RARE TOPIC DISCOVERY AND USER BEHAVIOR ANALYSIS ON DOCUMENT STREAMS IN SOCIAL MEDIA

[1]MINU T LALSON, [2]KISHORE SEBASTIAN

[1,2]Computer Science & Engineering Department, St. Joseph's College of Engineering & Technology, Palai
E-mail: [1]minulalson@gmail.com, [2]kishore.sebastian@sjcetpalai.ac.in

Abstract - Internet contains document streams that are published in various forms like posts in social media, news streams, chats etc. Among these, documents published in social media get more focus. People use social media to express their opinion about various events. These document streams are based on some topic. Many people can talk on same topic. Therefore sequential topics can be obtained from these documents. These topics are related to some rare social events, which can happen on a particular location. Also these topics can characterize user behavior. The proposed system is a text mining approach which analyses text data from social media and discover topics related to rare events. It then analyses user's behavior towards the topic. The system contains four modules: data collection, data preprocessing, rare topic discovery and user behavior analysis. The experiment is done on twitter data which show that this approach is useful in twitter like social media sites itself.

Keywords - Document Streams, Rare Events, Social Media, Topic Discovery, User Behavior.

I. INTRODUCTION

The global network, internet, has become the unavoidable part of human life today. In internet, document streams are created and distributed in various forms like research papers, blog articles, chatting messages, emails, social media posts, news streams etc. These documents have content which concentrate on any specific topic that reflects to any social events or user characteristics in real life. Documents published in social media sites shows this reflection more.

Social media sites are getting more and more attention in people's daily life. People use these sites to express their feelings and opinions with other people. Twitter and Facebook are the most popular social media websites which is used by people to express their opinion on any topic or share their ideas. These tweets and posts reflect to the behavior of a user. Based on a location, people can talk on some social but rare events which may be the current trending topic on that location. These rare social events can be used to analyze user behavior. A user can behave in different ways like, positively, negatively or in an abnormal way. The abnormal or illegal behavior of users can make threats to social security. So we need a system that monitors behavior of users. The proposed system is a text mining approach which analyses the document streams published in social media sites and finds the topics related to some rare events that happened on a particular location. It then finds the most reacted users towards the topic and then analyses the behavior of those users. This approach needs a data collection phase to collect document stream, a preprocessing phase to process documents, a discovery phase to find topics and an analysis phase to find user behavior. This approach can be used to analyze the abnormal user behavior and highlight rare trending events in Twitter like sites itself. The twitter data set is used as an input to the system.

II. RELATED WORK

There are lot of researches are occurring in the area of text mining, more specifically, in topic mining. LDA [1] is the most common probabilistic topic model used to extract information from document streams. But LDA doesn't work well on short documents like tweets. Prefix Span [2] is the sequential pattern mining method mainly used for sequential topic mining. It requires a large memory space. Twitter data has been used in many analytics. P. Paroubek, et.al used twitter as a corpus for sentiment analysis and opinion mining [3]. They build a sentiment classifier, which is able to determine positive, negative and neutral sentiments for a document. An approach which does both topic mining and sentiment analysis is rare. Our proposed system is the one that does both this job successfully.

III. PROPOSED SYSTEM

Here, we propose a text mining approach to extract information from document streams. The processing framework is shown in Fig 1. It consists of four phases: Data Collection, Data Preprocessing, Topic Discovery, and User Behavior Analysis. At first, document streams are collected from a social media site (twitter). Then, preprocessing is done to obtain a term level document, containing document with important terms only. Topics are discovered in topic discovery phase and user behavior is done in user behavior analysis phase.

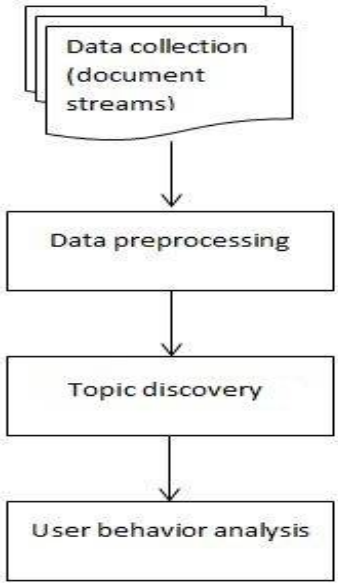

Fig.1. Processing Framework

3.1. Data Collection

The first phase is the data collection, in which textual documents are collected from sites and form a document stream. This document stream is given as the input to the system. To collect tweets of different users from twitter, an API is used. The users are chosen from a particular location, and their tweets which are published in a particular time period are collected. The time period can be a month or week.

3.2. Data Preprocessing

Preprocessing is the initial and important step in almost every text mining approaches. It converts the textual documents into a term level document which contains only the important words in the document. The frequency of terms in document set helps to determine the topics. Before finding the frequency we need to do some preprocessing steps.
The preprocessing steps are:

1. Tokenization
The text stream in each document in document set is converted into list of meaningful elements called tokens. The text stream is break by space.

2. Special Character Removal
The symbols like +, -, &, # are removed from documents. The tweets in languages other than English are filtered out. The tweets in English language are only our concern.

