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

Ali Akbar Jilani

Student No. 20078735
Department of Science and Computing,
Waterford Institute of Technology, Waterford, Ireland.
(Tel: 353-085-212-6940; e-mail: 20078735@mail.wit.ie).

Abstract. User Generated Content (UGC) over social media can be tapped to extract underlying public sentiment and take relevant actions. Proposed study uses sentiment analysis on twitter data to investigate its correlation with sentiments of same products as measured through Amazon user reviews. The study goes further into supervised machine learning techniques to train a binary classifier on the patterns of sentiment words occurring in the same twitter data to test its correspondence with online user reviews in the technology product domain.

Keywords: Sentiment Analysis, Binary Classification, Social Media, Twitter, Machine Learning, Text Mining, Technology Release, Product Success, Online Review.

1 Introduction

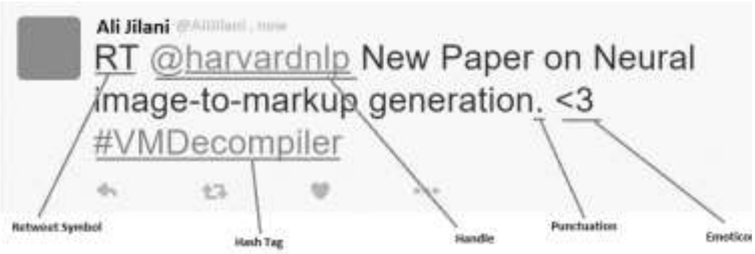
1.1 Background

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.

1.1.1 Structure of a Tweet:

A typical tweet may have the following components.



Ali Jilani

RT @harvardnlp New Paper on Neural image-to-markup generation. <3 #VMDecompiler

Figure 1: Components of a typical (re)-tweet.

A “tweet” primarily is a short text message that comprises up of 140 characters. A Hashtag is the # symbol followed by textual topic name. A hashtag is used to identify messages on a specific topic. A Twitter handle comprises up of @ symbol followed by a Username which could also be a pseudo-name. A handle is used to send direct messages to a user. When a tweet is re-shared by another specific user, an RT label is associated with it to represent its status as re-tweet. In addition to text, a tweet may also contain formations of special characters that combine to form emoticon.

1.2 Research Objective and Questions

This study is designed to achieve the objective of comparing social media buzz (on Twitter) with public sentiment about technology release caught from online review system (Amazon). 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. It is relevant to mention here that this research is likely to have a universal application.

According to research [1], 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 through an optimized machine learning technique 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 may comprise 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 sentence also contains just one phrase that in turn contains only one opinion.

1.3 Data Gathering

Below are three categories of names of technology products that can be used as Hashtags to harvest Twitter data related to products. The study is designed to pull 1000 tweets against each keyword that is fed to it and an overall 10 keywords out of below three categories of technology products will be used.

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

1.3.1 Data Filters

It is worth a mention that by design, there is no limitation on the date or time of data gathering. All data is gathered using Twitter APIs through a custom developed R Script that applies language and keyword filters along with an additional clause to exclude retweets from the resulting data corpus.

1.3.2 Time Constraints:

The data gathering procedure is constrained by the 9 days of free twitter data streaming range. Twitter APIs reject any request for historical archived data beyond this range. Access to all the twitter archives since the first tweet made in 2006, is available through twitter reseller accounts against a license fee. This study is independent of time series data in any way. Which means, that it is assumed that there will be no impact what so ever on results if the twitter corpus pulled for sentiment analysis is as current as 9 days old or as old as belonging to earliest days to Twitter launch as a service i.e. 2006. On top of that, results are also not impacted by time of the day or day of the week or even day of the month. It is for this reason that all twitter data will be progressively pulled using twitter live feeds as the project progresses. This ideally means that all data should be available in a matter of few minutes.

1.4 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.

1.5 Research Questions:

The purpose of this research is to measure public sentiments expressed by twitter users in the form of tweets posted about their experiences of technology, and to tally these sentiments with their equivalents of technology reviews found in sites such as Amazon. This research investigate the correlation between these sentiments as recorded through both the above sources. The underlying research question can therefore be defined as the following.

