Analysis of comments on twitter social networking sites using text-classification & k-means clustering through data mining.

[#1]Shweta Baviskar

[#1]Department of Computer Engineering,

JSPM's RSCOE, Tathawade, Pune.

***Abstract* :** Nowadays, review sites are more and more confronted with the spread of misinformation, i.e. opinion spam, which aims at promoting or damaging some target businesses, by misleading either human readers, or automated opinion mining and sentiment analysis systems. The analysis shows that the proposed method gives better clustering results and provides a novel use-case of grouping user communities based on their activities. Our approach is optimized and scalable for real-time clustering of social media data. a modified model for Naïve Bayes classifier for multinomial text classification has been proposed by modifying the conventional bag of words model .In order to understand the behavior and structure of a social network we need to study the network and this study is called social network analysis. There are various social networking sites available on internet like Linked Facebook, Instagram, Twitter, Google and many more. Interactions over such sites produces huge amount of data because billions of active users maintain their accounts .In this paper, we are concentrating on Twitter data. R language is used for acquisition, preprocessing, analyzing and visualization of the twitter data. Twitter data is extracted, preprocessed and then clustered based on the information.

Keywords: Twitter Comment, Mining, K-mean, Text Mining.

I. INTRODUCTION

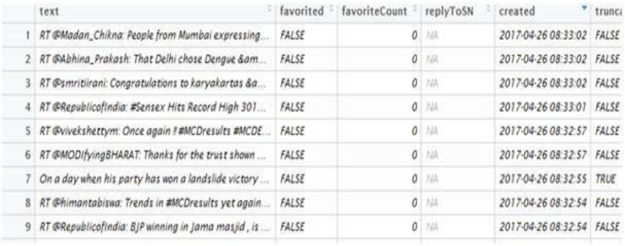
In the past few years, there was an immense growth of the social networks. Such networks provide the platform to users to express, share or discuss their ideas, opinions with their friends in social graph and also communicate with them. Social networks also provide the platform to the internet users to interact with both the technology and other people . There are many social media network sites, Twitter and Facebook remain the most well-know but other form being used are blogs, wikis and platforms with unfiltered text and information which remains of true focus for users. Slashdot [3] is a website that focuses on the publication news where the stories could be written by editors, or posted and commented by users.

A. Twitter

Twitter is a social networking service which allows the user to send and read the short message of 140 character called "tweets". There are two types of users for twitter account. One is registered users who can only read the tweets and another are registered users who can read and post the tweets. It is a public platform for all the people of different age categories all over the world. Data generated by twitter is heterogeneous in terms of content because user can post a text, image, video and audio in any format. Data is also big in size because hundred of thousand of tweets per day is generated [2].In the late 2009, twitter added a new feature which allows each tweet to be geo-tagged which is associated with longitude and latitude of specific location [7]. Fig 1. Shows some tweets extracted from twitter .In this paper, tweets are extracted, refined, analyzed and visualized in a geospatial representation. The main g oal of our research is to visualize the user's tweet in a particular area and visualize the clustered tweets according to the location information. The various packages and libraries are provided by R for extracting and processing the data and also for the visualization of clustered data.

***B.* Data format**

By default twitter data is extracted and presented in JSON format [4]. But for our analysis, we convert the JSON format into data frame format because handling of data is easier in this format. Fig 2. explains the structure of tweet in data frame format. The information contains tweet related data like tweet, latitude, longitude, date of creation, etc. This information also carries information of tweet creator, retweet.



	text	favorited	favoriteCount	replyToSN	created	trunc
1	RT @Madan_Chikna: People from Mumbai expressing...	FALSE	0	NA	2017-04-26 08:33:02	FALSE
2	RT @Abhina_Prakash: That Delhi chose Dengue &am...	FALSE	0	NA	2017-04-26 08:33:02	FALSE
3	RT @smritiirani: Congratulations to karyakartas &a...	FALSE	0	NA	2017-04-26 08:33:02	FALSE
4	RT @RepublicofIndia: #Sensex Hits Record High 301...	FALSE	0	NA	2017-04-26 08:33:01	FALSE
5	RT @vivekshettym: Once again ! #MCDresults #MCDE...	FALSE	0	NA	2017-04-26 08:32:57	FALSE
6	RT @MODIfyingBHARAT: Thanks for the trust shown ...	FALSE	0	NA	2017-04-26 08:32:57	FALSE
7	On a day when his party has won a landslide victory ...	FALSE	0	NA	2017-04-26 08:32:55	TRUE
8	RT @himantabiswa: Trends in #MCDresults yet again...	FALSE	0	NA	2017-04-26 08:32:54	FALSE
9	RT @RepublicofIndia: BJP winning in Jama masjid , is ...	FALSE	0	NA	2017-04-26 08:32:54	FALSE

