www.ijemr.net ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962

Volume-7, Issue-3, May-June 2017

International Journal of Engineering and Management Research

Page Number: 796-801

Trending Topics Detection using Machine Learning Approach

Palak Gour[1], Suneet Joshi[2]

[1,2]Department of Computer Science and Engineering, Medicaps University, Indore, INDIA

ABSTRACT

In this age the social media is one of such kind of application where easily data is obtainable by the users. According to these the test of user's can also be identifiable. Due to this the application of text mining is also increased in social media text mining and analysis. In this presented work social media text analysis based technique is introduced. That technique helps to identify the most trending topic from the micro-blog data. In this context three things are necessary first social media data, creation of topic model and evaluation of text data based on developed data model. Therefore for an effective social media data twitter data is considered. In further the pre-processing techniques are applied for quality improvement of raw data. In next the regular size of data is extracted as features of text twits using the term frequency and sentence formation probability. Finally the clusters of data are computed using FCM (fuzzy c means) clustering and the clustered data is used with the bay's classifier for assuring the topic name. The implementation of this topic model is performed on JAVA technology. After that on the basis of experimental evaluation the performance of topic model is computed. Additionally a comparative analysis with traditional topic model is performed. The computed outcomes show the proposed model is an efficient and accurate technique of data analysis.

Keywords-- social media, trending topic, text mining, data mining, fuzzy c mean, basiean classifier

I. INTRODUCTION

Social network is a term used to describe web-based services that allow individuals to create a public/semi-public profile within a domain such that they can communicatively connect with other users within the network. A significant power of social media is the rapidity with which new information is posted and shared. If a user is interested in a particular item, event, or topic, she can often provide a few relevant keywords to a social network's search function and track new developments by reading recent postings. For instance, one can track tweets mentioning "goal" on Twitter during the 2014 World Cup to follow when goals are scored [1]. Social networks are online applications that allow their users to connect by means of various link types. As part of their offerings, these networks allow people to list details about themselves that are relevant to the nature of the network. For instance, Facebook is a general-use social network, so individual users list their favorite activities, books, and movies. Conversely, LinkedIn is a professional network; because of this, users specify details which are related to their professional life (i.e., reference letters, previous employment, and so on.) Because these sites gather extensive personal information, social network application providers have a rare opportunity: direct use of this information could be useful to advertisers for direct marketing [2].

In this presented work the text data for microblog analysis is used for preparing the classification algorithm. Basically the microblogs are frequently used now in these days with small amount of communication data. But frequent use of this communication channel increases the amount of data for manual analysis [3]. In this research work we present the micro-blogs twitter data analysis for the trend topic detection using social network analysis. Extensive evaluations on a representative real-world dataset based on Twitter data demonstrate that our mechanism is able to characterize emerging topics well and detect them before they become hot topics.

II. PROPOSED WORK

Now in these days the use of social media is found in various places, sometimes it is also used as market strategy. Therefore finding the information about the trending topic and also discovering the hot topics in social media an effort is placed in this work. This chapter provides the detailed formulation of the proposed concept for hot topic discovery model.

A. System Overview

The new generation frequently usages the social media additionally a significant amount of time consumed with this platform. The utility of social media not only now limited for communication and connectivity with the friends and their circle. It is now a part of publicity and marketing

too. Due to this a significant amount of different subjective topics are flooded in this media. Identification of such kinds of topics and capturing trends on social media is a new domain of data analysis and text mining. In this presented work a new model for hot topic tracking and discovery is proposed and implemented. That helps to understand the current hot topic which is trending on social media. In order to perform such kind of data analysis the twitter is considered as the key data source with the big data environment.

The proposed model is a based on classification and clustering of the twits for finding the trending topic [4]. Both the techniques are applied in such manner by which first the subjects are grouped and then the most common twits are aggregated using the probability of subjects. For performing the clusters of twits first the fuzzy c means algorithm, because the accuracy of clustering is higher as compared to k-means algorithm [5]. Then the bays probability model is applied for finding the topic by using the prior and posterior probability. The advantage of this model is that it is light weight and efficient for computation point of view. Additionally the implementation of both kinds of data models improves the accuracy for topic discovery. This section provides the introduction of the proposed data model, in next section the detailed system design and their working is explained.

B. System Architecture

The proposed data model for hot topic discovery on twitter data is defined using figure 2.1. It contains different components as the functional blocks which produces some intermediate out comes for accepted data. This section includes the details of these participating components.