Research Question 1:

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

2 Literature Review

2.1 Machine Learning Techniques for Sentiment Analysis

There are multiple techniques for measuring sentiments, including lexical-based and supervised machine learning techniques. The supervised approach to binary classification assumes that there is a finite set of classes into which data should be classified and training data is available for each class. Neurocomputing[13] compared different machine learning techniques commonly used for sentiment analysis to categorize them as below.

2.1.1 Supervised Machine Learning Techniques

A supervised learning algorithm generates a model through a training process where the label or result of the data is known as priori. This model is used for classifying future instances in which the feature values are given as input, but the class label is unknown.

2.1.2 Unsupervised Machine Learning Techniques

An unsupervised learning algorithm divides data points into similarity groups called clusters. A cluster consists of a set of instances that are more similar to each other in some way than the other ones in other clusters. Since there is no need for output values in unsupervised methods, the class labels of the datasets used in this study are neglected for the training part of these methods.

2.2 Text Mining

A specialized research on text mining and sentiment analysis [12] suggests that classical data mining methods, text mining and sentiment analysis deal with unstructured data. Text mining is a specialized branch of Data mining. Data mining deals with mining hidden knowledge but in text mining, information is plainly present in text format and there is no concept of hidden information. Main objective of text mining is to get text in computer understandable form directly so that it can be processed without human intervention. Data mining works with structured data like databases, data warehouse, online shopping data, mobile usage data etc. whereas, text mining works with unstructured or semi-structured natural language data. Example of dataset for text mining is data generated by social media, which is natural language unstructured data. So biggest hurdle to text mining is Natural Language Processing (NLP).

2.3 Sentiment Analysis

Feldman [1] 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.

2.3.1 Challenges in Sentiment Analysis

Giachanou and Crestani [2] have explained below characteristics of twitter as the main challenges faced by sentiment analysts.

1. *Text Length:* Tweets have length limitation of 140 characters and are considered an informal medium. Both of above characteristics add complexity to sentiment analysis process thus making it more challenging.

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 other studies consider the hashtag symbol as a strong indicator of topic relevance. These approaches may be correct, but only to a certain degree, as commonly the sentiment does target the topic.

3. *Incorrect English:* Length Limitation and informality of communication make the language used in tweets very different from the one used in other genres (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 noise.

5. Compositional Sentiments: Feldman [1] 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: Feldman [1] specifies twitter as an informal mode of communication. Thus, 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.

2.4 Social Media

JISC [8] 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.

2.4.1 Social Media Data

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+).

3. *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.

2.5 Technology Release

Opportunity mining approach [14] to social media sentiment analysis is fast emerging as a 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.
4. Use the opportunity mining 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.

2.5.1 Sentiment Analysis Algorithms

Classification Based Algorithm:

In the territory of document level sentiment analysis as suggested by Feldman [1], Support Vector Machine (SVM), Linear kernel SVM, Naïve Bayes (NB), WSVM, C4.5 tree, AdaBoost, MaxEnt, Multi Naïve Bayes (MNB), CRF, Perceptron with Best Learning Rate, Voted Perceptron, Ensemble Method, Logistic Regression, or kNN are a few mentioned algorithms that can be used to perform sentiment analysis.

Lexicon Based Algorithms:

Giachanou and Crestani[2] have mentioned that Lexicon Based Algorithms have been extensively applied on conventional text such as blogs, forums and product reviews but have less been explored for Twitter Sentiment Analysis. SentiStrength, SentiCircles, Clustering-based Word Sense Disambiguation (WSD), and Lexicon-based classifiers are a few mentioned lexicon based algorithms. In addition, a three step technique referred for TSA [15] comprises up of preprocessing as step one, polarity detection as step two, and rule based classification as step three. Last two steps were based on WordNet and SentiWordNet.

SentiStrength being one of the most well-known lexicon-based algorithms developed for social media uses a list of emoticons, negations, and boosting words to effectively identify sentiment strength of informal text including tweets using a human-coded lexicon that contains words and phrases frequently confronted in social media.