Fig 1. Tweet in data frame format

II. LITERATURE SURVEY

In paper[1], Information filtering is the process of providing appropriate information to the people who need it. It significantly searches for what actually concerns the textual document, specifically web contents, and offers a user with classification mechanism to avoid the unnecessary information. This information filtering process is used in the online social network for insightful objective. To facilitate the content based filtering, this article introduces the filtered wall architecture. It will filter the incoming post based on the content. The main goal of this system is to provide customizable content based message filtering for online social networks, based on machine learning techniques. Information Filtering Systems are designed to categorize the information which are generated dynamically and offer the information to the user fulfil their requirement. In the content Based Filtering system, each user is assumed to operate separately. So the filtering system selects the information based on the correlation between the content of the items and user preferences. To support the content based filtering in online social network, Filtered wall architecture is introduced. In this architecture, text mining techniques are employed to categorize the incoming messages. Traditional text classification methods have major inadequacy in classifying the short text message. An automated system called filtered wall is designed in this paper to filter unwanted messages from user walls. Machine learning (ML) is used as text categorization techniques to automatically assign each short text message within a set of categories based on its content. In machine learning approach, the problem of classification is an activity of supervised learning because the learning step is supervised by the knowledge of the categories. Figure 3.2shows how admin add the short text word gd and full form of gd i.e. good day in database and whenever any user send that short text word i.e. gd, the full form of that word i.e. good day is display on the receivers wall. Short text classifier categorizes messages according to a set of categories. In this method short text word are set by admin in the data base. When any user sends any short text word which is set by admin the full form of that corresponding word are displays or show in the receiver wall. The machine learning mechanism is used to classify the short text.

III. ALGORITHM DESCRIPTION

In this section, we describe K-Means clustering algorithm ,Text classification.

A. K-means clustering.

The obtained (tweets x terms) D matrix easily form the base for clustering algorithms, according to the classical scenario employed in ML.

1) k-means: One of the older clustering algorithm is the well-know k-means algorithm [7]. Briefly, it works as follows: (a) k initial points are randomly generated in the textual vector space; (b) these points are assigned as the k cluster centers;(c) each tweet – a row in the D matrix – is assigned to the cluster with closest center; (d) the value of the cluster center is recomputed in order to consider this new element; (e) steps (c) to (d) are repeated until no changes occur in the labels assigned to each tweet [7].

Several metrics can be employed for distance computation. In the case of text document collections that follow the VSM.it is common to use one of two metrics: (a) the classical Euclidian distance, or (b) the cosine similarity measure, given, for documents d1 and d2, by:

$$dist(d1; d2) =< d1; d2 >$$

where <;> stands for the dot or scalar product, and d is the vector norm of the document d.

2) Text Preprocessing and Data Matrix

To obtain the relevant semantic elements from a text, we need to perform some preprocessing steps; the employed techniques are borrowed from the Information Retrieval (IR) area [1], [14].

1. The first activity is to identify the text units, which are, in our case, the tweets themselves. Then the text is treated by a series of procedures to perform: (a) case-folding and eventually additional UTF conversion; (b) stop-words removal; stop-words are textual elements that carry almost no semantics; they are very frequent in the text and can be eliminated; a stopword list include articles, prepositions and conjunctions (in some applications numbers are also eliminated); (c) stemming: a procedure that aims to connect textual elements of similar semantics, by obtaining their "root"; suffixes and prefixes are eliminated, plurals and verbal variations of the same term are reduced to an unique form. As a result of these steps, each tweet can be seen as a series of elements that carry the text semantics, called terms in the IR community. The set of these terms form the dataset schema, that is, each one of the root terms are one dimension on a huge dimensional vector space, according to the Vector Space Model (VSM) [1], or similarly, as a bag of words model.

