Digital Innovations & Contemporary Research in SCIENCE, ENGINEERING & TECHNOLOGY
Vol. 5 No. 3. September 2017

Article Citation Format
A.B. Adeyemo & E.O. Durodola (2017). A Text Mining Tool for Sentiment Analysis of Tweets using Machine Learning Techniques. Journal of Digital Innovations & Contemp Res. In Sc., Eng & Tech. Vol. 5, No. 2. Pp 21-.30

Article Progress Time Stamps
Article Type: Research Article
Manuscript Received: 29th July, 2017
Review Type: Blind
Final Acceptance:: 24th September, 2017
DOI Prefix: 10.22624

A Text Mining Tool for Sentiment Analysis of Tweets using Machine Learning Techniques

A.B. Adeyemo & E.O. Durodola
Department of Computer Science
University of Ibadan
Ibadan, Nigeria
sesanadeyemo@gmail.com>

ABSTRACT

Sentiment Analysis is a subfield of text mining which aims at determining the attitude of a speaker or writer with respect to some topic or the contextual polarity of a document. This study implements an efficient model for determining the sentiments expressed in texts. A software tool that identifies the sentiment polarity of customer tweets was developed using machine learning techniques. The application was used to predict the sentiments expressed in the tweets and results obtained can be used for strategic management decisions.

Keywords: Sentiment Analysis, Text Mining, Machine Learning Techniques.

1. INTRODUCTION

Sentiment Analysis is a task of text mining. It is the computational study of people's opinions, attitudes and emotions toward an entity (Pang et al., 2002). The entity can represent individuals, events or topics. These topics are most likely to be covered by reviews. It is a sub field of Text Mining and Natural Language Processing. Sentiment Analysis identifies the sentiment expressed in a text then analyzes it. Therefore, the target of sentiment analysis is to find opinions, identify the sentiments they express and then classify their polarity (Walaa et al, 2014). Sentiment analysis is a prominent and active area of research, spurred particularly by the rapid growth of web social media and the opportunity to access the valuable opinions of numerous participants on various business and social issues particularly in the developing countries where the increasing use of mobile devices and technologies has made it easy for people to have access to Internet based services (Ghiassi et al., 2013).

Sentiment analysis has mostly been performed for product reviews (including movie reviews, hotel reviews), forums, blogs, micro-blogs ("tweets"), news articles and social media in order to disclose the writer's opinions, attitudes and emotions toward individuals, events or topics (Pang & Lee 2008). Sentiment Analysis can be considered a classification process. There are three main classification levels in sentiment analysis:

Document-level: which aims to classify an opinion document as expressing a positive or negative opinion or sentiment. It considers the whole document as a basic information unit, that is, focused on one topic.
Sentence-level: which aims to classify sentiment expressed in each sentence. Its intended to identify whether the sentence is subjective or objective. If the sentence is subjective, Sentence-level sentiment analysis will determine whether the sentence expresses positive or negative opinions. Aspect-level: which aims to classify the sentiment with respect to the specific aspects of entities. It intends to identify the entities and their aspects. Opinion holders can give different opinions for different aspects of the same entity like this sentence "The restaurant's food tastes good but their customer service is poor".

Researches in sentiment analysis has focused on document-level classification of overall sentiment that distinguishes positive from negative reviews (Kang and Park, 2014). Sentiment analysis is a challenging interdisciplinary task which includes natural language processing, web mining and machine learning. It is a complex task which comprises of subjectivity classification, sentiment classification, opinion holder extraction and object /feature extraction. Sentiment classification techniques can be divided into: machine learning approach, lexicon based approach and hybrid approach which combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods.. The main problem with the lexicon-based approach lies in cross-domain adaptability. Words carrying a positive connotation in one domain may be negative or neutral in another. Thus, the lexicon has to be adapted to the domain. The machine learning approach (used in this study) requires a lot of human effort in document annotation and a good match between the training and testing data with respect to the domain (Taboada, 2016).

