




Query-oriented topical influential users detection for top-*k* trending topics in twitter

Sarmistha Sarna Gomasta¹ · Aditi Dhali¹ · Md Musfique Anwar¹ · Iqbal H. Sarker² 

Accepted: 3 April 2022 / Published online: 25 May 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Online Social Networks (OSNs) have become inevitable for any new methodology both for viral promoting applications and instructing the creation of inciting information and data. As a result, finding influential users in OSNs is one of the most studied research problems. Existing research works paid less attention to the temporal factors associated with the activities performed by the social users. Our motivation is to find influential users who show their most powerful interests towards a given query on various subjects (topics) at different time intervals by featuring more on users' most recent activities as well as their associations with different users. To address this problem, we propose a temporal activity-biased weight model that gives higher weight to users' recent activities and develops an algorithm to list the most effective influential users. In addition, our proposed model also considers the impacts of topical similarities both from direct and indirect neighbors of the users. Experimental results on two real datasets demonstrate that our proposed framework yields better outcomes than the baseline method.

Keywords Online social network · Trending topic · Common neighbors · Influential User

1 Introduction

Since the recent decades, internet technology has played an inevitable role in changing human civilization. Nonetheless, the idea of online social media has been developed based on the very beginning of human interaction. It is an

efficient way for changing the method of collaboration and communication all through the world. The number of online media sites and their accessibility to everyone has likewise accomplished diverse excellence that results in the continuous expansion of the web access to the mass people and the growing usage of mobile phones. Sharing new data, insights, learning, and finding new ways of human correspondence are far quicker than ever with real-time data using these online platforms. Around 3.96 billion individuals have invested their time on the internet everywhere in the world till the start of July 2020 [25]. There are numerous types of social organizing sites that have become popular among mass people. The shared information is represented in multiple forms in different media such as tweets or retweets, hashtags in Twitter, likes, comments, messages, videos, etc., in Facebook, uploading images in Instagram, etc. Each social networking site includes different gatherings where users can impart their insight, spread data, or convey it to others. This has created a new way to copy genuine human interaction.

Twitter has become one of the biggest and very well-known microblogging sites on the internet. The idea of microblogging contributes to a blog that licenses users to give brief message updates through various channels via web or mobile. Twitter is utilized to distribute and acquire

This article belongs to the Topical Collection: *Big Data-Driven Large-Scale Group Decision Making Under Uncertainty*

✉ Sarmistha Sarna Gomasta
sarmistha.stu2016@juniv.edu

Aditi Dhali
aditi.stu2015@juniv.edu

Md Musfique Anwar
manwar@juniv.edu

Iqbal H. Sarker
iqbal@cuet.ac.bd

¹ Department of Computer Science and Engineering,
Jahangirnagar University, Savar, Dhaka, Bangladesh

² Department of Computer Science and Engineering,
Chittagong University of Engineering and Technology,
Chittagong, 4349, Bangladesh