***Twitter*:**

Twitter is one of most famous social media platform of the world. A significant amount of users are consuming it's services regularly. In addition of that the world famous personalities are also carrying the account with twitter. A user can follow a number of persons in this platform additionally due to small amount of text communication it is also known as the micro blog. In this experimental scenario the twitter topics are analyzed and the hot topics are computed on live data. For targeting and experimental points of view some limited amount of subjective twit are subscribed and their data is utilized with the proposed model.

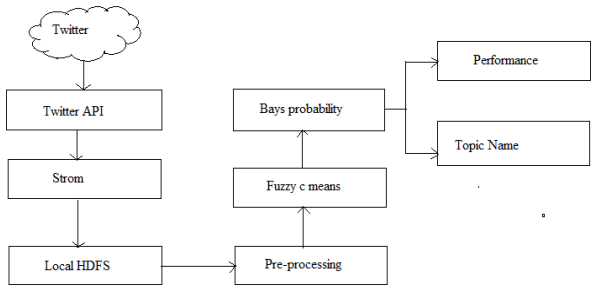

Figure 2.1 proposed system

Twitter API:

Due to significant amount of data generation on twitter it a good data source for text analysis and hot topic. In order to fetch data from such data source in a secure manner twitter application provides an API. This API help to authenticate the user's twitter account credentials and enable to capture the live twits from twitter account. Therefore a separate program is needed to be developed that authenticate the user and enable access to the twitter current account.

Apache Storm:

The apache storm is a kind of third party user library which is configured with the help of Hadoop. This library enables the user to extract the twits from the twitter authenticated account. Basically this API helps to pump twits from twitter database to any local Hadoop file system. Therefore after twitter account validation process the twits are extracted with the help of apache Strom.

Local HDFS:

The live twits from the twitter server are downloaded to the HDFS directory using a simple text file. Here the system provides both kinds of ability to use the text, user can input these tweets directly to the system for analysis or the downloaded twits are preserved separately in a text file to use these twits in offline mode for demonstration purpose.

Pre-processing

The pre-processing of data is a key action in machine learning. Using the pre-processing the quality is improved and the noisy contents are removed from the data. In order to perform these different techniques of pre-processing is used with the data based on the nature of data (i.e. structured or unstructured). In this work the micro blog data is used thus it is a text document which is obtained in form of unstructured data. Thus for pre-processing and refinement of data the three step process is used.

1. Remove stop words: in this process a list of frequently used words are prepared, this list of words are used with the find and replace function to remove all the words from the downloaded twits. The stop words are those words that are used for preparing the sentences but they not have much significance for identification of any twit's subject.

2. Remove special characters: this process is similar to the previous stop word removal process. Here also a list of special character list is prepared in a file and each character is searched on input data for removing all the contents as listed in the special character list from the tweets.

3. Feature selection: after refinement of data is used for feature computation therefore a term frequency and sentence formation probability is computed for each remaining words. The term frequency of a word in twits are computed using the following formula:

$$term\ frequency = \frac{total\ number\ of\ times\ a\ word\ appeared}{total\ amount\ of\ words\ in\ selected\ tweets}$$

Similarly the sentence formation probability is also computed for all the words. The probability is computed using the following formula:

$$sentence\ formation\ probability = \frac{a\ word\ appeared\ in\ number\ of\ sentence}{total\ sentences\ in\ downloaded\ tweets}$$

Here computed both the parameters are used to select the essential keywords from the input twits. But the length of twits are not regular in size thus a threshold is used as length of 20 to select the features from each twits. This process makes the entire twits in regular length twits.

Fuzzy C-means:

The fuzzy c means algorithm is used with the selected features from the previous step. FCM [6] is a representative algorithm of fuzzy clustering which is based on K-means concepts to partition dataset into clusters. The FCM algorithm is a ‘‘soft’’ clustering method in which the objects are assigned to the clusters with a degree of belief. Hence, an object may belong to more than one cluster with different degrees of belief. It attempts to find the most characteristic point in each cluster, named as the centre of one cluster; then it computes the membership degree for each object in the clusters. The fuzzy c-means algorithm minimizes intra-cluster variance as well. However, it inherits the problems of K-means, as the minimum is just a local one and the final clusters depend on the initial choice of weights.
FCM algorithm follows the same principle of K-means algorithm, i.e. it iteratively searches the cluster centers and updates the memberships of objects. The main difference is that, instead of making a hard decision about which cluster the pixel should belong to, it assigns a object a value ranging from 0 to 1 to measure the likelihood with which the object belongs to that cluster. A fuzzy rule states that the sum of the membership value of a pixel to all clusters must be 1. The higher the membership value, the more likely a pixel will belong to that cluster [7].

