# A Combined Approach For Sentimental Analysis Using Multiple Hashtags On Twitter

Rehee Mehta
*Department of Computer Science Engineering*
*Chitkara University, India*

Dr. Shaily Jain
*Department of Computer Science Engineering*
*Chitkara University, India*

**Abstract- From past few years several research organizations and businesses have started focusing on the social media and other web resources to have an insight into the market trends. The current research illustrates the detailed work done in developing a framework which can fetch the tweets regarding any specified topic using mutilple hashtag approach through twitter API. The system readily processes the captured tweets by preprocessing it and then classifying it into positive and negative categories depending upon the type of feedback or review presented by the users using Naïve Bayes Classifier. The main objective of the research is to automatically generate the overall opinion polarity for the tweets gathered using multiple hashtags in a certain duration, which appears to be a remarkable improvement from the past research. Further the fundamental details and the future scope of the research has been explained.**

**Keywords – Sentiment analysis, Opinion mining, Twitter, Multiple hashtag, Classification, Tweets.**

## I. INTRODUCTION

The extraction of invisible predictive information from huge databases is referred as data mining. It is an emerging dynamic technology with high potential, aiding numerous companies and organizations to learn about their customer's reviews and opinions in regard to their products and services. Data mining software is one of the analytical tools for examining the data. It grants users to analyze data from diverse dimensions, classify it, and summarize the connections identified.[23] These tools anticipate future trends and nature, granting the companies to make relevant, knowledge-driven decisions. They can also answer business related queries that traditionally were very time consuming to resolve.

Data Mining has number of applications, out of which Web Mining is an approach to identify patterns from the Web. According to the purpose of the analysis, web mining can be divided into three broad categories: Web content mining, Web structure mining and Web usage mining. In order to extract useful information from text, image, audio or video data from the web, Web Content Mining is used. It can also be referred as web text mining as the text content is the most widely researched area. Opinion mining is a sub discipline of web content mining which is also called as sentiment analysis. It is a process of finding users opinion about particular topic or a product. It aims to make a computer machine capable of understanding human emotions and sentiments in a way a human could understand and respond accordingly [6,8].

Nowadays, people are no longer restricted to their friends and families for seeking reviews or feedbacks of either a product or an idea. There are enormous user reviews and opinions regarding any topic in public forums on the Web.[22] Applications of sentiment analysis have expanded to almost every possible domain, from user products, services, healthiness, economic services, etc to social affairs and political elections. For an organization, conducting surveys, polling and focus groups, etc. are no longer necessary to gather public reviews because there is an abundance of such information publicly available online. In recent years, it has been witnessed that opinionated postings on social media have helped reform businesses, and influence public opinions and emotions, which have strongly impacted our social and political systems. It has thus become inevitable to gather and examine opinions on the Web.[11]

The main challenge in this section is the opinion classification in which the opinion may be a judgment, frame of mind or assessment of an entity namely movie, book, product, etc which can be in the form of document or sentence or feature that can be marked as positive or negative.[10] Classifying entire datasets according to the sentiments towards certain entities is called as sentiment or opinion classification. One form of sentiment mining regarding product feedbacks is to produce feature-based summary.[14] To create a summary on the features, commodity features are first determined, and then the positive and negative opinions are aggregated.[26] Though it is very time consuming and dreary for human users to categorize thousands of feature expressions that can be identified from the text for sentiment mining application into feature classes.[25] Thus some kind of automated assistance is necessary. For this, number of classification algorithms has been developed which aids in classifying several user reviews into different categories.

## II. MOTIVATION

From past few years businesses and research organizations have started focusing on social media and online forums. But the supply chain regulation has been left behind in the field of exploration and research. The outcome of the paper involved the tweets used by distinct organizations involving supply chain experts and groupings such as information services, computer companies, builders, producers, etc for data sharing, appointing experts interacting with stakeholders etc. Also several other topics such as logistics and corporate communal duties to uncertainty, manufacturing and even civilian powers were examined [1].

The social media gathers the data in structured and unstructured, formal and informal form as the users do not care about the spellings and grammatical formation of a sentence while exchanging information or ideas amongst themselves using different social networking websites such as Facebook, orkut, LinkedIn, instagram, etc.[13] The collected data consisted of sentiments and opinions of users which were processed using data mining techniques and were analyzed for capturing the useful information from it [3].

