Multidisciplinary Review Journal

Journal of Critical Reviews

ISSN- 2394-5125 Vol 7, Issue 4, 2020

Review Article

CLASSIFICATION AND RANKING OF TRENDING TOPICS IN TWITTER USING TWEETS TEXT

N. Umakanth[1], S. Santhi[2]

[1]Assistant Professor (Sr.Grade), Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India. umakanthn@mepcoeng.ac.in
[2]Assistant Professor (Sr.Grade), Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India santhicse@mepcoeng.ac.in

Received: 07.12.2019 *Revised: 16.01.2020* *Accepted: 11.02.2020*

Abstract
Every day millions of twitter user tweet their views on various topics using short messages of 140 to 280 characters length. Here we monitor the public opinion in twitter thereby identifying trending topics. To identify the current trending topics on Twitter, tweet classification is done. Tweet classification is a process of classifying the tweets based on the topics using the keywords of the tweets as a feature. Tweets are extracted from twitter using Twitter API. In the existing system, trending topics in twitter can be identified by hash tags. This work is going to categorize the tweets into seventeen classes and identify the trending topics by using words. Dataset used in this work is the list of tweets taken for two to three days from Twitter. By classifying tweets based on sensitive words for a certain time, top *k* most trending words from a topic will be obtained that convey the dynamic trend.

Keywords: Cosine Similarity, Frequency Distribution, Web Crawling, Trend Analysis, Text Classification, Topic Tracking.

DOI: http://dx.doi.org/10.31838/jcr.07.04.171

INTRODUCTION

In today's world, the development of information technology has increased the expression of public opinion on the internet. Social media is the most popular form of spreading public opinion. Social media such as facebook, twitter has an increase in the number of users and people started to express their own opinion on finance, politics, sports, entertainment etc.

In twitter, the user expresses their opinion in the form of short messages known as tweets. Millions of tweets are generated every day as an average of 6000 tweets per second. Due to this rapid use of the Internet, there may be a chance of spreading rumors. A user of Twitter can follow any other user, can share any tweet, can retweet others tweet and can enjoy reading tweets.

This huge data of tweets can have both the correct message and the wrong rumors. It is essential to pay attention to irrelevant messages of public and have to monitor the spreading of these messages. Monitoring of public opinion helps in discovering people's belief. This can be achieved by classifying the sensitive information and topic tracking of tweets to identify the useful and useless tweets.

Sensitive information means the words can be used more frequently by the user on the internet then the word becomes the trending topic. The emergence of sensitive words is due to suddenness. Sensitive words have to be tracked by analyzing public opinion and to provide a warning to the public regarding irrelevant information.

Our work is to identify the public opinion in twitter there by 1) classifying tweets based on text sensitive information classification and 2) tracking the trending topics (i.e.) topic tracking.

RELATED WORKS

In the internet, sensitive words of public monitoring exist in website crawling from Tibetan web pages [1]. Sensitive information classification and topic tracking on Tibetan web pages involved in three stages. The first stage includes establishing a sensitive vocabulary. It involves collection of relative words from a news website to built sensitive vocabulary.

The second stage involves the classification of text sensitive information. The articles from the web pages are crawled and they are classified under certain topics by using support vector machine [7]. The final stage involves topic tracking which includes analysis of trending topics in Tibetan web pages. This is useful for long documents such as article as it counts the frequency of words.

Classification of data can also be done through support vector machine (SVM) in machine learning [7]. SVM constructs a hyper plane. It separates the space into two half spaces. A good separation is achieved by the hyper plane that has the largest distance to the nearest training data points.

For short text, classification can be done under topic modelling via self-aggregation [2]. Short and sparse text can be modeled in two steps. In the first step, a short text is aggregated into a pseudo-document before topic inference using the K-means clustering algorithm. In the second step Latent Dirichlet Allocation (LDA) for topic extraction. Here topic Modeling is done for automatic discovery of thematic information.

Text category can be achieved through selection approach based on term frequency [3]. T-test classification can be used to measure the diversity of distribution of term frequency between specific categories. T-test selection algorithm helps in finding similarity between word and category.

