Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Sentiment Analysis on Tweet Dataset Using Data Mining Techniques

Dola Saha, Prajna Paramita Ray
Asst. Professor, CA Dept, GNIT,
WBUT, West Bengal, India

Abstract— Twitter, a popular micro blogging tool, is strong growing and widely using to expand information since its launch in October, 2006.There would be value to several domains in discovering and visualizing sentiments in online posts. Sentiment analysis is used for knowing voice or response of crowd for products, services, organizations, individuals, movie reviews, issues, events, news etc... In this paper we are going to discuss about exiting methods, approaches to do sentimental analysis for unstructured data which reside on web. Currently, Sentiment Analysis concentrates for subjective statements or on subjectivity and overlook objective statements which carry sentiment(s). So, we propose new approach classify and handle subjective as well as objective statements for sentimental analysis. This paper presents Sentiment Analysis to identify the topic/content of a tweet, the noun clause is considered as features for each tweet and K-Means clustering algorithm is applied to partition the tweets into two different clusters and finally the clusters are validated to measures the accuracy of the work.

Keywords— Microblog, Twitter, Sentiment, Sentiment Analysis, Sentiment Classification

I. INTRODUCTION

Microblog is a form of a blog where people send short messages of text or media (pictures, video, or sounds). These messages can be sent to a website and shown to either a small group or to the public. A "microblogger" (someone who uses a microblog) could send their messages from different sources including cell phone text messages, email, instant messagesor through a website.

Microblogging has quickly grown as the avatar of social interaction. Though many websites like Friend Feed, Daily booth,and Tumblr support microblogging, Twitter is the most favored and widely used website. Boasting more than 500 million registered users, about 1 million new accounts are added and over 400 million tweets are posted every day. Twitter's ability to propagate real-time information to a wide set of users makes it a potential system for disseminating vital information, and an invaluable source of news repository.

A **Microblog** is usually different from a normal blog. It has much smaller pieces. A microblog entry can be just one sentence, or a link to an image or short video. A person writing a microblog can use many ideas from "what am I doing now," to Race cars or politics or information about their business or personal life.

Micro blogs are in use in many ways on different websites. They became popular with sites like Twitter where people send messages to each other with text comments or links to other media. Newer sites are also available now which allow you to share media like videos, pictures or sounds directly instead of sending a link.

Many social networking sites (such as Facebook or MySpace) use a type of microblogging using status updates where the person tells their friends what they are thinking or doing.

Twitter is a microblogging service that was founded in early 2006 to enable people to share short textual messages—"tweets"—with others in the system. Because the system was originally designed for tweets to be shared via SMS, the maximum length of a tweet is 140 characters. Though the service evolved to include more uses besides SMS, such as web and desktop clients, this limitation persisted, and so was re-narrated as a feature. Twitter's Creative Director Biz Stone argues, "creativity comes from constraint". Twitter combines elements of social network sites and blogs, but with a few notable differences. Like social network sites, profiles are connected through an underlying articulated network, but these connections are directed rather than undirected; participants can link to ("follow") others and see their tweets, but the other user need not reciprocate. Like blogs, participants' Twitter pages show all of their tweets in reverse chronological order, but there is no ability to comment on individual posts. User profiles are minimal and public, but users can make their tweet stream public or protected (a.k.a. private); the default and norm is public. The central feature of Twitter, which users see when they log in, is a stream of tweets posted by those that they follow, listed in reverse chronological order. Participants have different strategies for deciding who they follow—some follow thousands, while others follow few; some follow only those that they know personally, while others follow celebrities and strangers that they find interesting.

Sentiment means a view or opinion, but it can also mean an emotion. Maybe you prefer tragic movies because you enjoy the **sentiment** of sadness. This meaning of **sentiment**s taken to an extreme in yet another version of the word, meaning something like "overdone, exaggerated feelings, especially of sadness or nostalgia."

Sentiment is a combination of beliefs and emotions that explains an action.
Sentiment analysis is done on three levels

- Document Level
- Sentence Level
- Entity or Aspect Level.

