Sentiment Analysis: Machine Learning Approach

Article *in* International Journal of Engineering and Technology · June 2017

DOI: 10.21817/ijet/2017/v9i3/1709030151

2 authors:

Dipak Kawade
Sangola College
14 PUBLICATIONS **44** CITATIONS
SEE PROFILE

Kavita Oza
Shivaji University, Kolhapur
75 PUBLICATIONS **83** CITATIONS
SEE PROFILE

Sentiment Analysis: Machine Learning Approach

Dipak R. Kawade[#1], Dr.Kavita S. Oza[*2]
[#] Department of ComputerScience, Sangola College, Sangola Dist-Solapur (MS) India
[1] dipakkavade@gmail.com
[*] Department of ComputerScience, Shiveji University, Kolhapur (MS) India
[2] skavita.oza@gmail.com

***Abstract*—Twitter is one of most popular social networking site where people are expressing their views, opinion and emotions liberally. These tweets are recorded and analysed to mine emotions of people related to a terrorist attack (Uri attack). Present study retrieve tweets about Uri attack and find emotions and polarity of tweets. To mine emotions and polarity in tweets, text mining techniques are used. Approximately 5000 tweets are recoded and pre-processed to create a dataset of frequently appearing words. R is used for mining emotions and polarity. Experimental result showed that 94.3% people were disgusted by Uri attack.**

Keyword-Twitter, text mining, sentiment analysis, polarity, R

I. INTRODUCTION

With increase in social awareness; popularity of social networking such as twitter is increased. Twitter is one of the important and popular social media where anyone can post tweets about any event. This is open platform where people may express their views/opinions or emotions freely. Due to less internet charges, less expensive portable devices and increase social importance; people have twitter account. Most of them tweet on different events. In the social networking age people express theirs opining and feelings through twitter. So twitter contains huge amount of data. We know that length of any tweets is not more than 140 characters so people can write tweets with correct sentiment/emotions for each word.

Sentiment analysis or opinion mining is nothing but analysis of opinions or emotions from text data. Sentiment analysis identifies opinion or sentiment of each person with respect to specific event. For sentiment analysis we need to pass document or text which can be analyzed and generates system or model which represent summarized form of opinion of given document.

Twitter sentiment analysis is one of recent and challenging research area. As social media like twitter contains huge amount of text sentiment data in the form of tweets it is useful to identify sentiments or opinion of people about specific event.

Sentiment analysis or opinion mining is useful for review of movies, products, customer services, opinion about any event etc. This helps us to decide whether specific item or service is good/bad or preferred or not preferred. It is also useful to identify opinions of people about any event or persons and also finds polarity of text whether positive, negative or neutral. Sentiment analysis is a type of text classification which can classify text into different sentiments.

II. RELATED WORK

Hybrid classification technique has been used for sentiment classification of movies reviews. Integration of different feature sets and classification algorithms such as Naïve Bayes, Genetic algorithm has been carried out to analyze performance on the basis of accuracy. The output of research works shows that hybrid NB-GA is efficient and effective than base classifier and comparing in NB and GA, GA is more efficient than NB. [1]

Polarity of document is also an important aspect in text mining. Future engineering with tree kernel has been discussed by [2]. This technique gives better result than other techniques. In the paper author has define two classification models namely 2-way and 3-way classification. In 2-way classification, sentiments are classified into either positive or negative and in 3-way classification, sentiments are classified into positive, negative or natural. Author considers Tree based representation of tweets in tree kernel method. Tree kernel based model achieved best accuracy and best feature based model. Experiment achieves 4% gain than unigram model. [2]

Hierarchical approach for sentiment analysis can be used for cascaded classification [3]. Author cascaded 3 classification- objective versus subjective, polar versus non-polar and positive versus negative to make hierarchical model. This model was compare with 4-way classification (Objective, Neutral, Positive, Negative) model. The output of comparison shows that hierarchical model outperform 4-way classification model. [3]

