Multimedia Tools and Applications (2023) 82:8943–8982
https://doi.org/10.1007/s11042-021-11855-3

1209: RECENT ADVANCES ON SOCIAL MEDIA ANALYTICS AND MULTIMEDIA SYSTEMS: ISSUES AND CHALLENGES

Topic adaptive sentiment classification based community detection for social influential gauging in online social networks

P. Kumaran[1] · S. Chitrakala[2]

Received: 21 September 2020 / Revised: 26 September 2021 / Accepted: 23 December 2021 /
Published online: 25 February 2022

Abstract

Online Social Networks (OSNs) such as Twitter, Facebook, Instagram, and WhatsApp are turned as a place for many of people in recent years to spend much of their time, due to their huge network structure and massive amounts of user-generated data in it. Those data's are widely used in various real-world applications such as online marketing, epidemiology, digital marketing, online product or service promotion, and online recommendation systems. Presently, the twitter has grown to become a mainstream medium for the dissemination of messages, which creates necessitated intensive research challenges in the field of social influential gauging, Influence Maximization Problems, alongside an information diffusion. First, to address the social influential gauging a novel Topic Adaptive Sentiment Classification based Community Detection (TASCbCD) algorithm is proposed to detect communities in twitter network based on the results of topic based sentiment classification using robust topic features. In the topic modelling, the initial topics of each extracted data and the robust topic features were used to classify using a multi-class support vector machine. The WordNet and SentiWordNet are benchmark data sets that are used for supporting those classification to achieve the desired results. The resultant communities give a better visualization of identifying the overlapping communities that helps to gauge the topic based social influential user in OSNs. However, from the experimental result, it is observed that the proposed algorithm achieves better results in RandIndex and Scaled Density metrics than state-of-the-art methods for communities detection.

✉ P. Kumaran
kumaran.0991@gmail.com

S. Chitrakala
chitrakala.au@gmail.com

[1] Department of Computer Science and Engineering, National Institute of Technology Puducherry, Karaikal, India

[2] Department of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai, India

Keywords Topic modeling · Sentiment analysis · Community detection · Influential spreader identification · Online social networks

1 Introduction

Today, OSN have piqued the interest of individuals from all walks of life, particularly in the previous decade [12, 46]. They have played, and continue to play, a critical role in the sharing of information in today's hyper-connected society [24]. Because information exchanges on social networks include not just text but also URLs, photos, audio, and video, they are difficult testbeds for data mining research [21, 58]. In particular, people more concerned about finding breaking news or discover a hidden market or an underground political movement in social networks [10]. Twitter is a popular online social networking site that provides a platform for the general public to express their opinions or feelings about social events and products in the form of tweets [11, 32]. Official statistics presented at the April 2010 conference [61] on Twitter and its users show that there were 106 million registered users in 2010 with 180 million new ones accessing it every month. Every day, 600 million queries received, along with 3,00,000 newly registered users through search engines and APIs. At the last count, 37% of users actively access Twitter for messaging via smartphones.

In recent times, businesses have been using social networking sites as sources of information for ideas to improve their marketing strategies. One such nugget of information is how information spreads through the network [63]. Such information can be used by businesses to determine the best way to promote their products, services, or campaigns [41, 45]. In this system, the estimation of the number of people reached carried out. By running it over different sets of sources, the best possible sources can also be identified [35]. In recent times, Twitter has grown to become a mainstream medium for the dissemination of messages and public discussion of news and events on its vast network [26]. Subsequently, the increased use of Twitter for information diffusion has necessitated intensive research in the Prediction of Information Diffusion (PID), with an effective influence spreader selection strategy on the network [31, 38].

The social networking service like twitter helps the in the current world, moment-by-moment (Schonfeld 2009). Twitter also claims that it is a potential medium for reflecting the thoughts and opinions of millions of people at a time, as individual users or as part of a group [49]. Compared to other forms of text information, Twitter data has a style of its own, and the limitations that go with it. Single tweets can accommodate a maximum of 280 characters, and Twitter supports 40 languages. For example, RT @bob has a new #car [36]. Here, RT refers to retweets of previous conversations; @ mentions the username a user will reply to, and # deals with the subject or topic discussed. Tweets offer users certain unique features to express their opinions or thoughts on twitter. Here are some sample tweets:

Example 1 I currently use the Nikon D90 and love it, but not as much as the Canon 40D/50D. I chose the D90 for the video feature. My mistake

Example 2 RT @cxrktree: where is your boy tonight i hope he is a gentlemen maybe he won't find out what i know you were the last good thing about this...

Example 3 RT @splcenter: "We have laws to protect civil rights. We'll enforce them, "says attorney general who just set up hate crime hotline https:/...

Example 4 RT @KateUpton: Hey @MLB I thought I was the only person allowed to fuck @Justin Verlander ?! What 2 writers didn't have him on their ballot?

Example 5 After a whole 5 hours away from work, I get to go back again, I'm so lucky!

Example 6 "I love being ignored."

Example 7 "It's Wednesday and it's freezing! It's raining! How better can this day be????????"

Example 8 #Rio2016 Olympics to open an hour later in Maracana Stadium!!!!. Worldwide media are ready to cover the sports gala.

In general, whenever there is a tweet collection, the following fields are collected to study a particular tweet in greater detail. Groups of tweets typically consist of 16 mandatory fields Tweet ID, User ID, Tweet text, Favorite, Favorite count, Retweets, Replies, Mentions, Created date, ID of the user, Screen name, Status source, Retweets count, Is re-tweeted, Longitude, and Latitude.

From a data mining perspective based on the features above, two fundamental processes can undertake with Twitter data:(a) Graph mining, based on user network links (b) Text mining, based on user-generated content [29]. The twitter network structure is involved in graph mining. It performs associated tasks such as influential user identification, expert prediction, provenance determination, and community detection [18]. Likewise, from a text mining perspective, twitter texts can be subjected to sentiment analysis, topic modeling, and opinion mining [25]. With this in view, a Robust Topic Features based Sentiment Classification for Overlapping Community Detection algorithm proposed for detecting communities, and its overlapping structure based on a different set of topics and its sensitivity level(polarity) to select influential users in the large-scale Twitter network [5, 53].

The selection of the influential spreader in OSN may not necessarily be a user with network centrality measures like outdegree, betweenness, closeness. The discovery of these users may also vary on other features like the topic of the user-generated content and its sensitivity level(polarity) [50]. The most critical component of influential spreader identification is determining the relevant features that describe the suitable social influencer [8]. However, the primary drawbacks addressed in this work is to operate most of the social network is their complex structure. In that, users may choose not to be swayed by the persuasive skills of an influential user, and those who do might not be entirely willing to be part of the information required. Another drawback is identifying the influential spreaders in a large-sized social network. Direct selection is a complex task; at best, community detection in online social networks helps select competent, influential users.

Community detection is the process for detecting and forming a group with the set of people who have the same interest or aim on some particular topic [21]. In the general community, detection techniques also help identify the different opinions of a particular topic by the people over the different periods [48]. Firstly, topic polarity differs daily in the real world due to the impact of different opinions of people in their group. Secondly, the diversity of topics culminates in overlapping community structures [16]. In general, prediction of information diffusion includes influence spreader identification, and influence maximization [7, 19]. However, This paper proposed community detection, which helps to enhance the selection of influence spreader according to the topic to improve the prediction accuracy of information diffusion in OSN [9].

Community detection in online social networks primarily relies on node and edge structure or content, or sometimes both. In such cases, the detected community fails to incorporate the content topic and its sensitivity feature to avert overlapping communities [37]. There is, therefore, a need for an approach that addresses these issues [4]. Extracting information from tweets is, in itself, a challenge since valuable data is present in limited characters is necessary to ensure that relevant information does not get eliminated as noise and that preprocessing steps designed accordingly, However, some users have difficulty trying to comprehend Twitter's recent updates, particularly when receiving overwhelming, difficult-to-recognize responses to input tweets [23, 38].

Finally, the rest of the sections in this paper formulated as follows. Section 2 discussed the extensive review of the literature survey of community detection algorithms. Sections 3 and 4 explained the detailed architecture of the proposed work and its algorithms to address problem statements. Section 5 describes the proposed system results, performance evaluation, and the discussion of comparative analysis with different test cases. At last, the paper concluded with future enhancement in Section 6.

2 Literature review

Online Social Networks (OSN) has become a perfect place to promote products and campaigns or services for worthy causes. Information diffusion in OSNs is gaining attraction today because of people's willingness to connect and communicate over the internet. Maximizing the spread of information in OSNs is a problem that has attracted the attention of promoters all over the world. Studying the diffusion process can help to make endeavors successful. However, community detection reduces the search space for identifying influence spreaders during the information diffusion process. The over all hierarchy of different categories of Community Detection is shown in Fig. 1. In the first category is node-centric community detection. Nodes in this category satisfy many features such as complete mutuality (cliques), members' reachability (k-clique, k-clan, k-club), nodal degrees (k-plex, k-core), and the relative frequency of Within-Outside Ties (LS sets, Lambda sets). Traditional social network analysis frequently use these categories but this category has certain limitations:

- Too strict, but can be used as the core of a community
- Not scalable, commonly used in network analysis with small-size network
- Sometimes not consistent with property of largescale networks.

The next type of community detection is group-centric community detection, which takes into account all of the relationships inside a group. However, some nodes can tolerate minimal connection in this circumstances. Recursive Pruning is a common strategy employed in this category.

Network-Centric Community Detection form a group, it need to consider the connections of the nodes globally. This partition of the network into disjoint sets is done based on following methods.

- Groups based on Node Similarity
- Groups based on Latent Space Model
- Groups based on Block Model Approximation
- Groups based on Cut Minimization

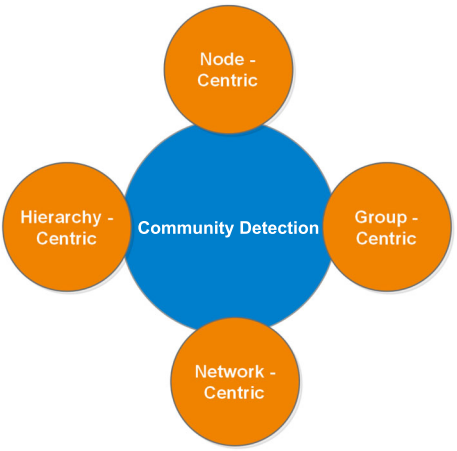


Fig. 1 Categories of Community Detection Methods

– Groups based on Modularity Maximization

Finally, the Hierarchy-Centric Community Detection build a hierarchical structure of communities based on network topology. To facilitate the analysis at different resolutions

– Representative Approaches
– Agglomerative Hierarchical Clustering

Atlast Communities in Heterogeneous Networks available in different form like Heterogeneous Network, Multi-Mode Network, Multi-Dimensional Network but Does Heterogeneity Matter?. In Social Media presents heterogeneity in networks so can we simply ignore the heterogeneity.

To observe from the above discussion about community detection category the paper classify its approaches in two main categoeies in OSNs namely network-based and content-based approaches. The above discussed all four categories are comes under network-based categories. In a network-based approach, node or edge centralities used to detect communities [6]. Next possible categoy is content-based method, which uses user-generated content is a decisive feature in community detection. This section examines the literature on the topics associated with community detection. Previous work on community detection can be roughly classified into two categories [40], as shown in Fig. 2.

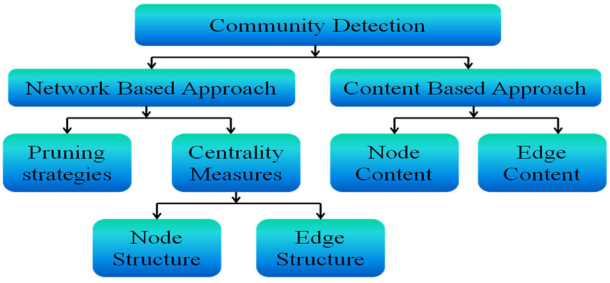


Fig. 2 Classifications of Community Detection Approaches

2.1 Network-based approaches

In a network-based approach, the incorporation of pruning strategies for community detection was proposed in [39]. Who used a clique-based algorithm to quickly identify the magnitude of community structures on large and sparse graphs with a comparable runtime. The algorithm and the new heuristic are well-suited for parallelization. They are also applicable for detecting overlapping communities in networks. Next, A bottom-up approach for community detection was proposed in [2] for discovering fine-grained communities to detect real network communities using a pruning strategy. Next, merging sub-communities using a hybrid approach, which combines both pairwise merging and multiple communities.