An opinion mining extraction algorithm to jointly explore the essential opinion mining elements was proposed. Particularly, the algorithm automatically creates kernels to join closely related words into new terms from word level to phrase level based on dependency relations and assured the certainty of opinion expressions and polarity based on fuzzy dimensions, opinion rate intensifiers, and opinion patterns. Some interesting observations were acknowledged like the negative polarity of video dimension was greater than the product usability dimension for a product. Still, increasing the dimension of product usability could effectively improve the product [4].

The information on current trends, applications of opinion mining, several areas where it could have been used and also lot of meaningful information on the recent research work that was being carried out in this field of data mining was provided.[17] Also, the primitive work plan of the sentiment analysis process, the challenges and the forthcoming research being planned in the area of sentiment analysis was explained remarkably [5].

An approach to extract and gather goods characteristics, reviews from numerous online sources related to a particular product in which a rule-based approach was enforced, that practiced semantic and opinion mining of messages to extract the pair of characteristics and sentiments that consists of statement-level occurrence in customer feedback records. The captured characteristic and sentiment combination were structured, categorized and represented accordingly [6].

A novel approach for contextualizing and enriching massive semantic knowledge bases for sentiment analysis with a focus on Web intelligence platforms and other highly efficient big data applications was presented. The method was not only relevant to traditional sentiment lexicons, but also to broader, complete, multi-dimensional affective resources such as SenticNet [7].

The tool of Opinion mining and Sentiment Analysis processes a set of search results for a given item based on the quality and characteristics. By analyzing customer review one can scale a particular product and provide opinions for it. Research has been carried out in this field to mine opinions in the form of document, sentence and feature level sentiment analysis. Various techniques and tools of Opinion Mining were discussed in this paper [8].

A framework for examining polarity for Twitter texts was introduced in this paper.[30] In order to evaluate the performance of the proposed framework, number of text datasets was considered. The research made use of number of distinct classifiers namely: Nearest Neighbors (KNN); Naïve Bayes (NB); Decision Trees (J48); Support Vector Machines (SVM). The conclusion defines a suitable framework to automate the polarity examining process which showed high accuracy and less false positive rates. [9]

## III. FRAMEWORK TO FETCH TWITTER DATASET

Twitter can be represented as a micro blogging service that permits users to connect with short, 140-character texts that roughly correspond to ideas or thoughts.[16] Apart from this, one can think of Twitter as being akin to an open, high-speed, universal text-messaging service. In other words, it's an excellent source that enables quick and convenient communication.[31]

3.1 Creating twitter API connection: Twitter has taken huge concern to develop an extremely straightforward RESTful API that is instinctive and trouble-free to use.[15] Prior to make an API request to Twitter, an application at https://dev.twitter.com/apps has to be created. Developing an application is the regular technique for developers to achieve an access to API and also for Twitter to supervise and communicate with another party platform developer as required. The method for developing an application is quite ordinary. The only constraint which is required is read-only connection to the API.

In order to log into third-party websites using one's twitter account without sharing their sensitive information like passwords, OAuth is used. It makes use of few key pieces of information that are retrieved from one's newly created application's settings which are: consumer secret, access token, consumer key, and access token secret. These four references contribute the entire useful information that an application needs to permit itself through a sequence of redirects along with the user granting permission. Hence it is of same sensitivity as a password. With an approved API association, one can now generate a request.

3.2 Using hashtags to fetch Tweets by keyword: Hashtags are used to extract the tweets corresponding to the topics one wants to fetch for further analysis through Twitter API.[28] Twitter provides a capability to search only one keyword at a particular time.[29] Thus the drawback of using single hashtag at a time for data collection resulted in less accuracy due to the analysis of low volume of relevant tweets. In order to overcome this limitation, we have introduced the use of multiple hashtags to capture more relevant tweets from the twitter through its API resulting in more accurate and precise results. To facilitate the implementation of multiple hashtags, we have made use of n-gram algorithm which searches for all the possible combinations of all the keywords entered by the user.

