ISSN No:-2456-2165

Review On Topic Detection Methods for Twitter Streams

Vivek Ranjan, Pragati Agrawal
Computer Science & Engineering Maulana Azad National Institute of Technology Bhopal, India

Vaibhav Poddar
Economics Nowrosjee Wadia College Pune, India

Abstract:- Among the various social media platforms that dominate the internet today, Twitter has established itself as a major player and has become a preferred choice for expressing one's opinion on almost everything. Ordinary citizens aside, it is used by the governments world over to connect directly with their citizens, by the media houses, companies, research organizations, and so on. Given the omnipresent role it plays, it has thus become imperative to aggressively pursue topic detection from Twitter so that its benefits can be implemented in a range of applications such as natural disaster warnings, fake news detection, and user opinion assessment among other uses. This paper outlines different types of topic detection techniques that are employed frequently such as exemplar based topic detection, clustering based topic detection, frequent pattern mining, two level message clustering, combination of k-means clustering methods and singular value decomposition. While exemplar based topic detection technique uses the most significant tweets to detect topics, the clustering based technique uses different clustering methods such as sequential k-means, spherical k-means, DBSCAN, and bngram to detect the topics. In frequent pattern mining, the FP-growth algorithm along with its variation can be used to detect the topics. In two level message clustering, topics are clustered using different phases. A blend of both algorithms is employed in the combination of k-means clustering methods and singular value decomposition. A detailed discussion of these topic detection techniques has been done in this paper.

Keywords:- Text Mining; Topic Detection; Clustering; Twitter.

I. INTRODUCTION

Due to the rapid growth of social media on the internet in recent years, a multitude of valuable information has become available. Twitter is one such social platform where a large amount of information is available. The majority of information available on Twitter is in the form of textual documents. Twitter allows its users to send the information in the form of textual data termed as tweets almost in real-time. Each user can express his or her feelings, thoughts, knowledge in a maximum of 280 characters which makes Twitter a very important source for detecting important trends throughout the world. Millions of tweets are tweeted on Twitter every day. Because of the enormous number of tweets, identifying relevant topics in real-time is a challenging job. Detecting topics can be beneficial in several areas, namely, detecting natural disasters as quickly as possible and aiding political parties and corporations in assessing public opinion, detecting fake news as soon as possible, and many more. In this way, topic detection can help to reduce the misuse of social media. This paper outlines some of these topic detection techniques such as exemplar-based topic detection which uses most descriptive tweets to find the topic, frequent pattern mining in which patterns are searched using word count, clustering-based methods in which clusters are used to detect topics. Other approaches include two-level message clustering and a blend of singular value decomposition and k-means clustering.

The remainder of the paper is organized as follows. Section II mentions topic detection techniques, Section III discusses the complete approach of exemplar-based topic detection. Section IV discusses frequent pattern mining and variations that can be used to detect topics from Twitter. Various clustering methods such as sequential k-means, spherical k-means, DBSCAN, and bngram to detect the topics from Twitter are discussed in Section V. Section VI discusses all the phases used in two-level message clustering. Section VII discusses the combination of singular value decomposition and k-means clustering approaches in a detailed way. Finally, Section VIII concludes the various approaches used in the paper.

II. TOPIC DETECTION TECHNIQUES

There are three different approaches used by topic detection methods to represent the topics. Feature pivot method [2] is the first one. In this, terms are clustered together according to their patterns of co-occurrence. There are many approaches used in this method. One interesting approach is to utilize social features for selecting terms to be clustered [5]. The second one is the document pivot method [2] in which documents are clustered using some measure of document similarity and group the similar documents based on this [21].

The last one is probabilistic topic models in which topics are discovered by calculating the probability distribution of words, two most common probabilistic topic models are Latent Dirichlet Allocation (LDA) [3] and Probabilistic Latent Semantic Analysis (PLSA) [13].

Every topic detection technique has its own way of representing the topics. The most common approach to describe the topic is by using a list of keywords. The keyword's word count reflects the keyword's relevance in the given topic. Furthermore, explicit tweets can represent topics, in which the tweet keywords act as the topic's representation, as in the exemplar-based technique [14]. We

ISSN No:-2456-2165

identify multiple topic detection strategies in this paper such as clustering-based topic detection, frequent pattern mining, exemplar-based topic detection, two-level message clustering, and the combination of singular value decomposition and k-means clustering approaches.

