TREND ANALYSIS ON TWITTER

Tejal Rathod[1], Prof. Mehul Barot[2]
[1]Research Scholar, Computer Department, LDRP-ITR, Gandhinagar
Kadi Sarva Vishvavidhyalaya
[2]Assistant Prof. Computer Engineering Department, LDRP-ITR, Gandhinagar

Abstract
Twitter is most popular social media that allows its user to spread and share information. It Monitors their user postings and detect most discussed topic of the movement. They publish these topics on the list called "Trending Topics". It show what is happening in the world and what people's opinions are about it. For that it uses top 10 trending topic list. Twitter Use Method which provides efficient way to immediately and accurately categorize trending topics without need of external data. We can use different Techniques as well as many algorithms for trending topics meaning disambiguation and classify that topic in different category. Among many methods there is no standard method which gave accuracy so comparative analysis is needed to understand which one is better and gave quality of the detected topic.
Keywords: Information Retrieval, social media, Twitter, Twitter Trending Topic, Topic Detection, Text mining, Trend analysis, Real time.

I. INTRODUCTION

Twitter has become a huge social media service where millions of users contribute on a daily basis. It exchange wide variety of local and real world event. Short text messages post by user which is called "Tweets". Tweets are limited by 140 character in length and it can be viewed by user's follower. Twitter has two features:

- The shortness of tweets, which cannot go beyond 140 characters, it facilitates Creation and sharing of messages in a few seconds
- Easiness of spreading message to a large number of user with in little time.

Twitter has standard syntax which listed follow:

- User Mentions: when a user mentions another user in their tweet, @-sign is placed before the corresponding username. Like @ Username
- Retweets: Re-share of a tweet posted by another user is called retweet. i.e., a retweet means the user considers that the message in the tweet might be of interest to others. When a user retweets, the new tweet copies the original one in it.
- Replies: when a user wants to direct to another user, or reply to an earlier tweet, they place the @username mention at the beginning of the tweet, e.g., @username I totally agree with you.
- Hashtags: Hashtags included in a tweet tend to group tweets in conversations or represent the main terms of the tweet, usually referred to topics or common interests of a community. It is differentiated from the rest of the terms in the tweet in that it has a leading hash, e.g., #hashtag.

Trending Topic:

As we know twitter allow user to spread and share information. It monitor on its user's activity their postings and find out most discussed topic of the movement which is called "Trending Topics". Trend shows what people are talking about and their opinion about particular topic.

Trend show the list of the topics which are immediately popular rather than topics which have been popular on daily basis. Trending topic consist short phrases, keyword and hashtags. To find out trending topic there must be needed Real time data. To handle real time data is the complex task in information retrieval. Twitter

handle real time data, analyze that data then by Appling machine learning techniques it find out trending topics. It use many techniques and algorithms.

Example of Trending Topic:

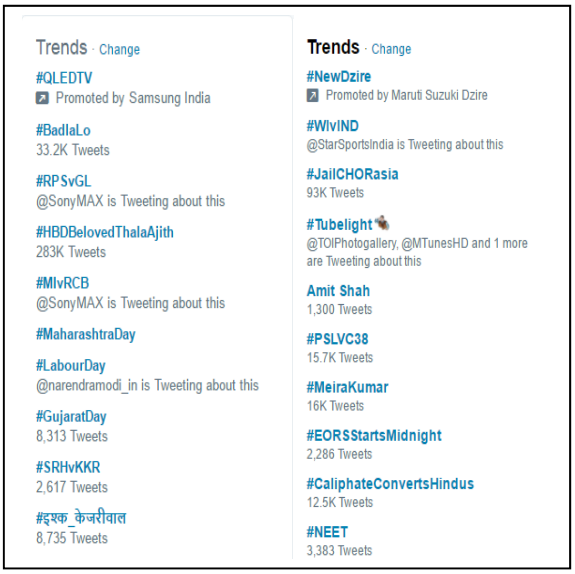

Fig 1 Trending topic list

II. LITERATURE SURVEY

Trend analysis is based on prediction by the Real time events. To monitor and detect all these aspects in real time, we need to extract relevant information from the continuous stream of data originating from such online sources. There are a wide variety of methods and algorithms they greatly affect the quality of results. Sampling procedure and the pre-processing of the data all greatly affect the quality of detected topics, which also depends on the type of detection method used [1].