The Max-Flow Min-Cut Theorem on pruning strategy-based community detection was attempted by [43]. This method validates a different variety of datasets ranging from synthetic graphs to real-world benchmark datasets. The output produces an optimal set of local communities. Finally, intermediate local communities merged to form an original community structure using a hierarchical clustering algorithm called the quasi-clique merger, without considering node content.

A new framework designed for community detection, based on the pruning strategy with a timestamp feature, was proposed by [3]. In this method, identifying automatic relationships between the nodes in a community is performed by the Markov clustering algorithm [15]. Older information on the relationship between nodes is used to form a future pattern of community structure over a timestamp using historical and snapshot quality features [27]. Likewise, another timestamp feature-based community formation was proposed by [17]. In this way, the latent space model-based community detection approach with a dynamic Mixed Membership Stochastic Block (dMMSB) model, is introduced.

The next method in the Network-based approach is centrality-based community detection. It uses either node centrality or edge centrality features, or both. A novel approach was proposed by [52], to differentiate between a communityś internal links with the external links of connected communities, based on link weights. In this way, unweighted networks are converted into weighted networks by assigning link weights. Finally, community detection is made possible, based on weak and active links [51].

A novel approach was introduced under the centrality-based approach and is edge centrality-based community detection [22]. This method comprehensively investigates the role and characteristics of edge centrality for community detection that develops a method is called the Edge Antitriangle Centrality (EACH). The anti-triangle property computes the centrality score for all edges in the network iteratively until the value of the edges becomes zero. Afterward, it determines the community structure independently.

Edge centrality-based community detection plays a significant role in disseminating information diffusion in OSNs with a proper selection of an initial set of nodes for its spread. This initial set of nodes considered for influential nodes selection that is identified based on edge centrality (Salton centrality) in the network's topology [1]. From the resultant influential nodes, the community structure formed, and the method needs no prerequisite knowledge of the number of communities to form in a given network.

In the centrality technique, vertex degree-based community detection proposed by [54] for user privacy in OSNs. The vertex degree uses the structural diversity of anonymity for community detection. In this case, the k-Structural Diversity Anonymization (k-SDA) was developed to ensure the number of vertices and degrees remains the same for formulating different sets of communities in social networks. Finally, large-scale communities formed over complex social networks.

In the centrality approach, community detection is carried out adaptively to establish similarity in the network topology. Lv H et al. [34] proposed closeness centrality with signal transmission for community detection. This work selects a center node for community formation from the final rank of closeness centrality and similarity between nodes computed using the signal transmission. Finally, small sets of communities are combined to form a resultant community adaptively with an iterative update of community center nodes.

In centrality-based extended community detection in undirected graph to a directed graph with bivariate distributions addressed in [9]. This method includes node centrality, relative centrality, and modularity features for node selection in the community formation process. In general, undirected graph-based community detection chiefly uses algorithm: partitional, fast-unfolding, and agglomerative. However, the partitional method provides excellent results compared to the other two, owing to the network structure's directional feature. Centrality measures in community detection play a vital role in understanding the structural features of any given network. Wang X et al. [59] advanced a method that uses structural centrality to uncover overlapping communities on Twitter. Structural centrality is applied to obtain the center nodes for community formation in the network, with a weighted strategy and local search procedure provides promising results obtained at the end of the overlapping community detection.

Centrality-based community detection creates increased demand on social networking sites to ascertain their network property, given the enormous time spent therein to generate different large-sized datasets. Rani S and Mehrotra M [44] used a Label Propagation Algorithm (LPA) with influence centrality to reduce running time complexity in community detection. The LPA works fast in a graph-based algorithm with a semi-supervised learning strategy. The LPA performs poorly in certain conditions in the absence of influence centrality. With the result that there is either a massive-sized community or no single community to deal with at all. The hybrid method of influential centrality is used to improve LPA performance. Next, centrality-based community detection with a timestamp feature was proposed by [55]. This method uses an evolutionary multi-mode clustering over a multi-mode social network instead of a single-mode network. In the multi-mode network, node and edge structures are considered essential for community detection in a dynamic environment.

2.2 Content based approaches

The second category (content-based approaches), the community detection is based on the availability of content either in the node or edge, or sometimes both.

In the content-based approach, edge contents are used to compute intra-centrality and inter-centrality for community detection over the weighted network [33]. The intra-centrality computation deals with a possible number of shortest paths between a pair of nodes in the same community. Inter-centrality computation deals with a potential number of shortest paths between a pair of nodes in two different communities. In both cases, edge contents are considered a feature that identifies the shortest path between nodes. As a result, the final community structure is widely used to deploy data forwarding in Delay Tolerant Networks (DTN) and the worm containment strategy in OSNs.

Discovering dynamic community structures in online social networks has become a trend. Content-based community detection works on the premise that rich information in networks is present in node and edge contents, rather than its structure. Significant topological populations are detected only by considering the content features of nodes and edges [56]. This method introduced a new transformation from a content-based network to a

Node-Edge Interaction (NEI) for detecting dynamic communities through a NEIWalk (NEI network-based random Walk).

Content-based community detection plays a vital role in social networks. A wealth of information is generated from multiple sources through which secret relationships between people are discovered to help form topologically significant communities [47]. This method detects communities based on assumptions governing relationships in groups and informal interactions with the people in the groups concerned. The results show that a single community may accommodate more than one topic. A person may be present in more than one community. Community detection, based on node content, is used to detect trending topics in online social networks from streamed data [13]. This method collects region-wise streamed data, taking into consideration the cohesiveness factor across content. A content similarity computation is carried out on the collected contents, and communities are formed using the following methods: eigenvector, fast greedy, and Walktrap. Finally, the resultant communities are used to find trending topics from the network.

In the content-based approach, a unified framework is designed to combine topic modeling and community detection as an interlinked work. This method uses edge connectivity and edge content to identify the community structure over the network [30]. The Bayesian hierarchical approach identifies topic modeling (using the LDA) and community detection (using network edge content) to identify overlapping communities better. This approach is tested on research paper citation data from CiteSeer.

Link structures and their contents are merged for practical community detection approach is proposed by [60]. The problems addressed in this method are community membership and hidden content in a community. These two problems are resolved by merging link and content using a discriminative analysis. First, link analysis is carried out using a conditional model with hidden variables. It is to be noted that irrelevant content features impact the analysis of the content. Finally, optimized service is ensured by using bound optimization and alternating projection for merging links and content for community detection.

Currently, most online social networks automatically impose edge content while sharing images, videos, user tags, and comments. Edge contents provide better supervision for detecting communities at different levels of weights in the contents concerned [42]. Researchers [62] proposed a method to combine features such as nodes, edges, contents, and structures to detect communities over an extensive network. Vertex connectivity and neighborhood similarity play a significant role in detecting densely-connected communities from the network. The system calculates the unified distance measure for the entire network. The resultant distances are updated automatically over time by adopting an automatic learning strategy to extract structural and attribute similarities.

Considering network structure for community detection could significantly impact community detection results, because two users with a deep relationship may not necessarily have strong network ties [57]. This method suggests that considering user content along with network structure may maximize effective community detection. Liu L et al. [28] also proposed the same approach for community detection by considering both node structures and content. However, node contents are collected dynamically based on the consequences of the interaction among the nodes. These interactions are modeled using content propagation and the random walk. Here, content propagation is modeled with relative influence, and the random walk modeled directly. Finally, this periodically-updated content is merged with node structures to form a community.

Community detection based on a merger of structure and content features does not always perform well, on account of a content mismatch with the topological structure. This

issue is resolved by introducing a generative probabilistic model derived from the nested expectation-maximization algorithm [14]. This model forms two generalized community structures for network structure and content separately. The correlation parameters perform well to interpret the community structure, based on the topic over a synthetic benchmark dataset.

Another pattern of user content-based community detection has evolved from user searches in social networks, which are inherently asymmetric and vary in strength [20]. Studies in the context of a global social search have been conducted. All the network information available to the search algorithm is used to make assumptions. Based on the assumptions made, a set of users with similar search patterns is selected to form a group for community detection.

The literature reviewed above makes it understandable that the following issues with some limitations are tabulated in Table 1:

- In pruning based community detection methods fails to consider node and edge centrality measures and content as features for community detection, leading to issues with overlapping communities
- The result of pruning based community detection could result in an inefficient selection of influence nodes for the information diffusion process
- The pruning based community detection methods are not scalable enough to handle large-scale networks
- The techniques mentioned in centrality based community detection fail to incorporate user-generated content features for community detection. As a result, the topic-wise influence node selection is not possible to identify the active node for the topic from the network's community structure
- The network-based community detection method does not focus on user tweets, suggesting that users' current opinions are not taken into consideration
- The content-based community detection method fails to notice the physical connection between two users in the network
- The other community detection methods fall short of addressing the problem of overlapping communities

Hence, a community detection system is required that identifies overlapping communities with a combination of network and content features, along with topic sensitivity level (positive, negative, neutral) in it. Such a community detection system will select efficient influence spreaders to predict information diffusion on Twitter.

3 Problem statements

The previous techniques examined above have several unresolved problems in building an efficient community detection system from online social networks, especially on Twitter. These problems produce unsatisfactory results in terms of the resultant community structure in real-world applications. A few issues identified in designing an efficient system for community detection in OSNs are as follows:

Problem 1: How to design a topic polarity-based community detection system in Twitter Network?

Table 1 Comparison of other community detection techniques and limitations

Authors	Algorithms	Approaches	Limitations
Bharath et al.	Clique-based Algorithm	Pruning Strategy	Fail to consider the user-generated contents for overlapping community detection
Xingqin et al.	Quasi-clique Merger Algorithm	Pruning Strategy	Its applicable for community detection with a limited network structure that too from the absences of a node or edge contents
Wenjie et al.	Mixed Membership Algorithm	Pruning Strategy	This method is not scalable enough to handle large-scale networks
Songwei et al.	Edge Antitriangle Centrality	Centrality-based Approach	This technique fail to in-corporate features that are generated from user content for community detection.
Ahajjam et al.	Salton centrality	Centrality-based Approach	The topic-wise influential nodes are not selected in this approach for possible identification of active node for the topic in any given network
Seemarani et al.	Label Propagation Algorithm	Centrality-based Approach	This method performs poorly in certain conditions in the absence of influence centrality. With the result that there is either a massive-sized community or no single community to deal with at all.
Zongqing et al.	Intra-centrality and Inter-centrality Algorithm	Centrality-based Approach	For unweighted network this approach not suited. because the edge weight is the core feature for this community detection
Paramita et al.	Fast greedy Algorithm	Content based Approach	A single community with more than one topic is present in their results. The cohesiveness factor across content are not determined in order to identify overlapping communities in it.
Yu-Ru et al.	Bayesian hierarchical	Content based Approach	Uses edge connectivity and edge content to identify the community structure over the network

– The literature discussed so far on community detection is either based on node and edge structures or content or sometimes both. No algorithm has so far worked on topic adaptive polarity-based community detection on Twitter.
– Presently available algorithms for community detection on networks are topology-oriented approaches such as the eigenvector, fast greedy, and Walktrap.

Problem 2: How to identify overlapping communities in Twitter Network?

– An array of algorithms detects communities with different parameters, such as nodes and edges. However, a few address overlapping communities at the bare minimum, but fail to use the topic adaptive feature and polarity to detect overlapping communities in highly dense social networks accurately.

Problem 3: How to design a community detection system with a combination of network and content features form dynamic social network like Twitter?

– Other methods work well over the benchmark and real-world datasets (in a static network); however, they give high false alarm rates when applied to a dynamic social network with combined network and content features, because of the network structure and content produced by the network's users change over time.

Significant contributions of a proposed Robust Topic Feature-based Sentiment Classification in Community Detection method include the following:

– Improved community detection with a high Rand index value, ensuring a combination of network and content features with topic sensitivity.
– Easy identification of overlapping community structures resulting from topic sensitivity level.

4 Proposed work

The proposed technique detects community structures on Twitter based on topic adaptive sentiment sensitivity(polarity) level features in tweet content, that helps to detect overlapping communities between two topics or the same topics with different sensitivity (positive or negative or neutral). Overlapping community detection results in better identification of influence spreader according to the topic to maximizes the diffusion rate of information. The proposed TASCbCD technique comprises the following processes: Twitter data collection and preprocessing, Topic modeling, Sentiment analysis, Network graph construction, and Community detection.