The text classification methods using machine learning approach can be roughly into supervised and unsupervised learning methods. In supervised learning, computer programs capture structural information and derive conclusions (predictions) from previously labeled examples (instances, points). Supervised learning classifiers are categorized into: Probabilistic classifiers (examples are Naïve Bayes classifiers and Maximum Entropy classifiers), Linear classifiers (examples are Support Vector Machine and Neural Network) and Rule based classifiers. The most popular unsupervised learning classifiers are clustering algorithms (examples are Kmeans and Self Organizing Feature Maps).

Twitter is one of the most popular micro blogging platforms on the Internet. Twitter which was launched in 2006 now plays an important role in the research of social network. People share their preferences on Twitter using free-format, limited-length texts, often called "tweets" which provide rich information for organizations that want to know public opinions about their products or services (Ghiassi et al., 2013). Researchers have developed various approaches to monitor tweets in real-time for the occurrences of major events, the outbreak of news stories, and the reactions of the users to events (Benhardus and Kalita, 2013; Bifet and Frank, 2010). Majority of researches in the area of Twitter research have employed sentiment analysis approaches to identify and evaluate the opinions of users expressed in their tweets. Tweets are very short units of text, at maximum 140 characters long, and characterized by casual, compact language with extensive usage of slang, abbreviations, acronyms, and emoticons. Tweets also contain hashtags, user references (mentions), and embedded links to other websites containing additional referenced information, further complicating the sentiment analysis. Given the character limitations of tweets, classifying the sentiment of twitter messages is similar to sentence-level sentiment.

Sentiment Analysis by analyzing the social media has become an alternative to carrying out user surveys, and it is being postulated that it could even be more effective than user surveys. Research works on applications of sentiment analysis include early works in this area by Turney (2002) and Pang et al., (2002). More recent works include those of Kouloumpis et al. (2011), Agarwal et al. (2011), Ghiassi et al. (2013), Vasu (2013), Jafari Asbagh et al. (2014), Afful-Dadzie et al. (2014), Kang and Park (2014), Adeyemo and Ojo (2014), Gamallo and Garcia (2014), Tutubalina et al. (2015), Kolchyna et al. (2015) and Fersini et al. (2015). In study a text mining tool (software) has been developed for sentiment analysis of tweets using machine learning techniques.

2. MATERIALS AND METHODS

The text mining model implemented in the software developed in this study used both clustering and classification techniques of machine learning for the sentiment analysis task. The clustering technique was used as a pre-processor for the classification task since the data was generated in a raw unlabeled format. The classification technique was used for identifying the polarity of the text. K-means algorithm was used for the clustering phase while Support Vector Machine and Naïve Bayes algorithm was used for the classification phase. The software tool developed was implemented in the Python language which supports natural language processing using the PyCharm community edition IDE.

The data used for the study was collected from Twitter (a social media) using the Twitter Streaming API. The Streaming APIs give access to (usually a sample of) all tweets as they published on Twitter. On average, about 6,000 tweets per second are posted on Twitter and normal developer users get a small proportion (about 1%) of the tweet. The Streaming APIs (which only sends out real-time tweets) is one of the two types of Twitter APIs available. The other one is called REST APIs is more suitable for singular searches, such as searching historic tweets, reading user profile information, or posting Tweets. In order to retrieve tweets, a user needs to have a Twitter account and obtain credentials (API key, API secret, Access token and Access token secret) on the Twitter developer site to access the Twitter API (Crockford, 2006). The Python Twitter Tools library was used to connect to TwitterAPI to download the data from Twitter. The tweets used for the study were about MTN, a telecommunications firm operating in Nigeria. The MTN Group, formerly M-Cell is a South Africa-based multinational mobile telecommunications company, operating in many African, European and Asian countries. MTN Nigeria is part of the MTN Group (MTN Nigeria, 2015). The following code was used to extract tweets relating to MTN in English language.

```
iterator = twitter_stream.statuses.filter(track="mtn", language="en")
```

The data returned by the twitter streaming API is a set of documents, one per tweet, in the JavaScript Object Notation (Crockford, 2006). These documents, apart from the text of the tweet, contain additional data like: tweet information (date, source of tweet, type); user information (profile, location and counters for favourites, friends, and followers); entities mentioned in the tweet text (urls, hashtags and user, among other information). The JSON tweet document contains attributes describing the tweet, user information, tweet relations with other tweets, a lists of urls, hashtags and user mentions contained in the tweet. In some cases information related with the location of the user is also provided in the document (Crockford, 2006). The dataset used for the study contained 17,824 tweets which were collected over a period of three months from July 7, 2016 to September 9, 2016. After filtering and processing the dataset, 6046 tweets were found to be meaningful and were used for the study