2.5.2 Sentiment Analysis Approaches

Giachanou and Crestani [2] mention document, sentence and entity levels as the three different levels at which Sentiment Analysis (SA) have been applied in literature. Document level sentiment analysis aims to identify sentiment polarity that is expressed in the whole document. Sentence level sentiment analysis on the other hand classifies each sentence as positive or negative and entity level SA detects sentiment polarity of specific entity/ target of a specific object.

Mainly due to size limitations imposed by twitter, most of tweets contain a single sentence. Therefore, there is no fundamental difference between document and sentence level when it comes to TSA. In case of tweets, SA can be applied on message/ sentence and entity level.

Supervised Learning:

As per research[1], supervised approach to document level sentiment analysis expects a finite set of classes with access to training data for each one of them. A common example would be to classify (tag) a document into either positive or a negative class. The case can further be extended to also include a third neutral class. A relatively advanced case could contain a discrete numeric scale to classify the document into (like the five-star system employed by Amazon).

Using the training data, the system learns a classification model by using one of the common classification algorithms. The model is then used to tag new documents into various sentiment classes. Regression can further be used to assign a numeric value (in a finite range) to a document. Feldman also refers to research that shows that good accuracy is achieved even when each document is represented as simple "bag of words". More advanced representations utilize TFIDF, POS (Part of Speech) information, sentiment lexicons, and parse structures.

Unsupervised Learning:

Unsupervised learning approach to document level sentiment analysis is based upon the determination of semantic orientation (SO) of specific phrases within the document. These phrases can be selected using either of POS pattern or a lexicon of sentiment words and phrases. Thus, if the average SO of these phrases falls above a specific threshold, it is classified as positive whereas, the average SO below the predefined threshold value classifies the document into negative class.

Another classic way of calculating SO of a given word or phrase is to calculate the statistical difference between the Pointwise Mutual Information: PMI(P,W) of the phrase ‘P’ with two sentiment words ‘W’ based on their co-occurrence in a given corpus or over the web. As an example quoted by Feldman [1], the SO measures whether the phrase P is closer in meaning to ‘excellent’ being positive word or ‘poor’ being negative word.

2.6 Sentiment Lexicon

An approach to sentiment analysis of short informal text [9] hints at NRC Emotional Lexicon [9], Bing Liu's Lexicon [10], and the MPQA Subjectivity Lexicon [11]. The NRC Emotion Lexicon is comprised of frequent English nouns, verbs, adjectives, and adverbs annotated for eight emotions (joy, sadness, anger, fear, disgust, surprise, trust, and anticipation) as well as for positive and negative sentiment. Bing Liu's Lexicon provides a list of positive and negative words manually extracted from customer reviews. The MPQA Subjectivity Lexicon contains words marked with their prior polarity (positive or negative) and a discrete strength of evaluative intensity (strong or weak). Entities in these lexicons do not come with a real-valued score indicating the fine-grained evaluative intensity.

2.7 Data Cleaning Activities

A step-by-step [3] approach to harvesting data via the Twitter Application Programing Interface (API), Cleaning of the data and the basic sentiment analysis code in R language provides value as a single and quick reference that still holds valid. Other closest references from Packt [4], de Vries approach [5], The Airline Consumer Sentiment Analysis [6] and Breen’s approach [7] are all outdated.

2.8 Conclusion:

State of the art suggests that a supervised machine learning technique using binary classification and the de Vries approach [5] for data pre-processing combined with a sentiment lexicon can offer the right set of tools required to accomplish the target of this study. R Studio with twitteR and tm packages (along with other packages) further provide the complete environment required to develop the target R script.

3 Theory

As part of this study, it is hypothesized that

3.1 **Hypothesis 1:** Aggregate sentiment analysis score on social media will correlate to aggregate online user reviews of technology.

Sentiment Analysis Algorithm employed as part of this study will make use of a set of sentiment lexicons [10] (positive and negative lexicons) of different sentiment words to gauge the aggregate sentiment of a corpus related to a specific technology product release. Average star rating as taken from online reviews site (e.g. Amazon) is as such a measure of public sentiment about the same technology product released but on a different scale. It is hypothesized that there is a correlation between both the sentiment scores.

3.2 **Hypothesis 2:** 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.