III. EXEMPLAR BASED TOPIC DETECTION

Exemplar based topic detection uses the most descriptive tweet to represent a topic [9]. In this, the arrangement of tweets is maintained due to which users can understand the topic in an easier way. This approach discards the keywords which are not correlated and that makes the topic domain smaller. This strategy is focused on the idea that a tweet that is identical to a group of tweets but differentiated from the majority of the tweets is a good prototype for the topics it addresses. Based on that theory, this approach computes the similarity matrix between each pair of tweets and then analyses the action of such distributions by categorizing them into three groups. These three groups are:

- The distribution similarity of tweet *j* will have a low sample variance if *j* is similar to many tweets.
- The distribution similarity of tweet *j* will have a high sample variance if *j* is similar to a particular group of tweets and less similar to other groups.
- The distribution similarity of tweet *j* will have a low sample variance if *j* is dissimilar to most of the tweets.

The tweets which belong to group two are chosen to be the best representative of topics. Since each tweet is equivalent to a series of tweets, the tweets in group two are selected to represent topics better. The tweets in category two are also different from the majority of the tweets, so it divides its topic from the rest of the topics.

The sample variance for the similarity distribution of each tweet is calculated as:

$$Variance(s_j) = \frac{1}{n-1}\sum_{k=1}^{n}(s_{j_k} - x_j)^2 \quad (1)$$

where, x_j is the mean of the similarity of tweet *j* and is given by:

$$x_j = \frac{1}{n}\sum_{k=1}^{n} s_{j_k} \quad (2)$$

In this method, tweets are sorted according to their variances, and the tweet having the highest variance is identified as the representation of the first topic. Following the first topic selection, all tweets that belong to the first topic are eliminated, and the methodology selects another tweet with high variance as the representative for the second topic, and so on until k topics are discovered.

IV. TOPIC DETECTION USING FREQUENT PATTERN MINING

Frequent pattern mining is used to find frequent patterns or associations from item sets found in various kinds of databases [11]. For a given set of transactions, this mining method is used to find association rules to predict the co-occurrence of items. Frequent pattern mining was first used in mining transactions to find items that co-occur in the transactions. There are several variations of frequent pattern mining. Some of them are DHP [18], DIC [4] and Apriori [1]. Frequent pattern mining uses FP-growth algorithm [12]. Steps for the FP-growth algorithm are given below.

- In the first step, the database is scanned to find occurrences of the item sets in that database. Frequency of words below a particular threshold is discarded.
- The FP-tree is constructed in the next step, the root is formed and is denoted by null.
- In the next step, the database is scanned again and transactions are examined to find out item sets in it. The item set is sorted in descending order due to which item with maximum frequency is taken at the top and so on. In other words, item sets generate the tree branch in descending order of frequency.
- In this step, the examination of the next transaction of the database takes place. The item sets are sorted in descending order of frequency. The transaction branch has the same prefix to the root if the item sets of this transaction are available in another branch. In other words, the repeated item set is connected to the new node of another item set.
- In this step, increase the frequency of the item sets by one and the common node and new node.
- The generated FP-tree is mined in this phase. First, the lowest node and the lowest node links are examined. The pattern scale of the lowest node is one. The route is traversed in the FP-tree from here. These are referred to as conditional pattern bases.
- In this step, a conditional FP-tree is formed due to the frequency of item sets in the path. In the conditional FP-tree, those item sets are considered which meets the threshold support.
- In this step, the conditional FP-tree is used to generate frequent patterns

The same method can be used to detect topics on Twitter. As the seen topic, the top k frequent patterns were discovered using this method. Soft frequent pattern mining (SFM) was also implemented as a variant of frequent pattern mining [19]. By measuring the similarities between the set and each term, SFM greedily extends the set, including just one term. This method is replicated until the next term's resemblance and the set is smaller than a certain threshold.

ISSN No:-2456-2165

V. TOPIC DETECTION WITH CLUSTERING TECHNIQUES

There are many clustering methods used for topic detection. Some of them are sequential k-means, spherical k-means, DBSCAN, bngram [14]. In clustering methods, clusters are used to define a particular topic. Centroids in each cluster are used as the representative for that cluster and top m-words of the cluster are used as a keyword of that topic. The total no of topics defined is the no of clusters formed. The TF-IDF scheme is used to represent the tweets. Some of the clustering methods used for topic detection are:

A. Sequential K-means

Sequential k-means works on two steps [16]. In step 1, each data point is allocated to that cluster that has the closest center with respect to the data point in terms of Euclidean distance. In step 2, the centroid of each cluster is recomputed by using all the members of every cluster. The Euclidean distance is calculated as:

$$\sqrt{(m_1 - m_2)^2 + (n_1 - n_2)^2} \quad (3)$$

where the first data point is (m_1,n_1) and the second data point is (m_2,n_2).

B. Spherical K-means

In spherical k-means, cosine similarity is used in place of euclidean distance [16]. The cosine similarity is calculated as:

$$\cos(t, e) = \frac{te}{\|t\|\|e\|} \quad (4)$$

where *t* and *e* are two feature vectors.

