International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 02 | Feb 2019 www.irjet.net p-ISSN: 2395-0072

IDENTIFICATION OF PREVALENT NEWS FROM TWITTER AND TRADITIONAL MEDIA USING COMMUNITY DETECTION MODELS

Dr. S.K. Jayanthi[1], K. Deepika[2]

[1]Head and Associate Professor, Department of Computer Science, Vellalar College for Women, Erode, Tamilnadu, India
[2]Research Scholar, Department of Computer Science, Vellalar College for Women, Erode, Tamilnadu, India

---

Abstract - *Information and communication technology has changed rapidly over the past 20 years with a key development being the emergence of social media which has the community based interactions, content sharing and collaborations. Twitter, a top microblogging service, has become an increasingly popular platform for web users to communicate with each other and also to spread and share breaking news, personal updates and spontaneous ideas. News media presents proficiently verified occurrences (or events). Twitter contains useful information that can hardly found in traditional media sources. The content analysis on Twitter and traditional media has not been well studied. This analysis can give the key point to know what kind of special and unique information may be in combined media sources, and to know how far the twitter media covers similar news information as like in traditional media, which is used to spread the low coverage news information to the society. This paper empirically compares the information of Twitter and traditional media, namely, BBC within a time span of one month. The three Community Detection namely, Girvan-Newman clustering, CLIQUE (CLuster in QUEst) and Louvain Community Detection Models were used to find the group of similar content.*

Index Terms — Microblog, Twitter, BBC, CLIQUE, Community Detection Model, Louvain model, Clustering, Modularity

I. INTRODUCTION

Social media activism brings an increased awareness about social issues and encourages people to use computers and mobile phones to express their concerns without actually having to engage actively with campaigns in real life. Data Mining techniques are utilized for mining the social media contents (i.e. Twitter, Facebook, Instagram) to shift through massive amount of raw data to discern the pattern. They also provide capabilities to predict the result of a future surveillance [2].

Social media is a type of online media that expedites conversation as opposed to traditional media, which delivers the useful content. Traditional media such as radio, books and network television is primarily designed to be a screen platform (one-to-many) whereas social media is designed to be a discussion platform (many-to-many). The social media interactions allow large groups of geographically dispersed users to produce valuable information resources, solving challenging problem by tapping into unique and rare expertise through discussion. Even though there is a broad variety of social media exist, Twitter is the widely used and familiar media that shares the opinion of the user.

Twitter is an electronic medium that allows a large number of users to communicate with each other concurrently. Twitter messages, called tweets can't exceed 140 characters. It is a medium of instantaneous feedback which means that any action in the real world usually receives a near immediate reaction or comment in terms of tweets expressing opinions or reactions to the action. The peculiarity of tweets is that sometimes they contain words prepended by \# which are referred to as hash tags which makes for a good candidate as a search key for the tweet [7].

While the availability of social content increases, the task of identifying high-quality content in sites based on user contributions becomes increasingly important which exhibits the additional source of information. Community detection is used to discover highly connected groups of individuals or objects inside the networks, which are called as Communities. The motivation behind the community detection is to understand the opinion of different groups towards the products, news or events. To ensure the graph-based clustering of Twitter and news media information, the keywords have to be identified. A graph represents the text and inter-connected words or other text entities with meaningful relations. The frequency criterion (TF-IDF) is used to extract the keywords, which is used to describe the news topics. The retrieval of keywords from the tweets and traditional media is accomplished by using Text mining techniques [9].

The rest of this paper is organized as follows: The work related to research is shown in Section II. Methodology is described to cluster the similar news information using Community Detection Models in Section III. Results and performance evaluation is shown in Section IV. Conclusion and future work is given in Section V.

