Twitter Sentiment Based Mining for Decision Making using Text Classifiers with Learning by Induction

Odusote Babafemi, Itakpe Jonathan D., Afolabi Ibukun T.
Department of Computer and Information Sciences, Covenant University

Abstract. The amount of data residing in social media currently untapped is certainly limitless as millions of people are constantly posting a message or the other to public forums on the internet. Twitter being one of the largest social media networks with over 336 million monthly active users has proven to be a fertile ground for harvesting opinion from multiple people. This work explores how opinion can be extracted from tweets to discover people’s view concerning a certain subject matter. It focuses mainly on overcoming the limitation of the current approach to social media sentiment based mining for decision making which is that opinions derived from multiple sources are limited to available connections on the social media platforms and lack of improved accuracy of mined opinions. In order to achieve this, the proposed framework provides a platform to mine opinions from more than the available friends and connections on the social media platform and in addition, improve the quality of the opinion mined by implementing supervised learning algorithms with learning by induction in Twitter data analysis.

In this research, three different supervised machine learning algorithms were applied to a dataset curated by graduate students at Stanford in order to accurately classify tweets into either positive or negative sentiment based on its content. It was discovered that Maximum Entropy had the highest accuracy of 83.5% among the three algorithms. The research has provided a web application which would enable users such as CEOs, Market Analysts, and random users make quality decision based on others’ opinions.

1. INTRODUCTION

Data is everywhere, in every aspect of our daily lives. Data is also being shared with or without knowledge or consent. From sharing that post on Facebook, to ordering that item from Amazon or even just visiting a website, data is collected and stored but what exactly is done with that data is what actually matters.

Over the years, hundreds of companies have made use of data collected from their customers to make decisions and predictions; these large datasets have been tagged as Big Data. As defined by the Oxford Dictionary in [1], big data alludes to a great degree, extensive datasets that might be broken down computationally to uncover examples, patterns, and affiliations, particularly identifying with human conduct and communications.
The measure of data in our reality increments at an exponential rate, and can be seen in all parts of our day to day lives. This ranges from data about buyers caught by organizations to posts by people on online networking [2]. There is need to maximize the potential of the availability of this data. “The emergence of social media (Facebook, YouTube, Twitter and many others) has made the World Wide Web a lively and interactive platform or medium through which billions of individuals around the world interact, share, post and conduct numerous activities” [3]. This in turn has removed the boundaries between real world and virtual world making it possible for people to co-depend and interrelate between the two worlds. This can be very advantageous for organizations looking to expand their sources of data for predictions and informed decision making.

Twitter is a free to use micro-blogging platform that empowers clients to send and read short 140-character messages called "tweets" tweets [4][5]. According to statistics from Internet Live Stats [6], “Every second, on average, around 6,000 tweets are tweeted on Twitter, which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year”. Twitter allows developers access to 1% of this streaming data for research purposes [6]. Twitter can serve as an important source of data for organizations and individuals alike as users are allowed to express their minds basically on any subject in a concise format due to the 140 character limit of tweets. This seemingly endless stream of data if properly harnessed, as already demonstrated by a number of researchers and scientists could help in real world scenarios.

The innate desire to get others opinion about a subjects matter is very pervasive among humans. This includes reviews about products on websites, retweets on twitter, likes and shares on Facebook. The average human assumes that if people say it’s good, then it is just that or vice versa as the case may be. Before purchasing a product or even selecting a life insurance plan, most people make relatively informed decisions based on what other ‘more informed’ people who must have been in their exact situation at a point in time did.

Organizations approach the issue of opinion gathering in a fairly similar method, by sending out marketers/market researchers to ask the users on a one-on-one basis or through the use of online opinion polls. More often than not, the opinions gathered are fairly limited to individuals known to the decision maker, or the few users responding to online polls by the organization. The use of social media has made this circle a bit broader but still relatively small when compared to the 336 Million monthly update active users on twitter constantly giving their opinions on a wide range of topics via tweets.