subtleties of information and data speedily. As a result, it is now one of the most used social platforms in the world [30]. This media permits twitterers (the users of Twitter) to issue tweets, and each tweet is limited to 140 characters because of quick read and compose. Each tweet offers a strong *social connection* among the users as indicated by the taste of the topic(s) they are interested in. This online social platform boasts monthly 330 million active users (as of 2020 Q1). Of these, more than 40 percent of users use the service daily (Twitter, 2021). Around 500 million tweets are sent each day which sums up to 200 billion tweets every year [41]. On average, 6 of every 10 (63%) Twitter users are between 35 to 65 years old around the world [30]. That depicts that Twitter has more mature users who can share thoughtful opinions on any topic. Twitter users' average time spent on it is around 3.39 minutes per session that is on top of any other online social platform [30]. Besides casual social interactions, almost 187 million daily active users use it for business purposes [22]. Around 67% of business-to-business (B2B) platforms use it as a digital marketing tool [45] because 50% of Twitter subscribers report buying something after seeing it on the forum. 93% of Twitter community members are open to a brand presence on Twitter [10, 22]. Twitter has also turned into a center of attraction for scientists and researchers, not only advertising products. They center around an enormous amount of data to create approaches on the best way to make this social site as a viral promoting stage [5]. As the social networking site (Twitter) coordinates a massive number of users and the most significant part of the cases is that the users send *following* request to others who have comparative similar topical interests with the user - the idea is known as homophily [32]. Above mentioned points can easily ensure that Twitter can reach out to a sizeable sensible community efficiently within a short time. This social platform can be used as a robust marketing platform undoubtedly. The efficiency of Twitter interactions acts as our motivation, and the plan is to discover the influential users. They can help by highlighting the products in viral marketing or conveying opinions on different social topics among their followers. We are interested to find out active, influential top users in many applications domains. For example, "Sam", a sports enthusiast, posts a lot about trending games, and all his posts are paid close attention by a large number of people. A sports company can easily target "Sam" to collaborate to achieve their marketing objectives. This is also called influencer marketing. Besides advertising the products, we have proposed to find influential researchers. So, there is a wide range of significant purposes in real life of our motivation. Some of them are:

- Remarkable impact in viral marketing and social decisions like election, donation, etc.

- Anticipating target customers and enhancing customers engagement
- Analyzing the trends and behavior of people towards any event
- Have a wide network which is essential for digital marketing
- Tracking down the collective identity and role models in social movements.
- Following influential researchers for reshaping scientific and technological advancements.

Our proposed strategy is fundamentally extensive approach of the TwitterRank method [44] by coordinating the *temporal* factor. We initially applied the Twitter-Latent Dirichlet allocation (Twitter-LDA) topic model [36] on the tweets distributed by the users to extract the concealed topics discussed by them. At that point, we narrowed down the trending topics list in a particular time frame and ranked those topics over-popularity. After that, we apply our proposed algorithm on the processed tweets to list the top influential users for a given inquiry comprising those trending topics. We also notice the dynamic change in the rank list over time by emphasizing users' latest activities (for instance, publishing tweets on Twitter) and their associations with other users. An academic Co-author network dataset (DBLP) [40] is an online reference for important computer science research papers. Our algorithm has also been applied in this dataset to identify promising researchers over time and collaboration. As a synopsis, our contributions are in the following list -

- We have used a Twitter dataset where verified twitterers can express their opinions through tweets and comments. We have applied Twitter-LDA to remove topics after preprocessing tweets to extract all related information about one particular event.
- We have proposed an approach for analyzing "trending topics" at each time interval. The idea of a trending topic is to calculate each user's freshness towards the activities.
- We have calculated connection trust between two users to find out the impact of common neighbors.
- We have emphasized users' topical degree of influence change over time, amplifying common neighbors effect and transition probability.
- We have performed Kendall's τ for measuring the correlation between the influential users' rank lists obtained by different algorithms.
- We have also assessed our algorithm effectiveness based on the recommendation task and compared it with other popular algorithms.
- Extensive tests are led on real datasets to show the adequacy of our proposed strategy

The rest of the paper is categorized as follows: Section 2 includes the relevant research works in this field and Section 3 establishes an extensive description of problem statements and our methods to find influential users in trending topics. Data preprocessing to find the rank of significant users are considerably described in this section. Section 4 covers experimental evaluation, performance analysis in recommendation task, and Section 5 summarizes our research work in the discussion part. Eventually, we conclude our work with the direction of future research scope in Section 6.

2 Related work

In contemporary network science, online social networks and social media analytics both are popular research areas. Javed et al. [23] has predicted the stock market using big data retrieved from social media like Yahoo!, daily newspaper, and Twitter. Even policing protests in the United Kingdom are analyzed by social media data [16]. Additionally, cyber risk management [39], mental health condition [37], suicide rate and causes [43], box office's profit [14] etc are predicted through social media analysis. The prominent Influencers' impact on consumer behavior is evaluated in the work of Pick et al. [35]. The effectiveness of sponsors and influential users' in multiple domestic and business sectors are analyzed in the work of Yang Feng et al. [19]. The study of N.I.M Anuar et al. [6] has found out the cause of being influenced by Instagram influencers in regards to purchasing intention of fashion items.