A domain specific feature based model for movies review has been developed by [4]. Here aspect based technique is used, which analyzes text movie reviews and assign sentiment label to it on the basis of aspect. Each aspect is then aggregated from multiple reviews to find sentiment score of specific movie. Author uses SentiWordNet based technique for feature extraction and to compute document level sentiment. The result obtained by algorithm is compared with Alchemy API result. The result of comparison shows feature based model result is better than Alchemy API technique. In short aspect wise sentiment result is better than document wise result. [4]

A huge collection of near about 300000 corpus tweets for sentiment analysis and opinion mining is collected by [5]. A sentiment classifier model is build which identifies tweets positive, negative or neutral. In this technique, collected corpus was divided into 3 sets namely positive emotions- happiness, amusement or joy; Negative emotions- sadness, anger or disappointment and Neutral-text doesn't contains emotions. Tree Tagger is used for POS-tagging for distribution of emotions.

Consumer marketing data is used for collecting sentiments about product and collected data is used for future prediction. Consumer review data is huge amount of data so author uses hadoop environment for sentiment analysis. Experimental work created hadoop clusters for analysis of data. Tweets were categorized as positive, negative and neutral [6].

Hadoop's FLUME and HIVE tools are also used for analysis of twitter data. FLUME tool extracts data and stores into HDFS form. HIVE tool is used to extract and analyze data from HDFS type storage. HIVE tool is helps in analysis of different topics by changing keywords. Author identifies sentiments and polarity of tweets from election voting data [7].

Twitter data is also automatically classified into positive, negative and neutral according to query term used in consumer review tweets. In the paper author uses Parts Of Speech (POS) polarity technique and tree kernel technique. Research work uses two types of resources such as hand dictionary of emotions and dictionary collected from web. Author used different types of classification and feature extraction algorithm.[8]

III. SENTIMENT ANALYSIS

Six different sentiments can be analyzed using sentiment package namely anger, disgust, fear, joy, sadness and surprise. By using word cloud frequently occurring words were recorded. A sentiment was added to these frequently occurring words. These new words and sentiments are added to sentiment file for sentiment analysis. Present uses bayes algorithm. Sentiment analysis algorithm compares each word with words in sentiment file and assigns count for each sentiment. Finally it can display count for each sentiment.

Present work also finds polarity of text. Polarity will be positive, negative or neutral. In this experiment new words were identified using word cloud and then polarity was assigned to them. Similar to sentiment analysis, it also compares each word with polarity word file and counts polarity of text file. Lastly it displays count for each polarity.

A. Dataset Creation

Dataset was created from retrieval of Uri Attack tweets. To retrieve tweets a tweeter application was generated to get ConsumerKey, Consumer_Secret, Access_Token and Access_Token_Secret. These keys are used to connect R Studio and tweeter application. Once the connection is done, providing search term as "Uri Attack" a dataset of 5000 tweets was created. This data set was pre-processed to eliminate duplicate tweets so final dataset contained 1788 tweets

B. Data Pre-processing

The final dataset consisted of raw tweets, which needed pre-processing to get good results. Tweets were processed to remove stop words, frequent usage words such as conjunctions, numbers, prepositions, names, base verbs, etc. These type of words do not play any important role in sentiment analysis. Following are pre-processing steps.

- Filtering-In this step tweets are cleaned by removing inks, some special words, emotions symbols, user names etc.
- Tokenization-In this step tweets are separated into different tokens.
- Stopwords Removal-Stop words are nothing but specific common words which have no analytic value are removed from the tweets.
- stemDocument-This step is used to remove common word endings such as "ing", "es", "s" etc.
- Remove White Spaces-Each text contains lots of white spaces. In this step white spaces are removes.
- Convert to lower-After removing all unnecessary terms from text, it is converted to lower case.