In a general scenario, the tweet's text is not directly processable. A set of preprocessing steps is required for the proper implementation of sentiment classifications. In this work, each tweet is first given to topic modeling for identifying the related topic of the tweet; afterward, the appropriate sentiment class label is identified based on the various extracted features like text, non-text, user content, tweet topic-adaptive keywords, and network parameters. Along with that above process, the user relationship network structure is also constructed parallelly for constructing an underlying base network for community formation.

In this paper, communities are formed on Twitter based on the topology and topic sensitivity(polarity) features. Distinct communities are formed for each topic separately in line with their sensitivity(polarity) nature. Finally, community structures identify overlapping

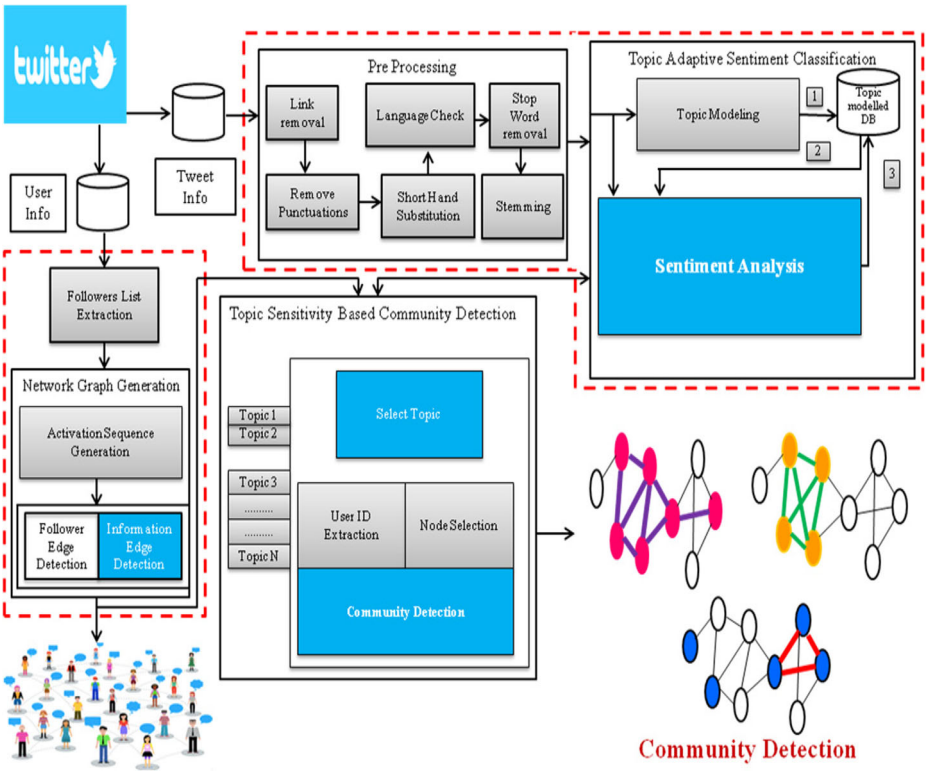


Fig. 3 A Proposed TASCbCD System Architecture

communities. The results are used to determine and rank influence spreaders for the particular topic with sensitivity for the different process shown in Fig. 3. Finally,the selected 'k' users are used to initiate information diffusion in online social networks. For example, the set of the user is who discussed or tweeted about the topic "Politics" positively for the particular period "t" they may or may not negatively discuss the topic in time "t". So, according to the type of diffusion needed for some applications for the topic with sensitivity, the proposed system helps select the initial set of influence users after the successful filtering of topic sensitive users and form the community on Twitter.

4.1 Preprocessing of tweets

The tweets' content is generally noisy and made up of mostly ill-formed words since users only have limited characters to work with. Therefore, cleaning and preparing meaningful data is essential for any analysis to be carried out. Tweet pre-processing involves the following tasks: link removal, shorthand substitution, language check, stop word removal, and stemming. Figure 3 shows the detailed steps involved in pre-processing one after another.

Tweet texts often contain links to related material on the internet. Additional information is tagged in the links and refers to videos, images, or web pages. Link removal involves removing words starting with "https.". Most tweets on Twitter use shorthand and acronyms because of the limitations of the number of characters used. So then, there is a need for

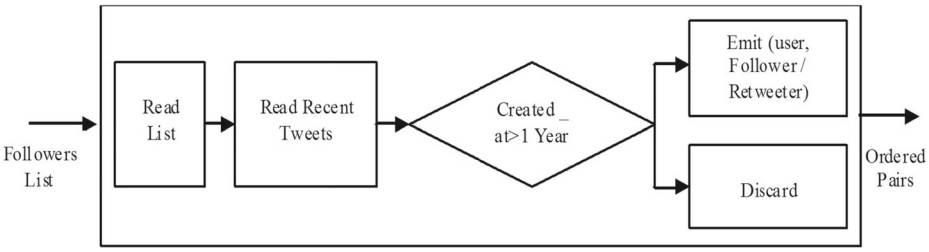


Fig. 4 Activation Sequence Generations

shorthand substitution, wherein a set of frequently occurring words is compiled from the internet and substituted with full, unabridged text. Only tweets in English are analyzed in this work, despite the availability of tweet texts in English that are transliterated from other languages. As a result, a tree is built with the depth-first search from an English dictionary to establish if a corresponding text is or is not an English word.

Stop words are articles and prepositions common in the English language, though not useful as processing information. A set of such words, obtained from GitHub with their references, is removed from the tweet text. Finally, stemming is applied in the preprocessing stage. The result is that the root word of any given word is identified. This identification is essential while processing words of all forms considered together. The famous Porter stemmer algorithm is used to reduce the number of words in 6 stages.

4.2 Network graph generation

4.2.1 Activation sequence generation

Building an activation sequence is essential for modeling the possible paths of information flow. The activation sequence determines the path along which information has traveled in the past. In general, the activation sequence is a subset of the edges in the network. However, this condition is relaxed here to make allowances for Twitter's retweeting facility, where a relationship between two users is no prerequisite for information to pass between them. The steps involved in the activation sequence generation are shown in Fig. 4.

Generating the activation sequence is relatively simple. The list of followers or retweeters already extracted from Twitter is read, checked, and the latest tweet of every user fetched. If the date goes back over a year, the account is assumed to be inactive and is discarded. For users who have been active in the last year, tweets predating this period are not collected. This happens because it is normal for a person's preferences to change. In case they have not, tweets from the period will be able to account for them as well. The user's profile biography may fill in any other gaps. A triple of the format (user A, relationship, user B) is used for all other users. To determine the kind of link between two users, two types of predicates are used. They are "knows" and "knows well." "Knows well" indicates that one user is the follower of the other, and "knows" indicates that there has been a flow of information between these two users. Both predicates are defined under the schema FOAF (Friend of a Friend) and used in user pages or articles about people. Both kinds of edges can exist between the same pair of users.

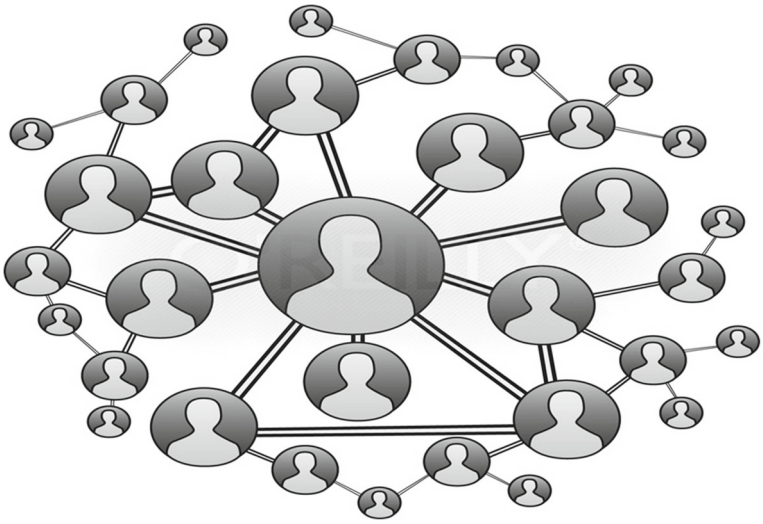


Fig. 5 Social Network Graph (https://goo.gl/images/eWaV2T)

4.2.2 Information Edge and Follower Edge Detection

The social network structure's best representation is a graph. Figure 5 shows the typical social network graph. This graph is generated from the activation sequence, which is added to the network's topology, and the few extra edges that exist to indicate the data flow minus implicit relationships. Each user is a node, and for every ordered pair of users, an edge exists for the corresponding nodes. The graph is not stored in a data structure but a graph database.

In addition to creating the graph database using the network structure available, a second essential task is carried out. Edges connecting users are differentiated into information edges and follower edges. Follower edges are generated as a product of the network structure when one user follows another. Information edges, on the other hand, which are established when one user retweets another. It is essential to take this into account such retweets on Twitter, as a user does not have to follow another user to view their tweets, subject to the condition that the user's profile has public settings. It is possible for an information edge and a follower edge to existing between two nodes. An information edge generally has a lower weight than a follower edge. The information edge is asserted mainly when the topic in question is also the retweeted tweet. The dash in Fig. 6 establishes the relationship, "knows" between users A and B, showing that there has been at least one instance where user A retweeted something originally tweeted by user B. The dotted line indicates the relationship "knows well" between users A and B, showing that user A follows user B.

Fig. 6 Information Edge Vs Follower Edge

4.3 topic adaptive sentiment classification (TASC)

Sentiment classification differs from one topic to another according to the sentiment words of each topic. For Example, the word extraordinary" gives positive sentiment in the field of "Sports," but it harms the topic or domain like "Medical" or "Mechanical." On Twitter, given that texts are both diverse and sparse, it is impossible to train a unified classifier for the entire topic. The sentiment classifier invariably turns in poor performance. To handle this situation, Topic Adaptive Sentiment Classification(TASC) is proposed for better sentiment analysis of tweets according to different topics. Before sentiment classification, the preprocessed tweets are topic-modeled using the well-known topic modeling algorithm i.e., Latent Dirichlet Allocation (LDA). The architecture of the TASC is shown in Fig. 7.

The topic modeling algorithm assumes that each sentence is created based on three generative processes: the number of words identified, a mix of the topics selected, and the number of words generated based on the topics. In this case, the pre-assumptions of no-of topics are 20 (the no.of initial topic selection depends on the researcher or author). During the learning phase of topic modeling, a topic is assigned randomly to each of the words in preprocessed tweets. This value is then iteratively refined several times until the assignments are relatively steady. The posterior probabilities were calculated to execute topic modeling results. Approximation and sampling are two conventional approaches for calculating the posterior probability in LDA. The Markov chain Monte Carlo (MCMC) falls undersampling techniques, while the variational Bayes are approximation techniques. Variational Bayes is found to be much faster than MCMC sampling and is just as accurate. Even so, it is computationally challenging to apply this technique to large datasets.

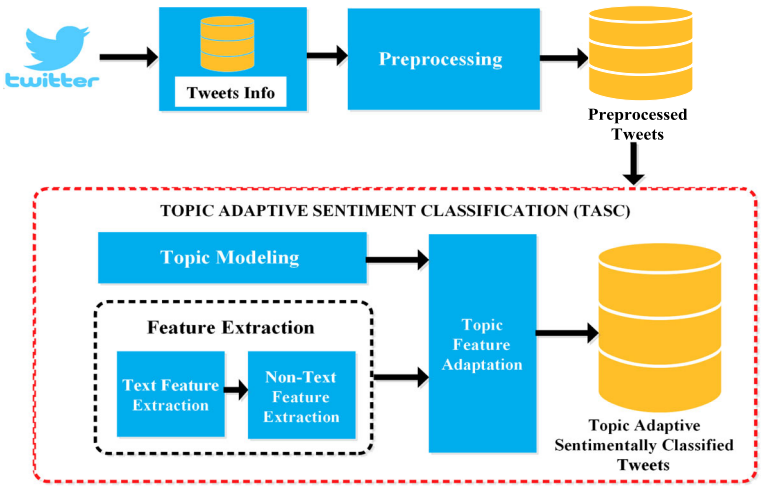


Fig. 7 Block Diagram of Topic Adaptive Sentiment Classification (TASC)

Algorithm 1 Topic adaptive sentiment classification (TASC).