More often than not, businesses find it relatively difficult to know what the masses think about a product or service that is being pushed to market by the company. Most of such businesses revert to sending out marketers/market researchers into the cities to ask questions from people (via direct interviews or questionnaire), the use of focus groups, use of ethnography or the creation of online opinion polls and spamming mailboxes of their customers with links to such polls [7]. These techniques of gathering opinion can be very time-consuming and frustrating for most business owners. The problem now being faced by the market experts is the collection of a bigger proportion of opinions from users of various backgrounds with the added benefits of ease of use and ultimately a faster opinion gathering technique.

The advent of social media and the ability to store large sets of data has made it possible to extract useful information which could be opinions from social media data, hence the term “Social Media Mining” – (SMM) [3]. “Social Media Mining is the process of representing, analyzing, and extracting actionable patterns from social media data. We are now faced with an exacerbated issue of big data – drowning in data, but thirsty for knowledge” [3], that is we are surrounded by so much data yet we have extracted little or no useful information (Knowledge) from it.

One of the most important technique in Social Media Mining is Sentiment-Based Mining. Sentiment-based mining, also known as “opinion mining”[8] is the field of study that breaks down individuals' assessments, feelings, evaluations, states of mind, and opinions towards elements, for example, items, administrations, associations, people, issues, occasions, themes, and their properties[9]. Therefore, sentiment-based twitter data analysis is the categorization of tweets based on
their contents as positive, neutral or negative with respect to a particular entity. [10]. Decision Support Systems are interactive, computer-based systems that help user in judgment and choice of activities. Decision Support Systems provide data storage and retrieval but enhance the traditional information access and retrieval function with support for model building and model-based reasoning [11]. Decision Support Systems has been used in various areas such as management and planning in business, the health care sector, the military, and also areas where management will encounter complex decision situations. Decision support systems are mainly used for tactical and strategic decisions which are faced in situations where the time taken to think through and model the challenges is required to be short.

Currently, the limitation of the current approach to sentiment based mining for decision making is that it involves the gathering of opinion from multiple sources which are limited to available connections on the social media platforms, this miniature fraction of the population can hardly serve as a justifiable basis to make a decision. Also previous methods for opinion/sentiment analysis have done little in implementing supervised learning algorithms with learning by induction in Twitter data analysis. This paper therefore aims at exploring the use of supervised learning algorithms such as Naive Bayes, Maximum Entropy and Support Vector Machines with learning by Induction for the purpose of sentiment based mining for decision making.