Document Level Sentiment analysis is performed for the whole document and then decide whether the document express positive or negative sentiment.
Sentence level sentiment analysis is related to find sentiment form sentences whether each sentence expressed a positive, negative or neutral sentiment. Sentence level sentiment analysis is closely related to subjectivity classification. Many of the statements about entities are factual in nature and yet they still carry sentiment. Current sentiment analysis approaches express the sentiment of subjective statements and neglect such objective statements that carry sentiment.
Entity or Aspect Level sentiment analysis performs finer-grained analysis. The goal of entity or aspect level sentiment analysis is to find sentiment on entities and/or aspect of those entities. For example consider a statement "My HTC Wildfire S phone has good picture quality but it has low phone memory storage." so sentiment on HTC"s camera and display quality is positive but the sentiment on its phone memory storage is negative. We can generate summery of opinions about entities. Comparative statements are also part of the entity or aspect level sentiment analysis but deal with techniques of comparative sentiment analysis.
Sentiment Analysis (also known as **opinion mining**) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.
Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation (see appraisal theory), affective state (that is to say, the emotional state of the author when writing), or the intended emotional communication (that is to say, the emotional effect the author wishes to have on the reader).
Sentiment classification: Classify topic/sentence/document/feature based on sentiments expressed by authors like positive, negative, neutral.
1) Topic based classification: topic words are important.
2) Sentence level classification: A sentence contain only one opinion.
Task 1:Identify if sentence is opinionated.
Task 2:Determine polarity of sentence.
3) Document level classification: Each document focuses on a single object.
Contains opinion from a single opinion holder
Task: Determine overall sentiment orientation in document.
4) Feature Level Classification: Produce a feature-based opinion summary of multiple reviews.
Task1: Identify and extract object features that have been commented on by an opinion holder.
Task2: Determine polarity of opinions on features.
Task3: Group feature synonyms.

II. STEP BY STEP SENTIMENT ANALYSIS ON TWITTER DATA WITH UTTARAKHAND FLOOD TWEETS, AUGUST 28, 2013

Goal: To identify the topic/content of a tweet, the noun clause is considered as features for each tweet and K-Means clustering algorithm is applied to partition the tweets into two different clusters and finally the clusters are validated to measures the accuracy of the proposed work.
Step 1: We will store these tweets in **uttarakhand.tweets** to a file for analysis and reference. Collected 2087 tweets in this file. We are going to convert the list of tweets to separate data by applying some techniques. A sample of this file is like
tweet-id | userid of user who posted tweet | time-stamp of tweet | text of tweet#
345367048655876096|403539516|2013-06-14 04:28:15|15,000 tourists stranded due to landslide in Uttarakhand http://t.co/x6H9xag4Te
345897263186472960|120742408|2013-06-15 15:35:08|Its raining heavily luvly weather in UTTARAKHAND
346295432680452097|17710740|2013-06-16 17:57:19|8 perish as rains lash Uttarakhand, Char Dham yatra suspended http://t.co/cH6pcRrD41
346426953433223168|72500509|2013-06-17 02:39:56|cc: @MrsGandhi look at this Anti-hindu rains "@timesofindia: Rains bring Chardham Yatra to a halt in Uttarakhand http://t.co/JU7dXy6eHS"
346480267256532992|400756107|2013-06-17 06:11:47|8 perish as rains lash Uttarakhand, Char Dham yatra suspended http://t.co/nqjemlY8lh
346480950936494080|235621342|2013-06-17 06:14:30|Uttarakhand: Nature's fury June 2013: http://t.co/l2NCVrPeGl via @youtube
346499347137060864|368346924|2013-06-17 07:27:36|Rains lash Uttarakhand, Char Dham yatra suspended.....big news
Step 2: Now the tweets and all the necessary information is extracted in this file and are available in the result.tweets data file. You can look some of the samples at below for the references.
15,000 tourists stranded due to landslide in Uttarakhand

Its raining heavily luvly weather in UTTARAKHAND
8 perish as rains lash Uttarakhand, Char Dham yatra suspended
look at this Anti-hindu rains Rains bring Chardham Yatra to a halt in Uttarakhand
8 perish as rains lash Uttarakhand, Char Dham yatra suspended
Uttarakhand: Nature's fury June 2013
Rains lash Uttarakhand, Char Dham yatra suspended.....big news
8 dead 3700 pilgrims stranded as incessant rains batter Uttarakhand:Eight persons were killed on Sunday
Uttarakhand monsoon rains 'kill 10': At least 10 people are killed in landslides and flooding.