Step1 : **for** each tweet check the relative topic T **do**

Step2 : **initialize** (Text Feature Extraction)

Step3 : **initialize** (Non-text Feature Extraction)

Step4 : **initialize** λ at random

Step5 : **initialize** $\gamma = 1$

Step6 : **for** each tweet **do**

Step7 : **while** $\gamma < .00001$ **do**

Step8 : $\varphi \propto (Eq[\log\theta] + Eq[\log\beta])$

Step9 : $\gamma = \alpha + \sum \varphi n$

Step10 : **end while**

Step11 : $\lambda = (1 - \rho) + \rho\tilde{\lambda}$

Step12 : **end for**

Step13 : Select Twitter Words ()

Step14 : **if** (textFeature U Non-Text Feature **U** Topic Feature)

Step15 : $calculateScore(t, nt, to > 0.0001)$

Step16 : **end if**

Step17 : UpdateTextFeature ()

Step18 : UpdateNonTextFeature ()

Step19 : **end for**

Step20 : **Train** multiclass SVM C* on features 't' and 'nt' using augmented L

Step21 : **return** L,t,nt and C*

A feature vector is constructed from the results obtained from topic modeling, based on the text features and the non-text features. In the text feature, topic-based sentiment words

and common sentiment words are considered for sentiment classification. The topic-based sentiment words and common sentiment words are collected from the standard sentiment dictionary like sentiWordNet and wordNet. Before that, the given input tweet undergoes POS (Part-Of-Speech) tagging, from which a standard set of adjectives, verbs, nouns, and adverbs is extracted as candidates for topic-adaptive sentiment classification.

Non-text features include @ network, and emotion and punctuation are extracted to capture tweets' unique nature. On Twitter, the @ symbol in a tweet refers to a user a person would tweet to, for instance, @abc, which means that the tweet refers to user ABC. This useful symbol feature can be extracted to eliminating the need to label the entire dataset. "@" is a widely used pattern in a tweet, e.g., user Ni mentions Nj via a tweet 't,' containing the symbol "$@Nj$". The '@' symbol reflects the relationship between nodes based on the same sentiments as tweet 't.' Thus, the proposed TASCbCD system uses the '@' network to build a graph model with tweets as nodes and '@' representing straight edges between nodes.

An essential feature of the topic modeling is the assumption of an infinite vocabulary. In a conventional system, the vocabulary is fixed, and the model generated remains static. However, in the case of a dynamic system, a vocabulary constraint is unreasonable. So, when a new word is observed, it is added to the model, and the appropriate statistical values are updated. Now, based on the results of topic modeling, text features, and non-text features, the topic-based sentiment class label is identified for individual tweets. In the TASC algorithm (Algorithm 1), 'T' represents individual tweets, 'λ' the topics, 'γ' the constant value, 'β' the posterior distribution of topics, 'θ' the topic proportions, 'φ' the variational Parameters, 'α' the hyperparameter, 't' the text features, 'nt' the nontext features, 'to' the topic features, '$C*$' the classifier and 'L' is a set of labels. In this text and non-text features are first initialized. Each text then undergoes sentiment classification, based on the three sets of features used in our system. Finally, the Support Vector Machine (SVM) finalizes the sentiment label for the tweets in question.

4.4 Topic adaptive sentiment classification based community detection (TASCbCD)

Community detection is essential to identify groups of people with similar interests. In this context, user input is a combination of a user query and its sensitive nature(polarity). Community detection is essential to identify groups of people with similar interests and sensitivity about some topics to gauge the social influence for information diffusion on Twitter. The TASCbCD algorithm receives a user input, checks it for the topic, examines the previous user history on the particular topic with sensitivity, and treats its tweet polarity positively or negatively. According to that sentiment analysis results, the corresponding user ID's of those tweets are extracted and figure it out in the previously constructed base twitter network graph. In the proposed system, the node selections and community formations are varied based on the topic selected, with its sensitivity(polarity).

Once the user query is received in the form of the topic and its sensitivity, the relevant user IDs are extracted from the resultant database of sentiment classification based on likes, interest in the topic, tweets, retweets, replies, mentions, and locations. Firstly, in this flow, likes for a particular topic are taken into consideration and corresponding to that the user IDs extracted. Secondly, user tweets on the topics in question are extracted. Figure 8 depicts the flow of community detection based on TASC.

Once the user IDs are extracted, they are stored and used to construct an initial graph structure from the original graph database, which combines information and follower edges. In the initial graph construction, all the adjacent nodes of the selected users are taken into

account. In this process, the set of initial vertices comprises selected user IDs. The set of edges is made up of nodes adjacent to the users. Using those vertices and edges, the initial graph structure is constructed for community formation. The proposed TASCbCD technique consists of the following processes: Twitter data collection and preprocessing, topic modeling, sentiment analysis, network graph construction, and community detection. A step-by-step procedure for the proposed TASCbCD method is shown in the sections above. The overall steps for the proposed TASCbCD method are illustrated in Algorithm 2.

Algorithm 2 TASC based community detection (Tw,G).

Input : – T_w (Tweet) , G - Network Graph

Output : – Community Structure C1, C2,..., Cn (Topic Sensitive)

Step 1: P= Preprocess (Tw) // **Tweet Preprocessing**

Step 2: T= LDA (P) // **LDA for topic modelling**

Step 3: S = senti (T, P) // **Topic Adaptive Sentiment Classification**

Step 4: $User\_id\_Extract(T, S, Tw)$ // **Extract user Id from sentiment database**

Step 5: if $Topic == T \&\& sentiment == S$ **do**

Step 6: **return** Users// **Set of user with same topic sensitivity**

Step 7: end if

Step 8: $G = Networkofuser G(V, E)$

Step 9: $Community_Detection(User, G)$ // **Detect community in network graph**

Step 10: for $i = 1 : n$ **do**

Step11: $detect = edge(G, User)$ // **Detect User connection in G network**

Step 12: end for

Step 13: return detected communities C1,C2,...,Cn (Topic Sensitivity)

5 Results and performance analysis

5.1 Experimental setup

This proposed technique has applied to the dataset collected manually through R Twitter API, with the total number of tweets in the dataset is 1.1 million. These tweets are collected at various periods in a 24/7 basis (January - October 2015, June - September 2016, and

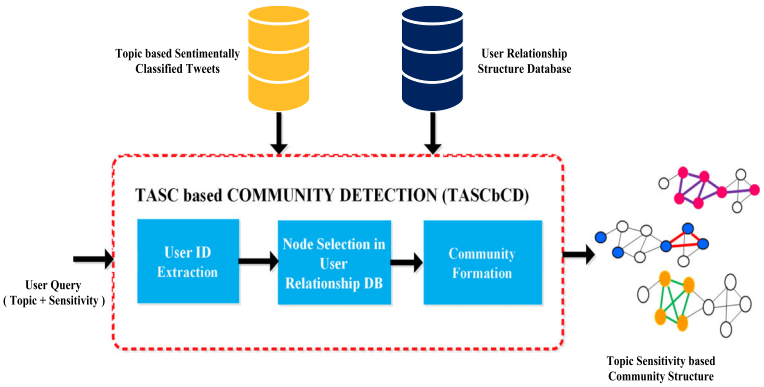


Fig. 8 TASC based Community Detection(TASCbCD)

January - March 2017), with different size of the user is in the networks, namely Network 1 with 6,000 users, Network 2 with 8000 users, and Network 3 with 12,000 users. The TASCbCD algorithm is implemented and the results compared with other methods to determine the accuracy of the proposed system for sentiment classification with different networks, different users, and the same topic. The collected dataset tweets has 16 features in each tweet which plays major impact on influential user selection strategy approach . The features of each tweet are tweet text, favorite, favorite count, reply to SN, created date, truncated, reply to SID, ID of the user, reply to UID, screen name, status source, retweets count, is retweeted, longitude and latitude. The use and explanation of each features are discussed below.

- Tweet ID - Identifies a particular tweet
- User ID - Identifies a user creating a tweet
- Tweet text - The content of a tweet
- Favorite - Indicates a well-liked or popular tweet
- Favorite count - No.of people who like a tweet
- Retweets - A forwarded tweet
- Replies - An answer to another user's tweet
- Mentions - Tweets that specify the topic anywhere in their contents
- Created date - Date on which a tweet is created
- ID of the user - The identification number of a user replied replied to
- Screen name - Name of a user in a tweet
- Status source - Hyperlink of a tweet
- Retweets count - No.of times a tweet is retweeted count
- Is re-tweeted - Checks whether or not a tweet is re-tweeted
- Longitude - Longitude of the user's location while creating a tweet
- Latitude - Latitude of the user's location while creating a tweet

This paper intends to handle the Twitter network in terms of both network structure and user tweets. Data collection was done via the R Twitter API. The collected anonymous data was preprocessed and stored in a database on MongoDB. This work is entirely coded in Python, running on IDLE v3.8 and v2.7 with a cluster of up to four machines. The cluster is set up using the Hadoop and MapReduce implementations, supported by the Python

package MRJob. External packages Tweepy, Twitter, Franz, PyQtGraph, Qt4, and NumPy provide the necessary support. AllegroGraph is used to store the network structure. The web interface, AGWebView, Linux application, and Gruff, are used to visualize intermediate stages in the graph. The relevant files are stored in the local machine to ensure that processing is still possible when disconnected from the AllegroGraph server. PyQt Graph is the package used to draw the graphs in the final stage of the system. This package depends on Qt4, which is mostly used to develop the GUI and NumPy for some serious crunching.

5.2 Performance analysis

For sentiment classification, the accuracy of the system is used to justify the appropriateness of the Support Vector Machine classifier and for community detection, the Rand index and Scaled Density metrics are used to evaluate performance.

Definition 1 The performance measure of tweet sentiment classification is justified by using the results of True Positives (TP), True Negatives(TN), False Positives (FP), and False Negatives(FN), as represented in (1).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

Definition 2 Rand index is a measure used to identify similarities between two communities. For a given pair, there is agreement when both nodes belong to the same community or different communities, for both community structures as represented in (2).

$$RandIndex = \frac{a + b}{a + b + c + d} = \frac{a + b}{n \mid 2} \tag{2}$$

$S \rightarrow V_1, V_2, \cdots V_n$ **//set of vertices V**
$X \rightarrow X_1, X_2, \cdots X_r$ **//a partition of S into r subsets**
$Y \rightarrow Y_1, Y_2, \cdots Y_s$ **//a partition of S into s subsets**
$a \rightarrow$ the number of pairs of elements in S that are in the same set in X and the same set in Y
$b \rightarrow$ the number of pairs of elements in S that are in different sets in X and different sets in Y
$c \rightarrow$ the number of pairs of elements in S that are in the same set in X and different sets in Y
$d \rightarrow$ the no.of pairs of elements in S that are in different sets in X and same set in Y

Definition 3 Scaled Density for a community 'C' is defined as the ratio of links that contains with nodes in the network, is mentioned in (3).

$$\rho'(C) = \rho(C)nc = \frac{2mc}{(n-1)} \tag{3}$$

$\rho \rightarrow$ Scaled Density
$nc \rightarrow$ number of nodes in community
$mc \rightarrow \frac{nc(nc-1)}{2}$ // for a CLIQUE

Table 2 Sentiment classification result of TASC algorithm on network 1

S.No	Tweet	Topic	UserName	Sentiment
1	YakubMemon burried im Mumbai amid tight security	Yakub Memon	Deepak Balamurali	Neutral
2	Worth reading and sharing... An open letter by a cop to those opposing death penalty to Yakub: via?	Yakub Memon	Rev. Jijo Varghese	Positive
3	Yakub Memon first in 31 years executed in Nagpur jail	Yakub Memon	Adarsh Mahalley	Neutral
4	By announcing the execution date of Yakub Govt has exposed the true Anti- National & Terrorist Sympathetic face of the s?	Yakub Memon	Lab Rat	Negative
5	I repeat not a single political party (other than BJP) has made an unequivocal statement supporting SC decision to hang terro?	Yakub Memon	TARUN SONI	Negative
6	Some Politicians are following Yakub Memon Bcos following Dr Kalam's Ideology is difficult.?	Yakub Memon	Chandrakant Nilewad	Negative
7	Full Update On Yakub Memon Case..	Yakub Memon	Filmy Hungama	Neutral

5.2.1 Experimental results

First, the TASC algorithm is implemented. The results compared with other methods to determine the accuracy of the proposed TASC system for sentiment classification with different networks, different users, and the same topic, implemented with results in Tables 2, 3, and 4. Here, the accuracy of the proposed TASC is compared with other classifiers such as AdaBoost, decision tree, Naïve Bayes, and random forest. Based on the number of True Positive, True Negative, False Positive, and False Negative, the classifier's accuracy is determined. The results of accuracy for different networks, as well as Network3 about other methods, are illustrated in Tables 5 and 6.