Various approaches exist to detect influential users in OSNs, starting from simply counting the immediate neighbors to more complex machine learning and message passing techniques. One of the first studies that attempted to find the parameters of this approach was taken by Zhang et al. [47]. They have considered users' retweet behavior patterns to investigate how friends in one's ego network influence retweet behaviors. In this model, the designs incorporate the social influence locality into a factor graph model, further leveraging the network-based correlation. Weng et al. [44] have suggested a measure named TwitterRank based on the idea of PageRank to compute users' topical influence in Twitter. This approach is based on the topic query set and it shows the relevance of link structure and the similar interest of users. Algaradi et al. [2] have calculated the users' interactions and modeled the social graph using the weighted k -core decomposition method to identify the influential spreaders in OSNs. To identify top- k significant users in social networks, Alshahrani et al. have proposed an efficient algorithm based on centrality measures [4]. Moreover,

Zareie et al. [46] have proposed a method to select the influential users based on the interest value of friends' interests and connected neighborhood. A new approach named Temporal Topic Influence (TTI) has been proposed by Wang et al. [42] states that analytical applications in online social networks can be generalized as the influence evaluation problem, which targets at finding most influential users. This model is dependent on time interval, content, and structure-aware. Juliana et al. [34] have investigated the effect of topic familiarity on listening comprehension and how far certain aspects of the language would likely be influenced by topic familiarity. The UIRank algorithm is based on the commitment of the user's tweet and the attributes of information dispersal in the microblog networks. It computes user influence score iteratively by user follower graph [24]. Most of them ignore the time factors in their work.

To discover highly reliable domain-based influencers at different time intervals, Bilal Abu-Salih et al. have suggested a framework with the help of semantic analysis and machine learning modules [1]. Again, there is an on-Demand Influencer Discovery (DID) model, which employs an iterative learning process incorporating the language attention network as a subject filter, proposed by Cheng Zang et al. that can identify influential users on any subject regardless of its demand on social media [49]. Their influence convolution network is built on user interaction. But they didn't suggest any rank for their influential users. The research of Divyani Mittal et al. has discovered and ranked significant users (topic wise) [33]. Their proposed Aggregation Consensus Rank Algorithm (ACRA) is applied on time intervals to generate top-ranked influential users' lists using different Twitter metrics. They analyze the connection between users and graph database to find this significant user. In a framework named Personalized PageRank that also identifies influential topical users based on both information gathered from the network and the data retrieved from user actions [3]. Additionally, fake influencers can be a threat to marketing and advertising. A trust-based method for identifying these inorganic users is proposed by Ayushi Dewan [17] using decision tree and machine learning.

Among the recent research, Li et al. [29] work on sensitive influence maximization on the different interesting topics of different users. Their proposed algorithm is based on graph pruning and a three-stage heuristic optimization strategy. Mandal et al. proposed Social Promoter Score (SPS)-based recommendation [31]. Kumar et al. [26] found Top- k influential nodes in a community using label propagation. They claimed their work using several real-life data. Another approach of influence maximization is proposed by Li et al. [28]. Their framework is based on

a meta-heuristic search algorithm. In a social network, Shi et al. [38] proposed a community detection algorithm established on Quasi-Laplacian centrality peaks clustering.

But, most of the existing approaches overlooked the combination of analyzing trending topics and the temporal factor, which significantly affects the ranking of the influential users. Our current proposed method is the extension of Topical Influential Users Detection (TIUD) algorithm [18]. This is also defined as finding significant users for a set of trending topics and listing the top significant users at different specific time intervals considering familiar neighbors.

3 Overview of research method

3.1 Problem formulation

We determine some fundamental ideas formally before presenting our concerns.

