

https://cran.r-project.org/web/packages/rvest/vignettes/selectorgadget.html

For Reviews, Amazon reviews were scraped using an open source Google Chrome extension called “Selector Gadget”. Selector Gadget integrates with Chrome to ease CSS selector generation and discovery on complicated websites. Once the CSS selector is determined, it can be used in R to extract values off a webpage. Is iPhone successful criteria of success, measure of sale,

decide on your own if it is successful under my criteria. goto twitter..

do tweets also support related to post release info/ pre release information..

Other options of review sites contain <http://www.trustedreviews.com/>

Product Name/ ID, Release date range, Hash tags related to them.

**SA algo to apply dplyr**

**what correlation algo..** **for smaller samples.. (Spiermen R)**

**what binary classifier.. (Perceptrons or SVM Support Vector Machines), What is the most suitable binary classifier**

**Available Online Resources:**

A number of sentiment lexicon are available to be used for sentiment analysis. Given below are a few references.

1. General Inquirer Lexicon: http://www.wjh.harvard.edu/~inquirer/ spread-sheet\_guide.htm.
2. Emotion Lexicon: http://www.purl.org/net/emolex
3. Financial Sentiment Lexicons: http://nd.edu/~mcdonald/Word\_Lists.html.
4. MPQA Subjectivity Lexicon: http://www.cs.pitt.edu/mpqa/subj\_lexicon.html
5. SentiWordNet: http://sentiword-net.isti.cnr.it/
6. Sentiment Lexicon: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

**Lexicon (D:\Ali\_Home\WIT\TSA\Learning R\sentiment\_scoring)**

<https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>

Also contains good sentiment analysis tutorial on amazon reviews.

**Description of the Experimental Design/ Verification Methodology:**

As per (Feldman, 2013), there are two main approaches to sentiment analysis, “Supervised” and “Unsupervised”. A single tweet may contain more than one sentence. It is important to note here that each sentence can have its own **Sentiment Orientation** (SO). Which simply means that, it is fair to treat it as a document to apply document sentiment analysis techniques to it or to treat it as a “bag of words”. A tweet level sentiment score can therefore be calculated (using linguistic resources) and associated with each tweet. The same tweet in another way can also be broken-down into sentences, which implies that sentence level sentiment scores can also be associated to each sentence. There are separate techniques/algorithms used for document sentiment calculation and sentence sentiment calculation. It should be possible to compare the two outcomes (document sentiment & sentence sentiment) with each other.

As part of this study, it is hypothesized that

**H1:** Aggregate sentiment analysis score on social media will correlate to aggregate online user reviews of

technology?

**H2:** Binary Classification could be used to classify social media sentiment as positive or negative in a way that

corresponds with online user reviews of technology. (here we don’t use the aggregate sentiment score). All the words become variables here and frequencies are the values. The model uses supervised learning i.e. it will be fed twitter data and told if it corresponds to success. Upon successful learning, given a new technology, it will predict. 10 tech products are not enouph of training model

get another fewtweets but not use algo on it. And let the model predict it. Then check star rating.

Pecetrons case: The model will position continuously move the hyperplane to position it self based upon the data.

Binary classifiers will work the same way.

Model may end up with .89 that is close enouph to one pole for us to live with.

Error is the difference between what it gives out and what we want it to give out.

**H1:** The document sentiment of a tweet and its aggregate sentence level sentiment are equivalents/ correlated.

**H3:** These two techniques (tweet level technique and sentence level technique) are not always the same for every set of tweets taken from the Twitter corpus.

**H4:** The overall success of a technology release will be likely if the aggregate of these document level sentiment scores is positive and unlikely if their aggregate is negative (falls below zero).