Figure 9 shows that the TASC with SVM outperforms other classifier models understudy and achieves an accuracy rate of 80% on average. This paper's primary focus is on topic adaptive sentiment classification based community detection, rather than traditional network measure-based community formation. The implementation is tested with all three networks, and the results showcase topics with sensitivity(polarity)-based community pairs. Here, the results consider both node data and the physical structure connecting nodes. From the topic modeling results, twenty different topics are identified with their sensitivity, based on their term frequency from a higher to a lower order. A few of the topics are listed here: sports, food, movies, TV shows, politics, education, books, Yakub Memon, Companies, and music. Each topic is passed over the community formation system to identify the nodes and their interest in actively discussing the topics with some sensitivity, such as positive or negative or neutral. Node selection is an interesting process that is made easy since each node's data in the network is preprocessed and clustered already based on the topic with its sensitivity.

The given topic is identified with its cluster, nodes belonging to the same self cluster, and marked in the overall network community structure. Out of 20 topics selected for this proposed system implementation of 3 different sensitivity levels, some example community

Table 3 Sentiment classification result of TASC algorithm on network 2

S.No	Tweet	Topic	UserName	Sentiment
1	media has made a spectacle of this Death Penalty! On a sad day instead of Focus being on A Yakub Memon Never again all Gloating	Yakub Memon	Dr. Fahad Samadi	Negative
2	APJ Abdul Kalam and Yakub Memon two people of the same religion is to be buried today. Choice is yours how to live your life. # YakubHanged	Yakub Memon	Dashmeet Singh	Positive
3	Worth reading and sharing ... An open letter by a cop to those opposing death penalty to Yakub: via?	Yakub Memon	Er. Kuldeep Kumar	Positive
4	Congress criticizes FM for his remarks that Congress is making irresponsible comments about execution of Ya?	Yakub Memon	Kailashsadani	Negative
5	42% Viewers on AajTak agreed W Salman That Yakub Memon Shud nt hang. Now VHP BJP SS MNS all Pigs come together send 42% In?	Yakub Memon	Salman's Rose	Negative
6	Reasons why Sonia- Rahul does not stop Digvijays Iyers to take pro-Yakub line!	Yakub Memon	Shashi Ranjan	Neutral
7	Yakub Abdul Razak Memon shouldn?t have been hanged — The Indian Express	Yakub Memon	Nauman's Butt	Negative
8	# YakubMemon burried in Mumbai amid tight security	Yakub Memon	Deepak Balamurali	Neutral

structures of 6 different topics -books, food, movies, music, Yakub Memon, and TV shows - with their sensitivity are illustrated in Table 7. Individual community structures are shown in Figs. 18, 19, 20, 21, 22, and 23 attached in Appendix.

This paper is also focused on identifying the structure of overlapping communities based on the TASCbCD algorithm. In general, detecting overlapping communities is a trivial task given the complex structure of Twitter's network. However, the proposed method quickly identifies overlapping communities using TASCbCD. The results above depict different sets of topic sensitivity based community structures using the proposed method. Overlapping communities are directly detected, and the results are shown in Fig. 10.

The evaluation of the proposed system's resultant community over different topics is analyzed using the Rand index and Scaled density metrics. The results show that the proposed TASCbCD method has effectively identified the topic-sensitive community and overlapping communities. The identification helps to construct the initial graph structure for influence spreader identification according to the given topic. Here, Table 8 illustrates the Rand index and Scaled density of nine resultant communities over different example topics with their sensitivity. The Rand index analysis for Table 8 is shown in Fig. 11 and Scaled density in Fig. 12. Here the sensitivity(polarity) of the topic is represented by the positive (+), negative (-), and neutral (N).

Table 4 Sentiment classification result of TASC algorithm on network 3

S.No	Tweet	Topic	UserName	Sentiment
1	Hanging #AfzalGuru was for Congress. Is hanging #Yakub is for the BJP : A vote bait for an intellectually challenged Indian	Yakub Memon	Raees	Negative
2	"media has made a spectacle of this Death Penalty! On a sad day instead of Focus being on "A Yakub Memon Never Again" "all Gloating"	Yakub Memon	Dr. Fahad Samadi	Negative
3	Congress critizes FM for his remarks that Congress is making making irresponsible comments about execution of Ya?	Yakub Memon	kailashsadani	Negative
4	42% Viewers on AajTak agreed W Salman That Yakub Memon Shud nt hang. Now VHP BJP SS MNS all Pigs come together and send 42% in?	Yakub Memon	Salman's Rose	Negative
5	Reasons why Sonia-Rahul does not stop Digvijays-Iyers to take pro-Yakub line!	Yakub Memon	Shashi Ranjan	Neutral
6	I don't know if Tiger Memon wiil b caught ever! But he will die everyday thinking of Yakub n that slow death is his punishment?	Yakub Memon	Mayank Pandey - IMJ	Negative
7	"Those who say" Terrorism has no "religion" should check the amount of crowd in Janaja of Yakub. Why do we..."	Yakub Memon	Amit Sharma	Negative
8	#Yakub issue is polarising. But CJI asking 3 SC judges to sit at 2 AM to give a condemned terrorist fair hearing is somet?	Yakub Memon	Tushar Joshi	Negative
9	The crowd built up for Yakub memon is because of the Govt. Why drag case for 22 Y?	Yakub Memon	STOP POSCO	Negative
10	its quite ironic as u were crying for Yakub Memon	Yakub Memon	Prafulla C Tiwari	Negative

It is observed from Fig. 11 that the maximum value of the Rand index lies between 0.9 and 1.0, showing that an average accuracy of 93% is achieved by the proposed TASCbCD system in predicting community detection.

From Fig. 12, it is observed that the edge count predicted for community detection is almost on par with the actual community in existence, demonstrating that the proposed TASCbCD method provides promising results in identifying community structures. The same experiment is executed with different networks and different topics over different periods. The performance of the proposed TASCbCD method is evaluated using the same Rand index and Scaled density. Table 9 illustrates community detection performance over different Twitter networks with different data collection times on three sample topics: sports,

Table 5 Compare the sentiment classification accuracy of proposed TASC algorithm with other networks in twitter

Network #	Accuracy
Network1	0.79
Network2	0.77
Network3	0.80

Table 6 Classifier accuracy of proposed TASC method with other methods

Classifier model	Accuracy
Ada-Boost	0.676
Decision Tree	0.710
Naive Bayes	0.713
Random Forest	0.722
TASC with SVM Classifier	0.800

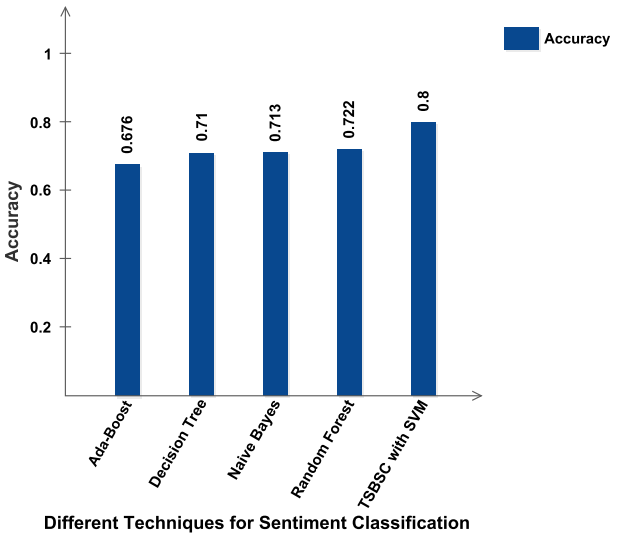


Fig. 9 Classifier Accuracy using SVM in proposed TASC

Table 7 Community detection of 6 example topics with its sensitivity on network 3

Topic (Tweets period - June to Sep 2016)	Sensitivity
Books	Positive
Food	Positive
Movies	Positive
Music	Positive
Yakub Memon	Negative
TV show	Negative

Fig. 10 Result of Overlapping Community Detection using Proposed TASCbCD Algorithm

Table 8 Rand index, scaled density – twitter network2 (July-Sep 2016)

Community (Topic-Sensitivity)	TASCbCD (Rand Index)	Network Based Approach (Rand Index)	Content Based Approach (Rand Index)
Sports (+)	0.9479166667	0.7916824789	0.8265847481
Food (+)	0.9793281654	0.6854688748	0.8158796421
Movies (N)	0.9392405063	0.71487845826	0.7154897452
TV show (N)	0.9062500000	0.79935892325	0.8325487691
Politics (N)	0.9896193772	0.71566874458	0.7758791254
Yakub Memon (-)	0.9077669903	0.74975645667	0.8425481596
Books (+)	0.9281914894	0.79745726487	0.8265847152
Company (+)	0.9891067538	0.74548727845	0.7785781568
Music (-)	0.9192825112	0.73754578127	0.8112548795

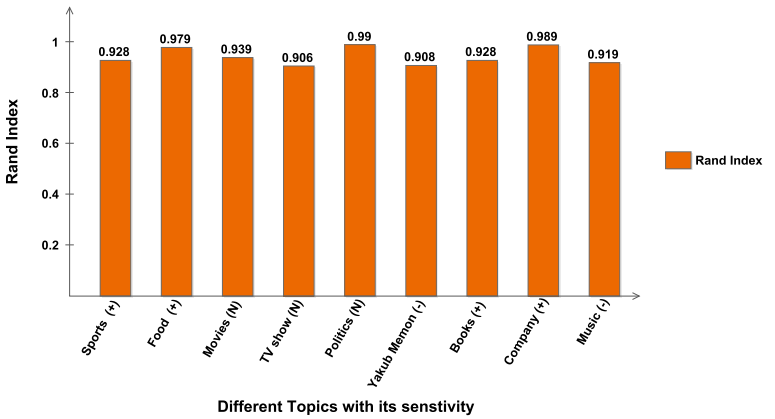


Fig. 11 Rand Index for Network 2 over (July-Sep 2016) collected data

food, and movies. The results of the graphical representation of the Rand index and Scaled density are shown in Figs. 13 and 14. Here, sensitivity is represented by the positive (+), negative (-), and neutral (N).

Figure 13 shows a comparison of the Rand index across different topics and is implemented with different networks. The results are tabulated in Table. The Rand index value for different topics with their sensitivity over different networks lies between 0.85 and 0.90. The Rand index of the proposed TASCbCD system achieves an average of 87% accuracy in predicting community detection over different networks with data collected over different periods. Figure 14 , derived from Table 9, shows that the proposed method is scaled density is quite close to the scaled density of links already available over different networks. The x-axis in Fig. 14 represents data from different networks with topics and sensitivity, as shown in column 1 of Table 9.