3. Stopword Removal
The articles like a, an, are, the, etc. are removed from documents. Stopwords are the common words that do not carry any significance in a sentence.

4. URL Removal
There will be some links in most of the tweets. It is better to remove those links from documents.

5. POS Tagging
Here the words are tagged by parts of speech of a sentence. The important parts of speech of English language are noun, verb, adverb and adjective. They carry the most essence of a sentence. The words other than these categories are removed from document. POS tagging can be done using a tagger. Stanford POS Tagger [4] is used here to obtain tagged words.

After preprocessing, we obtain documents carrying important words only. Therefore, we can say that a term level document set is obtained in this phase.

3.3. Topic Discovery

From the obtained term level document set, we have to find the topics related to the rare events. The term "rare" can be defined in the terms of type of the event people discussed on. The procedure is done for each document associated to each user. Here we use the TF-IDF algorithm. TF-IDF is a clustering algorithm used in many text mining applications. It is a term weighting scheme. For each term**t**, in each document**d**, the frequencyis calculated. This value shows how important a term is to the document and this value is called Term Frequency (TF). For the entire document set, the importance of a term is also calculated. This value is called Inverse Document Frequency (IDF).

$$TF(\mathbf{t}) = \eta/\psi \quad (1)$$

$$IDF(\mathbf{t}) = \log\_e(\tau/\omega) \quad (2)$$

Where, η is number of times term **t** appears in a document **d**, ψ is the total number of terms in the document **d**, τ is the total number of documents and ω is the number of documents with term **t** in it.
Then both these values are multiplied to obtain a weight for term **t**, called TF-IDF value. The value υ,

$$\upsilon(\mathbf{t}) = TF(\mathbf{t}) * IDF(\mathbf{t}) \quad (3)$$

A. Algorithm for calculation of TF-IDF value for each terms in document set

Input: Document set D, containing documents of tagged termsseparately for each user
Output: TF-IDF value for each **t**

1. For each **d** in D
2. For each **t** in **d**
3. Prepare list of all terms in D,***LIST***(without repetition of terms)
4. For each **d** in D
5. For each **t** in **d**
6. Prepare list of all terms ineach **d**for each **d** , ***list***(without repetition of terms)
7. For each ***list***
8. For each **t** in ***LIST***
9. if ***list*** contains **t**
10. then find TF(**t**)
11. else set TF (**t**)=0

12. For each ***list***
13. For each **t** in ***LIST***
14. Find IDF(**t**)
15. For each **t** in ***LIST***
16. Find υ (**t**) = TF (**t**) * IDF (**t**)

After this calculation, form a TF-IDF matrix like the Table 1 shown below in csv file format. Here we take an average value for each term in matrix, computed by taking average of all the TF-IDF values obtained by the term

Table 1: TF-IDF matrix format

	t1	**t2**	**t3**	**t4**…..
d1	υ11	υ12	υ13	υ14….
d2	υ21	υ22	υ23	υ24….
d3	υ31	υ32	υ33	υ34….
d4	υ41	υ42	υ43	υ44….
.	.	.	.	.
.	.	.	.	.
Average	A1	A2	A3	A4….

In Table 1,**t1**, **t2**, etc., represents all terms in ***LIST*** and **d1**, **d2**, etc., represents all the documents in D. υ11, υ12, etc., represents the TF-IDF value for each **t** in each document**d**. A1, A2, etc., represents average of TF-IDF values for each term.

The average value obtained by each term shows the importance of that term to the document set. The terms with highest average values will be the ones that have more importance in the document set. That means they are the topics related to the events discussed by the users. Thus by checking the average values we can find the topics.

3.4. User Behavior Analysis

This is the last phase of the system. The content of tweets reflects the user behavior. By analyzing the tweets we can find out what is the attitude of usertowards the topic. In this phase, first, we find out the most reacted users towards the topic and then find the reaction of those users. To find the reaction of users the sentiment analysis technique is used. Sentiment analysis is a method to find out the opinion or reaction of an author towards a topic without considering its content. We do this analysis to find out the reaction of users towards the discovered topics. This helps to find out whether there is any abnormal behavior or illegal behavior of users.

B. Most Reacted Users

After the topic discovery, the most reacted users are found. It is calculated by countingthe number of tweets of each user which contains the topics discovered.

At first, take the document of each user containing tweets (without preprocessing). Eachline in text file is a single tweet. Now, check whether the discovered topics are present in each tweet, if found, count the number of those tweets. The counting should be done for each user's file. The users with highest number of count can be considered as the most reacted users.

C. User behavior analysis

Here we found the reaction of most reacted users. Sentiment analysis is used to discover the reaction. A user can react in a positive way, negative way or in a neutral way. The level of positivity and negativity determine the nature of users. A lexical resource SentiWordNet [5] is used here for opinion mining. Each WordNet synset is associated with three scores: positive, negative and objective. It describes how the terms are contained synset are positive, negative and objective. The latest version SentiWordNet 3.0 [6] is used in our system. For each term in the document we calculate a score, senti score, using the scores in SentiWordNet. Based on this senti score we determine the reaction of users.