Instead of aggregate sentiment score, our binary classification model [1] such as SVM or Perceptron will be fed with twitter data. All the words in twitter corpus related to a specific technology product release will serve as variables while their frequencies of occurrence in the corpus will be treated as values. Our model will be trained using supervised learning [13] to be told if the overall set of words and values corresponds to success or to failure (calculated through the average online review on Amazon as a side process). Twitter corpora relating to a minimum of 10 technology products released will be used to develop a learning for our model using training data. Training data will be taken out of the harvested tweets and would represent 80% of the overall corpus size. Test data would be represented by rest of the 20% data. Upon successful learning, the same model will be utilized in predicting the outcome against a test data partition in terms of either of success or failure. Results will be compared with online user reviews. It is hypothesized that the binary outcome of our model actually corresponds with online user reviews.

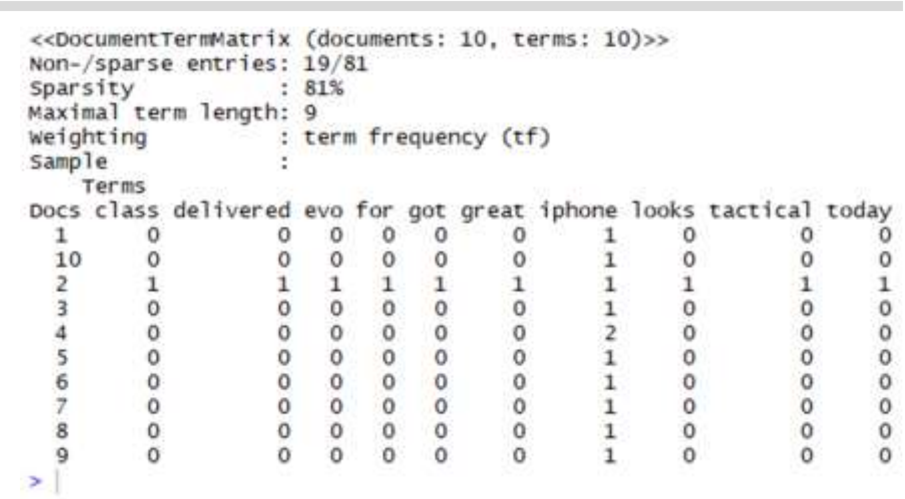
4 Methodology

Success of a technology product is a subjective matter that can be explained in lots of different ways. Market share of a technology product or number of items sold since product release could be considered different examples of success criteria. Online product reviews are discrete representations of public sentiments about a product and to some extent can also be taken as success criteria [14].

In order to be able to relate Amazon reviews to public sentiment about the same product as found on social media [8], we device a script using R programming language to procedurally [5] convert these public words into their equivalents of sentiment over a scale comparable with online user reviews. This study employs a binary classifier [1] to investigate the possibility of correlating social media buzz with public reviews on Amazon. This section outlines the methodology adopted and the research design used in order to test the hypotheses presented in section 3.

For the scope of this study, Tweets are considered a ‘bag of words’ [1] as mentioned in section 2.5.2. Similarly, document level sentiment analysis [2] is applied to each tweet to calculate aggregate sentiment score per tweet using a pair of lexicons. These sentiment scores are then aggregated to summarize all the 1000 rows of a technology product into a single row. Supervised Machine Learning technique can confidently be applied to train a binary classifier before using it for predictions.

Sentiment analysis algorithm in this study is applied to a corpus of 1000 tweets per technology product, pulled through live twitter stream based upon a criteria matching our specified keywords and defined language filter using twitter APIs called into our R Script through twitteR package using step-by-step[3] approach mentioned in section 2.7. These vanilla tweets are then converted into data frame (pre-process) being more suitable for the purpose of text analysis. de Vries approach [5] is further applied to clean the harvested tweets 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. The tweets representing rows of data frame are then extracted from this collection to form a corpus comprising Text vectors. 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/terms 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.



```
<<DocumentTermMatrix (documents: 10, terms: 10)>>
Non-/sparse entries: 19/81
Sparsity           : 81%
Maximal term length: 9
Weighting          : term frequency (tf)
Sample             :
    Terms
Docs class delivered evo for got great iphone looks tactical today
  1      0         0   0   0   0     0      1     0        0     0
  10     0         0   0   0   0     0      1     0        0     0
  2      1         1   1   1   1     1      1     1        1     1
  3      0         0   0   0   0     0      1     0        0     0
  4      0         0   0   0   0     0      2     0        0     0
  5      0         0   0   0   0     0      1     0        0     0
  6      0         0   0   0   0     0      1     0        0     0
  7      0         0   0   0   0     0      1     0        0     0
  8      0         0   0   0   0     0      1     0        0     0
  9      0         0   0   0   0     0      1     0        0     0
> |
```

Figure 2: A sample Document Term Matrix (DTM).