In the case of a large amount of data, cloud computing can be used for internet public opinion monitoring model [4]. Big data analysis can be achieved using Hadoop and HBase. Hadoop is an open source and distributed framework which supports HBase.

THE PROPOSED METHOD

System Design

The Sensitive Vocabulary is established by crawling news archives then tweets are collected from Twitter. Then it is pre-processed by removing stop words, punctuation, URL, numerals, hash and emoticon. The pre-processed tweets are classified by using Text Sensitive Information Classification. The classified tweets are sent to the topic tracking module. Then the frequent words are identified and tracked for some time which helps to calculate twitter trend. The entire flow of the proposed system is shown in Fig. 1.

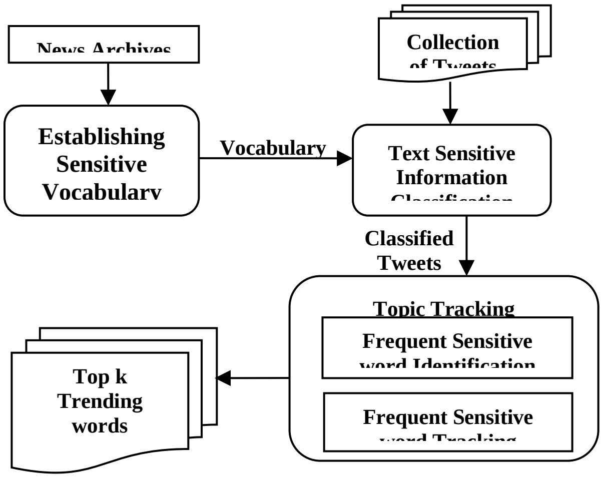
News Archives
Collection of Tweets
Establishing Sensitive Vocabulary
Vocabulary
Text Sensitive Information Classification
Classified Tweets
Topic Tracking
Frequent Sensitive word Identification
Frequent Sensitive word Tracking
Top k Trending words

Fig. 1: System Design

Establishing Sensitive Vocabulary

Sensitive words are words related to the current incident. The newspaper archives are crawled to collect sensitive words using a keyword. The pre-processing of the data includes the following steps:

a. Stop words removal

The Stop words present in the newspapers are removed because there is no meaning in having these words. It includes before, on, in, the, are etc.

b. Punctuation removal

The punctuation marks such as ', : , !, ...are removed from the collected data.

c. Special characters removal

The Special characters includes ", $, &, %, # are removed in pre-processing.

d. Numerals removal

The numbers present in the newspaper are also removed from the collected words since it also carries no information.

e. POS Tagging

The collected words are tagged using POS tagger and words with unrelated tags such as a preposition, adjective and conjunctions are removed.

For example, consider the news collected from the news website (www.thehindu.com) is represented in Table I.

Table I: Sample news Article before and after pre-processing

News	Preprocessed data
At least 40 Central Reserve Police Force (CRPF) personnel were killed on February 14, 2019 when a convoy in which they were travelling was attacked near Awantipora on the Srinagar-Jammu Highway. A suicide bomber of the Jaish-e-Mohammed (JeM) rammed an explosives-laden vehicle into one of the convoy's buses.	central reserve police force personnel killed convoy travelling attacked awantipora Srinagar Jammu highway suicide bomber jaish e mohammed jem

The vocabulary has many different categories and the sensitive words are spread out under these categories. The vocabulary also has the frequency and frequency distribution of sensitive words. These details are used in addition with words to enhance the classification process and to avoid data sparsity problem in short text. The Sensitive Vocabulary is used as the reference for the classification of tweets into their categories. The vocabulary is made up to date for a better result. Table II shows the classes and number of words per class in vocabulary.

Table II: Sensitive Vocabulary

Class Name	Number of words per class
Cricket	91
Football	89
Law	145
Politics	133
Accident	152
Disaster	123
Health	122
Drug	114
Finance	102
Environment	133
Army	131
Music	87
Technology	84
Internet	78
Education	90
Food	98
Movie	97
Total	1869

Frequency Distribution FD for a word W is calculated by,

$$FD(W) = F(W) + \sum_i^n FD(W_i) \qquad (1)$$

where F(W) is the Frequency of Word W and n is the number of words in the class.