2.1 Sentiment Analysis Model

The Machine Learning approach for text classification was used for the sentiment analysis task. It is a supervised algorithm that analyses data that were previously labeled as positive, negative or neutral; extracts features that model the differences between different classes, and infers a function that can be used for classifying new examples previously unseen. Figure1 presents the general process involved in sentiment analysis.

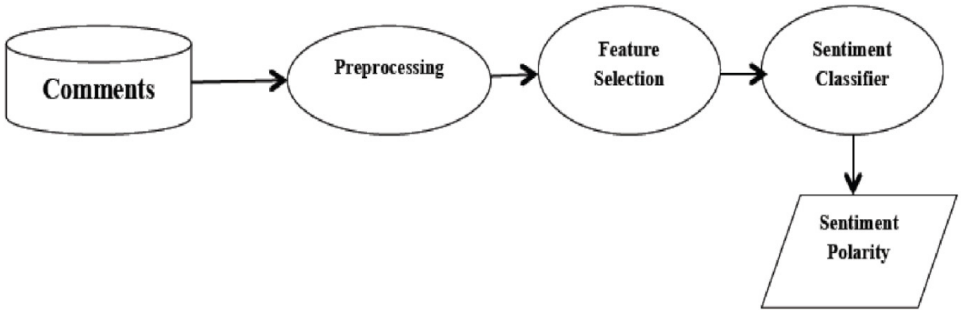

Figure 1: Generic methodology for sentiment analysis

The data generated is in text format. This may be product reviews, feedbacks, tweets etc. The data can be saved in different formats (csv, txt, arff, etc.). In this study, tweets are used as input to the system. They contain entities such as urls, hashtags, retweet (RT) and slangs which are noise to the system. The tweets need to be preprocessed to convey meaning to the system and help in proper analysis of the data. The preprocessing stage removes these entities and performs tasks such as removing stop-words (such as the, a, of, etc.) and stemming (converting words to their root form such as 'ate' being converted to 'eat'). The Feature Selection stage extracts relevant features from the preprocessed data and converts them into vectors. The Sentiment Classifier stage implements the classification algorithm used in the model. The vectors are passed into the trained model for predictions which are made by generating the polarity of the input (tweet).

The sentiment analysis model developed in this study uses both unsupervised and supervised machine learning techniques. The unsupervised clustering technique was used first because the tweets were generated in the raw format and were not labeled. For sentiment classification task, both the training and testing dataset needs to be labeled. Instead of manually labeling the dataset which is time-consuming and laborious, the K-means clustering algorithm was used to cluster and label the dataset. The labeled dataset was then divided into training set and testing set in the ratio 75% for training set and 25% for testing set. The training set goes through the preprocessing and feature extraction stage. Two supervised machine learning classifiers: Support Vector Machine and Naïve Bayes algorithm were trained using the features generated. The unlabeled test dataset were then passed into the training model for prediction purposes by determining the sentiment polarity. Figure 2 presents the schematic diagram of the sentiment analysis model.