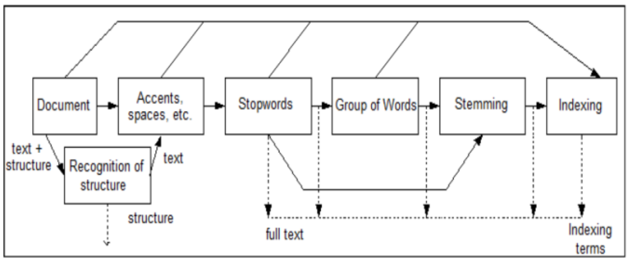


Fig 2. Text mining process

It is usual do consider this dataset as in a matrix form, that has the tweets as rows and the terms as columns. Let D = (texts units X terms), or (tweets x terms) in our case. The entry of D for a given (tweet, term) D(d; t) is the weight of the term t for the tweet d. The weights can be: (a) Boolean, 0 for absence and 1 for presence of the term t in the tweet d; (b) the frequency tf of the term t in the tweet d; (c) the so-called term frequency-inverse document frequency (tf -idf) schema, given by:

t -id(d; t) = t (d; t) _ log(ND=d(t))

IV. PROPOSED METHODOLOGY

They use cluster analysis and text-mining to extract patterns from a dataset composed by large collections of text messages. In this section we present several works that corroborate to one fundamental hypothesis: the social networks – and particularly microbloggers such as Twitter – are nowadays very important sources for real-time event detection and analysis. Also, these works have in common the fact that they employ clustering techniques, but a direct comparison of the available clustering algorithms cannot be found. Text classification, that can be considered as a very promising technique to perform clustering in text applications.

V. APPLICATION

- This application is useful for common people who don’t want to write any unwanted messages like vulgar, political, sexual messages on his\her own wall by any third person.
- Mostly, this type of activities are happen with some famous personalities, So if this facility will provide with OSN sites then people can protect his wall from this type of malpractices.

VI. CONCLUSION

This paper, propose a novel method to analyze social media data. Our method used K-Means algorithm along with Text classification method to cluster the social media community based on leadership, follower and attitude scores. With the empirical evaluation, the proposed algorithm outperforms other existing methods. It also presents a use-case of the method to further describe user community by getting more insights from clustering results and assigning self-explanatory labels to each cluster. For future work, the method is to be used with different domains further. Using this information it is possible to generate graphs associated to the emission and propagation of the specific topics that appear on this social micro blog. Also, time-dependent spreading of the event news can be analyzed.

VII. REFERENCES

[1] Andreas M Kaplan and Michael Haenlein. Users of the world, unite the challenges and opportunities of social media. Business horizons, 53(1):59–68, 2010.

[2] Bogdan Batrinca and Philip C Treleaven. Social media analytics: asurvey of techniques, tools and platforms. AI & SOCIETY, 30(1):89–116, 2015.

[3] J´erˆome Kunegis, Andreas Lommatzsch, and Christian Bauckhage. The slashdot zoo: mining a social network with negative edges. In Proceedings of the 18th international conference on World wide web, pages 741–750. ACM, 2009.

[4] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. Science (New York, NY), 323(5915):721, 2009.

[5] Claudio Cioffi-Revilla. Computational social science. Wiley Interdisciplinary Reviews: Computational Statistics, 2(3):259–271, 2010.

[6] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In Proceedings of the 19th international conference on World wide web, pages 591–600. ACM, 2010.

[7] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone’s an influencer: quantifying influence on twitter. In Proceedings of the fourth ACM international conference on Web search and data mining, pages 65–74. ACM, 2011.

[8] L. Bijuraj, “Clustering and its applications,” in Proceedings of National Conference on New Horizons in IT-NCNHIT, 2013, pp. 169-172.

[9] P. Garg, R. Rani, and S. Miglani, “Analysis and visualization of professionals Linkedin data,” in the proceedings of International Conference on Emerging Research in Computing, Information, Communication and Applications, Springer, 31 July - 01 August 2015, pp. 1–9.

[10] https://blog.twitter.com/2009/location-location-location

[11] https://dev.twitter.com/streaming/overview

[12] https://oauth.net/

[13] http://cran.rproject.org/bin/windows/base/

[14]http://ftp.iitm.ac.in/cran

[15] https://en.wikipedia.org/wiki/Tag_cloud