II. RELATED WORKS

Derek Davis et. al [2017], has proposed unsupervised framework - SociRank which identifies news topics prevalent in both social and news media, and then ranks them by prevalence using their degrees of Media Focus (MF), User Attention (UA), User Interaction(UI). The input from news media and social media were used. Key terms

are extracted and filtered corresponding to a particular period of time. The intersection of the keywords from Social and News media are found and then the correlated words with the intersection words are obtained. These words are used as graph vertices and the edges are obtained using three similarity measures namely Dice, Jaccard and Cosine. A graph is constructed from the previously extracted key term. The graph is clustered in order to obtain well-disjoint Topic Clusters (TC) using Girvan-Newman Clustering, which uses the betweenness and transitivity as metrics. The Topic Cluster is ranked using the factors MF, UA, and UI. It automatically discovers the hidden popular topic, to improve the quality of news recommender system [1].

Hongzhi Yin et. al [2013], has proposed user-temporal mixture model for detecting temporal and stable topics simultaneously from the social media data. It is proposed to distinguish temporal topics from stable topics. To improve this model's performance, they design a regularization framework that exploits prior spatial information in a social network, as well as a burst-weighted smoothing scheme that exploits temporal prior information in the time dimension. The experimental results verify that our mixture model is able to distinguish temporal topics from stable topics in a single detection process. It is enhanced with the spatial regularization and the burst-weighted smoothing scheme significantly outperforms competitor approaches, in terms of topic detection accuracy and discrimination in stable and temporal topics [4].

W.X.Zhao et al. [2011], proposed a Twitter-LDA model to discover topics from a representative sample of the entire Twitter. Text mining techniques were used to compare the Twitter topics with topics from New York Times. Twitter can be a good source of entity oriented topics that have low coverage in traditional news media. Although Twitter users show relatively low interests in world news, they actively help to spread news of important world events [5].

Mario Cataldi et al. [2010], have proposed a novel topic detection technique that permits to retrieve in real-time the most emergent topics expressed by the community. First, they extract the contents (set of terms) of the tweets and model the term life cycle according to a novel aging theory intended to mine the emerging ones. A term can be defined as emerging if it frequently occurs in the specified time interval and it was relatively rare in the past. They formalized a keyword-based topic graph which connects the emerging terms with their cooccurrent ones and allows to detect the emerging topics under user-specified time constraints [6].

Canhui Wang, Min Zhang, Liyun Ru, Shaoping Ma [2008], Topic Detection and Tracking (TDT) proposed to bring convenience to users who intend to see what is going on through the Internet. However, it is almost impossible to view all the generated topics, because of the large amount. So it will be helpful if all topics are ranked and the top ones, which are both timely and important, can be viewed with high priority. The main contributions are: (1) present the quantitative measure of the inconsistency between media focus and user attention, which provides a basis for topic ranking and an experimental evidence to show that there is a gap between what the media provide and what users view. (2) To the best of our knowledge, it is the first attempt to synthesize both media focus and user attention into one algorithm for automatic online topic ranking [8].

III. SYSTEM METHODOLGY

The interest of social media and their relationship is used to know what is going around the world. Twitter users often spread the breaking or trending news. It is different from news media in which it shares the opinion of the people. The distribution of information between the twitter and news media is totally different. The opinion of the different users towards the news information can used to discover the interest and emergence topic (the topic frequently occurs in the specified period of time) which helps to spread the information that has the low coverage in social media.

SYSTEM ARCHITECTURE

The current research work has sequence of phases to accomplish the objectives. The figure 1 shows the overall architecture of the current work. The phases involved in this process are as follows,

- Dataset collection
- Preprocessing
- Key Term Graph Construction
- Community Detection Model
  - Girvan-Newman Clustering
    - Cluster Ranking: User Attention(UA)
  - CLIQUE Community Detection Model
  - Louvain Community Detection Model
- Performance Evaluation

A. DATASET COLLECTION

NEWS DATA

The BBC news website (https://www.bbc.com/news) contains international news coverage, as well as British, entertainment, science, and political news. Many reports are accompanied by audio and video from the BBC's television and radio news services. It is providing the interdisciplinary fields of news such as world, sports, weather, travel, business, entertainment, health, science and technology. The sports news for certain period is downloaded from the website (https://www.bbc.com/news/sports).