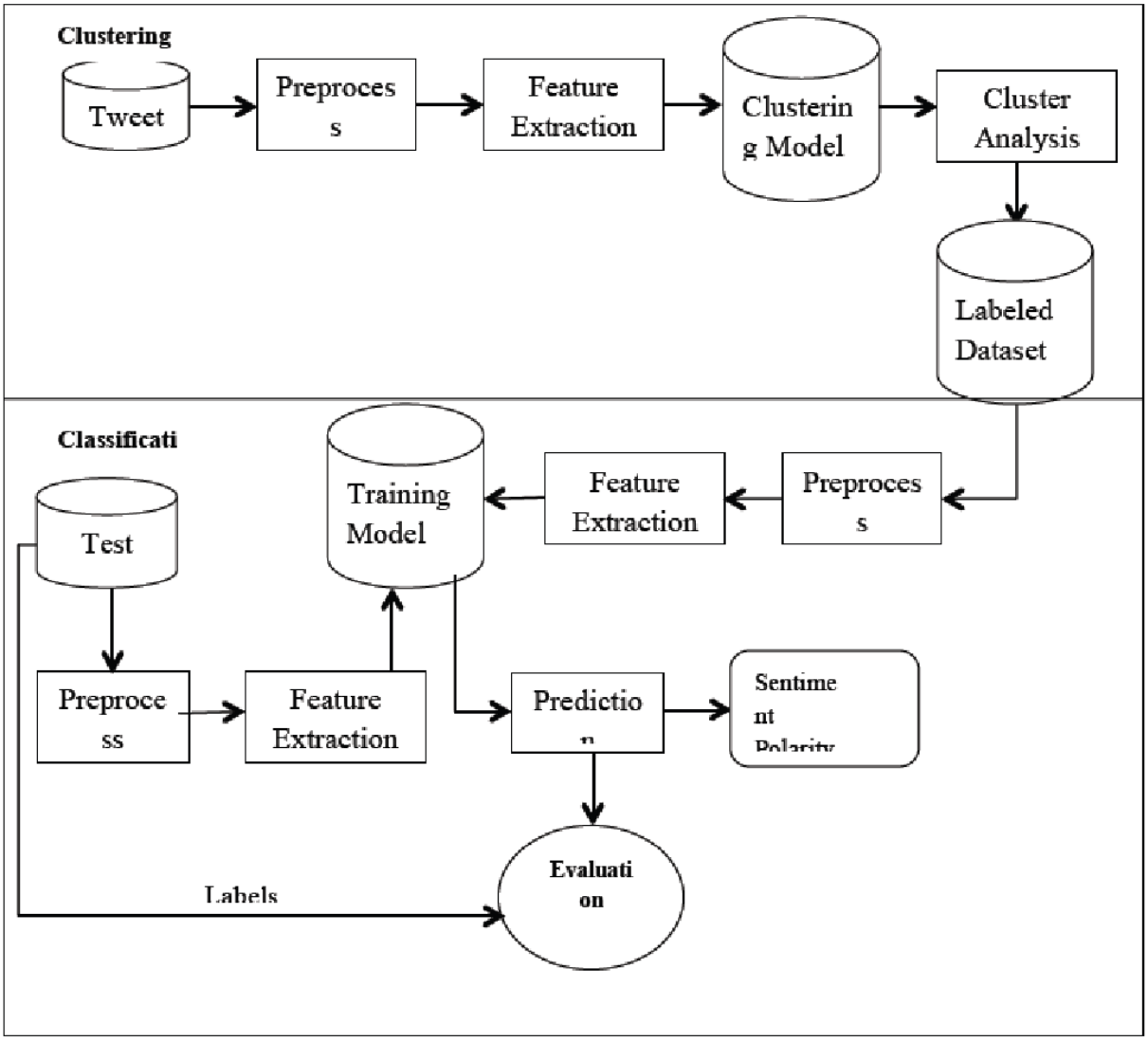

Figure 2: Developed Data processing model for Sentiment Analysis.

2.2 Sentiment Analysis Software Architecture

Figure 3 presents the Sentiment Analysis Software Architecture. The system architecture has three components. It was designed using the Django framework (Django, 2016) that supports the Python language. The First component is the web interface which accepts input text from the user. The Second component is the text mining engine which performs the preprocessing of text, feature extraction and classification. The Text mining engine was designed using the NLTK and SCIKIT-LEARN which are Python modules. The Third component is the view or interface that displays the result of the analysis.