III. TOPIC DETECION FORM TWITTER

How Trend Determine?

Trend are determined by an algorithm and by default are tailored for you based on who you follow, your interests, and your location. This algorithms identifies topics that are popular now, rather than topics that have been popular for a while or on a daily basis, to help you discover the hottest emerging topics of discussion on twitter.

The Number of Tweets that are related to the Trends is just one of the factor the algorithm looks at when ranking and determining trends. Algorithmically, trends and hashtags are grouped together if they are related to the same topic. Like #MondayMotivation and #MotivationMonday may be represented by #MondayMotivation.

We can choose to see trends that are not tailored for you by selecting a specific trends location on twitter.com, iOS, Android. Location trends popular topics among people in a specific geographic location.

Event Detection and Extraction in Twitter:

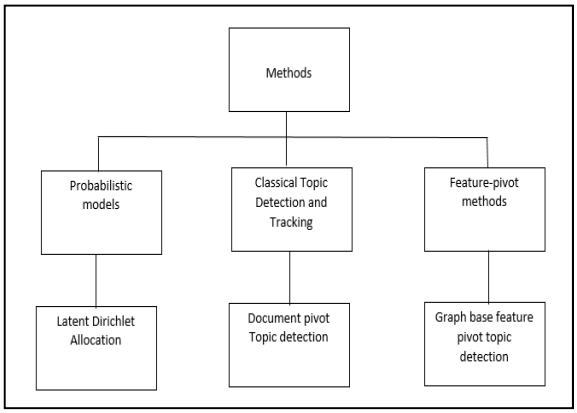

Fig 2 Trend Detection Methods [1]

1) Latent Dirichlet Allocation (LDA): Topic extraction in textual corpora can be addressed through probabilistic topic models. In general, a topic model is a Bayesian model that associates with each document a probability Distribution over topics, which are in turn distributions over words. LDA [9] is the best known and most widely used topic model. According to LDA, every document is considered as a bag of terms, which are the only observed variables in the model. The topic distribution per document and the term distribution per topic are instead hidden and have to be estimated through Bayesian inference. We use the Collapsed Variation Bayesian inference algorithm [10], an LDA variant that is computationally efficient, more accurate than standard variation Bayesian inference for LDA, and has parallel implementations already available in Apache Mahout 1. LDA requires the expected number of topics as input and in our evaluation we explore the quality of the topic for different values of. The estimation of the optimal, although possible through the use of non-parametric methods [11].

2) Document-Pivot Topic Detection (Doc-p): It is Topic Detection and Tracking method that uses a document-pivot approach. It uses LSH to rapidly retrieve the nearest neighbour of a

document and accelerate the clustering task. The principle behind this method is the same used for the near-duplicate detection in the similarity-based aggregation step of the pre-processing phase.

Work as follow:

- Perform online clustering of posts: Compute the cosine similarity of the tf-idf[13] representation of an incoming post to all other posts processed so far. If the similarity to the best matching post is above some threshold θtf-idf, assign the item to the same cluster as its best match; otherwise create a new cluster with the new post as its only item. The best matching tweet is efficiently retrieved by LSH.
- Filter out clusters with item count smaller than.
- For each cluster , compute a score as follows:

$$score_c = \sum_{i=1}^{|Docs_c|} \sum_{j=1}^{|words_i|} \exp\left(-p(w_{ij})\right) \quad \text{..... (1)}$$

Where wij is the th term appearing in the ith document of the cluster [11]

The merit of using LSH is that it can rapidly provide the nearest neighbours with respect to cosine similarity in a large collection of documents. An alternative would be to use inverted indices on the terms that appear in the tweets and then compute the cosine similarity between the incoming document and the set of documents that have significant term overlap with it. The use of LSH is much more efficient as it can directly provide the nearest neighbours with respect to cosine similarity.

