International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 12 | Dec 2019 www.irjet.net p-ISSN: 2395-0072

A Survey on Trend Analysis on Twitter for Predicting Public Opinion on Ongoing Events

V. D. Punjabi[1], Snehal More[2], Deepika Patil[3], Vinay Bafna[4], Shrushti Shah[5], Harshada Bachhav[6]

[1]Assisant Professor, Dept. of Information Technology, R. C. Patel Institute of Technology, Maharashtra, India
[2,3,4,5,6]Students, Dept. of Information Technology, R. C. Patel Institute of Technology, Maharashtra, India

---

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. 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.*

Key Words: Information Retrieval, social media, Twitter, Twitter Trending Topic, Topic Detection, Text mining, Trend analysis, Real time.

1. INTRODUCTION

Twitter has become a huge social media service where millions of users contribute on a daily basis. Two features have been fundamental in its success:

(1) The shortness of tweets, which cannot exceed 140 characters, facilitates creation and sharing of messages in a few seconds.

(2) The easiness of spreading those messages to a large number of users in very little time. Throughout the time, the community of users on Twitter has established a syntax for interaction with one another, which has become the standard syntax later officially adopted by its developers[1].

Most major Twitter clients have implemented this standard syntax as well. The standards in the interaction syntax include:

- **User mentions**: when a user mentions another user in their tweet, an at-sign is placed before the corresponding username, e.g., you should all follow @username, she is always abreast of breaking news and interesting stuff.

- **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 agree with you.

- **Retweets**: a retweet is considered a re-share of a tweet posted by another user, 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. Furthermore, the retweet attaches an RT and the @username of the user who posted the original tweet at the beginning of the retweet. For instance: if the user @username posted the tweet .

- **Hashtags**: similar to tags on social tagging systems or other social networking systems, hashtags included in a tweet trend to group tweets in conversations or represent the main terms of the tweet, usually referred to topics or common interests of a community. A hashtag is differentiated from the rest of the terms in the tweet in that it has a leading hash, e.g., #hashtag[2].

1.1 Trending Topic:

One of the main features on the homepage of Twitter shows a list of top terms so-called trending topics at all times. These terms reflect the topics that are being discussed most at the very moment on the site's fast-flowing stream of tweets. In order to avoid topics that are popular regularly (e.g., good morning or good night on certain times of the day), Twitter focuses on topics that are being discussed much more than usual, i.e., topics that recently suffered an increase of use, so that it trended for some reason.

Trending topics have attracted big interest not only among the users themselves but also among other information

consumers such as journalists, real-time application developers, and social media researchers. Being able to know the top conversations being discussed at a given time helps keep updated about current affairs, and discover the main concerns of the community. Twitter defines trending topics as "topics that are immediately popular, rather than topics that have been popular for a while or on a daily basis". However, no further evidence is known about the algorithm that extracts trending topics. It is assumed that the list is made up by terms that appear more frequently in the most recent stream of tweets than the usual expected[3].

1.2 Modules:

The system comprises of 4 major modules with its description as follows:

1. **Login:**

- User need to login first using valid credentials to access the system.

2. **Search for Latest Trends:**

- After successful login, user can search for latest trending tweets by entering the keyword in the search column.

3. **View Latest Trending Tweet:**

- Based on user-inputted keyword, the search results will be displayed in form of trending tweets.

4. **View Tweets:**

- User can click on respective trending tweet to view the message twitted by other users.

2. LITERATURE SURVEY

Trend analysis and based on that predicting public opinions. It plays important role, many researcher working on automatic technique of extraction and analysis of huge amount of twitter data. Luca Maria Aiello, compare six trend detection method and find that standard natural language processing technique perform well for social streams on particular topic. They conclude that n-gram give best performance other than state-of-art techniques.[4].

Ltawaier, M. M., & Tiun, S. (2016) have used three different machine learning algorithms Naïve Bayes, Decision Trees and Support Vector Machine for sentiment classification of Arabic dataset which was obtained from twitter. This research has followed a framework for Arabic tweets classification in which two special sub-tasks were performed in pre-processing, Term Frequency-Inverse Document Frequency (TF-IDF) and Arabic stemming. They have used one dataset with three algorithms and performance has been evaluated on the basis three different information retrieval metrics precision, recall, and f- measure[5].