If already existing words are collected while crawling, then the FD of that word needs to be updated. So, Frequency Distribution for a word W is updated by,

$$FD_{update}(W) = \frac{(F_{new}(W)*FD_{new}(W) + F_{old}(W)*FD_{old}(W))}{(F_{new}(W) + F_{old}(W))} \qquad (2)$$

where F_{new} (W) and FD_{new} (W) is the newly extracted Frequency and Frequency Distribution of the word,

F_{old} (W) and FD_{old} (W) is the already existing Frequency and Frequency Distribution of the word.

Text Sensitive Information Classification

The tweets dataset need to be made ready before entering the classification phase which classifies the tweets. The complete detailed procedure of classification is explained below.

a. Preprocessing of tweets

The tweets are collected from Twitter using Twitter API. In order to access Twitter API, we need to get four pieces of information from Twitter such as API key, API secret, Access token and Access token secret then the extracted tweets are pre-processed. The pre-processing involves removal of emoticons, URLs, user mentions, numerals, hashes, punctuations, special characters and stop words from tweets.

Emoticons removal

Tweets are the short text which includes feeling of users. It is expressed in the form of emoticon. These emoticons are removed from the tweets are represented as Table III.

Table III: An example for emoticons removed tweet

Example Tweet	Emoticons Removed Tweet
@Twitter See it? Tweet it! #feature Our updated15mpx camera is just a swipe away, so you get the shot fast. Rolling out to all of you over the next few days. 😀😀😀😀 pic.twitter.com/moOEFO2	@Twitter See it? Tweet it! #feature Our updated 15mpx camera is just a swipe away, so you get the shot fast. Rolling out to all of you over the next few days. pic.twitter.com/moOEFO2

URL removal

The external URLs present in tweet are removed since it carries no information is represented in Table IV.

Table IV: An example for URL removed tweet

Example Tweet	URL Removed Tweet
@Twitter See it? Tweet it! #feature Our updated 15mpx camera is just a swipe away, so you get the shot fast. Rolling out to all of you over the next few days. **pic.twitter.com/moOEFO2**	@Twitter See it? Tweet it! #feature Our updated 15mpx camera is just a swipe away, so you get the shot fast. Rolling out to all of you over the next few days

Mentions removal

Many tweets have mention of other user which is denoted by @username. This detail should be removed as mentioned in Table V.

Table V: An example for mention removed tweet

Example Tweet	Mention Removed Tweet
@Twitter See it? Tweet it! #feature Our updated 15mpx camera is just a swipe away, so you get the shot fast. Rolling out to all of you over the next few days.	See it? Tweet it! #feature Our updated 15mpx camera is just a swipe away, so you get the shot fast. Rolling out to all of you over the next few days.

Numerals removal

The numbers present in the tweet are removed from the tweet is represented in Table VI.

Table VI: An example for numerals removed tweet

Example Tweet	Numerals Removed Tweet
See it? Tweet it! #feature Our updated **15**mpx camera is just a swipe away, so you get the shot fast. Rolling out to all of you over the next few days.	See it? Tweet it! #feature Our updated mpx camera is just a swipe away, so you get the shot fast. Rolling out to all of you over the next few days.

Punctuation removal and Lowercasing

The punctuation marks and special characters are removed and the tweet is converted to lower case for better processing of tweets represented in Table VII.

Table VII: An example for punctuation removed and lowercased tweet

Example Tweet	Punctuation Removed Tweet (Lower case)
See it**?** Tweet it**!** #feature Our updated mpx camera is just a swipe away, so you get the shot fast. Rolling out to all of you over the next few days.	see it tweet it feature our updated mpx camera is just a swipe away so you get the shot fast rolling out to all of you over the next few days

Stop words removal

The stop words are removed and it is shown in Table VIII. A python package called NLTK plays a major role in pre-processing.