3) Graph-Based Feature-Pivot Topic Detection: This method has unique feature is that for the feature clustering step it uses the Structural Clustering Algorithm for Networks (SCAN) [14]. A property of SCAN is that apart from detecting communities of nodes, it provides a list of hubs, each of which may be connected to a set of communities. In a feature-pivot approach for topic detection, the nodes of the graph would correspond to terms and the communities would correspond to topics. The detected hubs would then ideally be considered terms that are related to more than one topic, something that would not be possible to achieve with a common partitioned clustering algorithm and would effectively provide an explicit link between topics. We select the terms to be clustered, out of the set of terms present in the corpus, using the approach in [15]. It uses an independent reference corpus consisting of randomly collected tweets. For each of the terms in the reference corpus, the likelihood of appearance p (w|corpus) is estimated as follows:

$$p(w|corpus) = \frac{N_w + \delta}{(\sum_u^n N_u) + \delta n} \quad \text{..... (2)}$$

Where, Nw is the number of appearances of term in the corpus, n is the number of term types appearing in the corpus, δ is a small constant that is included to regularize the probability estimate [15].

To determine the most important terms in the new corpus, we compute the ratio of the likelihoods of appearance in the two corpora for each term. That is, we compute:

$$\frac{p\left(w|corpus_{new}\right)}{p\left(w|corpus_{ref}\right)} \quad \text{..... (3)}$$

The terms with the highest ratio will be the ones with significantly higher than usual frequency of appearance and it is expected that they are related to the most actively discussed topics in the corpus. Once the high-ranking terms are selected, a term graph is constructed and the SCAN graph-based clustering algorithm is applied to extract groups of terms, each of which is considered to be a distinct topic.

The algorithm steps are the following:

- Selection: The top terms are selected using the ratio of likelihoods and a node for each of them is created in the graph G.
- Linking: The nodes of G are connected using a term linking strategy. First, a similarity measure for pairs of terms is selected and then all pairwise similarities are computed. Various options for the similarity measure are explored: the number of documents in which the terms co-occur, the number of co-occurrences divided by the larger or smaller document frequency of the two terms, and Jaccard similarity.

- Clustering: The SCAN algorithm is applied to the graph; a topic is generated for each of the detected communities.
- Cluster enrichment: The connectivity of each of the hubs detected by SCAN to each of the communities is checked and if it exceeds some threshold, the hub is linked to the community. A hub may be linked to more than one topic.

IV GENERAL APPROCHES FOR TRENDING TOPIC MEANING DISAMBIGUATION

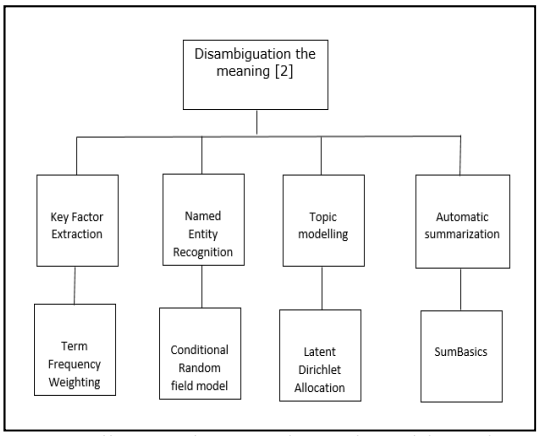

Fig 3 Trending topic meaning Disambiguation [2]

1) Key Factor Extraction: Term Frequency Weighting: TF (Term Frequency) weighting is a classic key factor extraction technique for automatic determination of term relevance [16]. The term frequency in the given tweets gives measure of importance of the term within the particular document. TF can be determined the exact values in various ways, such as raw frequency, Boolean frequency, logarithmically scaled frequency, and augmented frequency. We used raw frequency calculation, which is the most classical approach. The TF weighting tf(t,d) can be calculated by counting the number of times each term occurs in a document.

$$tf(t,d) = \frac{f(t,d)}{max\{f(w,d) : w \in d\}} \quad \ldots\ldots (4)$$

2) Named Entity Recognition: CRF Sequence Model: Named entity recognition (NER) is widely used for labelling the name of objects in documents. It labels sequences of terms, which are about the name of objects, such as person, organisation, or location. By recognising named entities, it can be easy for people to identify what kind of subject/topic the document is discussing about. We applied one of the most popular Named Entity Recognition approach, Conditional Random Field (CRF) sequence model. CRF-based NER are investigated by Stanford NLP lab [17] and it is widely used as a standard NER technique. CRF is a type of probabilistic sequence model, and it is applied for sequential data labelling.