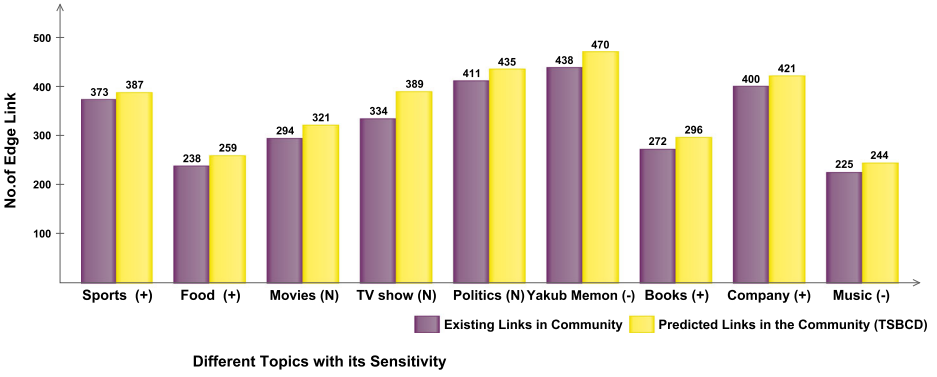


Fig. 12 Scaled Density of Network 2 over (July-Sep 2016) collected data

Table 9 Rand index, scaled density – across different twitter networks (different periods of data)

	Different Twitter Networks with various periods	Community (Topic-Sensitivity)	TASCbCD (Rand Index)	Network Based Approach (Rand Index)	Content Based Approach (Rand Index)
Network1	(Jan-Oct,2015)	Sports (+)	0.9276	0.7956	0.8495
	(June-Sep,2016)	Sports (+)	0.9588	0.7354	0.8425
	(Jan-Mar, 2017)	Sports (+)	0.8986	0.7474	0.8867
Network2	(Jan-Oct,2015)	Food (+)	0.9691	0.7297	0.8185
	(June-Sep,2016)	Food (+)	0.9793	0.6854	0.8158
	(Jan-Mar,2017)	Food (+)	0.9287	0.7365	0.8185
Network3	(Jan-Oct,2015)	Movies (N)	0.9234	0.7848	0.7856
	(June-Sep,2016)	Movies (N)	0.9191	0.7245	0.8382
	(Jan-Mar,2017)	Movies (N)	0.8975	0.7654	0.7912

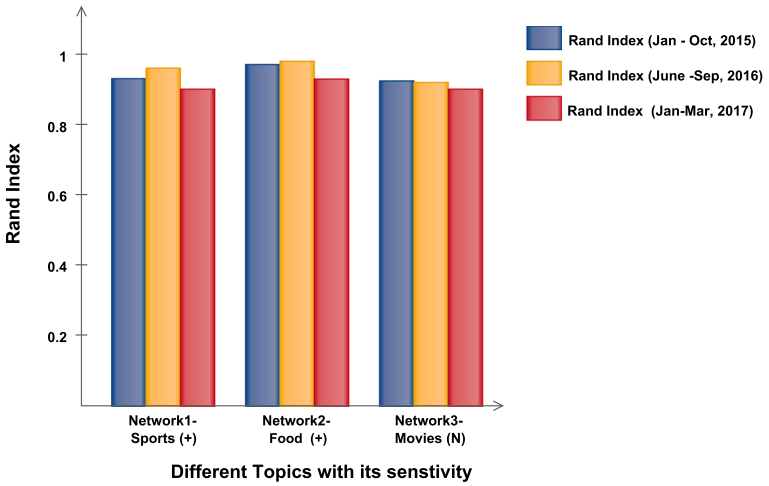


Fig. 13 Rand Index across Different set of Twitter Network

5.3 Comparison with other community formation method

To ascertain the average performance of the community detection system, the results of the proposed TASCbCD are compared against two other existing approaches for community detection on Network2: (a) Network-based ([39]) and (b) Content-based ([33]). The results show that the proposed TASCbCD method provides clear improvements in community detection for the topics and their sensitivity. The results of the community detection method establish the link between any two nodes. The resultant node pairs are used to carry out the final community formation visualized in graph form using a graphical visualization tool.

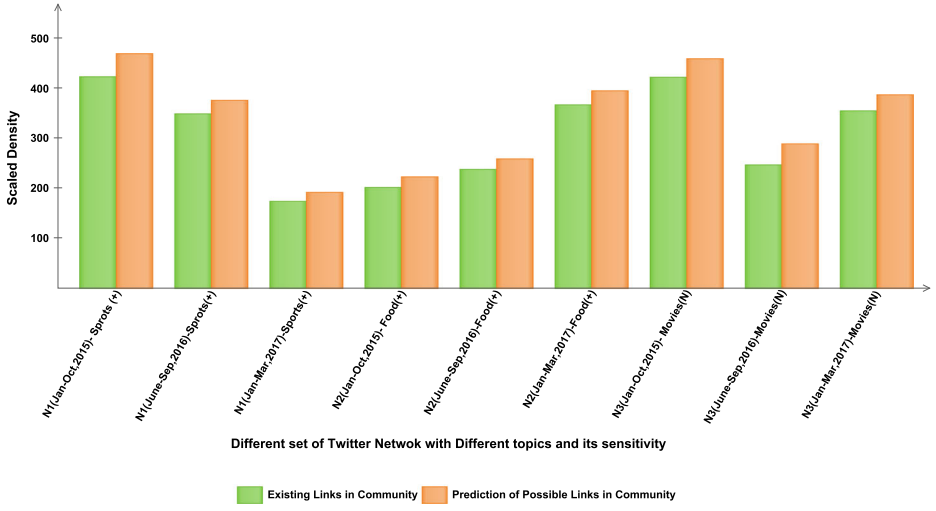
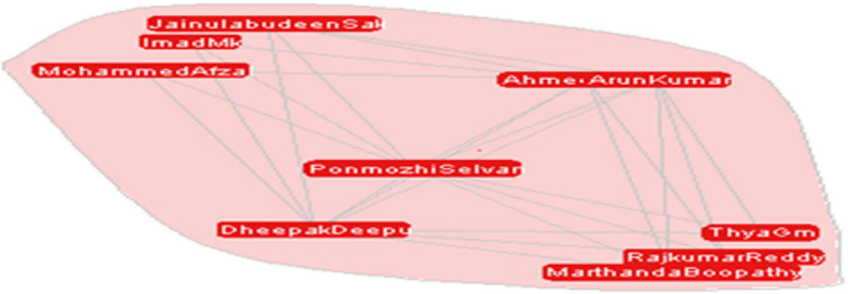
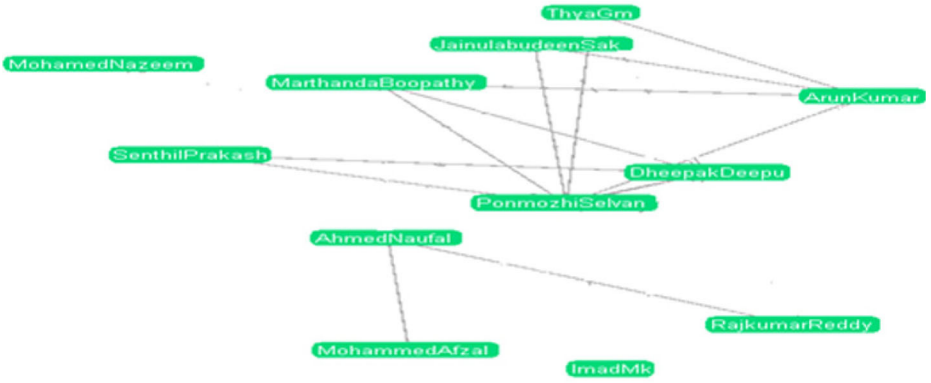


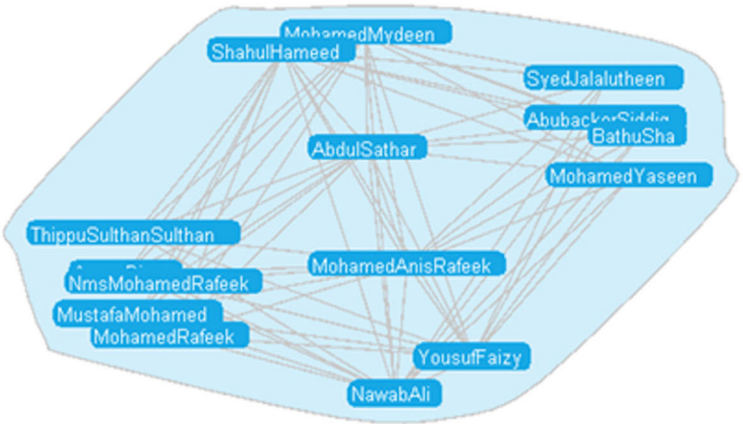
Fig. 14 Scaled Density across Different set of Twitter Network



a.Network-based Approach



b.Content-based Approaches



c.Proposed TASCbCD Approach

Fig. 15 Community Detection results for the topic 'Food' (+ve, sensitivity) in Network 2 for other approaches (a) Network-based (b) User Content-based (c) Proposed TASCbCD

Community detection in a network-based method is based on a topological approach based on people who have physical connections. For this category the topological structures play the principal role in performing community detection, using a variety of centralitarian metrics, such as degree centralization, closeness centrality, betweenness centrality, eigen

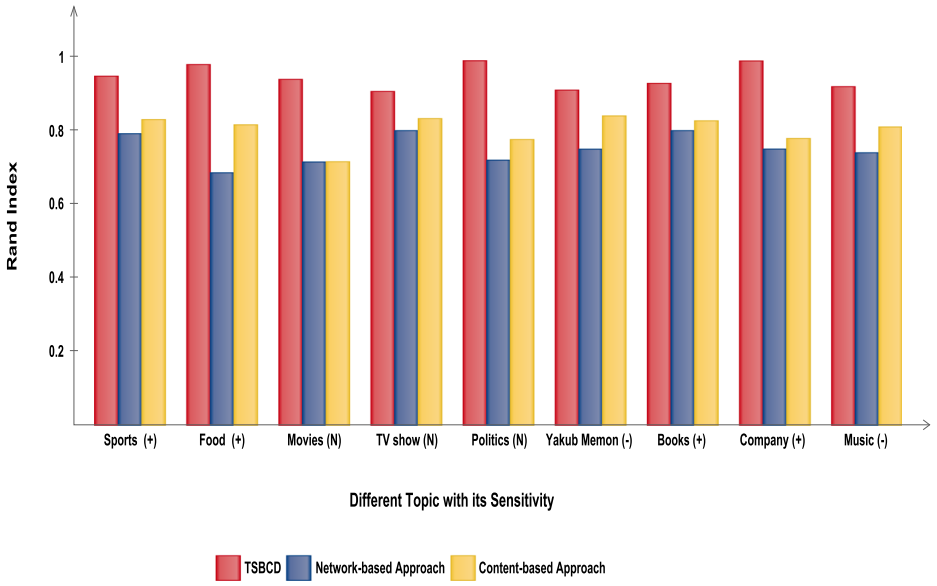


Fig. 16 Comparative Analysis Rand Index for proposed TASCbCD with other methods in Network2

vector centrality, etc. It does not consider the user features and user-generated content features to be a function for community detection, which makes an enormous impact in the ultimate outcome of influential user selection.

Following that, the content-based strategy for community detection uses node or user-generated content features as a main determinant in selecting a collection of node's or user's in community formation. However, it fails to recognise the physical network connections that exist between two nodes. As a result, there may be several discontinuous communities. To address these challenges, the proposed method combines network and content-based approaches as major elements for selecting a node for community identification.

Thus, the proposed TASCbCD method produces significant community detection results for different topics with its sensitivity(polarity). One such example topic 'food' with positive sensitivity is shown in Fig. 15. The results in Fig. 15 are observed because the proposed TASCbCD method detects overlapping communities more effectively compared with the other two methods on account of the incorporation of the topic sensitivity feature in the network.

Table 10 illustrates the Rand index result of the proposed TASCbCD method with the other method. Figure 16 shows the output of these comparisons graphically. Figure 16 shows that the network-based method produces an average Rand index rate of 74%. The content-based method produces 80%, while the proposed TASCbCD method produces 94% as the Rand index average for the nine topics in Table 10 over Network 2.