TWITTER DATA

Twitter account (https://twitter.com/Deepikakuppusa3) is used to create the Application Program Interface (API). The API provides the consumer secret key and access

token secret key for authenticated retrieval of tweets. The number of tweets related to sports news is collected.

B. PREPROCESSING

The collected news articles and the tweets are preprocessed in this phase.

News Keywords (N)

The meaningful keywords from the news articles are extracted by removing the numbers, punctuations, stop words and white spaces. N is a set which contains the keywords related to the news media information.

Twitter Keywords (T)

The tweet is collected by creating an API for twitter account. Sometimes the tweets may contain the Latin characters or other language words, so that the tweets with other characters have been discarded. The stop words, stem words are eliminated. T is a set which contains the keywords related to the twitter post.

C. KEY TERM GRAPH CONSTRUCTION

A graph G is generated, whereas the clustered nodes represent the prevalent news topic in both news and social media. The vertices in the graph G is the terms retrieved from N and T and the edges exhibit the relationship among the nodes. The following methods are used to find out the relationships between the words.

- **Term Document Frequency**

The document frequency of each term in News and Twitter is calculated accordingly. The set of news keywords (N) and set of twitter post keywords (T) is used to find the frequency of words. Here df(n) is the occurrence of the news term **n** and df(t) is the occurrence of twitter term **t.**

- **Relevant Key Term Identification**

To extract the topics that are prevalent in both news and social media, the following formula is used.

$$I = N \cap T \quad (1)$$

This intersection of N and T eliminates the terms from T that are not relevant to the news and terms from N that are not mentioned in twitter. Intersection words I are ranked based on their prevalence in both sources. The prevalence of a term is the combination of its occurrence in both N and T.

$$\forall i \in I : p(i) = \frac{df(n) \times \frac{|T|}{|N|} + df(t)}{2|T|} \quad (2)$$

Where, |T| is the total number of tweets chosen between dates d1 and d2 and |N| is the total number of news chosen in the same period of time.

The term in intersection word set (I) are ranked by their prevalence value, and only those in the top π-th percentile are selected. Using π value of 50 gives the best result in the current work.

$$I_{top} = \left\{ i \in I : \frac{|P_i|}{|I|} \times 100 > \pi \right\} \quad (3)$$

Where $|P_i|$ is the number of elements in subset P_i, which represents the term in I, with lower prevalence value than that of term and |I| is the total number of element in set I.

I_{top} represents the subset of top key term for the period from d1 to d2.

- **Key Term Similarity Estimation**

The perception behind the co-occurrence is the terms that co-occur frequently are related to the same topic and may be used to summarize and represent it when grouped. The co-occurrence for each pair (i, j) is found and defined as co(i, j). The term-pair co-occurrence is then used to estimate the similarity between terms. A number of similarity measures wereused namely Jaccard, Dice and Cosine similarity.

The Dice similarity between term i and j is calculated as follows,

$$dice\_QS(i,j) = \begin{cases} 0 & if co(i,j) \leq \vartheta \\ \frac{2 \times co(i,j)}{df_{top}(i) + df_{top}(j)} & otherwise \end{cases} \quad (4)$$

The Jaccard similarity between term i and j is calculated as follows,

$$jac_{QS(i,j)} = \begin{cases} 0 & if co(i,j) \leq \vartheta \\ \frac{co(i,j)}{df_{top}(i) + df_{top}(j) - co(i,j)} & otherwise \end{cases} \quad (5)$$

The Cosine similarity between term i and j is calculated as follows,

$$cosine_{QS(i,j)} = \begin{cases} 0 & if co(i,j) \leq \vartheta \\ \frac{co(i,j)}{\sqrt{df_{top}(i) \times df_{top}(j)}} & otherwise \end{cases} \quad (6)$$

where,

df_{top} (i) is the number of tweet that contain term i$\in I_{top}$

df_{top} (j) is the number of tweet that contain term j$\in I_{top}$

co(i,j) is the number of tweets in which terms i and j co-occur in I_{top}

ϑ is a threshold used to discard whose similarity that fall below it.