In the process, first, the tweets of most reacted users, on the discovered topic, are collected separately. It is stored as separate textual documents and does data preprocessing steps (Tokenization, Special character removal, Stop word removal, URL removal and POS (Part-of-speech) tagging) on it. Now we get a tagged document set and give this as input to score calculating process. The path to SentiWordNet 3.0 is given in the process. Then do the following steps:

For each tagged term **s** in document

1. Check the SentiWordNet file for **s**
2. Do V= Pos(**s**) - Neg(**s**)
3. Create a hash array to include terms and its vector
4. Add values V of each **s** as vector along with index, in array
5. If index is higher than current vector size then add the remaining values by 0.0
6. Else do step 5
7. For each **s** in array
8. DoS1=Sum(1/1*vector(1),1/2*vector(2),....,1/n* vector(n)) , where n - vector size
9. Do S2=SUM(1/1,1/2,...,1/n), where n - vector size
10. Do Score(**s**)=S1/S2

Pos(**s**)is the positive score andNeg(s) is the negative score of word s in SentiWordNet. For each term in our document set we calculate a score V by taking the difference of Pos(**s**) and Neg(**s**). Then in a hash array we store this value along with an index, in a vector format, where the term will be the key and its vector will be the mapping value. Then do the calculation in step 9, 10 and 11 to obtain senti score for a term.

We do this process for every term in document set. The score obtained will be in the range of -1 and 1. Here we set some threshold values and comparing the obtained score with this threshold value we predict

the reaction. For example, the score 0 can be said as neutral reaction, score greater than 0.5 can be said as positive reaction and score less than -0.25 can be said as negative reaction.

IV. EXPERIMENT AND RESULTS

Collected Twitter dataset as real document streams. First selected location, Kerala, and choose some users, about 20. Then selected a time period, March 2017. By using Twitter4jAPI collected tweets of these users of about 100 pages. The collected tweets are stored asseparate text files for each user in a folder.

In the preprocessing phase, the text stream in each file is converted into tokens. Thenthe special characters are removed by pattern matching method. A standard stop words list isprepared and stored as another text file. By comparing each word in each text files with thestopword list the stop words from the documents are eliminated. After stop word removal, URL removal is done by pattern matching method. Then POS tagging is done by using Stanford POStagger. The words other than noun, verb, adjective and adverb are eliminated and stored as textfiles.

In the topic discovery phase, TF-IDF algorithm is applied on the terms that we obtainedin the preprocessing phase. A TF-IDF matrix is created and average frequency for everyterm is calculated. Average frequency value obtained in a range of 0.0000288 and 0.172239.The terms with highest frequency can be considered as topics and based on a threshold valuewe selected the topics. A threshold value of 0.03 is selected. The terms with greater than or equal to this threshold value is selected as output. The topics got as output are:

justiceformishel 0.17223869678933
Prayers 0.05985392102455
Mysterious 0.09109956926813
Death 0.07235921587964

In User behavior analysis phase, first we find the most reacted users. It is done by counting the number of tweets on the discovered topics. The output obtained like:

Debasish0102
RajLashly
RCB AnilThomas
smithaattingal

These are the user names of users in twitter. Then sentiment analysis is done for these users. In Sentiment analysis, first, preprocessingis done then, SentiWordNet 3.0 is used. A Senti score is calculated for every terms and reaction of users are discovered. The output obtained like:

User: Debasish0102 Reaction: Negative
User:RajLashly Reaction: Neutral
User:RCBAnilThomas Reaction: Neutral
User:smithaattingalReaction: Negative

CONCLUSIONS

Twitter and Facebook are the most popular social media websites which is used by peopleto express their opinion on any topic or share their ideas. The proposed system is a text miningapproach which analyses the document streams published in social media sites and finds the topics on rare events discussed by the users. It finds the most reacted users and the behavior of thoseusers. The system consists of four modules: Data collection, Data preprocessing, Topicdiscovery and User behavior analysis. Twitter data is taken for analysis. POS-tagger, TF-IDFand Sentiment analysis are some of the main techniques used in the system. This system can beused for monitoring abnormal or illegal user behaviors in internet and analyzing rare trending topics. As a future work, we thought of using Hadoop for processing big set of data like the data from Facebook and other micro-blogs.

REFERENCES

[1] D. Blei, A. Ng, and M. Jordan, “Latent Dirichlet allocation”, J. Mach. Learn. Res., vol.3, 2003, pp. 993-1022.
[2] J. Pei, J. Han, et.al, “PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth”, in Proc IEEE ICDE01, 2001, pp. 215-224.
[3] 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
[4] Kristina Toutanova, et.al. 2003. “Feature-Rich Part-of-Speech Tagging with a CyclicDependency Network”. In Proceedings of HLT-NAACL 2003, pp. 252-259
[5] Andrea Esuli, FabrizioSebastiani, “SENTIWORDNET: A Publicly Available Lexical Resource forSentiment Analysis and Opinion Mining”, 2006
[6] Stefano Baccianella, et.al, “SENTIWORDNET 3.0:An Enhanced Lexical Resource forSentiment Analysis and Opinion Mining”, 2010

★★★