Next, the same experiment is carried out over different networks with a different topic and its sensitivity. Table 11 illustrates the implementation of community detection using the proposed TASCbCD with network and content-based methods. The graphical representation of the Rand index measure for the proposed TASCbCD method with other methods is shown in Fig. 17. It is observed that the average Rand index of the proposed TASCbCD method is

Table 10 Comparison of rand index measure for community evaluation over different topic in network 2 (july-sep 2016) for TASCbCD with other Methods

Three set of Twitter Networks with different period of tweets collections		Community (Topic-Sensitivity)	Rand Index	Scaled density	
				Existing Links in Community	Predicted links in the Community (TASCbCD)
Network1 (N1)	(Jan - Oct, 2015)	Sports (+)	0.92764	423	469
	(Jun - Sep, 2016)	Sports (+)	0.95886	349	376
	(Jan - Mar, 2017)	Sports (+)	0.89864	174	192
Network2 (N2)	(Jan - Oct, 2015)	Food (+)	0.96911	202	223
	(Jun - Sep, 2016)	Food (+)	0.97932	238	259
	(Jan - Mar, 2017)	Food (+)	0.92874	367	395
Network3 (N3)	(Jan - Oct, 2015)	Movies (N)	0.92348	422	459
	(Jun - Sep, 2016)	Movies (N)	0.91914	247	289
	(Jan - Mar, 2017)	Movies (N)	0.89754	355	387

Table 11 Comparison of rand index measure for community evaluation over different topic in different period of twitter network for TASCbCD with Other Methods

Three set of Twitter Networks with different period of tweets collections		Community Name with (Topic - Sensitivity)	TASCbCD (Rand Index)	Network Based Approach (Rand Index)	Content Based Approach (Rand Index)
Network1	(Jan - Oct,2015)	Sports (+)	0.9276	0.7956	0.8495
	(June - Sep,2016)	Sports (+)	0.9588	0.7354	0.8425
	(Jan - Mar,2017)	Sports (+)	0.8986	0.7474	0.8867
Network2	(Jan - Oct,2015)	Food (+)	0.9691	0.7297	0.8185
	(June - Sep,2016)	Food (+)	0.9793	0.6854	0.8158
	(Jan - Mar,2017)	Food (+)	0.9287	0.7365	0.8185
Network3	(Jan - Oct,2015)	Movies (N)	0.9234	0.7848	0.7856
	(June - Sep,2016)	Movies (N)	0.9191	0.7245	0.8382
	(Jan - Mar,2017)	Movies (N)	0.8975	0.7654	0.7912

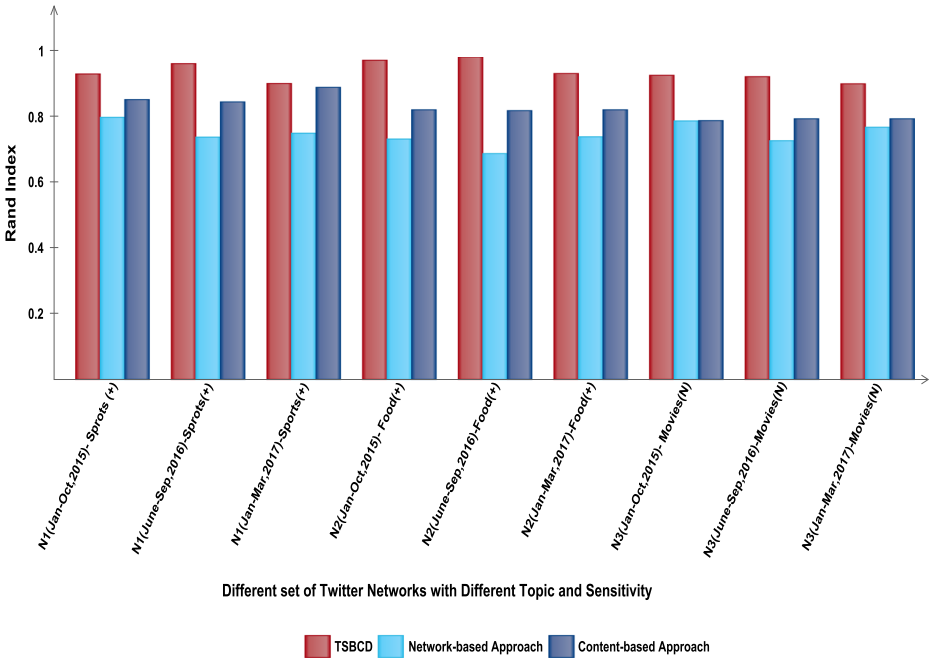


Fig. 17 Comparative Analysis Rand Index for proposed TASCbCD with other methods in Different set of Twitter Network

93% over that of different networks, as specified in Table 8. The other two methods produce an average Rand index value of 75% (network-based) and 83% (content-based).

Thus, it is concluded from the results above that in all cases, the proposed TASCbCD method is found to be better in terms of the Rand index and Scale density measure than network-based and content-based methods of community detection for different topics over the Twitter network. The overlapping communities are also easily identified in the proposed TASCbCD method due to topic sensitivity-based community detection. The results obtained from overlapping communities help easily distinguish topic-sensitive influence spreaders for information diffusion on Twitter.

In this paper, a novel TASCbCD method has been presented wherein topic-dependent features are used for sentiment classification and topic-sensitivity features for community detection. From the results of the different topics used in community detection, the overlapping community is easily identified in this system. The TASCbCD algorithm is applied over different Twitter networks with different topics and their sensitivity with data collected over different periods in time. The results of our proposed method are compared with the other methods and indicate that our methodology is most effective at identifying a topic-sensitive community over Twitter. During the experiments, it was observed that

– The TASCbCD technique produces a higher Rand index value, showing that a combination of network and content features with topic sensitivity increases community detection accuracy.
– The overlapping community structure is identified using topic sensitivity.

– The TASCbCD method produces promising results when compared to other sentiment classification methods.

The results of the proposed TSBCD method form the topology structure used to compute the topology score for influence spreader identification for information diffusion.

6 Conclusion and future work

This paper has designed a community detection based social influential gauging technique on Twitter. The topic-adaptive features are used for sentiment classification, and topic-sensitivity(polarity) features are used for community detection. The technique, based on topic sensitivity features in tweet content, identifies overlapping communities between two topics or the same topics, with different sensitivity indicators (positive, negative, and neutral). Overlapping community detection enhances influence spreader identification to maximize information diffusion. The TASCbCD algorithm is applied over three different networks on Twitter, with different topics and sensitivity, spread across data collected over different periods in time. The result of the proposed method is compared with other methods and indicates that our methodology effectively identifies topic-sensitive communities on Twitter. The result of the community structures discovered is used to identify and rank influence spreaders for information diffusion. For the proposed TASCbCD method and the SVM outperforms all the other classifier models studied to achieve an accuracy rate of 80%. For community detection, the overall average Rand index for communities of 20 different topics is 90%, 92% and 95% in Network1, Network2 and Network3 respectively. The scaled density of the proposed TASCbCD method produces 88% on Network1, 85% on Network2, and 95.7% on Network3, demonstrating a comparatively better performance than other methods.

There are certain limitations in the paper. In general, people may choose to keep certain elements of their life completely private and offline, away from the reach of social networks, therefore influence selection utilising community detection may provide fewer accurate results. Companies might be denied access to information relevant to their goods and marketing tactics if users' privacy settings allow them to control who has access to their accounts. The influence factor that affects the forwarding and non-forwarding of node prediction in the Prediction of Information Diffusion (PID) is not always reliant on the same network. It's tough to tell which accounts on different websites belong to the same person in these situations. Tweets have a number of flaws, including keyword ambiguity, a lack of keywords, and a lack of linguistic information, all of which reduce the algorithm's effectiveness.

In future work the diffusion process for a keyword can only be determined in future work if the word appears in a set that is recognised in the system's early phases. It is required to develop algorithms that make use of dictionary information in order to identify the diffusion process, even for terms that do not belong to the set. Knowing who will assist disseminate the information the most is an extension of understanding how information flows in a system.

Appendix

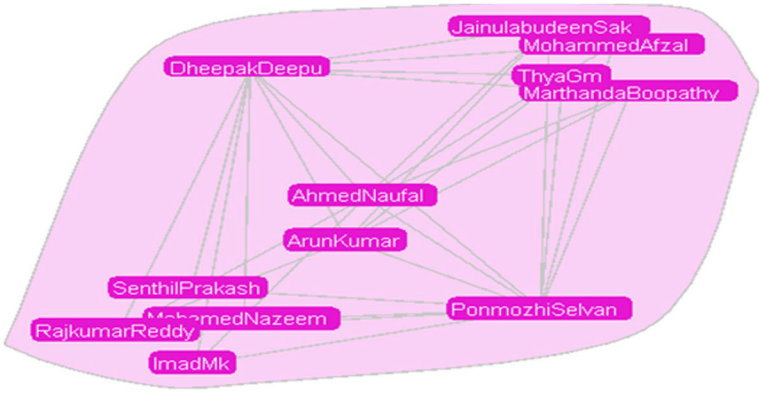


Fig. 18 TASCbCD based community result of Topic 'Books' (Positive Sensitivity)

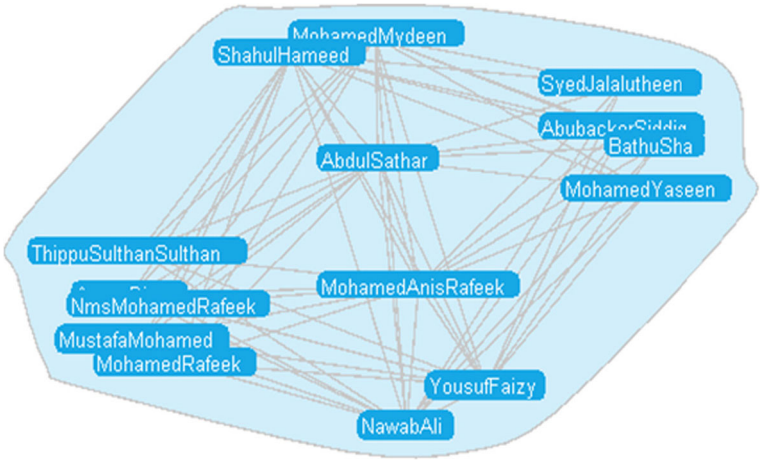


Fig. 19 TASCbCD based community result of Topic 'Food' (Positive Sensitivity)

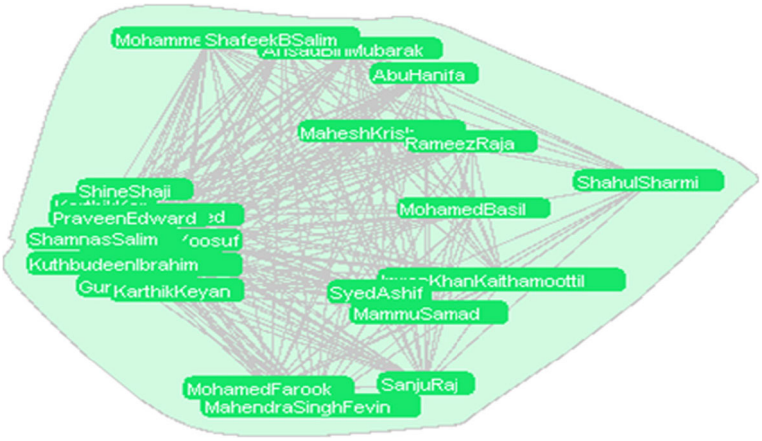


Fig. 20 TASCbCD based community result of Topic 'Movies' (Positive Sensitivity)

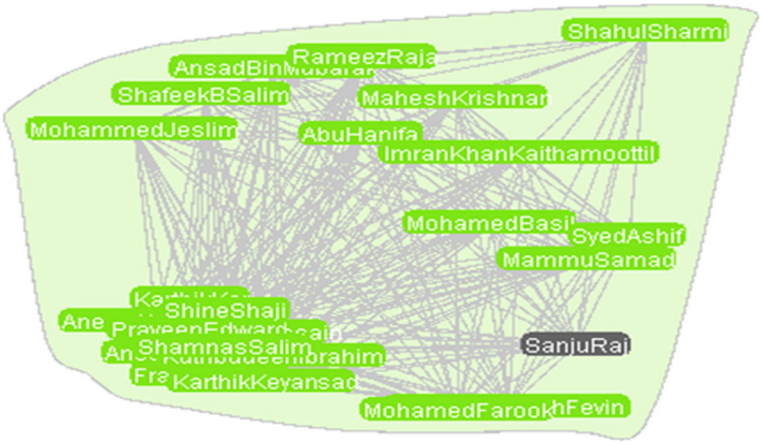


Fig. 21 TASCbCD based community result of Topic 'Music' (Positive Sensitivity)

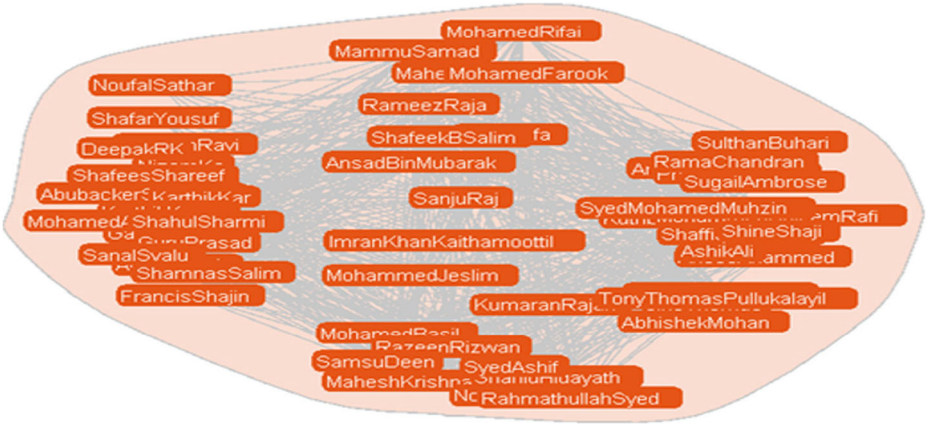


Fig. 22 TASCbCD based community result of Topic 'Yakub Memon' (Negative Sensitivity)

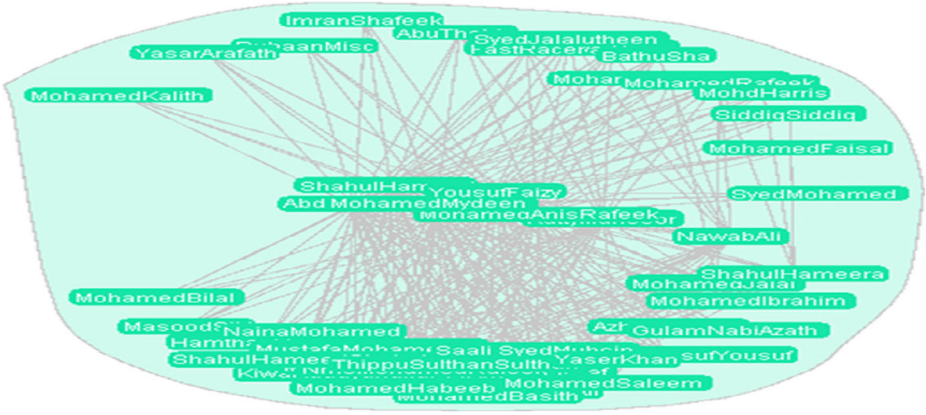


Fig. 23 TASCbCD based community result of Topic 'TV show' (Negative Sensitivity)

Data Availability Statement The dataset underlying this article will be shared on reasonable request to the corresponding author through mail.