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \parallel x_i - c_j \parallel, \qquad 1 \le m < \infty$$

Where m is any real number greater than 1, $u_{ij}$is the degree of membership of x_i in the cluster j, x_i is the i[th] of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left[\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right]^{\frac{2}{m-1}}}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

As mentioned earlier, this has an iterative process (FCM pseudo-code).

FCM pseudo-code:
Input: Given the dataset, set the desire number of clusters c, the fuzzy parameter m (a constant > 1), and the stopping condition, initialize the fuzzy partition matrix, and set *stop = false.*
Step 1: Do:
Step 2: Calculate the cluster centroids and the objective value J.
Step 3: Compute the membership values stored in the matrix.
Step 4: If the value of J between consecutive iterations is less than the stopping condition, then *stop = true.*
Step 5: While (! $stop$)
Output: A list of c cluster centres and a partition matrix are produced.

Table 2.1 FCM algorithm

The FCM algorithm creates groups of the similar twits, according to the selected domains of tweets initially.

Bay’s probability:

Now for finding the hot topics finally the cluster-wise bay’s classifier applied on twits features for improving the accuracy of the classification. The bays model is described as:

The Naive Bayes classification algorithmic rule is a probabilistic classifier. It is based on probability models that incorporate robust independence assumptions. The independence assumptions usually don't have an effect on reality. So they're thought of as naive. You can derive probability models by using Bayes' theorem (proposed by Thomas Bayes). Based on the nature of the probability model, you'll train the Naive Bayes algorithm program in a very supervised learning setting. In straightforward terms, a naive Bayes classifier assumes that the value of a specific feature is unrelated to the presence or absence of the other feature, given the category variable. There are two types of probability as follows [8]:

- Posterior Probability [P (H/X)]
- Prior Probability [P (H)]

Where, X is data tuple and H is some hypothesis. According to Baye's Theorem

$$P\left(\frac{H}{X}\right) = \frac{P\left(\frac{X}{H}\right)P(H)}{P(X)}$$

Performance:

The final outcomes of the system are defined in two parts in first the performance of classification and hot topic model is evaluated in terms of accuracy, error rate, and time and space complexity. Additionally the trending topic is concluded.

Topic name

That is final outcome of the system which provides the name of topic which is trending in twits.

C. Proposed algorithm

ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962

The entire system can be described using the step of processes; the table 2.2 contains the suitable algorithm steps that followed for extracting the twitted subjects.

Input: twitter dataset D, number of clusters N Output: twit's subject S
Process: $R = readDataset(D)$ $R_s = StopWordRemove(R)$ $R_c = RemoveChar(R_s)$ For each word in R_c $tf = \frac{total\ number\ of\ times\ a\ word\ appeared}{total\ amount\ of\ words\ in\ selected\ tweets}$ $S_p = \frac{a\ word\ appeared\ in\ number\ of\ sentence}{total\ sentences\ in\ downloaded\ tweets}$ $end\ for$ $Sel[\ \ ] = SelectFeatures(R_c, tf, S_p, 20)$ $C_{data} = FCM.ComputCluster(Sel[\ \ ], N)$ $S = Bay's.Classify(C_{data})$ $retrun\ S$

Table 2.2 proposed algorithm.

III. RESULT DISCUSSION

The given chapter provides the detailed understanding about the evaluated results of the proposed Trend Topic Detection and Classification using Fuzzy Membership and Bayesian Classifier. Therefore this chapter includes the different performance parameters and their description on which the proposed system is evaluated using different size of data.

A. Accuracy

In a classification technique the accuracy is measurement of accurately classified patterns over the total input patterns of online and offline tweets produced for classification. Therefore that can be a measurement of successful training of the classification algorithm. The accuracy of the classifier can be evaluated using the following formula:

$$\text{Accuracy} = \frac{\text{Total correctly classified patterns}}{\text{Total input patterns to Classify}} \text{X100}$$

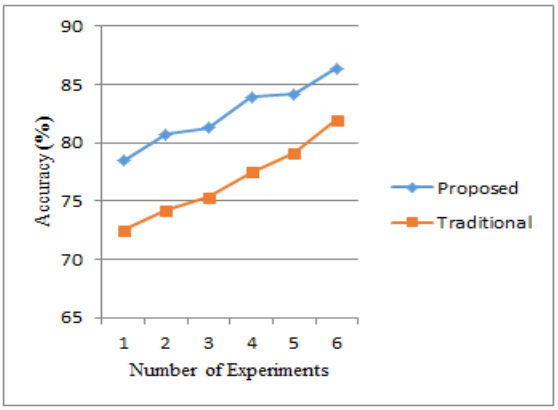

Figure 3.1 Accuracy

The accuracy of the implemented proposed algorithm of twitter trending is represented using table 3.1 and figure 3.1. The given graph figure 3.1 contains the accuracy of the implemented algorithms. The X axis of the diagram contains the amount of data during the training and testing and Y axis contains the obtained performance in terms of accuracy percentage. To demonstrate the performance of both the techniques the blue line is used for proposed model and red line shows the performance of traditional approach. According to the obtained results the performance of the proposed model provides more accurate results. Additionally the accuracy of the feature classification model is increases as the amount of instances for the learning of algorithm is increases.