2. LITERATURE REVIEW

A lot of work and research has gone into extracting meaningful information from twitter data, from forecasting of events by tweet data mining. Few of such work was carried out by [12], where he analyzed quantitative characteristics of frequent sets and association rules in tweets relating to various events, to using Twitter to predict football outcomes, [13] built predictive models based on tweets about the English premier league over a period of 3 months. Others include predicting events such as outcomes of elections [14], stock market prices [15], box-offices revenues of movies before release [16] and the spread of diseases [17], amongst others. Following the recent work which evaluated Twitter messages with influenza rates in the United States, Michael and Mark considered a wider range of public health applications for Twitter. They applied the recently presented Ailment Topic Aspect Model to more than 1.5 million wellbeing related tweets and found mentions of over twelve diseases, including hypersensitivities, weight and a sleeping disorder. They introduced expansions to incorporate earlier learning into this model and apply it to a few assignments: following diseases over times (syndromic observation), measuring behavioral danger components, limiting diseases by geographic district, and breaking down indications and medicine use. They demonstrate quantitative relationships with general wellbeing information and subjective assessments of model yield. The outcomes recommend that Twitter has wide relevance for general wellbeing research [17]. The researchers in [18] noticed that competition existed among Indonesian mobile phone providers to gain new clients through adverts, particularly on social networks. The issue rises on the most capable strategy to gauge the brand reputation of these suppliers taking into account singular response on their service quality. The paper addressed this issue by measuring brand reputation considering purchaser devotion through client's sentiment analysis from Twitter data. Test model is collected and extricated from 10,000 unrefined tweets from January to March 2015 of main three mobile phone suppliers in Indonesia (PT XL AxiataTbk, PT TelkomselTbk, and PT IndosatTbk). They analyzed a couple of feature extractions, algorithms, and the classification schemes. After data cleaning and data adjusting, the sentiments are grouped and analyzed utilizing three unique algorithms: Naïve Bayes, Support Vector Machine, and Decision Tree classifier method. [14] used Utilizing Linguistic Inquiry and Word Count (LIWC) content analysis programming on 104,003 tweets distributed in the weeks paving the way to the government race of the national parliament in Germany. Their research showed that twitter can be used for political consultation as analysis of the tweets' political sentiment showed close representation to the parties' and politicians' political positions demonstrating that the substance of
tweets practically mirrors the real political scene. [19] examined how twitter could be used to analyze public opinion, and therefore predict presidential election results. The research by [12] examined quantitative attributes of frequent sets and association rules in the posts of Twitter microblogs identified with various events discussion. For the analysis, they utilized a theory of frequent sets, association rules and a theory of formal concept analysis. They uncovered the frequent sets and association rules which describe the semantic relations between the concepts of analyzed subjects. The backing of some frequent sets comes to its global maximum before the expected occasion yet with some time delay. Such frequent sets might be considered as predictive markers that portray the importance of expected events for blogosphere users. [13] carried out a research to study whether data mined from Twitter can be utilized for the purpose of designing a predictive model for football (Soccer). They were able to develop a set of predictive models for the result of football games of the English Premier League for a 3 month period.

Some of the tool kits used in mining from social media data include but are not limited to the following: DatumBox Machine Learning Framework (http://www.datumbox.com/machine-learning-framework/), Waikato Environment for Knowledge Analysis (WEKA) [20], RapidMiner (https://rapidminer.com/), Sentiment140 [21], AlchemyAPI (https://www.ibm.com/watson/alchemy-api.html) and so on.

3. METHODOLOGY
This section gives an explanation of methods used in this research. 

Blinding applying data-mining methodologies (also known as data dredging in statistics) can be a dangerous activity, often ending in the discovery of meaningless and invalid patterns according to [22]. In this research, the machine learning techniques that was used for sentiment based mining for decision making includes; Naïve Bayes, Maximum Entropy, Support Vector Machines.

Sentiment analysis is a natural language processing research problem that involves the detection of positive, negative or neutral sentiments from people’s opinion, which in this case are tweets. An Opinion is a quintuple as stated by [9];

\[(E_i, A_{ij}, S_{ijkl}, H_k, T_l)\]

Where:

- \(H_k\) – The person with the opinion.
- \(E_i\) – Name of an entity of which the opinion holder (\(H_k\)) is referring to.
- \(A_{ij}\) – Specific attribute of the entity (\(E_i\)) in question.
- \(T_l\) – The specific time which the opinion was given.
- \(S_{ijkl}\) – The sentiment associated with the opinion given by the holder (\(H_k\)) about the attribute (\(A_{ij}\)) of an entity (\(E_i\)) at a given time (\(T_l\)).

The problem of sentiment analysis is similar to people’s opinions and how they are expressed. The use of lexicons has been the go to approach for Sentiment analysis, lexicons could be taken to mean the collection of sentiments that generally represent positive or negative opinion. Positive sentiments could be words or phrases such as lovely, exciting, very good while negative ones could be terrible, disgust. Such words are then examined through the portion of text to determine if they are positively or negatively inclined sentimental-wise, this technique however do not work for texts which are sarcastic in nature, and a plethora of others. The above algorithms were implemented using Python programing language together with Python Machine Learning Packages such as SciKit-Learn, NumPy, SciPy, Pandas, Flask and PyCharm.

