International Journal of Advanced Research in Engineering Technology & Science
ISSN: 2394-2819
Email: editor@ijarets.or Volume-4, Issue-6 June- 2017 www.ijarets.org

LARGE SCALE OPINION MINING FOR SOCIAL, NEWS AND BLOG DATA

Kumari Sarita
Department of C.S.E
GRD Institute of Management & Technology
Rajpur Road, Dehradun

Bal Krishna Yadav
Department of C.S.E
GRD Institute of Management & Technology
Rajpur Road, Dehradun

ABSTRACT:
This work presents a framework which allowed the identification and classification of opinions in short text fragments which is based on Twitter data. In this paper pre-processing of raw data for a data analysis approach is presented that extracts qualitative information from the social media text selected in the form of tweets. The selected dataset is then transformed into more useful structured data. Using Twitter, the most popular micro blogging platform, the proposed approach aims to complete the task of pre-process data for the purpose of opinion mining with the help machine learning classifier (support vector machine). In this work opinion mining is performed on around 2000 tweets about products. Through this analysis we get to know about the reviews and opinions of people on a brand that helped us to gain insight into how a brand is being perceived by the public. This is an effective technique, which will aspire to convert raw data into useful transformed form to be used for the political scenario analysis. Twitter is large source of data, which make it more attractive for performing opinion mining.

KEYWORDS – Opinion Mining Blog data. SVM

I. INTRODUCTION

Opinion mining is a standout amongst the most prominent patterns in this day and age. It is the technique by which data is extricated from the suppositions, evaluation and feelings of individuals concerning substances, occasions and their qualities. It is the computational strategy for removing, arranging, understanding, and evaluating the assessments communicated in different substance. These information discover its way on person to person communication locales like twitter, face book and so on. Conclusion Mining frequently called as Sentimental examination. It should be possible at report, expression and sentence level. In record level, outline of the whole report is taken first and after that it is break down whether the feeling is sure, negative or impartial.

In expression level, examination of expressions in a sentence is considered to check the extremity. In Sentence level, each sentence is arranged in a specific class to give the assumption. Supposition Mining has different applications. Web-based social networking is one of the greatest gatherings to express assessments. It is utilized to produce assessments for individuals of online networking by examining their sentiments or contemplations which they give educate of content. Sentiment Mining is space focused, i.e. aftereffects of one space can't be connected to other area.

Assessment mining is utilized as a part of numerous genuine situations, to get audits about any item or motion pictures, to get the monetary report of any organization, for expectations or showcasing. The postulation for the most part concentrates on small scale blogging informal communication webpage twitter for conclusion mining.

Twitter has given an extremely massive space to expectation of purchaser brands, motion picture surveys, just appointive occasions, securities exchange, and prominence of big names. On twitter, individuals impart their perspectives and insights in the types of "tweet", which are 140 characters long. These tweets are the primary part which decides supposition of the general population as these assessments are unique and specifically from clients. As the measure of tweet is constrained, it is anything but difficult to register feeling in tweets as opposed to long reports.

II. OBJECTIVE

The objectives of the work have been discussed in the following points:-

- To investigate, break down and concentrate the current opinion mining procedures in the online small scale blogging system.
- To think about how the tweets (identified with point # tags) can be produced from the twitter with the Twitter API.
- Collection of related dataset from twitter with the assistance of Twitter API.
- To execute and characterize the dataset into positive, negative or unbiased subsequent to applying administered machine learning classifier (Support Vector Machine).

Subsequently, the objective of this proposal is to perform conclusion mining in brand setting. Popular suppositions on brand are mined from Twitter and after that characterized into positive, negative or unbiased by utilizing administered machine learning classifier (SVM). These outcomes will tell us about the surveys and conclusions of individuals on the brand. With regards to the supposition digging being completed for this reason, the outcomes will permit us to pick up knowledge into how a brand is performing in the market.

III. LITERATURE REVIEW

String and lee in their paper titled "Thumbs up? Estimation characterization utilizing machine learning methods", [1] proposed the framework where a conclusion can be certain or negative was discovered by proportion of positive words to aggregate words. Later in 2008 the creators built up a procedure in which tweet result can be chosen by terms in the tweet. Contrast with baselines that are created by people, the outcomes are quite great when machine learning strategies are utilized. SVM gave best outcome as contrast with Naïve Bayes. Despite utilizing diverse sorts of elements the creators did not achieved fancied correctness over point based classification.