Table VIII: An example for stop words removed tweet

Example Tweet	Stop Words Removed Tweet
see **it** tweet **it** feature **our** updated mpx camera **is just a** swipe away **so you** get **the** shot fast rolling **out to all of you over the** next few days	see tweet feature updated mpx camera swipe away get shot fast rolling next days

b. Classification of tweets

Text Sensitive Information Classification is done by referring Sensitive Vocabulary. The pre-processed tweets are given as input in the classification module. The Text Sensitive Information Classification uses cosine similarity to find similarity between a tweet and classes. If the word is in Sensitive Vocabulary then the tweet will be assigned a weight which is the frequency distribution of the word else the weight will be zero. So, the tweet will be represented as a collection of weights.

$$Cosine\ Similarity(c\ , D) = \frac{\sum_i W_{Wi}}{\sqrt[2]{\sum_i {W_{Di}}^2} \ * \ \sqrt[2]{|c|}} \qquad (3)$$

where c is the class in Classes, D is the tweet and W_D is the tweet's weights and W_w is the weight of the word in both tweet and Vocabulary.

The similarity will be calculated between a tweet and each class in Vocabulary. The class with the highest similarity is assigned to the tweet. The tweets with zero similarity are labelled as 'Others'

Topic Tracking

The classified tweets are taken to the next stage called topic tracking omitting the tweets labelled as 'Others'. Topic tracking can be done in two ways. In the first method, the overall trending in the tweets is being calculated. In the second method, the class wise trending in the tweets is being calculated.

a. Overall Trend Analysis

In topic tracking, first the number of tweets in each class is counted and top three classes with the highest number of tweets are selected. Secondly, the tweets in selected classes are grouped by their dates and the numbers of files in the selected classes are counted and a class that toped in the count is selected.

Lastly, the frequency of words in the tweets belonging to that class is calculated and the sensitive words are identified. Those words are selected and tracked for a certain time to identify the trend. The sensitive words are tracked instead of topics to gain better results.

b. Class wise Trend Analysis

In this method, the trending words under each class are identified. The words frequency for each class is counted and the top ten sensitive words for each class are identified. This shows the concern of Twitter users in each class. It helps to identify the dynamic trend on Twitter

EXPERIMENT AND ANALYSIS

Dataset Description

The experimental data are collected from news websites like The Hindu (www.thehindu.com) and social media like Twitter (www.twitter.com) for a certain time. The sample data for two days in different timestamp will be given in Table IX.

Table IX: Sample Data taken from Twitter

Date	Time	Tweet	Name
14/11/2019	10:13:40 AM	RT@narendramodi: Productivity of the 16th Lok Sabha is exceptionally high, with the average being over 85%. The House took landmark decision	Anil Kumar
14/11/2019	10:10:17 AM	RT @rajnathsingh: The @BJP4India today launched a unique outreach programme named Bharat Ke Man Ki Baat, Modi Ke Saath to take feedback	dhanrajpinu choudhary
14/11/2019	9:47:25 AM	RT@Raffles_Statue: Every time I take a sip, I remember how lucky we were to colonise India. https://t.co/9Z3WoBKTtI	Kotchka Babushka
15/11/2019	10:30:16 AM	Google Maps Marketing Agency in Mumbai India https://t.co/v1ZXkpkJbF	SEO Agency

Establishing Sensitive Vocabulary

Sensitive Vocabulary can be built from news websites which hold the current sensitive words. The articles of a certain time in www.thehindu.com are crawled using a keyword. The keyword used in the classes is mentioned in Table II. The collected data are pre-processed. The preprocessing includes removal of stop words, punctuation and special characters.

The collected words are tagged using POS tagger and words with unrelated tags such as a preposition, adjective and conjunctions are removed. By counting the frequency, words related to the keyword (class) along with Frequency F and Frequency Distribution FD are identified and high-frequency words are added to Vocabulary then it is updated regularly. Some of the words along with their corresponding frequency and frequency distribution under the topic Law in the vocabulary are listed below in Table X.

Table X: Sample Vocabulary

Word	Frequency Distribution (FD)	Frequency (F)
court	0.06514	2861
lawyers	0.016826	739
justice	0.025887	1137
petition	0.006876	302
bench	0.013866	609
advocate	0.0051	224

Text Sensitive Information Classification

The dataset is made ready by extracting tweets from Twitter using Twitter API with the help of tweepy package in python. Tweets are extracted for the dates February 14 and February 15 2019. After extracting the tweets, they are preprocessed. The preprocessing involves removal of emoticons, URLs, mentions and stop words from tweets.