**H5:** A dominant pattern is visible when all these sentiments are presented using a visualization tool.

**Special Resources Required:**

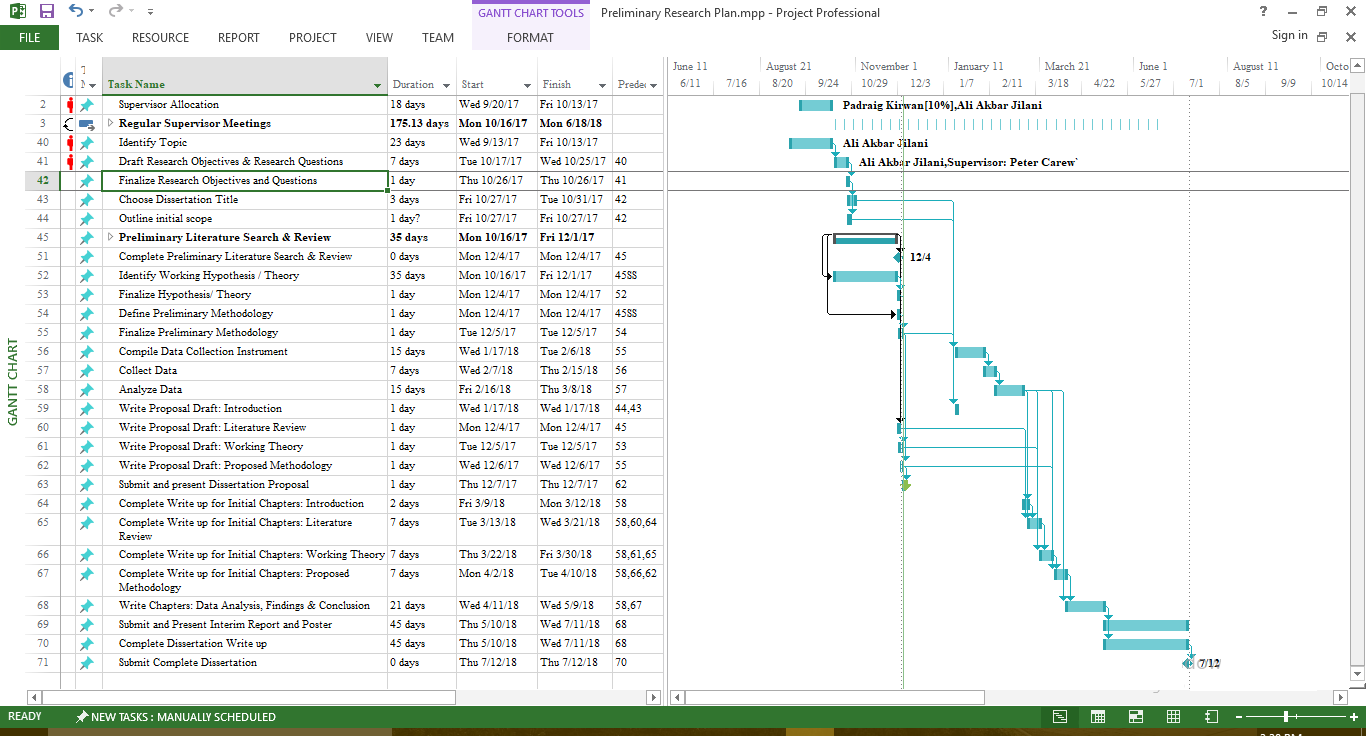
Since the Software development tool is to be developed using **“R”** as a programming language, An IDE would be required. A number of freely downloadable IDEs are available online. The most suitable in terms of features seems to be “**RStudio**” which has a Desktop and Server edition. Open Source versions of both, the RStudio Desktop and RStudio Server are freely available for download (https://www.rstudio.com/products/rstudio/), A number of R Packages for Text mining and Sentiment Analysis are available out of which **TM** (Text Mining) and **RSentiment** initially seem to be interesting. For visualization and graphics, **Shiny** (https://shiny.rstudio.com/) will be used.

|  |
| --- |
| Sure, you can easily examine complex formulas on a spreadsheet. But it's not nearly as easy to run multiple data sets through spreadsheet formulas to check results as it is to put several data sets through a script, he explains. |
| https://www.computerworld.com/article/2497143/business-intelligence/business-intelligence-beginner-s-guide-to-r-introduction.html |