C. Experimental Work

Experimental work is carried out on Windows XP operating system. Configuration used for working environment is Intel i3, 3.3 GHz core processor with 2 GB RAM. For experimental purpose R studio version 0.99.486 and R version 3.2.2 is used.

R is open source statistical programming language and software environment. It is mostly useful for data manipulation, data analysis, calculations and visualization of result in graphical format. Some important characteristics of R are; it is vast capabilities, wide range of statistical, graphical, data mining, data analysis techniques or functionalities, excellent community support, extensible functionality and lots of help is available. It is mostly useful for educational and research purpose. It is open source and freely available [9].

Present study uses R for sentiment analysis of Uri Attack. R has rich set of built in packages such as tm, sentiment, wordcloudetc [10] are used for sentiment analysis.

A sentiment analysis algorithm is applied on pre-processed dataset of tweets which gave 6 different sentiments with its counts and polarity of each tweet positive, negative or neutral. Table 1 show sentiment score for each emotion whereas table 2 shows polarity count.

Table 1. Sentiment score for each emotion

Emotions	**anger**	**disgust**	**fear**	**joy**	**sadness**	**surprise**
Count	330	351	994	85	11	17
Percentage	18.46	19.63	55.59	4.75	0.62	0.95

Table 2. Polarity count for each emotion

Polarity	positive	negative	natural
Count	320	1201	267
Percentage	17.90	67.17	14.93

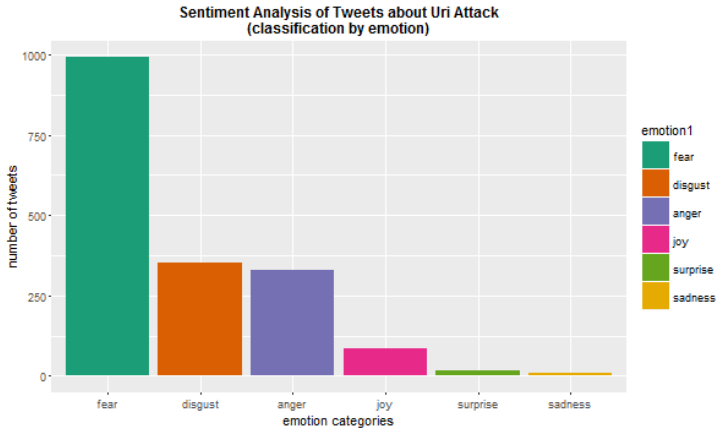

Figure 1: Sentiment analysis of Uri Attack based on emotions

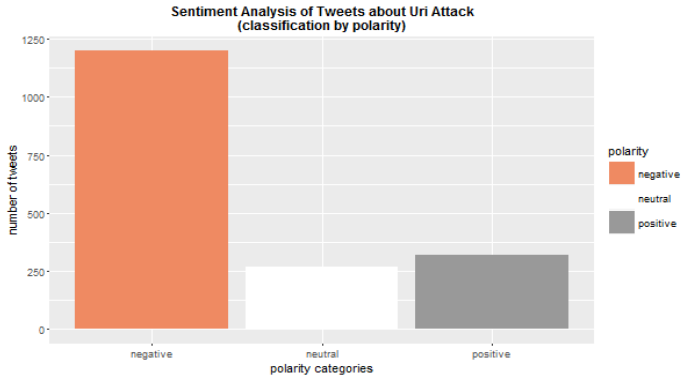

Figure 2: Sentiment analysis of Uri Attack based on Polarity

IV. OBSERVATIONS

Present study classifies tweets into six different emotions namely anger, disgust, fear, joy, sadness and surprise. Also we classified these tweets into three polarities namely positive, negative and neutral. Table 1 shows count of sentiment for different sentiments and table 2 shows polarity count for each type of polarity.Figure 1 and figure 2 shows sentiment analysis about Uri attack according to emotions and polarity.