Kathy Lee, proposed supervised learning techniques to classify twitter trending topic for that they use text based and network based classifier and conclude best performance. In A. Hernandez-Suarez propose model which predict public opinion on political event by Appling different classifier which predict that whether mood is positive or negative Kathy Lee, proposed a way to get the pre labeled data from twitter which can be used to train SVM classifier. They used the twitter hash tags to judge the polarity of tweet. To analyze the accuracy of proposed technique, a test study on the classifier was conducted which showed the result with the accuracy of 85%[6].

T Go, A. Bhayani, R., & Huang, L. (2009) introduced a new technique to classify the sentiment of tweets as positive or negative. They presented and discussed the results of machine learning algorithms for twitter sentiment analysis by using distant supervision[7].

Training data, the authors used consisted of tweets with emotions which were used as noisy labels. According to authors, the machine learning algorithms such as Naive Bayes, Maximum Entropy and SVM when trained with emotion tweets can have accuracy more than 80%. The study also highlighted the steps used in preprocessing stage of classification for high accuracy. In sentiment analysis perform using SVM in that two pre classified datasets of tweets are used then do comparative analysis, they use measures Precision, Recall and F-Measure[8].

3. PROPOSED METHODOLOGY

3.1 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. 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 [9], 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 [10].

3.2 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 representation of an incoming post to all other posts processed so far. If the similarity to the best matching post is above some threshold θ tf-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[11].

3.3 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). 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. Ina feature-pivot approach for topic detection, the nodes of the graph would correspond to terms and the communities would correspond to topics[12].

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. It uses an independent reference corpus consisting of randomly collected tweets. [13]

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.
- **Linking**: The nodes of Graph 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. Ahub may be linked to more than one topic[13].

4. CONCLUSION

Tweet having short message it use that for predicting public opinions on sports, Economy, ongoing events etc. It is finding keyword in tweet and predict whether it is having weightage positive or negative by applying machine leaning algorithms. It can apply multi classification algorithms like SVM, Naïve Bayes, Logistic classification, KNN and Decision tree. The Information retrieval measures like precision, recall and F- measure. In future it can test with python coding and find best classifier. 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] 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.

[2] Soyeon Caren Han, Hyunsuk Chung, Do Hyeong Kim, Sungyoung Lee, and Byeong Ho Kang "Twitter Trending Topics Meaning Disambiguation" Springer International Publishing Switzerland 2014.

[3] https://www.google.co.in/url?sa=t&rct=j&q=&esrc= s&source=web&cd=6&cad=rja&uact=8&ved=0ahUK EwiR2bbL_I_TAhVGu48KHUQ0BzEQFggyMAU&url=h ttp%3A%2F%2Fcucis.ece.northwestern.edu%2Fpub lications%2Fpdf%2FLeePal11.pdf&usg=AFQjCNFo8_ u6YiJFhnwOKV1RDN_yVIDT Q&sig2=Fq3_jJ34fq6iDkQW3PVbNw&bvm=bv.15217 4688,d.c2I.

[4] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker, Ioannis Kompatsiaris, Senior Member "Sensing Trending Topics in Twitter" IEEE, and Alejandro Jaimes IEEE Transactions On Multimedia, Vol. 15, No. 6, October 2013.

[5] Altawaier, M. M., & Tiun, S. (2016) "Comparison of Machine Learning Approaches on Arabic Twitter Sentiment Analysis" International Journal on Advanced Science, Engineering and Information Technology, 6(6), 1067-1073.

[6] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, Alok Choudhary, "Twitter Trending Topic Classification" 2011 11th IEEE International Conference on Data Mining.

[7] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(2009), 12.

[8] Munir Ahmad, Shabib Aftab, Iftikhar Ali "Sentiment Analysis of Tweets using SVM" International Journal of Computer Applications November 2017.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.