C. DBSCAN

DBSCAN (Density-based spatial clustering of applications with noise) is a density-based algorithm, which predicts fixed density clusters. DBSCAN works on two parameters [10]:

- Eps: The neighborhood around a data point is termed as eps. In other words, if the length between two points is lesser than or equal to eps, the points are termed as neighbors. The major portion of data will be considered as outliers if the eps value is too small. The clusters combine and the majority of data points will lie in the same clusters if the eps value is too high.
- Minpts: It is the least no of data points within the eps radius. The Minpts must be large for larger datasets. Minpts must be greater than or equal to 3.

DBSCAN follows the following step:

- All the data points within the eps are searched and core points are identified.
- If any core point is not assigned to a cluster, new clusters are created.
- All the density connected points of the core point are searched recursively and they are assigned to the same cluster as the core point.
- If there exists a point c that has enough points in the neighbor of point a and points b and both a and b are within the eps then points a and point b are called density connected points.
- The process is repeated for the remaining points of the dataset which are unvisited.

D. Bngram

Bngram was proposed to detect topics from twitter streams [2]. Instead of using single words for topic detection, it uses ngrams. This method employs a new time-based scoring measure for each ngram. It is defined as:

$$df - idf_t = \frac{df_j + 1}{log(\frac{df_{prev}}{t} + 1) + 1} \quad (5)$$

where *df* is the number of tweets that appeared in this ngram in time slot *j*. df_{prev} is the number of tweets that were used in this ngram during the previous time slots. The frequencies which are increased during the present time slots in comparison with the previous time slots are chosen by this score measure. Following that, ngrams are ordered according to this score scale, and the top ngrams are chosen. A group average hierarchical clustering of ngrams is done in this process, with the similarities between two ngrams defined as tweet fragments in which both ngrams occur. This process is terminated when the correlation between the nearest unmerged clusters falls below a specific threshold value. At last, the ranking of clusters is done and topics are returned as the top clusters.

VI. TWO-LEVEL MESSAGE CLUSTERING FOR TOPIC DETECTION

This topic detection approach also involves various factors of topic enrichment and presentation like ranking of topics, title extraction, keyword extraction, representative tweets selection, and relevant image extraction [20]. All the phases used by this approach are given below:

A. Pre-processing Phase

Duplicate items are assembled and language-based filtering is carried out in this phase. Duplicate items are retweets and copies of previous tweets. It is done by hashing the text of each tweet and in one bucket, the only text of a tweet is allowed. Thus, only a single item of duplicate copies is processed while also keeping all other duplicate items. After this, language-based filtering is done in which only english tweets are kept and non-english words are discarded.

B. Topic Detection

After pre-processing, the modified document pivot approach is used for detecting topics in which tweets containing the same URL and a tweet long with its reply are clustered because they refer to the same topic.

In first-level clustering, the grouping of items is done. Union-Find structure [6] is used in grouping of items in the first level. A Graph containing one node for each tweet is created. Pair of tweets containing the same URL are connected and the set of the connected component is obtained. Now, those components which have more than one tweet are those first-level groups that will be also used in the second level. Those tweets will be also considered

which are only members of a component containing a single element. In second level clustering, the algorithm used is incremental threshold-based clustering and locality-sensitive hashing(LSH), but with some modifications. In this, all the first level clusters become members of second-level clusters. Moreover, the second-level cluster also has those tweets which were not the members of the first-level.

C. Ranking

In this phase, the ranking of topics produced in the second level is done in order to choose the most important.

D. Title Extraction

In this phase, the text of each clustered tweet is split into sentences for obtaining a particular set of titles. For this, Levehnstein's distance for each pair of candidate titles is computed for reducing the number of actual candidates. after that, candidate titles are ranked using their textual features and frequency. The score of the title is its frequency multiplied by the average likelihood of the appearance of those terms contained in an independent corpus.

E. Keyword Extraction

In this phase, keyword extraction is done. It is similar to the title extraction process but instead of complete sentences, either verb phrases or noun phrases are examined. After that, keywords are ranked similar to that of the title extraction but keyword extraction is not limited to selecting a particular candidate. To do this, the score difference between two succeeding candidates is computed and then after ranking them, the position in the ranked list with the largest difference is discovered and all terms until that position are selected.

F. Representative Tweets Selection

In this phase, representative tweets are selected. Moreover, all replies from the topic's cluster are added to the group of representative tweets.

G. Relevant Image Extraction

In this phase, relevant images are retrieved using a simple procedure. The most frequent image is found and returned if the tweets related with a topic have the URL of a few images.

VII. COMBINATION OF SINGULAR VALUE DECOMPOSITION AND K-MEANS CLUSTERING METHODS FOR TOPIC DETECTION

In this approach, the blend of singular value decomposition(SVD) and k-means clustering is used to detect Twitter topics [17]. Singular value decomposition(SVD) is used for reducing the dimensions. In the case of textual data, Singular value decomposition (SVD) is also known as latent semantic analysis(LSA) [8] [7]. At first, SVD reduced the dimensions of the tweets due to which its computational time becomes faster and after that, the k-means clustering method is used to form clusters of tweets from which topics are detected. There are two phases used in this approach.