The implementation process of this research work is the continual repetition of three steps which includes; Gathering the data, in this case recent tweets from twitter; Perform the mining operation using the specified machine learning techniques and the result presentation. Figure 1 gives the overall architecture of the methodology of the developed system.
3.1. Data Description
The training set of tweets in this study were extracted from the [21] Sentiment140 project work. The Distant Supervision method was used in gathering the tweets over a 3 month period and it serves as one of the most reliable source of Twitter training Corpus. It is made up of a total of 1.6 million tweets shared evenly among the two polarity types (Negative & Positive).

3.2. Data Processing
Data processing also known as preprocessing in the Text mining sphere refers to the removal of noise from text. This is performed before the actual classification or learning takes place, this is essentially important especially for text that is unstructured. Some preprocessing techniques applied are:

a. Removal of duplicates: All repeated tweets or retweeted tweets (RT) are removed.

b. Case Folding: All words are converted to lowercase.

c. Removal of Punctuation and Special Characters: Web Addresses, RT, punctuations, usernames are removed.

d. Stopword Removal: Common English words known as stopwords for example (‘and’, ’the’, ’us’, etc.) are removed from the text.

3.3. Feature Selection (n-grams)
During the feature selection, the technique of selecting n-grams was used. An n-gram is a contiguous succession of n things from a given text or speech. The things can be phonemes, syllables, letters, words or base sets as indicated by the application. The n-grams commonly are gathered from a text or speech corpus. At a point when the things are words, n-grams may also be called shingles. An n-gram of size 1 is alluded to as a "unigram"; size 2 is a called "bigram"; size 3 is a "trigram". Bigger sizes are...
here and there alluded to by the estimation of n, e.g., "four-gram", "five-gram", and so on. In this system, the use of both ‘Unigrams’ and ‘Bigrams’ were employed.

### 3.4. Classification
In this classification step, Naïve Bayes, Maximum Entropy, Support Vector Machines algorithms were modeled with the same data source. 1.6 million Tweets were used to build the classifier.

### 3.5. Classifier Evaluation
This is the process of determining the accuracy of the classifier by splitting the initial test set into both test set and training set hereby calculation how accurate the classifier is based on how much of the data was classified correctly.

### 3.6. Store Classifier
The model of the classifier is stored in a file such that when it would be used, it doesn’t have to be trained again.

### 3.7. Use Case Narratives of the developed system
Table 1 – Table 3 give the use case narrative of the developed system which include User authentication, User Query system and used extending the training set respectively.

#### TABLE 1: User Authentication

| Use Case 1 | User Authentication |
|------------|---------------------|
| Brief Description | This singular module is for getting access into the system by the user. |
| Actor(s) | Users |
| Flow Of Events | **Basic Flow:** The use case begins when the user accesses the webpage  
1. The user enters the URL to the page.  
2. The user grants the application access to his/her Twitter account.  
3. System Displays Homepage  

**Alternative Flow:** If Twitter Access is denied, user is not granted access to the System. |
| Level | User use case |
| Parameters | **Input:** Twitter Login details via Twitter API  
**Output:** The system’s Homepage |
| Pre-Conditions | All users must:  
1. Have a valid Twitter account.  
2. Have a working Internet connection. |
| Post-Conditions (Success End) | If use case is successful, user is granted access to the System. |
| Post-Conditions (Failed End) | If use case is not successful, access is not granted |
| Trigger | Loading the URL into the web browser |
| Extension points | None |

#### TABLE 2: User Query System

| Use Case 2 | User Search Query |
|------------|-------------------|
| Brief Description | This module is for receiving user query and displaying output. |
| Actor(s) | Users |
Flow Of Events | Basic Flow:  
The use case begins when the user is successfully authenticated and has accessed the Homepage  
1. User enters search query (Person, Place, Product or Company)  
2. The user can decide to give advanced options.  
3. System displays results.  
Alternative Flow:  
If Query is not found, returns an error page.