These terms comprise up of both positive as well as negative words frequently occurring in customer reviews as suggested in Section 2.6. 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.

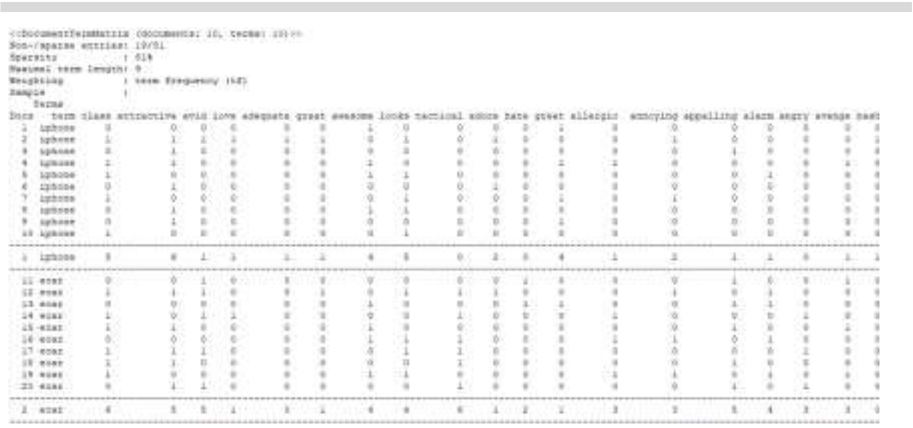


Figure 3: Term frequencies summed up to form a single data row.

The same process is repeated per technology keywords 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.

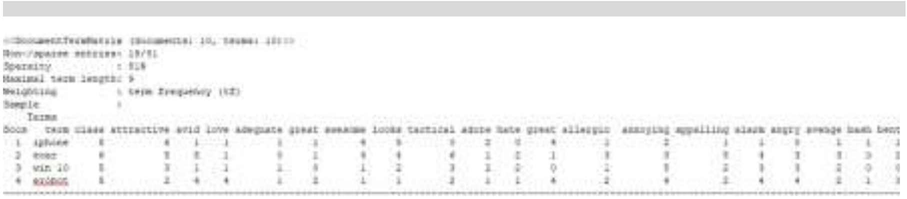


Figure 4: DTM combining the summary rows representing each of technology products.

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. A Sentiment Analysis Algorithm will further be employed to formulate the aggregate sentiment score per technology product as a function of sentiment scores of happy words and the sad words (say on the scale of 1 to 10).

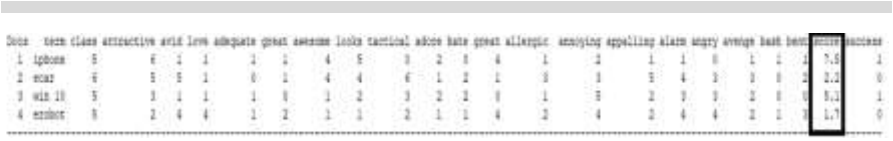


Figure 7: DTM with aggregate sentiment scores calculated.

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.