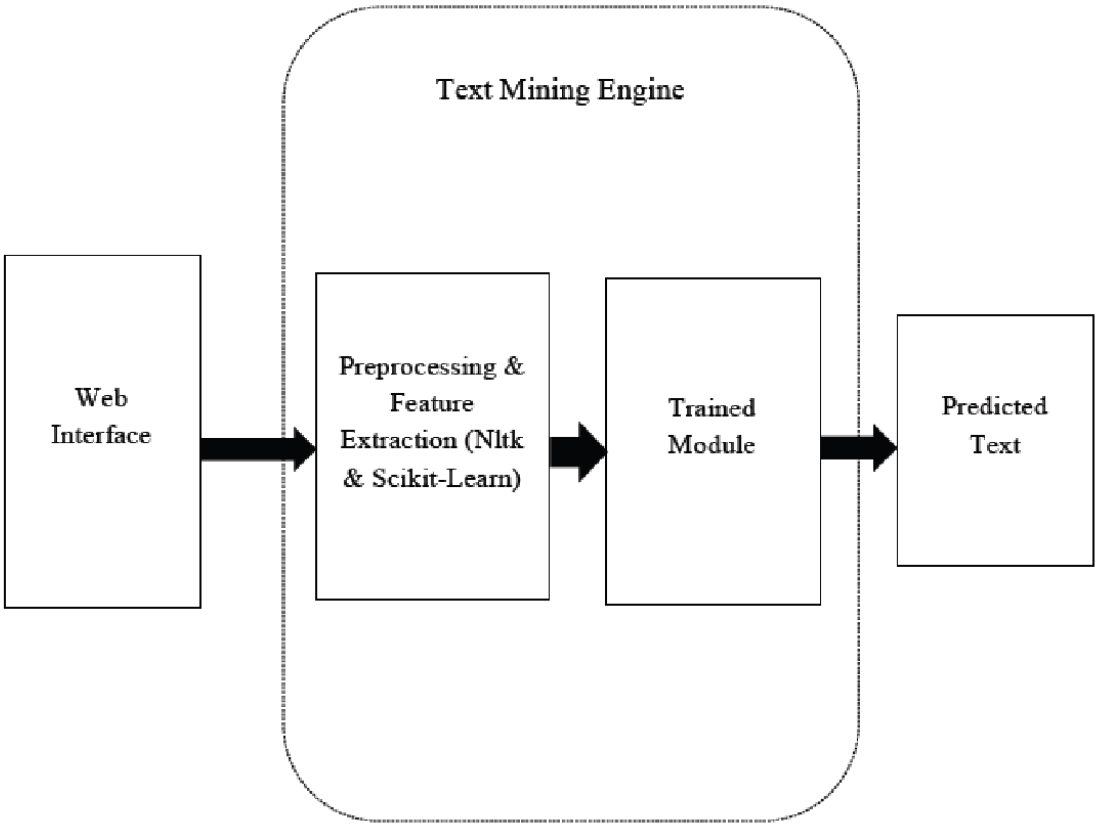

Figure 3: Sentiment Analysis Software Architecture.

The text mining software developed for sentiment analysis from tweets is called Sentalyzer. It is menu driven and has user friendly features. The user enters the tweets to be analyzed and clicks on the 'Analyze' button, and the user is then redirected to the result page which displays the class the input text belongs to. The application is able to predict the class the input text belongs to, based on a trained model. The trained model was achieved by learning a classifier on labeled tweets whose classes were obtained during the cluster analysis phase of the processing. Cluster analysis grouped the tweets into three classes. The PyCharm IDE was used for the cluster analysis. The following performance metrics were used to evaluate the performance of the clustering and classification algorithms that were implemented in the software: The time to build the model, the Precision, Recall, F-measure and Accuracy.

3. RESULTS AND DISCUSSION

In the pre-processing phase (tokenization) the tweets were transformed into a suitable representation for either clustering or classification task. Common twitter slangs (rt, hashtags, url, etc.) and stop words were removed to aid analysis. In the feature extraction phase both word frequency and inverse document frequency parameters were used in extracting terms. A comparative analysis was performed using unigrams, bigrams or trigrams during the clustering phase. The unigram model proved to more effective because the result it yielded proved to be more meaningful compared to other models. In pruning, words that appeared in less than 5% of all tweets as well as those words that appeared in more than 20% of all tweets were ignored.

In the clustering phase the dataset was clustered using the K-means algorithm. The clustering procedure was used to create three clusters. Cluster 1 contained 3279 tweets, Cluster 2 contained 1528 tweets while Cluster 3 contained 1239 tweets. The extracted terms in the three clusters are presented in table 1.