All of the formerly described similarity measures generate a value between 0 and 1.

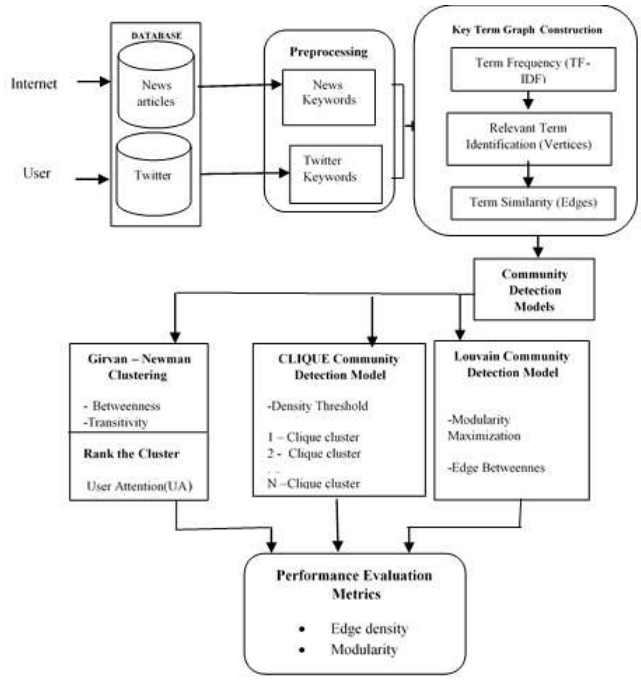

Fig.1 Overall Architecture of Community Detection Models

D. GIRVAN-NEWMAN CLUTERING

The Girvan–Newman algorithm detects communities (word cluster) by increasingly removing edges from the original network. The associated components of the remaining network are the communities. For any node, vertex betweennes is defined as the number of shortest paths between pairs of nodes that run through it. If there is additional shortest path between a pair of nodes, each path is assigned equal weight such that the total weight of all of the paths is equal to unity. Hence the edges connecting communities will have high edge betweenness (at least one of them). By removing these edges, the groups are divided from one another and so the underlying community structure of the network is exposed.

Steps for community detection algorithm is summarized as follows

Step 1: The betweenness values of all edges in graph G are calculated.

Step 2: The initial average betweenness of graph G is calculated.

Step 3: The high betweenness values are iteratively removed in order to separate clusters.

Step 4: The edge removing process is closed when removing additional edges yields no gain to the clustering quality of the graph. Once the process has been topped, the last detached edge is added back to G.

The betweenness centrality must be recalculated with each step. The reason is that the network adapts itself to the new conditions set after the edge removal. For instance, if two communities are connected by more than one edge, then there is no guarantee that all of these edges will have high betweenness. By recalculating betweenness after the removal of each edge, it is ensured that at least one of the remaining edges between two communities will always have a high value.

CONTENT SELECTION: USER ATTENTION (UA)

The User Attention (UA) represents the number of unique Twitter user related to the selected tweets. The tweets related to that topic are selected and then the number of unique users who created those tweets are counted. The equation for finding the UA is given below.

$$\forall TC \in GUA(TC) = \frac{|U_{TC}|}{\sum_{TC \in G} |U_{TC}|} \quad (7)$$

(or)

$$\text{UA} = \frac{Totalnumberofuniqueusers}{sumoftotalnumberofuniqueusers} \quad (8)$$

where U_{TC} is the number of unique users related to TC and G is the entire graph. This equation produces a value between 0 and 1.

E. CLIQUE DETECTION ALGORITHM

The CLIQUE algorithm was one of the first subspace clustering algorithms. It identifies dense clusters in subspaces of maximum dimensionality. The algorithm combines density and grid based clustering and uses an APRIORI style search technique to find dense subspaces. The algorithm then finds adjacent dense grid units in each of the selected subspaces using a depth first search. Clusters are formed by combining these units using a greedy growth scheme. The algorithm starts with an arbitrary dense unit and greedily grows a maximal region in each dimension until the union of all the regions covers the entire cluster. Redundant regions are removed by a repeated procedure [3].