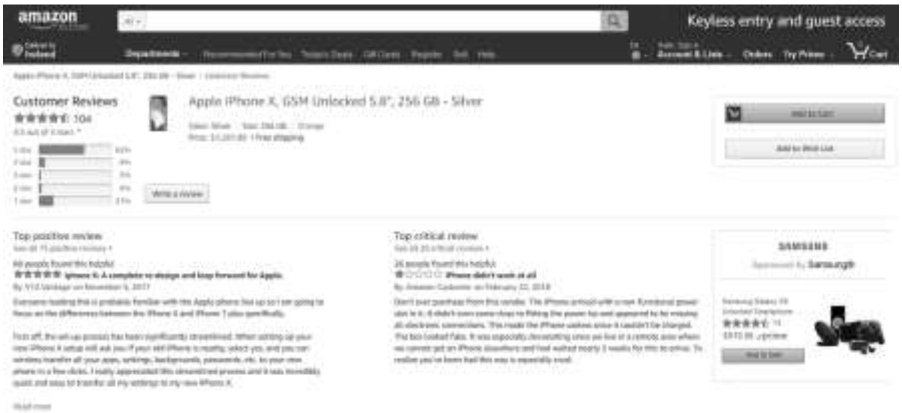
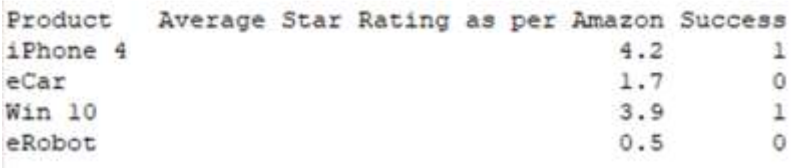


Figure 8: Amazon online reviews page for iPhone X.

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 and a representation of positive sentiments towards that technology. Below table represents a technology product, its average star rating taken from Amazon and a column representing success/ failure (+1/-1).



```
Product   Average Star Rating as per Amazon  Success
iPhone 4                                4.2        1
eCar                                    1.7        0
Win 10                                  3.9        1
eRobot                                  0.5        0
```

Figure 11: Product, Amazon Star Rating and Success Indicator.

In order to test our first hypothesis, Pearson Correlation Coefficient (r) will be used to measure the strength of a linear association between our aggregate sentiment scores and Amazon review star ratings on the other hand. Pearson Correlation Coefficient, r can take a range of values from +1 to -1. A value of zero will mean that there is no association between our two variables. A value of r greater than zero will indicate a positive association. That is, as the value of one variable increases, so does the value of our second variable. Similarly, a value of r less than zero indicates a negative correlation and will mean that, as the value of one variable increases, the value of the second variable decreases,

	Coefficient, r	
Strength of Association	Positive	Negative
Small	0.1 to 0.3	-0.1 to -0.3
Medium	0.3 to 0.5	-0.3 to -0.5
Large	0.5 to 1.0	-0.5 to -1.0

To test our second hypothesis, a machine learning (binary classification) model will treat our tweet corpus (in data frame format and relating to a separate technology product) as a bag of words. Twitter data will be converted into data frame for better text analysis. A set of data cleaning and the pre-processing activities defined under section (2.7) will be performed before converting the data frame into document term matrix (DTM) where terms will represent columns and their respective frequencies of occurrence as rows. The entire DTM rows will be combined into a single row (summary row) representing sum of term frequencies per column. The above mentioned steps will be performed with all data belonging to all the technology products being considered in the study and their resulting rows of DTM terms will be combined into a single DTM.

A binary classification model will treat these terms as features and their respective frequencies as values to be trained on the process of calculating success as an outcome. As such, the model will be classifying twitter sentiment as positive or negative. The binary classifier will learn by continuously accommodating the hyperplane to redefine a decision boundary by also including the newly fed data. Each row of data adds to the generalization capabilities of the model until it is able to generalize the overall behavior. A new data frame with another similar technology product will then be fed to the learned binary classier and asked to predict success of technology. The resulting prediction will be compared with the Amazon average star rating to calculate error which will be taken as the discrete difference between what the classifier is expected to predict and what it actually predicts. It is hypothesized that the trained binary classifier will be able to classify sentiments as positive or negative in a way that corresponds with online user reviews of technology.

4.1 Project Plan:

Given below is the updated project plan with progress update (%Complete).

