Topic Discovery and Opinion Mining from Twitter Data

Minu T Lalson

PG Student
Computer science & Engineering
St. Joseph's College of Engineering & Technology, Palai, India

***Abstract*—Twitter is one of the most visited social media site by people. People 'talk' on different trending social events. They share their opinion with other users. The document streams published in twitter are commonly called 'tweets'. The content of tweets reflects to some topics. Some topics discussed among a group of people may be based on any event happened on a particular location. Therefore twitter can be used as a source of data for topic discovery and opinion mining. The idea is to build a system that do both process: Topic discovery and opinion mining. This idea can be implemented in twitter like micro blogging sites for rare event discovery and user behavior analysis. TF-IDF and SentiWordNet can be used for topic discovery and opinion mining respectively.**

***IndexTerms*—Sentiment analysis, text mining, topic discovery, twitter, user behavior**

I. INTRODUCTION

Social media sites have become an unavoidable part of people's life now. Twitter, Facebook, Flickr are some important services among them. People use these sites to share their opinion on different events happening around them. Twitter is one of the most used microblog where people share their feelings and ideas in small sentences, called tweets.Some tweets may go viral or they can make a way to big discussions among people. The tweets can be called as documents published by users. The content of each document will be concentrated on any topic. This topic can reflect to real life of user or they can reflect to the personality or attitude of the user. Some users can behave in an abnormal way. Generally, the attitude of a person towards a topic can be classified as positive, negative or neutral. It depends on the topic they discussed.

There exist different systems that do either topic discovery or opinion mining on twitter data. The idea is to build a system that does both topic discovery and opinion mining. Here we find topics on rare events that are discussed only among some people or group of people. Then, finds the attitude of user towards the event. The system consists of mainly three modules: dataset collection, topic discovery and opinion mining.

II. RELATED WORK

Twitter data has been used in many research studies. Since twitter is a very much popular social media site, the studies are done to identify the user approaches. This helps in various fields like business and politics to know what is the approach of user towards a product or how much support the political parties get etc.,. Pak and Paroubek [1] used twitter as a corpus for sentiment analysis and opinion mining. They build a sentiment classifier using the multinomial Naïve Bayes classifier. Using this classifier they found positive, negative and neutral sentiments. Zhu et.al [2] shows how to mine rare sequential topic patterns from documents published in internet. They use LDA for data preprocessing. This topic model is not efficient in small documents like tweets of twitter. Gimpel et.al [3] address the problem of part-of-speech tagging for English data from the popular microblogging service Twitter. They develop a tagset, annotate data, develop features, and report tagging results. But the results obtained is not highlyaccurate. A system is needed which do both topic discovery and opinion mining in a better efficient way.

III. DATASET COLLECTION

Twitter data is the corpus for the system. Using twitter API collected tweets (only text posts) of some active users from a particular location. The posts within a particular time period (within a month or week) are collected and store as text files. And this is given as input to the system.

An API is a program that can be integrated with our program to collect data. Inorder to retrieve data using twitter API we need to register our application in twitter. They will provide us some access keys that we need to use in our program. After that we can access twitter and collect needed data. Tweets of a group of active users from a particular location is collected and stored as set of textual documents. This is given to the next phase of system.

IV. TOPIC DISCOVERY

Thetopic discovery is the phase where we find the topics related to some rare events happened in a location. The topics discussed in different documents on a single event show some correlation to each other. Most of the topics discussed bypeople show a sequential relation. The topics will be repeatedly used in many tweets. Topics are the meaningful terms that are present most in documents. Therefore we can find it using any clustering approaches in text mining. Term weighting methods are useful in these situations. So here we use TF-IDF (Term frequency-Inverse document frequency) technique. Before applying TF-IDF to the textual documents it is needed to do some preprocessing to the textual documents. Preprocessing is the initial step in almost all text mining approaches. So in topic discovery phase we do two steps: Preprocessing and TF-IDF technique.

1. Preprocessing

In preprocessing the textual documents are converted into term level document where the TF-IDF can be applied easily. The preprocessing converts the input to a proper form from which correct output can be obtained. The techniques used in preprocessing are: tokenization, filtering, stopword removal and POS (Part-of-Speech) tagging.

a. Tokenization: It is the process where the text stream is converted into stream of words called tokens. The text is split up by white space and comma (,).

b. Filtering: The tweets may consist of special characters and links to other pages or sites. Some tweets may be written in other languages than English. Our concern is only for English language. Therefore by using pattern matching method all these unwanted links, special characters or symbols are removed.

c. Stopword removal: Stopwords are the most common words that don't have higher significance in a sentence. A, an, and, are, the, but, so etc. are some common stopwords. Since they have no higher importance we remove it from our documents. For that first, prepare a standard list of stopwords and compare it with the words in the document set. Remove those words that get matched. The stopword removal process helps to reduce the usage of memory.

d. POS Tagging: This is the toughest technique among all the preprocessing techniques. Parts-of-Speech is the parts in a standard sentence. In POS tagging, the parts-of speech of a sentence are tagged or marked with their names as tags. This helps to identify the important words in the document. Noun, verb, adverb, adjective are some of the parts-of-speech. To tag the words in a document we can a use a tagger. A tagger is a program that does the process of POS tagging. In the system, Stanford POS tagger [4] is used to do this job. This tagger use dictionaries and rules to do tagging. Noun, verb, adverb and adjective are the important parts-of-speech which carry most of the essence in a sentence. In order to make the processing easier and efficient, those important words are only kept. All other words are removed.