CLIQUE, consists of the following steps:

- Identification of subspaces that contain clusters – which determine the dense units in all subspaces of interest
- Identification of clusters which determines the connected dense units in all subspaces of interests
- Generation of minimal description for the clusters which determines the maximal regions that cover a cluster of connected dense units for each cluster and minimal cover for each cluster.

Step 1: Divides each dimension in to equal-width intervals, save the intervals where the density is greater than the Density Threshold (T) as Clusters. The parameter

T is the minimum number of data points in unit space in dense region.

Step 2: If the number of data points in the region is not less than T then the region is defined as "dense area" or else that is defined as "sparse area".

Step 3: Each set of two dimensions are examined, if there are any intersecting intervals in these two dimensions and the density in the intersection of these intervals is greater than T. Then the intersection is again saved as Clusters.

Step 4: Repeat the steps for every set of dimensions.

Step 5: After each step adjacent clusters are replaced by a joint cluster. The overlapped clusters are the output at the end.

F. LOUVAIN METHOD

The Louvain method is a simple, efficient and easy-to-implement method for identifying communities in large networks. The method has been used with success for networks of many different types and for sizes up to 100 million nodes and billions of links.

The method consists of two phases.

Phase 1: It looks for "small" communities by optimizing modularity in a local way.

Phase 2: It aggregates nodes of the same community and builds a new network whose nodes are the communities.

The above phases are repeated iteratively until a maximum of modularity is attained. The partition found after the first step typically consists of many communities of small sizes. At succeeding steps, larger and larger communities are found due to the aggregation mechanism. This process naturally leads to hierarchical decomposition of the network.

The steps for the Louvain Algorithm are given below,

Step 1: Louvain algorithm is an iterative algorithm repeating until there is no additional improvement in modularity. It starts by initializing each node with its own community.

Step 2: Randomly choose the community of the neighbor node instead of considering all the communities. For each node in a graph, it computes the modularity gain ΔQ for all neighboring communities if the node moves.

Q indicates the modularity gain and is defined as

$$\Delta Q = \left[\frac{\sum_{in} k_{i,in}}{2m} - \left[\frac{\sum_{tot}+k_i}{2m}\right]^2\right] - \left[\frac{\sum_{in}}{2m} - \left[\frac{\sum_{tot}}{2m}\right]^2 - \left[\frac{k_i}{2m}\right]^2\right] \quad (9)$$

where. $\sum_{in}$ is the sum of the weights of the links inside the community to which the node i is assigned, $\sum_{tot}$ is the sum of the weights of the links incident to nodes in the community, and $k_{i,in}$ is the sum of the weights of the links from i to nodes in the community which is same with the community of node i.

Step 3: All communities are collapsed to the vertices, the edges are collapsed into a single self-looping edge. Multiple edges between every two communities are collapsed into a single edge, and the weight is the sum of the edges between them. Then the new graph is constructed.

Step 4: The process is repeated until the modularity recalculation process doesn't give any improvements in community structure.

IV. RESULTS AND DISCUSSIONS

A. EXPERIMENTAL ANALYSIS

The Process of existing and proposed work contains the following steps:

Step 1: The input news data is first downloaded from the BBC news portals (http://www.bbc.com/). The sports news such as Cricket, Rugby, Tennis, Football, Golf, Athletics and Cycling are considered in the current work. Tweets are collected by using the Twitter API, which provides the authentication to retrieve the posts for research oriented analysis. The Twitter API uses the twitter account (https//twitter.com/Deepikakuppusamy) for creating the authentication interface.

Step 2: The keywords of the news and twitter media are found separately using the TF-IDF. The intersection of the two media (Twitter, News media) keywords is found and ranked using their prevalence.