From table 1, it is observed that most of the people have fear about this event. Fear count was highest among all the emotions which are equal to 55.59%. Table also shows that anger and disgust emotions are little bit similar and are 18.46% and 19.63% respectively. Some tweets are related to sadness and are equal to 0.62%. Table also shows that some people were surprised about Uri attack. But shameful fact is that 4.75% people were in joy about Uri attack. This group of people may be terrorist or terrorist supporter. Overall we say that near about 94.3% people have fear, anger, sadness and disgust emotions, and as 5.7% people were surprised and joyed about Uri attack.

Polarity of tweet's aim is to identify overall conceptual polarity of writer. Table 2 shows polarity of the tweets. From table 2, we observe that 67.17% tweets express negativety, 14.93% express neutral stance and 17.90% were positive.

According to human psychology, any attack or terrorist activity generates panic. Present work accurately classifies people's emotions about Uri attack.

V. CONCLUSION

Uri attack was an attack by four terrorist on 18-sep-2016 on Indian military camp located near Uri town in Jammu And Kashmir State. This attack was condemned by the world. People all over the world tweeted on Uri attack. Tweets were extracted using Twitter API. This text data is pre-process by different technique to extract features which were helpful to identifies emotions and polarity of tweets. Present study accurately classifies emotions as per human psychology. It is useful to discover opinion/ sentiments of people when they tweet. It also helps in identifying polarity of tweets. All this is carried out using R studio and its text mining packages.

All over the world many netizens have tweeted on this event but only 5000 tweets were considered for analysis due to restriction of space and processing capability, remaining tweets with their emotions and polarity were not considered. In future big data analysis technique can be used to classify all emotions for large volume of tweets.

REFERENCES

[1] M.Govindarajan, Sentiment Analysis of Movie Reviews using Hybrid Method of Naive Bayes and Genetic Algorithm , International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970), Volume-3 Number-4 Issue-13 December-2013
[2] Apoorv Agarwal, BoyiXie Ilia Vovsha, Owen Rambow, Rebecca Passonneau, Sentiment Analysis of Twitter Data
[3] Apoor v Agarwal, Jasneet Singh Sabharwal, End-to-End Sentiment Analysis of Twitter Data, Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data, pages 39–44,COLING 2012, Mumbai, December 2012.
[4] V.K. Singh, R. Piryani, A. Uddin,P. Waila, Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification, Conference Paper March 2013, DOI: 10.1109/iMac4s.2013.6526500
[5] Alexander Pak, Patrick Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining
[6] Ajinkya Ingle, Anjali Kante, ShriyaSamak, Anita Kumari, Sentiment Analysis of Twitter Data Using Hadoop, International Journal of Engineering Research and General Science Volume 3, Issue 6, November-December, 2015, ISSN 2091-2730, www.ijergs.org
[7] Sangeeta, Twitter Data Analysis Using FLUME & HIVE on HadoopFrameWork,Special Issue on International Journal of Recent Advances in Engineering & Technology (IJRAET) V-4 I-2 For National Conference on Recent Innovations in Science, Technology & Management (NCRISTM) ISSN (Online): 2347-2812, Gurgaon Institute of Technology and Management, Gurgaon 26th to 27th February 2016
[8] VarshaSahayak, VijayaShete, ApashabiPathan, Sentiment Analysis on Twitter Data, International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163, Issue 1, Volume 2 (January 2015)
[9] http://www.rstudio.com accessed 10-Feb-2016
[10] https://cran.r-project.org/web/packages/ accessed 15-Feb-2016

AUTHOR PROFILE

Mr. Dipak R. Kawade is working asassistant professor in department of Computer Science at Sangola College affiliated to Solapur University (MS) India. Research Interest are Data mining, Text Mining, Digital Image Processing, Pattern mining etc.

Dr. Kavita S. Oza working asassistant professor in department of Computer Science at Shivaji University, Kolhapur (MS) India. Her area of research is machine learning and big data analytic.