N-grams of the content are widely used in natural language processing tasks and text mining. They are generally a set of co-existing words within a given dataset or sentence. The computation of n-grams conventionally involves the movement of one word forward. For N=1, it is termed as unigrams (individual words in a given statement). Similarly, N=2, N=3 are referred to as bigrams and trigrams and so on. If Z=number of words in an existing statement A, the no. of n-grams for the statement A would be:

$N_{gramsA} = Z - (N - 1)$

## IV. RESEARCH METHOD

4.1 Preprocessing: The data gathered from twitter in the form of tweets might be noisy (containing spelling errors, spams, outliers etc) and inconsistent (containing some type of discrepancies). Before initiating the process of classification, it is important to preprocess the data in order to get rid of such defects.[18] Most common data errors include the data entry mistakes, missing terms, spelling errors, etc. For this we have developed a dictionary to which the words in each tweet can be compared. In case of existence of an error, the incorrect word is highlighted to be rectified manually.[12] This would assist in enhancing the performance and efficiency of our research. After the preprocessing stage, the removal of stop words and stemming words is conducted to facilitate the classification process.

4.2 Stop Words Removal: The words which are drained out prior to processing of natural language data are referred as stop words. Though stop words generally refer to the most general words in a language, there is no unique universal list of stop words consumed by all natural language processing tools. The list of words that are to be filtered out, to conveniently aid the process of classification is called a stop list. Stop words are considered inappropriate for searching purposes because they appear frequently for which the indexing engine has been adjusted. In order to save both space and time, these words are removed at indexing time and then neglected at search time.[19]

4.3 Keyword Stemming: Stemming is the term used in lingual morphology and knowledge retrieval to illustrate the process for minimizing skewed or inflected words to their word stem, pure or basic form. It usually refers to a crude heuristic process that removes the ends of words in the hope of attaining this motive correctly most of the time, and often includes the chopping off the derivational affixes.

In our research we have used KMP (Knuth Morris Pratt) pattern searching algorithm for removal of the stemming words. This algorithm makes use of the degenerating feature (pattern having same sub-patterns appearing more than once in the pattern) of the pattern, O(n) as the worst case complexity. The fundamental objective behind KMP's algorithm is: whenever a mismatch is detected (after some matches), some of the characters in the text are already known (since they matched the pattern characters earlier to the mismatch). Hence the advantage of this information can be taken to avoid matching the characters that will anyway match.

4.4 Classification Process Using Naïve Bayes Classifier: Naive Bayes classifier is one of the most favorable algorithms for classifying the text documents. In our research we have used this algorithm for classifying various tweets into positive, negative and neutral categories.[21] By categorizing huge amount of user's opinions, feedbacks and suggestions provided on the web sources, it assists in analyzing and systemizing these reviews for better decision making. The approach is based on Bayesian theorem and is specifically used when the dimensionality of the inputs is high. The Bayesian Classifier is proficient of calculating the best possible output depending upon the input. The Bayesian classification is used as a probabilistic learning approach (Naive Bayes text classification).[24] The algorithm works as follows,

To caluculate the posterior probability $P(a|z)$, with respect to the Bayes Theorem, from $P(a)$, $P(z)$, and $P(z|a)$, can be as follows: The classifier predicts that the impact of the value of a predictor ($z$) on a given class ($a$) is self reliant of the values of other predictors.

$$P(a|z) = (P(z|a)P(a))/P(z)$$
$$P(a|z) = P(z_1|a)\times\cdots\cdots\times P(z_n|c)\times P(a)$$

P(a|z): The posterior probability of class (target) when a predictor (attribute) of class is given.
P(a): prior probability of class.
P(z|a): The likelihood which is the probability of predictor of given class.
P(z): prior probability of predictor of class.