Table 1: The Three Clusters with their extracted terms

Cluster 1	Cluster 2	Cluster 3
Data	Airtel	Data
Apologize	Etisalat	Airtel
Dear	Champion	Plans
Caution	Nigeria	Airtime
Disappearing	Airtime	Etisalat
Chat	Data	

For Cluster 1 (Class one), the important features extracted contained terms such as AIRTEL, champion, airtime, Nigeria, data. The cluster revealed customers tweets comparing MTN services with other network service providers such as AIRTEL, ETISALAT etc. Customers complained about excessive call rates and there was support or appraisal for the MTNMobile money plan. News on MTN introducing 4G, listing on the Nigeria Stock Exchange was mentioned. It was noticed that MTN has a strong presence on twitter through prompt response to customers' tweets and there was also the giving out of airtime to lucky customers.

For Cluster 2 (Class two), the important features extracted contained terms such as data, apologize, disappearing. The cluster revealed customers tweets complaining about poor internet experience, data disappearing despite not surfing the internet. MTN's customer care representatives apologized to customers on these complaints.

For Cluster 3 (Class three), the important features extracted contained terms such as plans, airtime, etisalat. The Cluster revealed promotions on MTN data bundles such as night surfing plans, free twitter data etc. Customers complained about MTN bombarding their phones with network messages. They compared ETISALAT internet plan with MTN's in terms of ETISALAT being more expensive and better compared with MTN.

Classification was carried out to determine the sentiment polarity of incoming or new tweets. The two classifiers (Support Vector Machine and Naïve Bayes) were used. The Support Vector Machine classifier gave the better result due to its accuracy in prediction. Table 2 presents the summary of the performance of both classifiers. Tables 3 and 4 presents the confusion matrix for the classifiers used in building the models.

Table 2: Summary of classifier performance

Metric Value	Support Vector Machine	Naïve Bayes
Time taken to generate features using the Tf-idf Vectorizer	1.58s	1.66s
Time taken to build model	0.58s	0.003s
Accuracy	91%	71%
Precision	0.919	0.805
Recall	0.912	0.714
F1-Score	0.912	0.639

Table 3: Confusion matrix for Support Vector Classifier

	Class one	Class two	Class three
Class one	518	43	12
Class two	0	272	2
Class three	24	17	227

Table 4: Confusion matrix for Navies Bayes Algorithm

	Class 1	Class 2	Class 3
Class 1	561	0	12
Class 2	255	17	2
Class 3	50	0	218

From table 2, it can be seen that the time taken for Navies Bayes algorithm to build the trained model is faster compared to the Support Vector Classifier. But in terms of accuracy, precision, recall and F1-score, the Support Vector Classifier performed better than the Navies Bayes algorithm. From table 3, the classifier made a total of 1115 predictions. In terms of True positives; 518 tweets belong to Class one, 272 tweets belong to Class two 2 and 227 tweets belong to Class three. From table 4, the classifier made a total of 1115 predictions. In terms of True positives; 561 tweets belong to Class one, 17 tweets belong to Class two and 218 tweets belong to Class three. Based on these results, it can be seen that the Support Vector Classifier gave the better result and was used to build the model for determining the sentiment polarity of customers' tweet.

4. FUTURE WORK

Sentiment Analysis is still considered as being in its' developing stage. More researches are carried out in respect to finding better and accurate methods to determine meaning or information in layers of words. Regarding this research work, future work is directed towards using lexicon-based and natural language processing (NLP) techniques to extract deeper context in sentences. It helps to detect the syntactical structure and semantic relations between words.

5. CONCLUSION

In this study, a sentiment analysis software application was developed using machine learning techniques. The study captured customer's tweets in form of opinions and feedbacks from the popular social media site, Twitter. The process involved retrieving tweets in a raw format and preprocessing it into a structured format. Text mining techniques using machine learning methods were applied on the structured text to extract relevant features from the text data. The extracted features were clustered to discover meaningful information from the data and classified for prediction purpose. The information obtained from the clustered text data can be used for decision making in appealing to customers' preferences. This will help to improve better network services and yield more profit to organizations.