Text Sensitive Information Classification is done by referring to Sensitive Vocabulary. The preprocessed tweets are given as input in the classification module. In this module, the preprocessed tweets are labeled with the classes shown in Table I using cosine similarity. This is achieved by finding the similarity of each class and each tweet. Sort these similarity values, the tweet with maximum similarity values are assigned to that class and tweet with zero similarity with these classes will have the value of 0. So, these tweets are labeled as ‘Others'.

In our experiment, the tweets on February 14 2019 and February 15 2019 are preprocessed. Then by applying cosine similarity, we can find the class for corresponding tweets**.** Table XI shows that sample Tweet is similar to ‘‘Law’’ category by 0.076. In the top three similarity values, the similarity of the tweet and ‘‘Law’’ class is the highest. Therefore this tweet is categorized as ‘‘Law” class. Similarly, the process is repeated for all tweets.

Table XI: Similarity between Tweet and Class

Tweet	Class	Similarity
RT@ANINewsUP:PM Modi in Jhansi: Conspirators of Pulwama attack will be punished, our neighbouring country has forgotten that this is a new India	Army	0.04697461037543795
	Law	0.07550643032685067
	Politics	0.03296511506912657

Topic Tracking

The classified tweets are given as input in this module. Topic tracking can be done in two ways. In the first method, the overall trending in the tweets is being calculated by the following phases.

In the first phase, the classified tweets are grouped under each class. As we are having seventeen classes, the number of tweets which are grouped under each class is calculated and the top three classes are identified from 17 classes. February 14 and February 15 2019 tweets are taken then count the number of tweets in each class. Sort the calculated count in descending order and select the top three classes which have the maximum number of tweets. This indicates that these classes are mostly used by the public and identify people opinion. The top three classes for our experiment include Army, Politics, and Food. Table XII shows the classes and their count of tweets.

Table XII: Classes and Number of tweets per class

Class Name	Number of tweets
Cricket	317
Football	318
Law	675
Politics	2574
Accident	237
Disaster	195
Health	210
Drug	69
Finance	634
Environment	573
Army	**3157**
Music	284
Technology	334
Internet	1214
Education	482
Food	2137
Movie	97

In the second phase, the identified top classes are selected and the tweet which belongs to those classes is grouped according to date wise. Here sort the result among them and Army class has more tweets and has more public opinion is represented in Table XII. The purpose of this work is used to identify variation of trending under each class during February 14 and February 15 2019. The selected top 3 classes are taken and analysis these class for daily trend.

Table XIII: Overall Trending Words

Top 10 Trending Words on Feb. 14, 2019	Top 10 Trending Words on Feb. 15, 2019
Happy valentine’s day	India
India	pulwama
Muslim	Pakistan
peace	attack
sad	nation
phirekbaarmodisarkar	kashmir
Anti-Indian	War
Anti-Muslim	terror
defence	soldiers
elected	Pulwama terror attack

In the next two phases, top ten sensitive words are taken from the Army class. It is processed based on the frequency and outputs the top three trending words. This method helps to identify the overall trending words. The frequency of the word is taken and sorted to find the trending words. The trending words obtained for February 14 and February 15 2019 are shown in Table XIII. The trending words on February 14 and February 15 2019 are valentines and Pulwama.

In the second method, the trending words under each class are identified. The words frequency for each class is counted and the top five sensitive words for each class are identified. This shows the concern of Twitter users in each class and helps to identify the dynamic trend on Twitter.

Table XIV: Trending Words in Army Category

Top 5 Trending Words on Feb. 14, 2019	Top 5 Trending Words on Feb. 15, 2019
Muslims	attack
antimuslims	India
Imran	War
mirage	terror
bomb	Kashmir

Table XIV shows the trending words for Army category that are mostly used by users on February 14 and February 15, 2019.