The basic idea of CRF sequence model is as follows.

- Assume X is a random variable over data sequences to be labelled, and Y is a random variable over corresponding label sequence. The nodes in the model are separated into two different sets, X and Y. A conditional distribution p(Y—X) with an associated graphical structure will be modelled.

3) Automatic Summarisation: SumBasics: Automatic Summarisation was introduced for people to save the document reading time by providing a summary that retains the most important points of the documents. There are two main approaches, extraction and abstraction, in automatics summarisation. According to the evaluation conducted by Inouye and Kalita [18], most extraction approaches produced better performance; especially SumBasic had the highest scores in ROUGE metrics.

Sum Basic is a frequency based summarisation system uses the following algorithm:

- First, it calculates the probability distribution over the words in the input data. For each sentence in the input, assign a weight equal to the average probability of the words in the sentence.
- Then, select the highest scored sentence that contains the best probability word. For each word in the chosen sentence, update the probability. If the desired summary length has not been reached, go back to the first step.

V TRENDING TOPIC CLASIFICATION

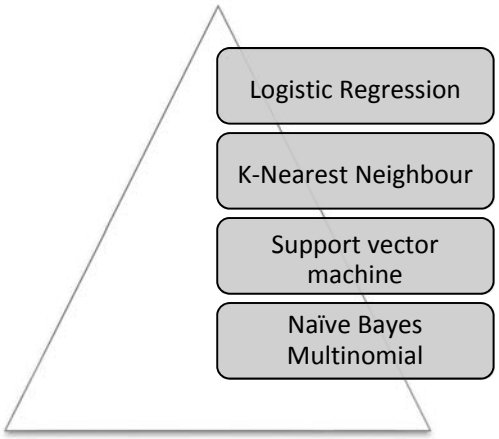

Fig 4 Trending topic Classification

We can classify trending topic in following category[20]:

News: breaking news tend to make it to Twitter early on, having even shown that on many occasions news break on We can define that a trending topic can be categorized as news when it is produced by a newsworthy event that major news outlets either had reported it by the time the trend popped up or will report it soon after it broke on Twitter.

Ongoing events: In type of trending topic we identified was produced by a community of users tweeting about an ongoing event as it unfolds. The practice of live-tweeting an event as it is taking place has become essential as twitter has gained importance as a real-time information sharing media.

Memes: It is viral ideas initiated by either an individual or an organization, who were usually popular enough to be able to spread something widely. We can define as the event that, without being apparently newsworthy or a mainstream event that a large audience is following, makes it to a large community of users for being funny or attractive to them.

Commemoratives: We define a trending topic as triggered by a commemorative when users are congratulating a celebrity for their birthday, celebrate the anniversary of a certain event or person, or it is a memorial day.

Classification Techniques:

Nearest-Neighbor Learning Algorithm: The k-nearest neighbor algorithm (k-NN) is a non-parametric method used for classification and regression [19]. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.it can be useful to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones.

Support Vector Machine: SVM which is called by support vector machine is machine learning technique used for classification. SVM use for finding best feed. It find closest data point from the line which maximize margin. SVM having less mean square error compare to linear and logistic regression. So by these many features we can use support vector machine in Trending Topic classification.

VI CHALLENGES & USEFUL TO SOCIETY

Challenges:

- It consist Phrases, Keyword or hash tags - Term Ambiguity
- Topic Classification
- Handle Real time data

Useful to Society:

- Health and Safety: Twitter data to better Predict flu. During the 2012-13 flu epidemic, researchers formulated a method to extract relevant data, based on tweets, to help them correlate the spread of the disease with a view to reducing its impact. Their system was able to "detect the weekly change in direction (increasing or decreasing) of influenza prevalence with 85% accuracy, a nearly twofold increase over a simpler model".[8]
- Stock Market
- To identify Breaking news.
- Election
- Entertainment.