Step 3: The relationships between the keywords are found by using three similarity measures namely Dice, Jaccard and cosine similarity measures. So the vertices are the text words which are connected by the edges, which create the graph.

Step 4: The vertices and edges form the clusters that are obtained using the Girvan-Newman clustering method. The betweennness concept is used to find the shortest path between the pair of nodes, so that it creates the topic clusters. The User Attention (UA) of the resultant cluster is calculated.

Step 5: The resultant graph obtained in step 4 is fed into the CLIQUE community detection; it considers the graph as the dimension. The number of dimension is divided and the dimensions are processed, starts from single region and then grows up to higher region. After that the overlapping regions are found.

Step 6: The outcome generated in step 4 is fed in to Louvain community detection model. Finally, the overlapped dense region is found.

Step 7: Finally, the edge density and modularity metric value is calculated to evaluate the quality of community structure. The highest value of metric value reveals the finest community, which is highly connected.

B. PERFORMANCE EVALUATION

Edge density

According to thesis, the exact expression for the number of densities of small clusters in two dimensional percolation models is

$$N(s, p) = \sum_{t=1}^{\infty} g(s,t)(1-p)^t P^s \quad (10)$$

The edge density range should fall within 1 to 2.

- Edge density 0.1 to 0.5 - Low edge density
- Edge density 0.5 to 1 - Medium edge density
- Edge density 1 to 2 - High edge density

Modularity

Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. Modularity is often used in optimization methods for detecting community structure in networks. The expected number of edges shall be computed using the concept of Configuration Models. The configuration model is a randomized realization of a particular network.

In the network n, where each node v has a node K_v degree, the configuration model cuts each edge into two halves, and then each half edge, called a stub, is rewired randomly with any other stub in the network even allowing self loops. Thus, even though the node degree distribution of the graph remains intact, the configuration model results in a completely random network.

$$l_n = \sum_n k_v = 2m \quad (11)$$

Randomly select two nodes *v* and *w* with node degrees K_v and K_w respectively and rewire the stubs for these two nodes, then,

$$\text{Expectation of full edges between v and w is } \frac{\text{(Full edges between v and w)}}{\text{(Total number of rewiring possibilities)}} \quad (12)$$

(or)

$$Q = \frac{1}{2m} \times \sum_{i,j} \left[A_{i,j} - \frac{k_i k_j}{2m}\right] \times \delta\,(c_i c_j) \quad (13)$$

where,

m = the number of edges, *Aij*= the weight of an edge between nodes *i*and *j*

ki= a degree of node *i*, *ci*= the community to which *i*belongs to, δ= a function, if $u = v$ then $(u, v) = 1$ or else$(u, v) = 0$

The modularity should fall in the range between 1 and 2.

- Modularity 0.1 to 1.5 - Medium
- Modularity 1.5 to 2 - High

Based on the number of twitter and news keywords, different community structure has been obtained. The performance is evaluated by increasing the number of news, check whether the methods yield finest community even for large graphs. Among the three methods, the CLIQUE method yields better (strengthen) community structure with the high rate of evaluation metrics, even the news and tweets are increased to some extent. The results for different number of input are explained in graphs.

Table 1 Number of Clusters for Twitter and News post using Community Detection Models

Number of Twitter and News Post	100 N + 400 T	200 N + 500 T	250 N + 1000 T	400 N + 1500 T	600 N + 2000 T
Algorithm	**Number of Clusters**				
Newman Cluster	3	4	5	6	12
Louvain Cluster	3	2	3	3	6
Clique Cluster	1	24	34	46	57