References

1. Ahajjam S, El Haddad M, Badir H (2016) Influentials identification for community detection in complex networks. In: 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt). IEEE, pp 111–115
2. Arab M, Afsharchi M (2014) Community detection in social networks using hybrid merging of sub-communities. J Netw Comput Appl 40:73–84
3. Asur S, Parthasarathy S, Ucar D (2009) An event-based framework for characterizing the evolutionary behavior of interaction graphs. ACM Trans Knowl Discov Data (TKDD) 3(4):16
4. Baek J-W, Chung K-Y (2020) Multimedia recommendation using word2vec-based social relationship mining. Multimed Tools Appl:1–17
5. Banik A, Shamsi Z, Laiphrakpam DS (2019) An encryption scheme for securing multiple medical images. J Inf Secur Appl 49:102398
6. Baroi SJ, Singh N, Das R, Singh TD (2020) Nits-hinglish-sentimix at semeval-2020 task 9: Sentiment analysis for code-mixed social media text. arXiv:2007.12081
7. Capuano N, Chiclana F, Fujita H, Herrera-Viedma E, Loia V (2017) Fuzzy group decision making with incomplete information guided by social influence. IEEE Trans Fuzzy Syst 26(3):1704–1718
8. Capuano N, Chiclana F, Herrera-Viedma Ex, Fujita H, Loia V (2019) Fuzzy group decision making for influence-aware recommendations. Comput Hum Behav 101:371–379
9. Chang C-S, Lee D-S, Liou L-H, Lu S-M, Wu M-H (2018) A probabilistic framework for structural analysis and community detection in directed networks. IEEE/ACM Trans Netw (TON) 26(1):31–46
10. De Maio C, Fenza G, Gallo M, Loia V, Parente M (2018) Social media marketing through time-aware collaborative filtering. Concurr Comput Pract Exper 30(1):e4098
11. De Maio C, Fenza G, Gallo M, Loia V, Parente M (2019) Time-aware adaptive tweets ranking through deep learning. Futur Gener Comput Syst 93:924–932
12. Derbas N, Dusserre E, Padró M, Segond F (2018) Eventfully safapp: hybrid approach to event detection for social media mining. J Ambient Intell Human Comput:1–9
13. Dey P, Chatterjee A, Roy S (2018) Knowledge based community detection in online social network. In: 2018 10th International Conference on Communication Systems & Networks (COMSNETS). IEEE, pp 637–642
14. Di J, Wang X, He D, Lu W, Fogelman-Soulié F (2017) Jianwu Dang. Identification of generalized communities with semantics in networks with content. In: IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, pp 1182–1189
15. Dongen S (2000) Graph clustering by flow simulation [ph. d. dissertation]. Centers for Mathematics and Computer, Science. University of Utrecht
16. Dou K, Guo B, Li K (2019) A privacy-preserving multimedia recommendation in the context of social network based on weighted noise injection. Multimed Tools Appl 78(19):26907–26926
17. Fu W, Le S, Xing EP (2009) Dynamic mixed membership blockmodel for evolving networks. In: Proceedings of the 26th annual international conference on machine learning. ACM, pp 329–336
18. Guidi B, Michienzi A, De Salve A (2019) Community evaluation in facebook groups. Multimed Tools Appl:1–20
19. Hajarian M, Bastanfard A, Mohammadzadeh J, Khalilian M (2019) Snefl: Social network explicit fuzzy like dataset and its application for incel detection. Multimed Tools Appl 78(23):33457–33486
20. Hangal S, MacLean D, Lam MS, Heer J (2010) All friends are not equal: Using weights in social graphs to improve search. In: Workshop on Social Network Mining & Analysis, ACM KDD
21. Ji P, Zhang S, Zhou Z (2020) A decomposition-based ant colony optimization algorithm for the multi-objective community detection. J Ambient Intell Human Comput 11(1):173–188
22. Jia S, Gao L, Gao Y, Wang H (2014) Anti-triangle centrality-based community detection in complex networks. IET Syst Biol 8(3):116–125
23. Kumar A, Sangwan SR, Nayyar A (2019) Rumour veracity detection on twitter using particle swarm optimized shallow classifiers. Multimed Tools Appl 78(17):24083–24101
24. Kumaran P, Chitrakala S (2017) Social influence determination on big data streams in an online social network. Multimed Tools Appl 76(21):22133–22167

25. Laiphrakpam DS, Khumanthem MS (2017) Cryptanalysis of symmetric key image encryption using chaotic rossler system. Optik 135:200–209
26. Li W, Ye Z, Xin M, Jin Q (2017) Social recommendation based on trust and influence in sns environments. Multimed Tools Appl 76(9):11585–11602
27. Lin Y-R, Chi Y, Zhu S, Sundaram H, Tseng BL (2009) Analyzing communities and their evolutions in dynamic social networks. ACM Trans Knowl Discov Data (TKDD) 3(2):8
28. Liu L, Xu L, Wangy Z, Chen E (2015) Community detection based on structure and content: A content propagation perspective. In: 2015 IEEE International Conference on Data Mining. IEEE, pp 271–280
29. Liu T, Xue F, Sun J, Sun X (2019) A survey of event analysis and mining from social multimedia. Multimed Tools Appl:1–18
30. Liu Y, Niculescu-Mizil A, Gryc W (2009) Topic-link lda: joint models of topic and author community. In: Proceedings of the 26th annual international conference on machine learning. ACM, pp 665–672
31. Loia V, Parente D, Pedrycz W, Tomasiello S (2018) A granular functional network with delay: some dynamical properties and application to the sign prediction in social networks. Neurocomputing 321:61–71
32. Loia V, Tomasiello S, Vaccaro A, Gao J (2020) Using local learning with fuzzy transform: application to short term forecasting problems. Fuzzy Optim Decis Making 19(1):13–32
33. Lu Z, Sun X, Wen Y, Cao G, Porta TL (2014) Algorithms and applications for community detection in weighted networks. IEEE Trans Parallel Distrib Syst 26(11):2916–2926
34. Lv H, Tao L, Xianglin H, Hongxiao G, Zengfeng B (2017) Detection algorithm based on closeness rank and signal transimission. In: IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, pp 443–447
35. Meetei LS, Singh TD, Bandyopadhyay S (2019) Wat2019: English-hindi translation on hindi visual genome dataset. In: Proceedings of the 6th Workshop on Asian Translation, pp 181–188
36. Nesi P, Pantaleo G, Paoli I, Zaza I (2018) Assessing the retweet proneness of tweets: predictive models for retweeting. Multimed Tools Appl 77(20):26371–26396
37. Ouvrard X, Le Goff J-M, Marchand-Maillet S (2020) Exchange-based diffusion in hb-graphs. Multimed Tools Appl:1–36
38. Pang J, Huang J, Zhang W, Huang Q, Yin B (2017) Justify role of similarity diffusion process in cross-media topic ranking: an empirical evaluation. Multimed Tools Appl 76(23):25145–25157
39. Pattabiraman B, Md MostofaAP, Gebremedhin AH, Liao W-k, Choudhary A (2015) Fast algorithms for the maximum clique problem on massive graphs with applications to overlapping community detection. Internet Math 11(4-5):421–448
40. Plantié M, Crampes M (2013) Survey on social community detection. In: Social media retrieval. Springer, pp 65–85
41. Porcel C, Ching-López A, Lefranc G, Loia V, Herrera-Viedma E (2018) Sharing notes: an academic social network based on a personalized fuzzy linguistic recommender system. Eng Appl Artif Intell 75:1–10
42. Qi G-J, Aggarwal CC, Huang T (2012) Community detection with edge content in social media networks. In: 2012 IEEE 28Th International Conference on Data Engineering. IEEE, pp 534–545
43. Qi X, Tang W, Wu Y, Guo G, Fuller E, Zhang C-Q (2014) Optimal local community detection in social networks based on density drop of subgraphs. Pattern Recogn Lett 36:46–53
44. Rani S, Mehrotra M (2017) Hybrid influential centrality based label propagation algorithm for community detection. In: 2017 International Conference on Computing, Communication and Automation (ICCCA). IEEE, pp 11–16
45. Rathore S, Loia V, Park JH (2018) Spamspotter: an efficient spammer detection framework based on intelligent decision support system on facebook. Appl Soft Comput 67:920–932
46. Ruta M, Scioscia F, Pinto A, Gramegna F, Ieva S, Loseto G, Sciascio ED (2019) Coap-based collaborative sensor networks in the semantic web of things. J Ambient Intell Human Comput 10(7):2545–2562
47. Sachan M, Contractor D, Faruquie TA, Subramaniam LV (2012) Using content and interactions for discovering communities in social networks. In: Proceedings of the 21st international conference on World Wide Web. ACM, pp 331–340
48. Sani NS, Manthouri M, Farivar F (2020) A multi-objective ant colony optimization algorithm for community detection in complex networks. J Ambient Intell Human Comput 11(1):5–21
49. Shadang M, Saharia N, Singh TD (2020) Towards the study of morphological processing of the tangkhul language. arXiv:2006.16212
50. Singh TD Addressing some issues of data sparsity towards improving english-manipuri smt using morphological information. Monolingual Mach Transl:46
51. Singh TD, Solorio T (2017) Towards translating mixed-code comments from social media. In: International Conference on Computational Linguistics and Intelligent Text Processing. Springer, pp 457–468

52. Sun PG (2014) Weighting links based on edge centrality for community detection. Physica A: Stat Mech Appl 394:346–357
53. Swain AK, Balabantaray BK, Rout JK, Satpathy S An optimal deep learning approach for classification of age groups in social network
54. Tai C-H, Philip SY, Yang D-N, Chen M-S (2013) Structural diversity for resisting community identification in published social networks. IEEE Trans Knowl Data Eng 26(1):235–252
55. Tang L, Liu H, Zhang J, Nazeri Z (2008) Community evolution in dynamic multi-mode networks. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 677–685
56. Wang C-D, Lai J-H, Philip SY (2013) Neiwalk: community discovery in dynamic content-based networks. IEEE Trans Knowl Data Eng 26(7):1734–1748
57. Wang C, Tang W, Wang Y, Fang J, Yao S (2017) Local community detection algorithm based on links and content. In: 2017 IEEE 2Nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, pp 1805–1808
58. Wang D, Long S (2019) Boosting the accuracy of differentially private in weighted social networks. Multimed Tools Appl 78(24):34801–34817
59. Wang X, Liu G, Li J (2017) Overlapping community detection based on structural centrality in complex networks. IEEE Access 5:25258–25269
60. Yang T, Jin R, Chi Y, Zhu S (2009) Combining link and content for community detection: a discriminative approach. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 927–936
61. Yarow J (2010) Twitter finally reveals all its secret stats. Business Insider SAI
62. Zhou Y, Cheng H, Jeffrey Xu Y (2009) Graph clustering based on structural/attribute similarities. Proc VLDB Endowment 2(1):718–729
63. Zhuang K, Shen H, Zhang H (2017) User spread influence measurement in microblog. Multimed Tools Appl 76(3):3169–3185