EVALUATION

In case of existing method [1] where Vocabulary with only words is used, the accuracy of classification is 73 %. In Proposed method the vocabulary with frequency and frequency distribution is used which has an accuracy of 89%. Accuracy can be calculated by dividing number of correct prediction and total number of prediction. Table XV shows the comparison between trending words on February 15, 2019 for existing and proposed system. The trending words obtained using proposed system is better than existing system. The trending words collected using proposed system is more relevant to the day because both the words related to Valentine's Day and Pulwama attack are present.

Table XV: Top 10 Trending Words on Feb. 15, 2019

Using Existing System (By Frequency)	Using Proposed System (By Frequency Distribution)
Vande mataram	Happy valentine's day
force	India
defence	Muslim
chowkidar	peace
report	Sad
India	phirekbaarmodisarkar
Pakistan	antiindian
launch	antimuslim
terror	defence
war	Pulwama terror attack

By involving the Frequency and Frequency distribution in vocabulary the classification of tweets under classes are improvised. Better results can be obtained by continuous updation of vocabulary.

CONCLUSION AND FUTURE WORK

In this system, the trend in twitter is identified by processing and analyzing the tweets posted by twitter users on twitter. The sensitive Vocabulary is built from a news website for finding sensitive words. The tweets in Twitter are classified by the classification phase with the help of vocabulary and tracked for a certain time to identify the dynamic trend in twitter. The trend is identified in two aspects. The first trend shows the Twitter users' concern in a particular class and the second trend shows their concern on a particular day.

In our future work, we will consider processing of hash tags and try to cut the computation time while calculating the similarity by using different similarity measures. Other neural network like RNN, CNN and LSTM can be explored in order to increase the accuracy for the larger dataset. This work can be further extended by accurately mentioning the subtopic of the tweet.

REFERENCES

1. Guixian Xu, Ziheng Yu, and Qi Qi, "Efficient Sensitive Information Classification and Topic Tracking Based on Tibetan Web Pages" in IEEE Access, vol. 6, pp. 55643-55652, 2018.
2. S. Santhi, N. Umakanth and T. Hemalatha. "Determining Hateful and Offensive Terms from Twitter Using Hate Speech Detection" in International Journal of Pharmaceutical Research, vol. 11, issue 1 pp. 867-871, December 2019.
3. X. Quan, C. Kit, Y. Ge, and S.J. Pan, "Short and sparse text topic modeling via self-aggregation", in Proc. Int. Conf. Artif. Intell., Buenos Aires, Argentina, pp. 22702276, 2015.
4. D. Wang, H. Zhang, R. Liu, W. Lv, and D. Wang, "t-test feature selection approach based on term frequency for text categorization", Pattern Recognit. Lett., vol. 45, pp. 110, Aug. 2014.
5. Z.T. He, X.Q. Zhang, F.W. Zhao, and T. K. Ji, "Internet public opinion monitoring model based on cloud computing", Appl. Mech. Mater., vol. 404, pp. 744747, Sep. 2013.
6. Tweepy Documentation v3.7.0 by Joshua Roesslein on Nov 30, 2018 http://docs.tweepy.org/en/v3.5.0/.
7. Dighe NS, Nirmal SA, Musmade DS, Dhasade VV. "Herbal Database Management." Systematic Reviews in Pharmacy 1.2 (2010), 152-157. Print. doi:10.4103/0975-8453.75067
8. https://developer.twitter.com/en/docs.html
9. Kamini Nalavade, B.B. Meashram. "Data Classification Using Support Vector Machine", National Conference on Emerging Trends in Engineering & Technology, pp. 181, Mar 2012.
10. Singh, K., Singh, G.Alterations in some oxidative stress markers in diabetic nephropathy(2017) Journal of Cardiovascular Disease Research, 8 (1), pp. 24-27.
DOI: 10.5530/jcdr.2017.1.5
11. Karankumar Sabhnani and Ben Carterette, "Real-time Topic Detection and Tracking in Microblog: Towards A Comprehensive Tweet Recommendation System", 23rd Text Retrieval Conference (TREC), 2015.
12. Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. "Twitter Trending Topic Classification", IEEE Conference Paper, 2011.