After preprocessing, now the system has documents consist of only important words or can be called as terms.

2. TF-IDF

It is the complete vector space model. TF-IDF (Term frequency-Inverse document frequency)[5] measure is a most commonly used text retrieval method in text mining approaches. It is a term-weighting method. A weight is calculated for each term in the document set. This shows how important the word is to the document.The term frequency is the number of times a term is present in a document. The inverse document frequency counts the number of document which contains the term. The term frequency and inverse document frequency are multiplied to get the measure TF-IDF. This measure is found for every term in the document set. And this value shows the important words or the topics from the document set.
For term t in document d:

TF (t)=number of times t present in d/total number of terms in d (1)
IDF (t) = log_e (total number of documents/ number of documents with t) (2)
Measure=TF (t)*IDF (t) (3)

The tweets collected are stored as textual documents and this document set is converted in to a term level document set, in preprocessing phase, in which documents consists of important terms only. Upon this document set the TF-IDF is applied to find out topics. For each term in document set the measure is calculated. Then a TF-IDF matrix is formed where each row represents documents, each column represents terms in document set and each entry contains the measure. After entering values in matrix, the average of measure for each term is calculated. The terms with highest average is selected as the topics. For that a threshold can also set. Based on this threshold value the topics can be selected. By analyzing the topics the rare event can be discovered.

V. OPINION MINING

This is the last phase of the system where the user's reaction towards the discovered topic is found. Opinion Mining or Sentiment Analysis is the process of identifying the attitude or opinion of an author towards a topic. Mainly, the reaction is categorized into positive, negative and neutral.The sentiment analysis helps to find out abnormal or illegal users. Here the system uses SentiWordNet 3.0 [6] for opinion mining. A SentiWordNet is a lexical resource used in support of sentiment analysis. The synset s in WordNet is associated with three scores: pos(s), obj(s) and neg(s) in SentiWordNet, to show how these synset are positive, negative and objective.
After topic discovery the important topics from the document set is obtained. Using these topics the most reacted users is found. This is done by counting the number of tweets that contains the topics. The users who publish more tweets on the topics are selected as the most reacted users. The sentiment analysis is done on these selected users. For that the tweets published by users on

the topics are collected separately and stored as textual files. The preprocessing (described in section IV) is again done on this document set to obtain a term level document. Now the system have important terms in document set. For each term in the document set, a sentiment score is calculated. The sentiment score helps to determine the reaction. For each synset sin SentiWordNet file calculate a score:

$$pos(s)-neg(s) \qquad (4)$$

Create a hash array which store synset as key and a vectoras value. The vector consist of index of s and its calculated score. Usually the synset are adjective, noun, verb words.Aword can appear more times in WordNet according to its usage. The index determines the number of appearances of a word in either adjective, noun or verb category. If index is higher than current vector add the remaining vector values by zero. Then for each s calculate:

$S=\Sigma(1/1*v(1),1/2*v(2),1/3*v(3),\ldots......,1/n*v(n))$, where v is vector and n is vector size (6)

$V=\Sigma(1/1, 1/2,1/3,\ldots..........,1/n)$, where n is vector size (7)

$$Score(s)=S/V \qquad (8)$$

After doing this calculation for all synset save this score like a dictionary. The scores obtained is in a range of 1 and –1. Now the terms in the document set is matched with the dictionary of scores. Each term will get three scores. From that the highest score is selected. It can be either highest positive value, highest negative value,zero or null. Depending on this score the reaction on a particular term is discovered. If a null value is obtained it shows the word is not in the dictionary.By analyzing the reactions got for each term, the overall reaction of a user is predicted. Morepositive values shows a positive attitude, morenegative value shows a negative attitude and more zero indicates a neutral reaction.

VI. CONCLUSION

Twitter has become a popular social media site used by people to share their ideas and opinions. Many studies are done on twitter data to study the user reactions like review the customer reactions on a particular product or the discover the attitude of users towards any social events. There will be some rare events discussed by users. Finding the topics related to this events helps to identify user attitude towards the event. This helps in identifying abnormal or illegal users. The proposed system does both topic discovery and opinion mining. TF-IDF technique is used in topic discovery and a lexical resource called SentiWordNet is used in opinion mining. This system can be used in identifying rare trending topics among a group of people.

References

[1] A. Pak, P. Paroubek, “Twitter as a Corpus For Sentimental Analysis and Opinion Mining”,in Proc 7th Conf. Int. Lang. Res. Eval (LREC10), May 2010.

[2] J. Zhu, et.al, “Mining User-Aware Rare Sequential Topic Patterns in Document Streams,” in IEEE Trans. Knowledge and Data Eng, 2016.

[3] K. Gimpel, N. Schneider, et. al, “Part-Of-Speech Tagging For Twitter: Annotation,Features and Experiments”, in Proc 49th Annual Meet of ACL, Vol 2, June 2011.

[4] Kristina Toutanova, et.al,“Feature-Rich Part-of-Speech Tagging with a CyclicDependency Network”, inproceedings of HLT-NAACL 2003, pp. 252-259, 2003.

[5] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", second edition, Elsevier, 2010.

[6] Stefano Baccianella, et.al, “SENTIWORDNET 3.0:An Enhanced Lexical Resource forSentiment Analysis and Opinion Mining”, 2010.