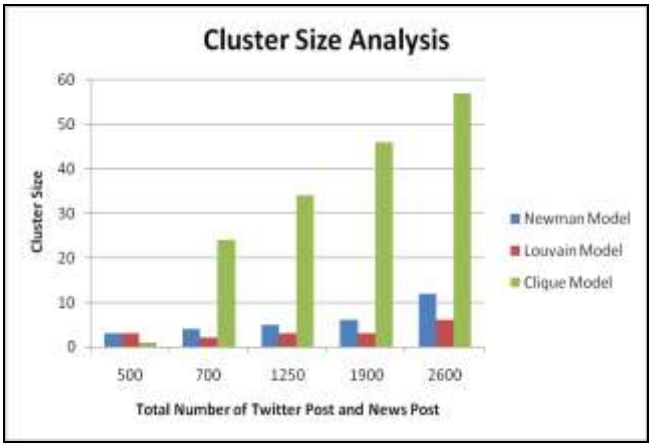

Fig.2 Resultant Graph based on Cluster size

Table 2 Modularity Values for Twitter and News Post using Community Detection Models

Number of Twitter and News Post	100 N + 400 T	200 N + 500 T	250 N + 1000 T	400 N + 1500 T	600 N + 2000 T
Algorithm	**Modularity**				
Newman Cluster	1.29	1.34	1.34	1.61	1.83
Louvain Cluster	1.34	1.02	1.05	1.34	1.78
Clique Cluster	1.34	1.43	1.45	1.63	1.89

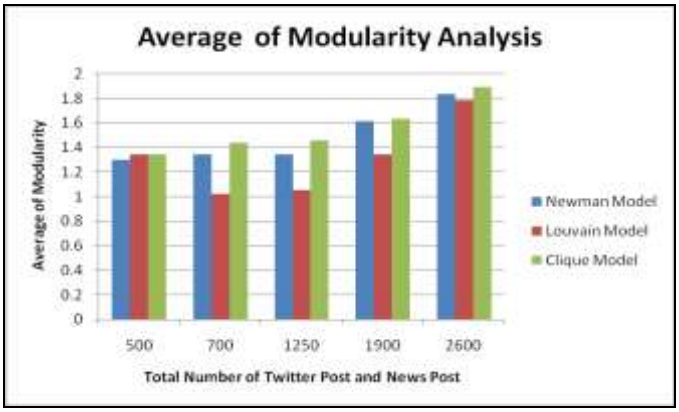

Fig.3 Resultant Graph based on Modularity Value

Table 3 Edge Density Value for Twitter and News post using Community Detection Model

Number of Twitter and News Post	100 N + 400 T	200 N + 500 T	250 N + 1000 T	400 N + 1500 T	600 N + 2000 T
Algorithm	**Edge Density**				
Newman Cluster	0.76	0.79	0.78	0.82	0.92
Louvain Cluster	0.56	0.65	0.67	0.70	0.78
Clique Cluster	1.02	1.72	1.23	1.34	1.38

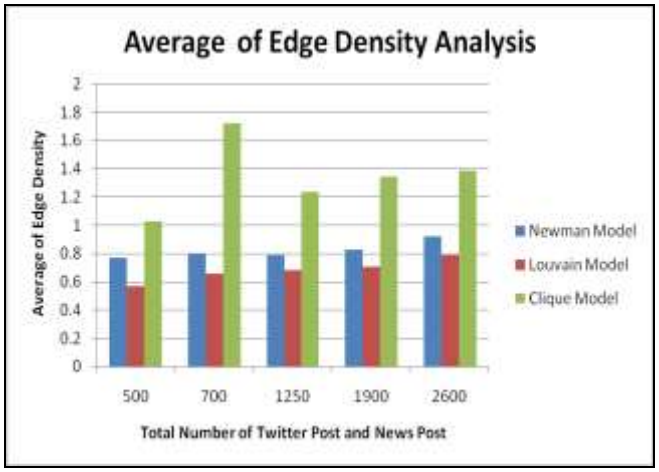

Fig.4 Resultant Graph based on Edge Density

To know the quality of the communities inside the news term graph, the evaluation metrics are used for both the existing and the proposed work. It ensures that the CLIQUE method detects the communities with high edge density and modularity (maximum number of relationship exist in a graph) than the other two methods (Girvan-Newman and Louvain method). The CLIQUE method gives better results even for the large number of input, finds the overlapping of clusters (internal clusters or Community).