Level | User use case  
Parameters | Input: Search Query (Person, Place, Product or Company)  
Output: The Result in form of Charts, Graphs  
Pre-Conditions | The user must have been successfully authenticated.  
Post-Conditions (Success End) | If use case is successful, result is displayed to the user.  
Post-Conditions (Failed End) | If use case is not successful, an error page is returned.  
Trigger | Entering the search term  
Extension points | None

### TABLE 3: User Extending Training Set

| Use Case 3 | User Extends training set.  
| Brief Description | This module gives the user the ability to extend the training set.  
| Actor(s) | Users  
Flow Of Events | Basic Flow:  
The use case begins when the user has received the result of the query.  
1. User gets the result of the query.  
2. The user can decide to correct the classifier in case of a wrong classification.  
3. System displays the tweets and the user can reclassify the wrong ones.  
Alternative Flow:  
If Query is not found, returns an error page.

Level | User use case  
Parameters | Input: Correct sentiment for the tweet  
Output: Success message  
Pre-Conditions | The user must have been successfully authenticated and sent in a query.  
Post-Conditions (Success End) | If use case is successful, a success message is sent to the user and is asked to check back in 24 hours.  
Post-Conditions (Failed End) | If use case is not successful, an error page is returned.  
Trigger | Entering the correct sentiment for a tweet  
Extension points | None

### SYSTEM DESCRIPTION

The developed systems has been able to provide users easy to understand abstracts of information mined from the tweets, accurately classify the tweets as positive or negative and provide users with an improved approach to sentiment based mining by allowing users to retrain or correct the classifier using learning by Induction. The following are some of the interfaces of the webapplication developed as proof of concept for the sentiment based mining for decision making.
When a user logs into the system, the Twit Sense landing page states the fact that the user needs to log in to twitter in order to access the application. After this hurdle is crossed, then figure 2 is revealed, which is search page of the application. This interface enables the users to input the query they are interested in getting opinion, for example Person, Product or Brand.

![Twit-Sense](image)

**Figure 2:** The Homepage / Search Page of the Application

The interface in figure 2 is followed by an interface which enables the user to set desired options such as the number of tweets to collect, the classifier to use, and various tweet preprocessing options. Figure 3 then displays the result of the previously given query which is a view containing a couple of charts and a table showing the final set of tweets used for the classification along with their weights.

![Result page with Bar chart showing the ratio of positive and Negative Tweets for the search query “Samsung Galaxy S7”.](image)

**Figure 3:** Result page with Bar chart showing the ratio of positive and Negative Tweets for the search query “Samsung Galaxy S7”.

User can also have access to the Word Cloud showing the most popular words used alongside a search criteria and a cross section of the tweets used for classification and their corresponding weights. Finally, since the classifiers used in this research are not 100% accurate, hence the results are prone to
errors and the users have the opportunity to correct the wrong classifications in order to improve the prediction accuracy. The interface responsible for this is displayed in figure 4. The classifier can then be retrained using the new data once every 24 hours.

![Modal form for correction of wrongly classified tweets](image)

Figure 4: Modal form for correction of wrongly classified tweets

5. CONCLUSION AND FUTURE WORK

This study is of immense significance because it gives us valuable insight to the amount of knowledge that can be extracted from the seemingly endless streams of data available from social media. The study also covers the use of various data mining and machine learning algorithms, both supervised and unsupervised learning, also taking into consideration the use of clustering, classification, feature selection and feature extraction.

The data mining approach to opinion gathering as shown in this project work is highly innovative, and is of tremendous benefit to the computing world, market researchers, CEOs and the regular person in search of people’s opinion. The project has demonstrated that valuable insights can be gotten from unstructured text, in this case from twitter. It also serves as a launching pad for building a decision making system that is based entirely on what people or consumers think or their opinions. The developed system will also provide an ease, reduction in time and cost, in the approach used by market researchers and CEOs in order to get specific, unbiased opinion on products through the social media platform. For further work, the field of Sentiment analysis and twitter mining still has a lot of issues that are not completely resolved which includes the following; integration of social media text across other platforms such as Google+, Facebook, Instagram and YouTube and so on. Also there is need for further text preprocessing such as sarcasm Detection, slang detection, key phrase extraction and so on.