Task #	Task Name	Duration	Start	Finish	% Complete
51	Complete Preliminary Literature Search & Review	0 days	Mon 12/4/17	Mon 12/4/17	70%
52	Identify Working Hypothesis / Theory	35 days	Mon 10/16/17	Fri 12/1/17	100%
53	Finalize Hypothesis/ Theory	1 day	Mon 12/4/17	Mon 12/4/17	100%
54	Define Preliminary Methodology	1 day	Mon 12/4/17	Mon 12/4/17	100%
55	Finalize Preliminary Methodology	1 day	Tue 12/5/17	Tue 12/5/17	100%
56	Compile Data Collection Instrument	15 days	Wed 1/17/18	Tue 2/6/18	100%
57	Collect Data	7 days	Wed 2/7/18	Thu 2/15/18	On-Demand
58	Analyze Data	15 days	Fri 2/16/18	Thu 3/8/18	30%

64	Complete Write up for Initial Chapters: Introduction	2 days	Fri 3/9/18	Mon 3/12/18	50%
65	Complete Write up for Initial Chapters: Literature Review	7 days	Tue 3/13/18	Wed 3/21/18	50%
66	Complete Write up for Initial Chapters: Working Theory	7 days	Thu 3/22/18	Fri 3/30/18	50%
67	Complete Write up for Initial Chapters: Proposed Methodology	7 days	Mon 4/2/18	Tue 4/10/18	50%
68	Write Chapters: Data Analysis, Findings & Conclusion	21 days	Wed 4/11/18	Wed 5/9/18	Pending
69	Submit and Present Interim Report and Poster	45 days	Thu 5/10/18	Wed 7/11/18	Poster Pending
70	Complete Dissertation Write up	45 days	Thu 5/10/18	Wed 7/11/18	Pending
71	Submit Complete Dissertation	0 days	Thu 7/12/18	Thu 7/12/18	Pending

References

1. FELDMAN, R. 2013. Techniques and applications for sentiment analysis. *Commun. ACM,* 56**,** 82-89.
2. GIACHANOU, A. & CRESTANI, F. 2016. Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Computing Surveys,* 49**,** 28-28:41.
3. STEPHEN, H. & REBECCA SCOTT (2017) 'Developing an Approach to Harvesting, Cleaning, and Analyzing Data from Twitter Using R', Information Systems Education Journal (ISEDJ), v15 n3 (May 2017), P42-54.
4. DANNEMAN, N., & HEIMANN, R. (2014). Social media mining with R. Packt Publishing Ltd.
5. DE VRIES, A. (2016) Text Analysis 101: Sentiment Analysis in Tableau & R. The Information Lab. Retrieved 30 May 2016, from http://www.theinformationlab.co.uk/2016/03/02/text-analysis-101-sentiment-analysis-in-tableau-r/.
6. BREEN, J. (2011). Mining Twitter for Airline Consumer Sentiment. Inside-r.org. Retrieved 30 May 2016, from http://www.inside-r.org/howto/mining-twitter-airline-consumer-sentiment.
7. Bryl, S. (2014). Twitter sentiment analysis with R. AnalyzeCore.com. Retrieved 30 May 2016, from http://analyzecore.com/2014/04/28/twitter-sentiment-analysis/.
8. CAROL, T., RACHEL, V. & DONALD, W. K. 2013. Social media and scholarly reading. *Online Information Review,* 37**,** 193-216.
9. KIRITCHENKO S, ZHU XD & M., M. S. 2014. Sentiment Analysis of Short Informal Text.
10. Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, pp. 168{177, New York, NY, USA. ACM.
11. WIEBE, J., WILSON, T., & CARDIE, C. (2005). Annotating expressions of opinions and emotions in language. Language resources and evaluation, 39 (2-3), 165-210.
12. OZA, K.S., NAIK, P.G., 2016. Prediction of online lectures popularity: a text mining approach. Procedia Comput. Sci. 92 (2016), 468–474.
13. DENIZ, A., COSAR AHMET (2017). 'Robust multiobjective evolutionary feature subset selection algorithm for binary classification using machine learning techniques', Neurocomputing, 241, pp. 128-146.
14. 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*.
15. ORTEGA, R., ADRIAN, F., and Andres Montoyo. 2013. SSA-UO: Unsupervised twitter sentiment analysis. In Proceedings of the 7th International Workshop on Semantic Evaluation - 2nd Joint Conference on Lexical and Computational Semantics (SemEval'13). Association for Computational Linguistics, 501–507.