Feeling examination is whether the announcements focuses negative or positive conduct towards the subject. The creators expressed that it is fundamental to unmistakably discover the semantic connections between the subject and the notion expressions to build the precision for the examination of supposition. With a specific end goal to recognize the suppositions in news articles and website pages, their proposed framework acquired high accuracy of 75-95%.

Kim and Hovy in their paper titled "Deciding the Sentiment of Opinions", [2] expressed that the recognizable proof of an assumption was testing issue. The creators built up a framework for a specific subject that naturally seeks the clients who posts their perspectives on that point and the estimation of each view. The frameworks comprise a module for portraying slant of a word and other for blending the assumptions into an announcement. The creators experimented with various models of characterization and consolidating the assumption at sentence and word level, given better outcomes. For the change of acknowledgment of Holder, the creators utilized parser to join regions that are more solid with Holders. The learning methods that are utilized is this framework are Support vector machines and choice rundown.

Wilson et al. in their paper titled “Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis”,[3] presented a method which first describes whether a statement is polar or neutral to phase-level sentiment evaluation and then ascertain the polarity of polar statements. By applying this methodology, the system was capable to identify automatically the contextual polarity of sentiment statements for huge subsets, obtaining results which are greater than baseline.

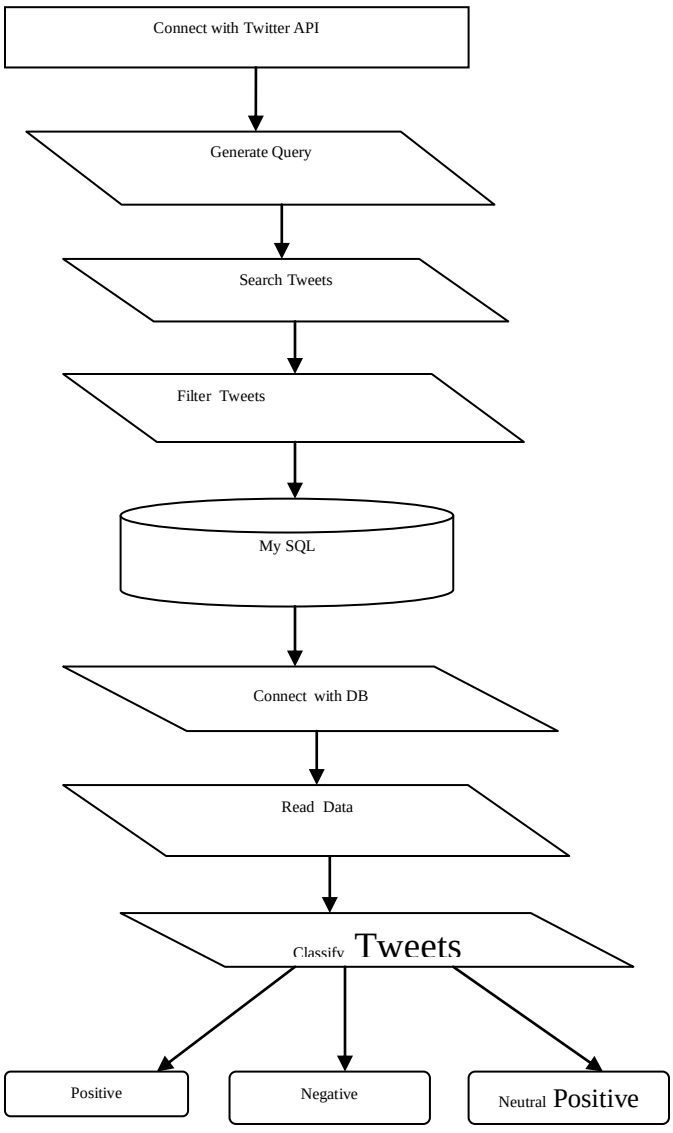
Kopel and Schler in their paper titled "the Importance of Neutral Examples for Learning Sentiment", [4] clarified that it is vital to utilize nonpartisan messages to get great information of extremity. The creators likewise expressed that the positive and negative messages alone won't give legitimate comprehension about unbiased messages. Thinking about impartial messages clear the distinction amongst positive and negative messages. The creators found that in one of the corpus having the vast majority of the unbiased records gives no opinion which can be utilized as counter to test both energy and pessimism of an archive.