**Main Milestones Anticipated:**

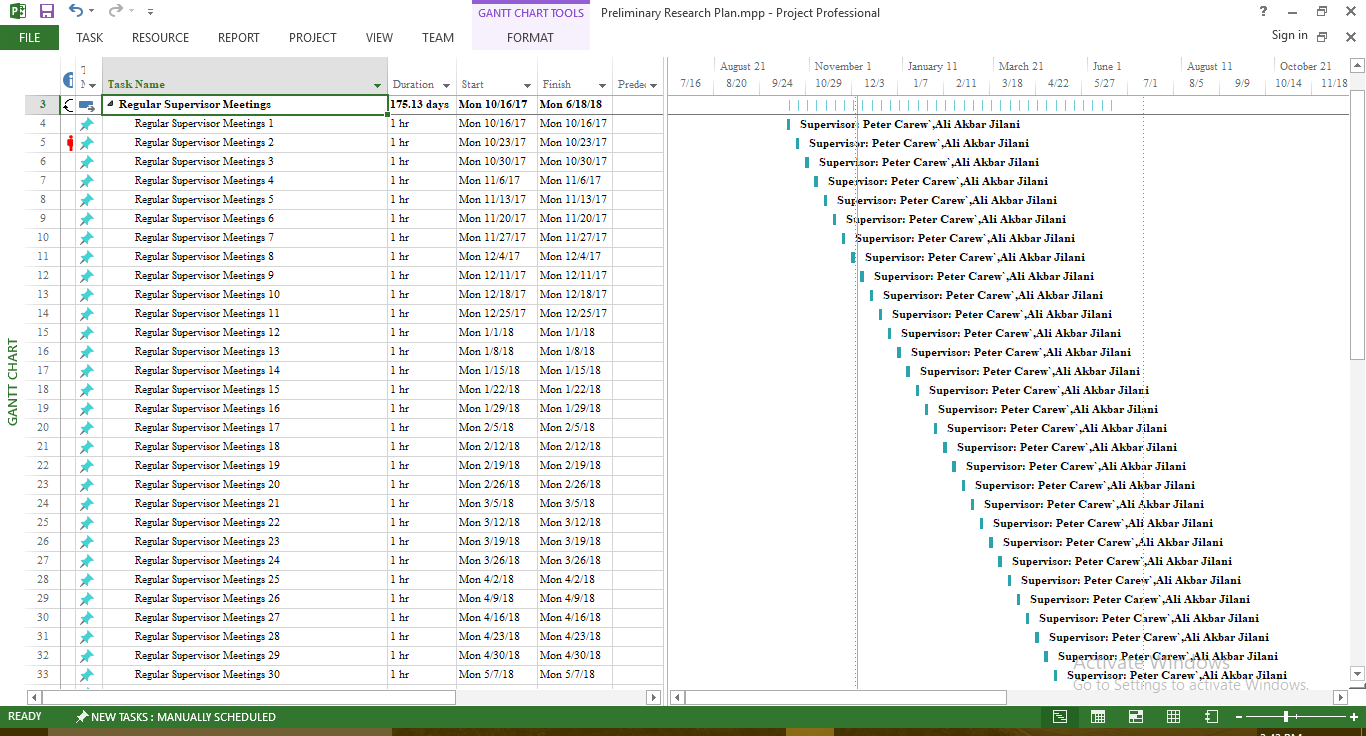
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Task #** | **Task Name** | **Duration** | **Start** | **Finish** | **Predecessors** |
| 2 | Supervisor Allocation | 18 days | Wed 9/20/17 | Fri 10/13/17 |  |
| 3 | **Regular Supervisor Meetings** | **175.13days** | **Mon 10/16/17** | **Mon 6/18/18** |  |
| 40 | Identify Topic | 23 days | Wed 9/13/17 | Fri 10/13/17 |  |
| 41 | Draft Research Objectives & Research Questions | 7 days | Tue 10/17/17 | Wed 10/25/17 | 40 |
| 42 | Finalize Research Objectives and Questions | 1 day | Thu 10/26/17 | Thu 10/26/17 | 41 |
| 43 | Choose Dissertation Title | 3 days | Fri 10/27/17 | Tue 10/31/17 | 42 |
| 44 | Outline initial scope | 1 day | Fri 10/27/17 | Fri 10/27/17 | 42 |
| 45 | **Preliminary Literature Search & Review** | **35 days** | **Mon 10/16/17** | **Fri 12/1/17** |  |
| 46 | Literature Search & Review : Social Media | 35 days | Mon 10/16/17 | Fri 12/1/17 |  |
| 47 | Literature Search & Review : Twitter | 35 days | Mon 10/16/17 | Fri 12/1/17 |  |
| 48 | Literature Search & Review : Sentiment Analysis | 35 days | Mon 10/16/17 | Fri 12/1/17 |  |
| 49 | Literature Search & Review : Technology Release | 35 days | Mon 10/16/17 | Fri 12/1/17 |  |
| 50 | Literature Search & Review : Advertising Campaign | 35 days | Mon 10/16/17 | Fri 12/1/17 |  |
| 51 | Complete Preliminary Literature Search & Review | 0 days | Mon 12/4/17 | Mon 12/4/17 | 45 |
| 52 | Identify Working Hypothesis / Theory | 35 days | Mon 10/16/17 | Fri 12/1/17 | 45SS |
| 53 | Finalize Hypothesis/ Theory | 1 day | Mon 12/4/17 | Mon 12/4/17 | 52 |
| 54 | Define Preliminary Methodology | 1 day | Mon 12/4/17 | Mon 12/4/17 | 45SS |
| 55 | Finalize Preliminary Methodology | 1 day | Tue 12/5/17 | Tue 12/5/17 | 54 |
| 56 | Compile Data Collection Instrument | 15 days | Wed 1/17/18 | Tue 2/6/18 | 55 |
| 57 | Collect Data | 7 days | Wed 2/7/18 | Thu 2/15/18 | 56 |
| 58 | Analyze Data | 15 days | Fri 2/16/18 | Thu 3/8/18 | 57 |
| 59 | Write Proposal Draft: Introduction | 1 day | Wed 1/17/18 | Wed 1/17/18 | 44,43 |
| 60 | Write Proposal Draft: Literature Review | 1 day | Mon 12/4/17 | Mon 12/4/17 | 45 |
| 61 | Write Proposal Draft: Working Theory | 1 day | Tue 12/5/17 | Tue 12/5/17 | 53 |
| 62 | Write Proposal Draft: Proposed Methodology | 1 day | Wed 12/6/17 | Wed 12/6/17 | 55 |
| 63 | Submit and present Dissertation Proposal | 1 day | Thu 12/7/17 | Thu 12/7/17 | 62 |
| 64 | Complete Write up for Initial Chapters: Introduction | 2 days | Fri 3/9/18 | Mon 3/12/18 | 58 |
| 65 | Complete Write up for Initial Chapters: Literature Review | 7 days | Tue 3/13/18 | Wed 3/21/18 | 58,60,64 |
| 66 | Complete Write up for Initial Chapters: Working Theory | 7 days | Thu 3/22/18 | Fri 3/30/18 | 58,61,65 |
| 67 | Complete Write up for Initial Chapters: Proposed Methodology | 7 days | Mon 4/2/18 | Tue 4/10/18 | 58,66,62 |
| 68 | Write Chapters: Data Analysis, Findings & Conclusion | 21 days | Wed 4/11/18 | Wed 5/9/18 | 58,67 |
| 69 | Submit and Present Interim Report and Poster | 45 days | Thu 5/10/18 | Wed 7/11/18 | 68 |
| 70 | Complete Dissertation Write up | 45 days | Thu 5/10/18 | Wed 7/11/18 | 68 |
| 71 | Submit Complete Dissertation | 0 days | Thu 7/12/18 | Thu 7/12/18 | 70 |