In order to classify the dataset into positive and negative categories, we have created a dictionary involving all possible negative sentiments, to which the entire dataset can be compared.[27] This aids the classifier in separating all the negative tweets from the positive ones by analyzing the frequency of occurrences of the negative sentiments.[32]

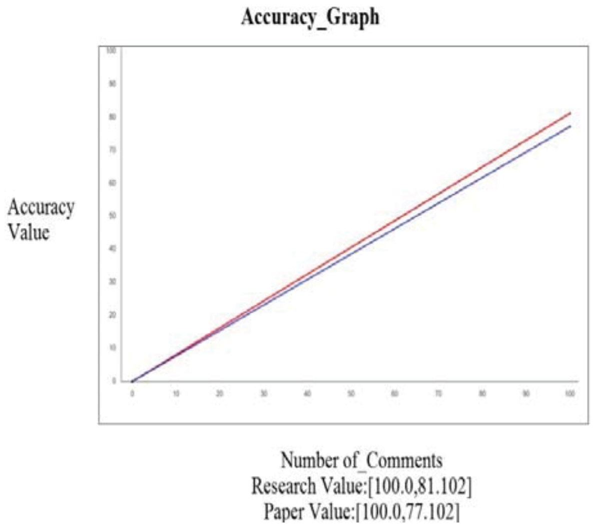
4.5 Representation: After the classification process is completed, the outcome is presented as a decision tree using decision tree algorithm. A decision tree is a classifier signified as a recursive division of the instance spread. It can also be illustrated as the blend of logical and analytical techniques to aid the explanation, classification and generalization of a given dataset. The decision tree forms a rooted tree consisting of number of nodes, implying it is a supervised approach with one of the object as "root", having zero incoming edges. In this, every inner object divides the sample space into number of partitions according to a discrete function of the given feature value. Every leaf is accredited to individual class exhibiting the most relevant target value. Alternatively, the leaf may possess a probability vector representing the probability of the target attribute having a particular value. Instances can be categorized by directing them from the parent node of the tree down to a leaf, according to the conclusion of the analysis along the path.

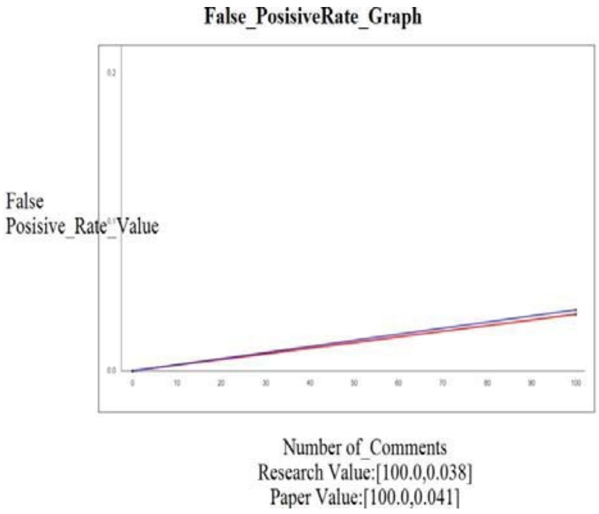
## V. OUTCOME ANALYSIS

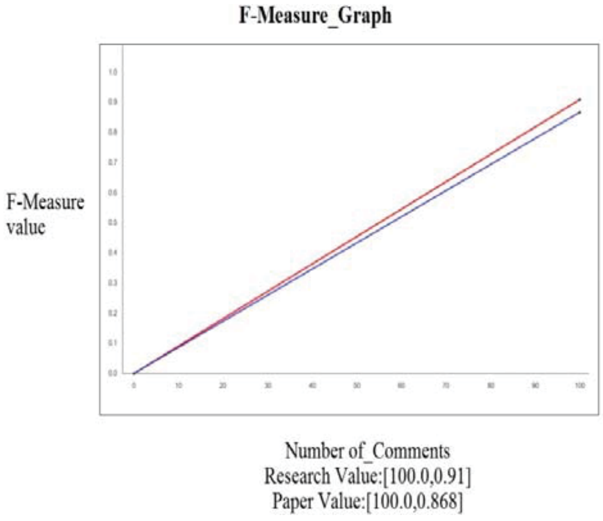
This part of the paper demonstrates the results acquired by implementing the projected framework to the twitter datasets recorded. The training sets were developed using a stratified approach, evaluating the count of positive and negative tweets for entire datasets.[20] The positive class is assumed to be composed of the tweets with positive sentiment whereas the negative class is assumed to be composed of the negative tweets. The polarity distribution is tested with a standard machine-learning classification algorithm i.e. Naïve Bayes Classifier. The results of our system have been compared to the past research results [15] as follows:

Table 1. Performance Analysis of Algorithm

| | **ACCURACY** | **FALSE POSITIVE RATE** | **F-MEASURE** |
|---|---|---|---|
| **Past research values [15]** | 77.102 | 0.041 | 0.868 |
| **Current research values** | 81.102 | 0.038 | 0.91 |







## VI. CONCLUSION AND FUTURE WORK

Use of single hashtag for gathering data resulted in less accurate outcomes due to the analysis of low volume of suitable tweets. In advance, we have introduced the use of multiple hashtags to extract more appropriate tweets from twitter through its API for improved results. Polarity classification of the tweets is another crucial task for number of reasons, involving the pace with which the tweets are obtained, the huge amount of tweets obtained around number of subjects, the syntactic ambiguous nature of the posts, and the informal structure and behavior of the messages. In extension to this, these tweets do not come with a specific polarity attached. Thus, to automate polarity categorization of messages, it is required to label part of the tweets as positive or negative to develop a classifier. But this labeling method is quite time consuming and intuitive when done manually, hence prone to inappropriate and inaccurate results. In order to overcome such complications, the current research proposed a framework which integrates a knowledge-based categorization with machine learning algorithms to automatically attach relevant polarity to twitter posts. The knowledge base consists of opinion-based words. The procedure involves searching of opinion-based words in all the captured tweets, extracting these tweets for training the classifier, and then describing the unlabeled tweets using a classifier. The analysis depicts better results indicating an improved method to automatically execute polarity examination of Twitter posts, with refined accuracy levels.

Future works can include the collection of data for a longer duration in order to develop more complete picture of the use of twitter to examine the market trends. Also one of the major challenges is that an opinion might be interpreted differently by different individuals resulting in inaccurate results.[33] Thus work can be done in this context to overcome this limitation.

## REFERENCES

[1] "Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research" by Bongsug (Kevin) Chae, International Journal of Production Economics, Vol. 165, pp. 247–259, (July 2015).

[2] "Reprint of: Computational approaches for mining user's opinions on the Web 2.0" by Gerald Petz, Michał Karpowicz, Harald Furschus, Andreas Auinger, Vaclav Stritesky Andreas Holzinger, Information Processing and Management, Vol. 51, Issue 4, pp. 510-519, (2015).

[3] "A Review: Text Classification on Social Media Data" by Ms. Priyanka Patel, Ms. Khushali Mistry, IOSR Journal of Computer Engineering, Vol. 17, Issue 1, pp. 80-84,( Jan – Feb 2015).

[4] "jointly identifying opinion mining elements and fuzzy measurement of opinion intensity to analyze product features" by Haiqing Zhang, aichasekhari, yacineouzrout, abdelazizbouras, Engineering Applications of Artificial Intelligence, (29 June 2015).

[5] "Procedure of Opinion Mining and Sentiment Analysis: A Study" by Rushabh Shah, Bhoomit Patel, International Journal of Current Engineering and Technology, Vol.4, No.6, pp. 4086-4090, (1 Dec 2014).

[6] "Identifying and Summarizing Features from Web Opinion Sources in Customer Reviews" by Nidhi R. Sharma, Vidya D. Chitre, Mining, International Journal of Innovations & Advancement in Computer Science (IJIACS),Vol. 3, pp. 8-14(7 Sep 2014).

[7] "Enriching semantic knowledge bases for opinion mining in big data Applications" by A.Weichselbraun , S. Gindl , A. Scharl, Knowledge-Based Systems, Vol. 69, pp. 78-85(October 2014).

[8] "An Analysis on Opinion Mining: Techniques and Tools" by G. Angulakshmi, Dr. R. Manickachezian, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 7, pp. 7483-7487(July 2014).

[9] "A polarity analysis framework for Twitter messages" by Ana Carolina E.S. Lima, Le and ro Nunesde Castro, Juan M. Corchado, Applied Mathematics and Computation, Vol. 270, Issue C, pp. 756-767(November 2015).

[10] "Methodological Study of Opinion Mining and Sentiment Analysis Techniques" by Pravesh Kumar Singh, Mohd Shahid Husain, International Journal on Soft Computing (IJSC), Vol. 5, No. 1, pp. 11-21(February 2014).