Godbole et al. in their paper titled "Huge Scale Sentiment Analysis for News and Blogs", [5] proposed a framework which contains period of recognizable proof in which conclusion for a specific point was shown and scoring stage which was utilized to score relative substances in a similar class. Finally the creators examined the significance of scoring strategies over enormous dataset of web journals and news. The creators underlined on the way that slants can differ as per the geographic area, news source or statistics gather.

Benamara et al. in their paper titled "Opinion Analysis: Adjectives and Adverbs are superior to Adjectives Alone", [6] expressed that the majority of the work done in past was finding the quality.

Subjective explanations inside an archive or expressions utilize the unique grammatical form, for example, things, verbs and descriptive words. The creators said that until their commitment there was not a solitary identified with qualifiers in neither slant investigation nor utilization of intensifier modifier blends (AACs) in conclusion examination. The creators proposed a conclusion examination strategy which depends on AACs which utilizes a phonetic assessment of degrees of intensifiers.

IV. METHODOLOGY



Flowchart of Methodology

The methodology adopted in proposed work focuses on performing opinion mining task effectively and efficiently. The Proposed work uses the following methodology:

- Collection of data related to the Brand from Twitter with the help of Twitter API
- Pre-processing of related data collected from Twitter so that it can be fit for mining.
- Collecting and organizing training datasets.
- Training the supervised machine learning classifier using training datasets.
- Classifying the testing datasets into positive, negative or neutral after applying supervised machine learning classifier (Support Vector Machine).
- Computing the result of classifier using datasets collected from Twitter.
- Plotting a graph that shows the trend of positive and negative opinion about the brand.

A. Data Extraction

It is the initial step of this entire procedure. In this Java is utilized as a Scripting dialect to extract data utilizing a Twitter API named as Twitter4j utilizing Net Beans IDE (Integrated Development Environment).The information was separated utilizing prominent hash labels utilized for Brand, for example, #Samsung, #galaxy4, #Samsung mobiles, #faulty Samsung etc. In that specific time span to guarantee that the tweets removed are applicable and as indicated by the requirement. In twitter hash labels turns into the important image to discover about something and it gives client breaking point of 140 words to express their perspectives and attitude. Extracted dataset comprised of content of the tweet alongside the date and time of tweet

B. Data Pre-Processing

It is found that to get fancied outcomes from the classifier we need to ensure that the tweets can be handled properly. As tweets can be in client dialect, so we need to clean every information which are unimportant to the data. Mostly tweets comprises of message alongside username as, exhaust spaces, exceptional characters, stop words, hash labels, time stamps, URL's ,and so forth. Along these lines to make this information fit for mining it is required to pre-prepare this information. When we are finished with it, we are prepared with handled tweet which is given to classifier to required results.

C. Database Creation

To plan stockpiling for sparing a lot of gushing information and to have the capacity to effortlessly reuse required data the database was made to encourage the means required in the following periods of feeling mining process. For this situation MySQL Workbench database is utilized for putting away information.

D. Collecting and Organizing Training Data

The fundamental wellspring of the preparation information is chosen, separated and labeled twitter information covering real assessments about the brand. The twitter information are labeled utilizing lexicon containing positive words, positive descriptors, negative words and negative modifiers. 70% of tweets removed are dealt with as preparing information, which is utilized to make machine learning classifier (Support Vector Machine) learn for future investigation and expectations while remaining percent is utilized as test information.

Classification of Data

Once the training data is collected, it is passed through the machine learning classifier to make classifier learn for analysis and prediction. In this work, Support Vector Machine (SVM) classifier is used for the classification of tweets**.**

V. IMPLEMENTATION

Parameters

The following Table describes the parameters used to perform opinion mining.

Parameters	Values
Number of Tweets	2000
Hash tags used	20
Number of Positive words in Dictionary	2006
Number of Negative Words in Dictionary	4783
Number of Nouns in Dictionary	4412

Data Extraction

The data is extracted using twitter API named as Twitter 4j.It consist various numbers of libraries that are used in the extraction. To use Twitter API we must first have a twitter account. It can be easily created by filling the sign up details in twitter.com website. After this we will be provided with a username and password which is use for login purpose. Once the account is created, we can now read and send tweets on any topic we want to explore.