Number of Experiments	Proposed	Traditional
1	78.43	72.49
2	80.73	74.24
3	81.28	75.32
4	83.91	77.48
5	84.15	79.11
6	86.38	81.98

Table 3.1 Accuracy

B. Error Rate

The amount of data misclassified samples during classification of algorithms is known as error rate of the system. That can also be computed using the following formula.

$$Error\ Rate\ \% = \frac{Total\ Misclassified\ Patterns}{Total\ Input\ Patterns} X100$$

Number of Experiments	Proposed	Traditional
1	21.57	27.51
2	19.27	25.76
3	18.72	24.68
4	16.09	22.52
5	15.85	20.89
6	13.62	18.02

Table 3.2 Error Rate

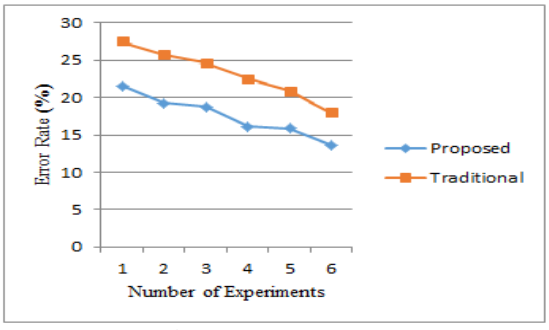

Figure 3.2 Error Rate

The figure 3.2 and table 3.2 shows the comparative error rate of implemented classification algorithm. In order to show the performance of the system the X axis contains the amount of data used for training and the Y axis shows the performance in terms of error rate percentage. The error rate of the traditional method is given using the red line and the performance of the proposed classification Technique is given using the blue line. The performance of the proposed classification is effective and efficient during different execution and reducing with the amount of data increases. Thus the presented classifier is more efficient and accurate than the traditional approaches of text classification.

C. Memory Usage

Memory consumption of the system also termed as the space complexity in terms of algorithm performance. That can be calculated using the following formula:

$$Memory\ Consumption = Total\ Memory - Free\ Memory$$

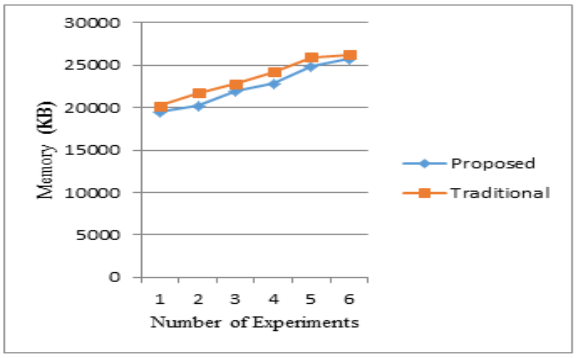

Figure 3.3 Memory Consumption

The amount of memory consumption depends on the amount of data reside in the main memory, therefore that affect the computational cost of an algorithm execution. The performance of the implemented classifier for topic classification is given using figure 3.3 and table 3.3. For reporting the performance the X axis of figure contains the amount of data required to execute using the algorithms and the Y axis shows the respective memory consumption during execution in terms of kilobytes (KB). According to the obtained results the performance of algorithm demonstrates similar behavior with increasing size of data, but the amount of memory consumption is increases with the amount of data.

Number of Experiments	Proposed	Traditional
1	19438	20199
2	20194	21734
3	21947	22817
4	22843	24174
5	24836	25942
6	25715	26261

Table 3.3 Memory Consumption

D. Time Utilization

The amount of time required to classify the entire test data is known as the time consumption. That can be computed using the following formula:

$$Time\ Consumed = End\ Time - Start\ Time$$

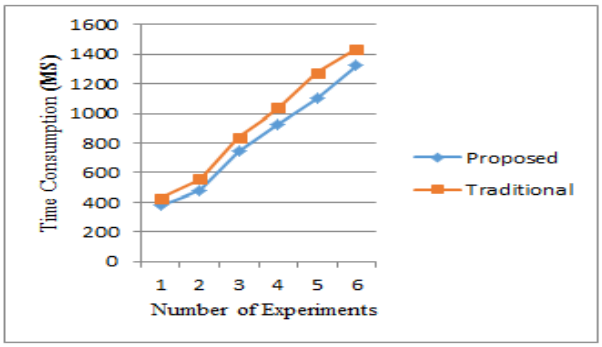

Figure 3.4 Time Consumption

The time consumption of the proposed algorithm is given using figure 3.4 and table 3.4. In this diagram the X axis contains the size of dataset and the Y axis contains time consumed in terms of milliseconds. According to the comparative results analysis the performance of the proposed technique minimize the time consumption. But the amount of time is increases in similar manner as the amount of data for analysis is increases.