[11] "Opinion Mining, Analysis and its Challenges" by Nidhi R. Sharma, Vidya D. Chitre, International Journal of Innovations & Advancement in Computer Science (IJIACS), Vol. 3, pp. 59-65(1 April 2014).

[12] "Opinion Mining about a Product by Analyzing Public Tweets in Twitter" by T. K. Das, D. P. Acharjya and M. R. Patra, International Conference on Computer Communication and Informatics (ICCCI), pp. 1-4(03 – 05 Jan 2014).

[13] "Classification of Opinion Mining Techniques" by Nidhi Mishra, C.K. Jha, International Journal of Computer Applications, Vol. 56, No.13(October 2012).

[14] "Applying Supervised Opinion Mining Techniques on Online User Reviews" by Ion Smeureanu, Cristian Bucur, Informatica Economica, Vol. 16, No. 2, pp. 81-91(2012).

[15] "Opinion Mining, Analysis and its Challenges" by Nidhi R. Sharma, Vidya D. Chitre, International Journal of Innovations & Advancement in Computer Science (IJIACS), Vol. 3, pp. 59-65(1 April 2014).

[16] "TOM: Twitter opinion mining framework using hybrid classification scheme" by F.H.Khan, S.Bashir, U.Qamar, Decis. Support Syst. 57, pp. 245–257(2014).

[17] "Mining and summarizing customer reviews" by Hu M, Liu B., In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 22, pp. 168-177(Aug 2004).

[18] "Identifying noun product features that imply opinions" by Zhang L, Liu B, In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, pp. 575-580(19 Jun 2011).

[19] "Twitter as a Corpus for Sentiment Analysis and Opinion Mining" by Pak, Alexander, and Patrick Paroubek, In LREc, vol. 10, pp. 1320-1326(2010).

[20] "Recognizing contextual polarity in phrase-level sentiment analysis" by Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann, In Proceedings of the conference on human language technology and empirical methods in natural language processing", pp. 347-354(2005).

[21] "Machine learning algorithms for opinion mining and sentiment classification" by Khairnar, Jayashri, and Mayura Kinikar, International Journal of Scientific and Research Publications 3, no. 6, pp: 1-6(June 2013).

[22] "Large Scale Sentiment Analysis for News and Blogs" by Namrata Godbole, Manjunath Srinivasaiah, Steven Skiena, ICWSM 7, no. 21, pp: 219-222(26 March 2007).

[23] “Survey of Classification Techniques in Data Mining” by S. Archana, Dr. K. Elangovan, International Journal of Computer Science and Mobile Applications 2, no. 2, pp: 65-71(Feb 2014).

[24] "A Comparative Study of Classification Techniques in Data Mining Algorithms" by Sagar S. Nikam.

[25] "Opinion mining and analysis: a survey" by Arti Buche, Dr. M. B. Chandak, Akshay Zadgaonkar, International Journal on Natural Language Computing (IJNLC), Vol. 2, No.3,(June 2013).

[26] "Clustering product features for opinion mining", In Proceedings of the fourth ACM international conference on Web search and data mining, 2011, pp: 347-354.

[27] "Enhanced sentiment learning using twitter hashtags and smileys." In Proceedings of the 23rd international conference on computational linguistics: posters, pp. 241-249,(2010).

[28] "Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach" by Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, Ming Zhang, In Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 1031-1040(24 Oct 2011).

[29] "Twitter sentiment analysis: The good the bad and the omg!" by Efthymios Kouloumpis, Theresa Wilson, Johanna Moore, Icwsm 11, pp. 538-541(17 July 2011).

[30] "Sentiment analysis of twitter data" by Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, In Proceedings of the workshop on languages in social media, Association for Computational Linguistics, pp. 30-38(23 June 2011).

[31] "Target-dependent twitter sentiment classification" by Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, Tiejun Zhao, In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Vol. 1, pp. 151-160(19 June 2011).

[32] “Identifying Sarcasm in Twitter: A Closer Look” by Roberto González-Ibáñez, Smaranda Muresan, Nina Wacholder, In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, Association for Computational Linguistics, Vol. 2, pp. 581-586(2011).

[33] "Twitter sentiment classification using distant supervision" by Alec Go, Richa Bhayani, Lei Huang, CS224N Project Report, Stanford 1, pp. 1- 12(Dec 2009).