References
[1] Oxford University. (2000). *Oxford English Dictionary*. Oxford: Oxford University Press.
[2] McKinsey Global Institute. (2011). Big data: The next frontier for innovation, competition and
productivity. McKinsey Global Institute.

[3] Zafarani, R., Abbasi, M. A., & Liu, H. (2014). Social Media Mining: An Introduction. Cambridge: Cambridge University Press.

[4] Kwak, H., Lee, C., Park, H. and Moon, S. (2010) ‘What is Twitter, a social network or a news media? Proceedings of the 19th International Conference on World Wide Web, 26–30 April, Raleigh, NC, doi:10.1145/1772690.1772751

[5] Afolabi, I.T., Ezenwoke, A.A., Ayo, C.K.(2017) Competitive analysis of social media data in the banking industry in International Journal of Internet Marketing and Advertising 11(3), pp. 183-201

[6] Twitter Developers. (2015, December 16). The Streaming APIs | Twitter Developers. Retrieved from Twitter: https://dev.twitter.com/streaming/overview

[7] Myers, C. (n.d.). Consumer Behavior Research Methods. Retrieved from Chron: http://smallbusiness.chron.com/consumer-behavior-research-methods-70706.html

[8] Pang, B., & Lee, L. (2008). Foundations and Trends® in Information Retrieval: Vol. 2. In B. P. Lee, Opinion Mining and Sentiment Analysis (pp. 1-135). Cornell: Now Publishers

[9] Liu, B. (2012). Web Data Mining. In B. Liu, Data-Centric Systems and Applications (pp. 459-526). Berlin Heidelberg: Springer.

[10] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment Analysis of Twitter Data. Proceedings of the Workshop on Language in Social Media (LSM 2011) (pp. 30-38). Portland, Oregon: Omnipress, Inc.

[11] Wu, Y., and Chan, C. W. (2009). A data analysis decision support system for the carbon dioxide capture process. Expert Systems with Applications, 36(6), 9949–9960

[12] Pavlyshenko, B. (2013). Classification analysis of authorship fiction texts in the space of semantic fields. Journal of Quantitative Linguistics, 20(3), 218–226.10.1080/09296174.2013.799914

[13] Kampakis, S., & Adamides, A. (2014). Using Twitter to predict football outcomes. arXiv. Retrieved from http://arxiv.org/abs/1411.1243

[14] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, (pp. 178-185). Munich.

[15] Bollen, J., Mao, H., & Zeng, X.-J. (2011). Twitter mood predicts the stock market. Journal of Computational Science (pp. 1-8). arXiv. doi:10.1016/j.jocs.2010.12.007

[16] Asur, S., & Huberman, B. A. (2010). Predicting the Future with Social Media. CoRR. doi:10.1016/j.japenergy.2013.03.027

[17] Paul, M. J., & Dredze, M. (2011). You are what you Tweet: Analyzing Twitter for public health. ICWSM, 265-272

[18] Vidya, N. A., Fanany, M. I., & Budi, I. (2015). Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers. The Third Information Systems International Conference (pp. 519-526). Elsevier.

[19] Shi, L., Agarwal, N., Agrawal, A., Garg, R., & Spoelstra, J. (2012). Predicting US Primary Elections with Twitter.

[20] Hall M. Eibe Frank, Holmes G., Pfahringer B. (2008). The WEKA Data Mining Software: An Update in SIGKDD Explorations 11(1) 10-13

[21] Go, A., Bhayani, R., & Huang, L. (n.d.). Twitter Sentiment Classification using Distant Supervision.

[22] Fayyad, U., Piattetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Database. American Association for Artificial Intelligence, (pp. 37-54). Rhode Island.