CONCLUSION

Topic detection from social media streams is a complex process that has to deal with all the interleaved dimensions that characterize the emergence of a story on a social network. The textual content of the user-generated posts, the distribution of the messages in time and the nature of the events around which the crowd is commenting are the three most important aspects

to consider. Many tools available for finding twitter trend. There is no standard topic detection technique has been established yet, comparative analysis is needed to understand to what extent these dimensions determine the quality of the detected topics.

REFERENCES

[1] Sensing Trending Topics in Twitter,Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker, Ioannis Kompatsiaris, Senior Member, IEEE, and Alejandro Jaimes IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 15, NO. 6, OCTOBER 2013.

[2] Twitter Trending Topics Meaning Disambiguation,Soyeon Caren Han1, Hyunsuk Chung1, Do Hyeong Kim2, Sungyoung Lee2, and Byeong Ho Kang1 1 School of Engineering and ICT University of Tasmania, Sandy Bay, 7005, Tasmania, Australia Kyung Hee University, Giheng-gu,Youngin, Korea. Y.S. Kim et al (Eds.): PKAW 2014, LNCS 8863, pp. 126–137, 2014. Springer International Publishing Switzerland 2014.

[3] Real-Time Classification of Twitter Trends Arkaitz Zubiaga1, Damiano Spina2, Raquel Mart´ınez2, V´ıctor Fresno2 1 Dublin Institute of Technology DIT Focas Institute, Camden Row Dublin 8, Ireland 2 NLP & IR Group at UNED C/ Juan del Rosal, 16 28040 Madrid, Spain .Journal of the American Society for Information Science and Technology copyright @ 2013.

[4] DETECTING TRENDS IN TWITTER TIME SERIES Tijl De Bie1;2, Jefrey Lijffijt1 C´edric Mesnage2, Ra´ul Santos-Rodr´ıguez2 2016 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 13–16, 2016, SALERNO, ITALY.

[5] Trend Analysis of News Topics on Twitter Rong Lu and Qing Yang, International Journal of Machine Learning and Computing, Vol. 2, No. 3, June 2012.

[6]https://www.slideshare.net/bpedro/information-retrieval-challenges

[7]http://www.socialmediatoday.com/content/8-excellent-twitter-analytics-tools-extract-insights-twitter-streams

[8]http://www.socialmediatoday.com/social-networks/heres-why-twitter-so-important-everyone

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003.

[10] Y. W. Teh, D. Newman, and M. Welling, "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation," in Adv. Neural Inf. Process. Syst., 2007, vol. 19.

[11] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," J. Amer. Statist. Assoc., vol. 101, no. 476, pp. 1566–1581, 2006.

[12] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to Twitter," in Proc. HLT: Annual Conf. North American Chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 181–189.

[13] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval.New York, NY, USA: McGraw-Hill, 1986.

[14] X.Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger, "SCAN: A structural clustering algorithmfor networks," in Proc. KDD: 13th ACM Int. Conf.Knowledge Discovery and Data Mining, New York, NY, USA, 2007,pp. 824–833.

[15] B. O'Connor, M. Krieger, and D. Ahn, "TweetMotif: Exploratory search and topic summarization for Twitter," in ICWSM, W. W.Cohen, S. Gosling, W. W. Cohen, and S. Gosling, Eds. Palo Alto, CA, USA: AAAI Press, 2010.

[16] Russell, M.A.: Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. O'Reilly Media, Inc. (2013)

[17] Ritter, A., Clark, S., Etzioni, O.: Named entity recognition in tweets: An experimental

study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2011)

[18] Inouye, D., Kalita, J.K.: Comparing twitter summarization algorithms for multiple post summaries. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom). IEEE (2011)
pp. 384-394.

[19] Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 46 (3): 175–185.

[20] Albakour, D. ; Macdonald, C. , and Ounis, I. . Identifying local events by using microblogs as social sensors. In Proceedings of OAIR 2013, the 10th Conference on Open Research Areas in Information Retrieval, Lisbon, Portugal, 2013. Springer.