Data Preprocessing using Java

Information acquired from twitter is not fit mining. Generally tweets comprises of message alongside usernames, purge spaces, exceptional characters, emoticons, re-tweets, rehashed words, hash labels, time stamps, URL's, and so forth. In this progression gathered information is pre handled. In preprocessing we first concentrate our primary content from the tweet, then we evacuate every single exhaust space, re-tweets, hash labels, rehashing words, URL's, and so on Java has been used for pre-processing the data.

Steps of Pre-processing

The steps used in preprocessing are

1. Basic cleaning, where emojis and URLs are removed
2. Parse the data for readability
3. Remove special characters using pattern
4. Lexical analysis to retain only numbers and alphabets

Data Storage

To prepare storage for saving large amounts of streaming data and to be able to easily reuse required information, the database is created to facilitate the steps involved in the next phases of opinion mining process. In this case MySQL Workbench database is used for storing data. Different tables are created in the database

Classification Process

The preprocessed information from MySQL database is stacked into MATLAB. The preprocessed information is put away in tangle documents (.tangle augmentation). MATLABSVM Toolbox is utilized for grouping of information. 70% of preprocessed information is utilized as preparing dataset. The preparation dataset is utilized to prepare the SVM classifier while the rest of the percent is utilized as test information.

Steps Used

Setup
Create Project in Twitter Develop
Generate Access token and Key
Initialize required variables
Create interface to access Tweets
Start
Step1. Conn ← create connection with mysql
Step2. Query ← #code
Step3. Search ← query result
Step4. Filtered Tweet ← Remove Re-tweets, Emoticons and URLs
Step5. Stmt←Sql query to insert data in tables
Step6. .R ← execute query
Step7. Repeat steps 2 through 6 till user exit
Step8. Db← create connection in Matlab
Step9. Data ← read data from database
Step10.for i←1 to length of tweets
RT← read tweet
not C← search for not in tweet
if found
not Count ← +1
end if
but← search for but in tweet
if found
but Count ← +1
end if
but for← search for but for in tweet
if found

but for Count ← +1
end if
for j ← 1 to length of negative words

Parameters	Base paper	Proposed Scheme
Positive words	486	493
Negative words	907	917

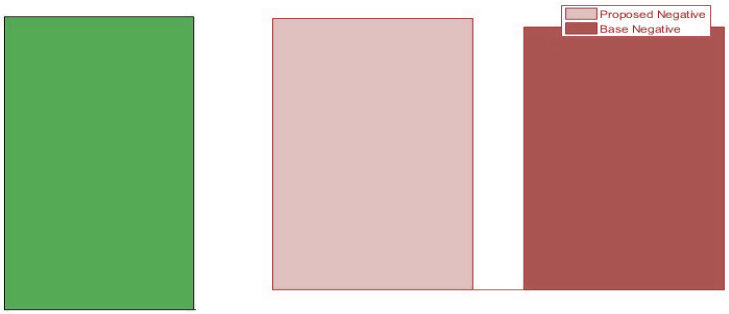
Nw← read negative word
Nw← search Nw in tweet
if found
Nw Count ← +1
end if
end for
for j ← 1 to length of positive words

Pw ← Read Positive words
Pw ← search Pw in tweet
if found
Pw Count ← +1
end if
end for
for j ← 1 to length of Neutral words
Nw← Read Neutral words
Nw← search Nw in tweet
if found
Nw Count ← +1
end if
end for
end for
Step 11. Svm Train ←train svm with negative and positive data
Step12. Result ← classify data using svm
Step13. Svm Train N← train svm with negative and nouns
Step14. Results2 ← classify data using svm
Step15. rP← calculate number of positive occurrences in classified data
Step16. rN← calculate number of negative occurrences in classified data
Step17. rNn←← calculate occurrences of noun in classified data
Step18. ifrP>rN&&rPn>rNn
Positive Response is considered
elseif rN2 >rNn
Negative Response is considered
else
Neutral Response is considered
end if
end

VI. RESULT AND ANALYSIS

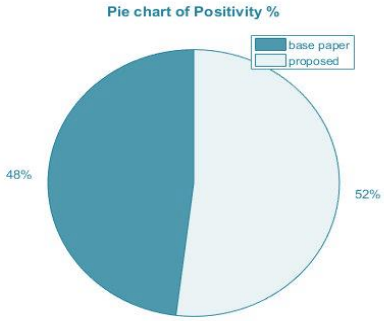
This chapter provides description about the results obtained from the implementation of the algorithm on the data collected from the tweets about products. Different # tags were considered for the study. A dataset of 2000 tweets was created for the purpose. We stored all these tweets and pre-processed them so that they can be fit for mining. The results of the analysis reveal how well people have taken the new models of mobile phones.