Step 3: Remove the stop words from the Data Set.

Stop words are words which are filtered out before or after data (text). There is not one definite list of stop words which all tools use and such a filter is not always used. Some tools specifically avoid removing them to support phrase search. Any group of words can be chosen as the stop words for a given purpose. For some search engines, these are some of the most common, short function words, such as *the*, *is*, *at*, *which*, and *on*. In this case, stop words can cause problems when searching for phrases that include them, particularly in names such as 'The Who', 'The The', or 'Take That'. Other search engines remove some of the most common words—including lexical words, such as "want"—from a query in order to improve performance.

List of Stop Words:

a,able,about,across,after,all,almost,also,am,among,an,and,any,are,as,at,be,because,been,but,by,can,cannot,could,dear,did,do,does,either,else,ever,every,for,from,get,got,had,has,have,he,her,hers,him,his,how,however,i,if,in,into,is,it,its,just,least,let,like,likely,may,me,might,most,must,my,neither,no,nor,not,of,off,often,on,only,or,other,our,own,rather,said,say,says,she,should,since,so,some,than,that,the,their,them,then,there,these,they,this,tis,to,too,twas,us,wants,was,we,were,what,when,where,which,while,who,whom,why,will,with,would,yet,you,your

After removing the stop words Getting a new file. Some of the samples of the file looks like
15,000 tourists stranded due landslide Uttarakhand
raining heavily luvly weather UTTARAKHAND
8 perish rains lash Uttarakhand, Char Dham yatra suspended
look Anti-hindu rains Rains bring Chardham Yatra halt Uttarakhand
8 perish rains lash Uttarakhand, Char Dham yatra suspended
Uttarakhand: Nature's fury June 2013
Rains lash Uttarakhand, Char Dham yatra suspended.....big news
8 dead 3700 pilgrims stranded incessant rains batter Uttarakhand:Eight persons were killed Sunday.

Step 4: Stemming

Stemming is the term used in linguistic morphology and information retrieval to describe the process for reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. The stem needs not to be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. Algorithms for stemming have been studied in computer science since the 1960s. Many search engines treat words with the same stem as synonyms as a kind of query expansion, a process called conflation.

A stemmer for English, for example, should identify the string "cats" (and possibly "catlike", "catty" etc.) as based on the root "cat", and "stemmer", "stemming", "stemmed" as based on "stem". A stemming algorithm reduces the words "fishing", "fished", and "fisher" to the root word, "fish". On the other hand, "argue", "argued", "argues", "arguing", and "argus" reduce to the stem "argu" (illustrating the case where the stem is not itself a word or root) but "argument" and "arguments" reduce to the stem "argument".

After stemming ,Getting a new file. Some of the samples of the file looks like
15,000 tourist strand due landslide Uttarakhand
8 perish rain lash Uttarakhand, Char Dham yatra suspend
look Anti-hindu rain Rain bring Chardham Yatra halt Uttarakhand
8 perish rain lash Uttarakhand, Char Dham yatra suspend
Uttarakhand: Nature fury June 2013
Rain lash Uttarakhand, Char Dham yatra suspend.....big news
8 dead 3700 pilgrim strand incessant rain batter Uttarakhand: Eight person kill Sunday
Uttarakhand monsoon rain 'kill 10': 10 people kill landslide flood
National Disaster Response Force send 12 team #Uttarakhand, 10,000 tourist strand Badrinath
Thousand pilgrim report stuck the hilly region Uttarakhand. way various Hindu shrine
Monsoon fury India, 10 dead Uttarakhand.