A. Singular Value Decomposition

Let M be a $r \times s$ matrix. A factorization $M=USV^T$ is said singular value decomposition for M, where U is a $r \times r$ orthogonal matrix, S is a $r \times s$ pseudo-diagonal matrix whose elements non-negative, and V^T is a $s \times s$ orthogonal matrix. The diagonal elements of the matrix S are called singular value of M [15].

In this phase, tweets dimensions are reduced using SVD. Suppose X be $s \times r$ tweet word matrix. This matrix is decomposed into $X=USV^T$, where U is a $s \times x$ orthogonal matrix. S is a $x \times x$ pseudo-diagonal matrix and V^T is an $x \times r$ orthogonal matrix. In this way, the dimension of the word tweet matrix is reduced and the reduced matrix SV^T will go to the next phase.

B. K-means Clustering

The following steps are used to conduct k-means clustering on the reduced matrix:

- Tweets in the reduced form are clustered.
- The identified centroids are translated into the tweet's original dimensions.
- The particular topic is described by the top-weighted words lies in each centroid.

VIII. CONCLUSION

In this study, we have examined five strategies for detecting topics in Twitter streams. To sum it up, we have discussed exemplar-based topic detection, which uses the most descriptive tweet to catch the topic. We have discussed frequent pattern mining, in which recurring patterns are produced using the FP-growth algorithm. We have also discussed soft frequent pattern mining (SFM), a variant of frequent pattern mining. We have also addressed several clustering techniques such as sequential k-means, DBSCAN, spherical k-means, and bngram, in which clusters are first discovered using the appropriate algorithms and then used to define a topic. We have discussed two-level message clustering which is explained using different phases. We have also explored the use of singular value decomposition (SVD) with k-means. Singular value decomposition (SVD) reduces the tweet's dimension, which reduces processing time, and k-means is then used to detect topics.

REFERENCES

[1.] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In VLDB, 1994.
[2.] L. M. Aiello, G. Petkos, C. J. Mart´ın-Dancausa, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, Y. Kompatsiaris, and A. Jaimes.¨ Sensing trending topics in twitter. IEEE Transactions on Multimedia, 15:1268–1282, 2013.
[3.] Blei, A. Ng, and M. I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022, 2003.
[4.] S. Brin, R. Motwani, J. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In SIGMOD '97, 1997.

[5.] M. Cataldi, L. D. Caro, and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In MDMKDD ’10, 2010.
[6.] T. H. Cormen, C. Leiserson, R. Rivest, and C. Stein. Introduction to algorithms, second edition. 2001.
[7.] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. Journal of the Association for Information Science and Technology, 41:391–407, 1990.
[8.] S. Dumais. Latent semantic analysis. Annual Review of Information Science and Technology, 38:188–230, 2005.
[9.] Elbagoury, R. Ibrahim, A. K. Farahat, M. Kamel, and F. Karray. Exemplar-based topic detection in twitter streams. In ICWSM, 2015.
[10.] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD, 1996.
[11.] Goethals. Frequent Set Mining, pages 321–338. 07 2010.
[12.] J. Han, J. Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery, 8:53–87, 2006.
[13.] T. Hofmann. Probabilistic latent semantic indexing. In SIGIR ’99, 1999.
[14.] R. Ibrahim, A. Elbagoury, M. Kamel, and F. Karray. Tools and approaches for topic detection from twitter streams: survey. Knowledge and Information Systems, 54:511–539, 2017.
[15.] B. Jacob. Linear algebra. 1990.
[16.] K. Jain, M. Murty, and P. Flynn. Data clustering: a review. ACM Comput. Surv., 31:264–323, 1999.
[17.] [17] K. Nur’Aini, I. Najahaty, L. Hidayati, H. Murfi, and S. Nurrohmah. Combination of singular value decomposition and k-means clustering methods for topic detection on twitter. 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), pages 123–128, 2015.
[18.] J. S. Park, M. Chen, and P. S. Yu. An effective hash-based algorithm for mining association rules. In SIGMOD ’95, 1995.
[19.] Petkos, S. Papadopoulos, L. M. Aiello, R. Skraba, and Y. Kompatsiaris. A soft frequent pattern mining approach for textual topic detection. In WIMS ’14, 2014.
[20.] Petkos, S. Papadopoulos, and Y. Kompatsiaris. Two-level message clustering for topic detection in twitter. In SNOW-DC@WWW, 2014.
[21.] S. Phuvipadawat and T. Murata. Breaking news detection and tracking in twitter. 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 3:120–123, 2010.