REFERENCES

1. Adeyemo A.B. and Ojo A.K. (2014). Classification of Social Blogs Comments Using Text Mining. International Journal of Computer Science Issues, Volume 11, Issue 6, No 1, pp. 54-58.
2. Afful-Dadzie, E., Nabareseh, S., Komínková, O. and Klímek, P. (2014). Enterprise Competitive Analysis and Consumer Sentiments on Social Media Insights from Telecommunication Companies. Available from: https://www.researchgate.net/publication/268669101. Retrieved on July 18, 2016.
3. Agarwal A., Xie B., Vovsha I., Rambow O., and Rebecca P. (2011). Sentiment Analysis of Twitter Data. Technical Report," Department of Computer Science Columbia, University New York, USA.
4. Benhardus, J. and Kalita, J. (2013). Streaming trend detection in Twitter. International Journal on Web Based Communities, Volume 9, number 1, pp. 122–139.
5. Bifet, A. and Frank, E. (2010). Sentiment knowledge discovery in Twitter streaming data. In Proceeding of 13th international conference on Discovery Science Conference, pp. 1– 15.
6. Crockford, D. (2006). RFC 4627 -The application/json Media Type for JavaScript Object Notation (JSON). Technical report, IETF, 2006. Available from: http://tools.ietf.org/html/rfc4627. Retrieved on July 27, 2016.
7. Django. (2016). Meet Django. Available from: https://www.djangoproject.com/. Retrieved on September 28, 2016.
8. Fersini, E., Messina, E. and Pozzi, F. (2015). Expressive signals in social media languages to improve polarity detection. Journal of Information Processing and Management."DOI: 10.1016/j.ipm.2015.04.004."Retrieved on July 27, 2016.
9. Filho, P.B. and Pardo, A.S. (2013)."NILC USP: A Hybrid System for Sentiment Analysis in Twitter Messages. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 568–572.
10. Gamallo, P. and Garcia, M, (2014). Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, pp.171–175.
11. Ghiassi M., Skinner J. and Zimbra D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with Applications, Volume 40, pp. 6266–6282.
12. Jafari Asbagh, M., Ferrara, E.,Varol, O., Menczer, F. and Flammini, A.(2014). Clustering memes in social media streams."Technical Report, School of Informatics and Computing, Indiana University, Bloomington IN (USA).
13. Kang, D. and Park, Y. (2014). Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. Expert Systems with Applications, Volume 41, pp. 1041–1050.
14. Kolchyna, O., Souza, T., Treleaven, P. and Aste, T. (2015). Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination." Available from:" https://arxiv.org/pdf/1507.00955. Retrieved on July 27, 2016.
15. Kouloumpis E., Wilson T., Moore J. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.
16. MTN Nigeria (2015). About MTN. Available from: http://www.mtnonline.com/aboutmtn. Retrieved July 28, 2016.
17. Pang B., Lee L., and Vaithyanathan S., (2002). "Thumbs up? Sentiment classification using machine learning techniques," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86.

18. Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, Vol. 2, Nos. 1–2, pp. 1–135. http://dx.doi.org/10.1561/1500000011.
19. Taboada, M. 2016. Sentiment Analysis: An Overview from Linguistics. Annual Review of Linguistics, Vol. 2, pp. 325–347. http://dx.doi.org/10.1146/annurevlinguistics-011415040518.
20. Turney P., (2002). "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in Proceedings of the Association for Computational Linguistics (ACL), pp. 417–424.
21. Tutubalina E., Zagulova M., Ivanov V. and Malykh V. A. (2015)." A Supervised Approach for SentiRuEval Task on Sentiment Analysis of Tweets about Telecom and Financial Companies. A Technical Report based on the Sentiment Evaluation for Russian (SentiRuEval-2015) initiative.
22. Vasu J. (2013). Prediction of Movie Success using Sentiment Analysis of Tweets. The International Journal of Soft Computing and Software Engineering [JSCSE], Special Issue: The Proceeding of International Conference on Soft Computing and Software Engineering, Volume 3, No. 3, pp: 308-313.
23. Walaa M., Ahmed H., and Hoda K. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, Volume 5, pp.1093–1113.