Step 5: Tagging the Each tweet of the above result using POS Tagger.

A **Part-Of-Speech Tagger** (**POS Tagger**) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained **POS** tags like 'noun-plural'.

In corpus linguistics, **part-of-speech tagging** (**POS tagging** or **POST**), also called **grammatical tagging** or **word-category disambiguation**, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context—i.e. relationship with adjacent and related words in a phrase,

sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.
Once performed by hand, POS tagging is now done in the context of computational linguistics, using algorithms which associate discrete terms, as well as hidden parts of speech, in accordance with a set of descriptive tags. POS-tagging algorithms fall into two distinctive groups: rule-based and stochastic. E. Brill's tagger, one of the first and most widely used English POS-taggers, employs rule-based algorithms.
15/CD ,/, 000/CD tourist/NNP strand/VB due/JJ to/TO landslide/VB Uttarakhand/NNP
rain/VBN heavily/RB luvly/RB weather/RB in/IN UTTARAKHAND/NNP
8/CD perish/NN as/IN rains/NNS lash/VBP Uttarakhand/NNP ,/, Char/NNP Dham/NNP yatra/NN suspended/VBD look/VB at/IN this/DT Anti-hindu/NNP rains/VBZ Rains/NNP bring/VBG Chardham/NNP Yatra/NNP to/TO a/DT halt/NN in/IN Uttarakhand/NNP
8/CD perish/NN rains/NNS lash/VBP Uttarakhand/NNP ,/, Char/NNP Dham/NNP yatra/NN suspended/VBD Uttarakhand/NN :/: Nature/NNP 's/POS fury/NN June/NNP 2013/CD
Rains/NNP lash/NN Uttarakhand/NNP ,/, Char/NNP Dham/NNP yatra/NNP suspended/VBD .../: .../: .../: .../: .../: .../: .../: .../: .../: .../: .../: big/JJ news/NN
8/CD dead/JJ 3700/CD pilgrims/NNS stranded/VBD as/IN incessant/NN rains/NNS batter/NN Uttarakhand/NNP :/: Eight/CD persons/NNS were/VBD killed/VBN on/IN Sunday/NNP
Step 6: Noun Clause is considered as features for each tweet. Since maximum occurances among all the parts of speech is Noun. Extracting the noun list from the above file
The Noun list is given below:
tourist
Uttarakhand
UTTARAKHAND
perish
rain
Uttarakhand
Char
Dham
yatra
Anti-hindu
Rain
Chardham
Yatra
halt
Step 7: Find the list of unique nouns from the above result.
Total Number of Unique Nouns are: 1398. Some of the samples are
tourist
UTTARAKHAND
perish
rain
Anti-hindu
Chardham
Yatra
halt
Nature
fury
June
lash
news
pilgrim
incessant
Step 8: Creating the table for entire set of the tweets. The rows of the table represent each individual table and the column of the table represents each unique noun of the noun list.
Step 9: Now calculate the frequency of each noun, within the dataset. Taking only those nouns having the occurrences 4% and above.
List of those Nouns:
Uttarakhand:2210
rain:117
pilgrim:102
people:271
India:180
flood:392
helpline:92

http:1173
relief:255
rescue:216
Kedarnath:407
Army:133
@:218
CM:102
Govt:127
RT:434
operation:113
victims:119
\E2\80:113
Disaster:121

Step 10: Now creating a new database for the entire datasets with respect to the columns
The database contains the frequency of each noun in each tweets. Some of the samples are given below:

ID	Uttarakhand	rains	pilgrims	people	India	flood	helpline	http	relief	rescue	Kedarnath	Army	@
2	1	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	0	0	0	0	0	0	0	0	0	0	0
4	1	1	0	0	0	0	0	0	0	0	0	0	0
5	1	1	0	0	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	0	0	0	0	0	0	0
7	1	1	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	1	0	1	0	0	0	0	0	0	0	0	0	0
9	1	0	0	0	0	1	0	0	0	0	0	0	0
10	0	1	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0