Number of Experiments	Proposed	Traditional
1	377	429
2	482	559
3	746	837
4	928	1037
5	1102	1274
6	1324	1436

Table 3.4 Time Consumption

IV. CONCLUSION

This chapter reports the conclusion and the future extension of the proposed working model. The conclusion and future work is developed on the basis of facts recognized in experiments and the observations during different experiments.

A. Conclusion

In this age of digital communication the internet based activities are increased in much rapid speed. Not only using the work stations and personal computers, even mobile phones are also enabled to use the internet based applications. Among the various different kinds of applications based on internet the social media platforms are attracting a significant amount of traffic and number of online users. Additionally users of social media consume a significant amount of time on these platforms. Therefore

companies, organizations, sport mans, politicians and others frequently usages these platform for their advertisements and publications. Due to this a similar kinds of twits are flooded in this network, on which the users are shows their interest. The analysis of such kind of topics is termed as trending topics.

In this presented work for extracting the subjects which are trending on social media a data model is presented. In this context the experimental data is used directly from twitter accounts using the Hadoop, Strom and twitter API. The model is designed in such manner the data can be used online and offline. In further the data is pre-processed and the essential features are extracted from the raw twits. The pre-processing and feature extraction includes the stop word removal, special character removal and computation of term frequency and sentence formation probability. Additionally to make the data in regular size the limited amount of features are selected using the term frequency and sentence formation probability. The computed features are used with first with the FCM (fuzzy c means) and the bay's classifier for computing the most trending topic among the downloaded twits. Basically FCM is an unsupervised approach of data analysis and bay's is a supervised approach of analysis. Both the models are used for improving the performance and accuracy of data analysis.

The implementation of the proposed concept is given using the JAVA developed technology. After implementation the performance is also computed in terms of accuracy, error rate, memory consumption and the time consumption. The summary of the computed performance is given in table 4.1.

S. No.	Parameters	Proposed technique	Traditional technique
1	Accuracy	High	Low
2	Error rate	Low	High
3	Memory usages	Low	High
4	Time consumption	Low	High

Table 4.1 performance summary

According to the obtained performance of the proposed system the model is efficient and accurate. Additionally in fewer amounts of resources provides the optimal performance.

B. Future Work

The main aim of the proposed work is accomplished successfully, the model is implemented and their performance is also effective. In near future the proposed data model can be extended for user interest tracking, product feedback evaluation and others.

REFERENCES

[1] L. Cipriani, Goal! detecting the most important World Cup moments, Technical report, Twitter, 2014.

[2] Jayant S. Rohankar, "A Study on Advanced Security Techniques to Provide Security for Social Networking as Data Mining", International Journal of Advance Foundation and Research in Computer (IJAFRC) Volume 2, Special Issue (NCRTIT 2015), January 2015.

[3] Zhao, Zhongying, et al. "Collecting, managing and analyzing social networking data effectively", 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), IEEE, 2015.

[4] Mohammed J. Zaki and Wagner MeiraJr, "Data Mining and Analysis Fundamental Concepts and Algorithms", Cambridge University Press Hardback, 2014 [Book]

[5] Neelamadhab Padhy, Dr. Pragnyaban Mishra, "The Survey of Data Mining Applications and Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), PP. 43-58 Vol.2, No.3, June 2012.

[6] J. C. Bezdek, R. Ehrlich, and W. Full, ''FCM: The fuzzy c-means clustering algorithm,'' Comput. Geosci., vol. 10, nos. 2–3, pp. 191–203, 1984

[7] Fahad, Adil, et al. "A survey of clustering algorithms for big data: Taxonomy and empirical analysis." IEEE transactions on emerging topics in computing 2.3 , 267-279,2014.

[8] Roshani Choudhary and Jagdish Raikwal, "An Ensemble Approach to Enhance Performance of Webpage Classification", International Journal of Computer Science and Information Technologies, Volume 5, Number 4, PP. 5614-5619, 2014.