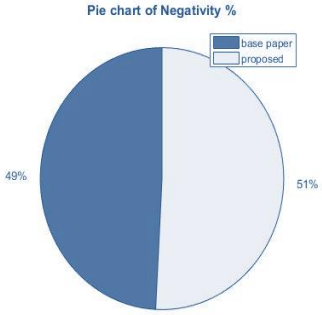
POSITIVE AND NEGATIVE WORDS OBSERVED FROM THE DATA



A. No. of positive words Extracted
B. No. of negative words Extracted

Opinion Analysis





a. PIE CHART OF POSITIVE TYPE PERCENTAGE
b. PIE CHART OF POSITIVITY VS. NEGATIVITY BASE PAPER

VII. CONCLUSION

The work presented a framework which allowed the identification and classification of opinions in short text fragments which is based on Twitter data. In this paper pre-processing of raw data for a data analysis approach is presented that extracts qualitative information from the social media text selected in the form of tweets. The selected dataset is then transformed into more useful structured data. Using Twitter, the most popular micro blogging platform, the proposed approach aims to complete the task of pre-process data for the purpose of opinion mining with the help machine learning classifier (support vector machine). In this work opinion mining is performed on around 2000 tweets about products. Through this analysis we get to know about the reviews and opinions of people on Samsung Mobile that helped us to gain insight into how a brand is being perceived by the public.

The results shows that people have not welcomed new Samsung Mobile models. The analysis has shown that people tend to be negative towards new brands of Samsung. This is an effective

technique, which will aspire to convert raw data into useful transformed form to be used for the business scenario analysis. Twitter is large source of data, which make it more attractive for performing opinion mining

VIII. FUTURE SCOPE

In future work, more efforts can be done on improving the machine learning Classifier .We can improve our system that can deal with sentences of multiple meanings. Future work can also be done on the extraction of data from other social media platforms like Face book and Instagram. The focus can also be given to multimedia data along with the textual information. There are various challenging aspects of opinion mining such as use of different languages, dealing with negation expression, comparison handling and complexity of sentences. More future research could be dedicated to these challenges.

IX. REFERENCES

[1] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up: sentiment classification using machine learning techniques", Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79--86, 2003.

[2] SM. Kim and E. Hovy, “Determining the Sentiment of Opinions”, Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics, 2004.

[3] T. Wilson, J. Wiebe and P. Hoffman, “Recognizing Contextual Polarity in Phrase- Level Sentiment Analysis”, Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics, pp. 347—354, 2005.

[4] M. Koppel and J. Schler, "THE IMPORTANCE OF NEUTRAL EXAMPLES FOR LEARNING SENTIMENT", Computational Intell, vol. 22, no. 2, pp. 100-109, 2006.

[5] N. Godbole, M. Srinivasaiah and S. Skiena,” Large-Scale Sentiment Analysis for News and Blogs”, ICWSM, vol. 7, no. 21, pp. 219—222, 2007.

[6] F. Benamara, C. Caserano, A. Picariello, DR. Recupero and VS. Subrahmanian,” Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone”, ICWSM, 2007.

[7] K. Denecke,”Using SentiWordNet for Multilingual Sentiment Analysis”, Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on IEEE, pp. 507—512, 2008.

[8] A. Harb, M. Plantie, G. Dray, M. Roche, F. Trousset and P. Poncelet, “Web opinion mining: How to extract opinions from blogs?” Proceedings of the 5th international conference on soft computing as transdisciplinary science and technology, ACM, pp. 211—217, 2008.

[9] A. Go, R. Bhayani and L. Huang, "Twitter sentiment classification using distant supervision", CS224N Project Report, Stanford, vol. 1, p. 12, 2009.

[10] J. Martineau and T. Finin, “Delta TFIDF: An Improved Feature Space for Sentiment Analysis”, Proceedings of the Third International ICWSM Conference, vol. 9, 2009

[11] T. Nasukawa and J. Yi, “Sentiment Analysis: Capturing Favorability Using Natural Language Processing”, Proceedings of the 2nd international conference on Knowledge capture, ACM, pp. 70—77, 2003.