Step 11: K-Means clustering algorithm is applied to partition the tweets into two different clusters
The Final Clusters By K-Means Clustering are as follows:
K1{
1 2,3,4,5,6,7,8,9,10,11,12,13,
14,15,16,18,18,19,20,21,22,23,
26,27,28,30,32,33,34,35,36,37,
……,
……,
……..,
……..,
}
K2{
24,25,29,31,38,40,87,91,98,106,
115,119,127,138,146,147,156,159,
189,192,198,208,211,231,242,289,
……,
……..,
……..,
……..,
}

Step 12:
Finally the clusters K1 and K2 are validated
Output:

Topic/Content	# of Tweets	Accuracy
Flood related Tweet	2172	76.84%
Non Flood Tweet	655	23.16%

Accuracy of the work =76.84%

III. RELATED WORK

There are a number of research studies that focus on predicting the users' location by either mining the content of their tweets, or by using the twitters' network information. For example, Cheng et al. [1] aim to solve this problem using the

textual content of tweets to estimate the location of users at city level, while [2] predicts the user's point of interests such as club or hotels by considering tweet's content and temporal information. Geo-tag information from Twitter data is utilized by [3] to build language models of locations at various levels of granularity. The work in [4] studies the Twitter network to analyze the impact of geography on user interactions; The authors in [5] infer states, cities, and time zones of the twitter users by using an ensemble of content based statistical and heuristic classifiers. In [6], the authors propose an unsupervised measure for evaluating the usefulness of tweet words for location prediction. [7] analyze Twitter network to study the impact of geography on user interactions.

IV. CONCLUSION

In this paper I performed sentiment classification on a novel collection of Tweet dataset related to Uttarakhand Flood on August 28, 2013 which are the comments about the flood of "Twitter" users. Tagging Each of tweet Using E. Brill's POS Tagger. For identifying the topic/content of a tweet, the noun clause is considered as features for each tweet and K-Means clustering algorithm is applied to partition the tweets into two different clusters and finally the clusters are validated to measures the accuracy of the proposed work.

REFERENCES

[1] "Opinion Mining and Sentiment Analysis" Bo pang and Lillian Lee, 2008

[2] G.Vinodhin,RM.Chandrasekar, an Sentiment Analysis and Opinion Mining, IJARCSSE

[3] Bai Wang. Community Detection in Large-Scale Social Networks Beijing University of Posts and Telecommunications, China*Review E*, 70(066111), 2004.

[4] A. Clauset, M. Newman, and C. Moore. Finding local community structure in networks. Physical Review E.

[5] Huang Sui, You Jianping, Zhang Hongxian, Zhou Wei Department of Computer Science Jinan University Guangzhou, China, 2nd International Conference on Computer Science and Network Technology.

[6] A. Clauset, M. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(066111), 2004.

[7] Ali Ghobadi, Maseud Rahgozar, "An Ontology -based Semantic Extraction Approach for B2C eCommerce", The International Arab Journal of Information T echnology, Vol. 8, No. 2, April2011.

[8] Ali Ghobadi, Maseud Rahgozar, "An Ontology -based SemanticExtraction Approach for B2C eCommerce", The InternationalArab Journal of Information T echnology, Vol. 8, No. 2, April 2011.

[9] Kang Wu, Bofeng Zhang, Jianxing Zheng, Haidong Yao School of Computer Engineering & ScienceShanghai University Shanghai, China." 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber" bfzhang@shu.edu.cn "**Sentiment Classification for Topical Chinese Microblog Based on Sentences' Relations**"

[10] "Text Mining Facebook Status Updates for Sentiment Classification"Jalel Akaichi Computer Science DepartmentInstitut supérieur de gestion de Tunis (ISG) 41, rue de la Liberté 2000 Le Bardo Tunisia, Zeineb Dhouioui Computer Science DepartmentInstitute supérieur de gestion de Tunis (ISG) 41, rue de la Liberté 2000 Le Bardo Tunisia, dhouioui.zeineb@hotmail.fr