V. CONCLUSION

Twitter is the top microblog network; its popularity makes it being a great platform to share the world events. Nowadays, the social media has become more popular to communicate rapidly millions of information with millions of users efficiently and effectively. The microblog messages contain the news information of traditional media, so the incorporation of two media sources provides stronger information about the current news. Here the problem is to detect the community structure; the resultant community reveals the most frequent topic spoken by the media.

In the existing scenario Girvan-Newman algorithm is used to identify the community cluster. It has issue with time complexity and accuracy. It takes more time to find the communities from huge network. Hence to overcome these issues, the CLIQUE detection and Louvain method is utilized in the current work. A key feature of CLIQUE detection is that it operates on multidimensional data, processing a single dimension and then grows up to high one. It also considers the overlapping of clusters. A key feature of the Louvain method is to maximize the modularity value to obtain the finest community. Hence the cluster of the news information has been done properly.

The current work is experimented on one month of BBC sports news; it has the collection of sports news topics such as football, cricket, rugby, tennis, golf, athletics and cycling. It reveals that the football news is frequently talked by the social media and traditional media.

The system is capable to handle the large graph and is more suitable to produce finest community with high rate of metric values. From the performance evaluation of the Modularity and Edge density, it is concluded that the CLIQUE algorithm could detect the community structure efficiently, even for the increased number of tweets and news information.

This research work can be enhanced in the future with the following scopes:

- All news topics can be considered (i.e. Politics, World events, Entertainment, Health, Business, Art and Travel)
- Other languages can be utilized
- Number of microblogs (i.e. Facebook, Twitter, MySay, Tumblr) can be incorporated
- The symbols in the tweets can be considered

VI. REFERENCES

1) Derek Davis, Gerardo Figueroa, and Yi-Shin Chen, SociRank: Identifying and Ranking Prevalent News Topics Using Social Media Factors, IEEE Transaction on Systems, Man, and Cybernetic Systems, vol. 47777, No. 6, JUNE 2017

2) Pang-Ning Tan Michael, Steinbach, Vipin Kumar, Introduction to Data Mining, ISBN - 978-93-325-1865-0, Pearson Education Inc., Third Impression, 2015.

3) Jyoti Yadav, Dharmender Kumar, Subspace Clustering using CLIQUE: An Exploratory Study, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) vol 3, Issue 2, FEBRUARY 2014.

4) Hongzhi Yin, Bin Cui, et al, A Unified Model for Stable and Temporal Topic Detection from Social Media Data, IEEE ICDE Conference, 978-1-4673-4910-9/13/$31.00, 2013.

5) Wayne Xin ZHAO, Jing JIANG, et al, Comparing Twitter and Traditional Media usingTopic Models, Institutional Knowledge at Singapore Management University, In Advances in Information Retrieval: 33rd European Conference on IR Research, ECIR2011, Dublin, Ireland, APRIL 18-21, 2011.

6) Mario Cataldi, Luigi Di Caro, Claudio Schifanella, Emerging Topic Detection on Twitter based on temporal and Social Terms Evaluation, ACM 978-1-4503-0220-3 ...$10.00, JULY 25, 2010.

7) JaganSankaranarayanan, HananSamet, et al, TwitterStand: News in Tweets, ACM GIS '09, Seattle, WA, USA, ACM, ISBN 978-1-60558-649-6/09/11 ...$10.00, NOVEMBER 4-6, 2009.

8) Canhui Wang, Min Zhang, Liyun Ru, Shaoping Ma, Automatic Online News Topic Ranking Using Media Focus and User Attention Based on Aging Theory, Napa Valley, California, USA, Copyright 2008 ACM 978-1-59593-991-3/08/10...$5.00, OCTOBER 26–30, 2008.

9) Rada Mihalcea and Paul Tarau, TextRank: BringingOrder into Texts, in Proc. EMNLP, vol.4, Barcelona, Spain, 2004.