**Gantt chart:**

****

**Weekly Supervisor Meetings:**

The task is arranged to occur recursively for 36 instances as scheduled below.



**References:**

BERMINGHAM, A. & SMEATON, A. F. 2010. Classifying sentiment in microblogs: is brevity an advantage? Ireland, Europe: Association for Computing Machinery.

CAROL, T., RACHEL, V. & DONALD, W. K. 2013. Social media and scholarly reading. *Online Information Review,* 37**,** 193-216.

FELDMAN, R. 2013. Techniques and applications for sentiment analysis. *Commun. ACM,* 56**,** 82-89.

GIACHANOU, A. & CRESTANI, F. 2016. Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Computing Surveys,* 49**,** 28-28:41.

JEONG, B., YOON, J. & LEE, J.-M. 2017. Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal of Information Management*.

KRILL, P. 2015. *Why R? The pros and cons of the R language* [Online]. Available: <https://www.infoworld.com/article/2940864/application-development/r-programming-language-statistical-data-analysis.html> [Accessed 08/12/2017 2017].

PENG , E. A. May 13, 2014. *Systems, methods and devices for generating an adjective sentiment dictionary for social media*

*sentiment analysis: [US Patent & Trademark Office, Patent Full Text and Image Database]*. United States, CA, Sunnyvale. patent application 13/082,963.

RIBEIRO, F. N., ARAÚJO, M., GONÇALVES, P., ANDRÉ GONÇALVES, M. & BENEVENUTO, F. 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science,* 5**,** 23.

<http://www.citethisforme.com/>

<https://www.ukessays.com/referencing/harvard/generator/>

https://www.qub.ac.uk/cite2write/harvard3l.html

http://www.harvardgenerator.com/references/conference-paper

[Mark McGuire, Constance Kampf (2015) *Using social media sentiment analysis to understand audiences: A new skill for technical communicators?*, 2015 edn., Limerick, Ireland: 2015 IEEE International Professional Communication Conference (IPCC).](http://ieeexplore.ieee.org.ezproxy.wit.ie:2048/document/7235801/)

Lexicon:

Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web." Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.

To be referenced.

Data Cleaning techniques: PacktPub Book of R

Not used yet.

Text Mining Research References: https://sites.google.com/site/miningtwitter/references

Good Sources of papers

<https://www-sciencedirect-com.ezproxy.wit.ie/search?qs=binary+classifier&origin=article&zone=qSearch>

Plotting different sentiments using R

https://sites.google.com/site/miningtwitter/questions/sentiment/sentiment