Social Graph: An attributed social graph manifests as $G = (U, E, A)$, where U symbolizes the set of social users or twitterers (nodes), E means the group of links (edges) in the users and $A = \{T_1, T_2, \dots, T_n\}$ is the set of topics contemplated by the social users in G [9].

Topic: Any specific keyword or a set of related words which illustrates equal thought can be assessed as the topic [7]. For instance, when *health* is a topic, words linked to health are like doctors, hospital, pandemic, corona, etc.

Trending Topic: A trending topic is a concern that meets an inundation of popularity, often advancing around

widespread contemporaneous phenomena. For example, celebrities' pronouncements, breaking news, social affairs, etc., for a limited range of time.

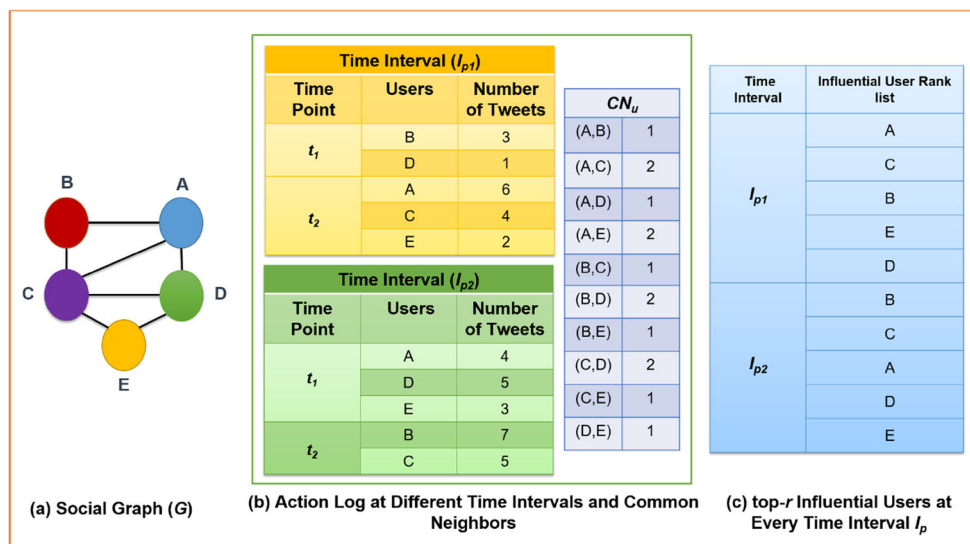
Activity: Social users execute actions (e.g., posting tweets on Twitter) at various time points. This activity is recorded as an activity tuple (u_i, Ψ_{u_i}, t_n) where Ψ_{u_i} exemplifies the set of the high-ranked trending topics which is covered by the working that is conducted by u_i at time t_n [8, 15].

Query: An input query $Q = \{\Psi_q\}$ assimilates a set of query topics Ψ_q .

Problem Definition: A social graph G is given and an input query Q contains a set of topmost trending topics, a time interval set \mathcal{I} . The function of Trending Topical Influential Users Detection (TTIUD) is to deliver a record \mathcal{R}^Q of top- r influential users at every time interval I_p .

A social graph (G) including five nodes where each node represents a twitterer has been illustrated in Fig. 1(a). Different users can display significant impact at different time intervals (I_p) for a specific query (Q). Here, 1(b) represents those users action log and users' number of common neighbors (CN_u). For each time interval, the action log represents the number of tweets posted by any user for a specific query at different time points, indicating a unit of time (hourly). At each time point, the user with the higher recency score is on the list. For instance, first-time interval I_{p1} is divided into two time points t_1 and t_2 where $t_2 > t_1$. At the time point (t_2), the user (A) has posted the most number of tweets (6) for a specific query. So, the recency score of the user (A) is the highest among all other users. It secures the first position in the rank list because of its highest recency score as well as connection strength in this time interval (I_{p1}) (Fig. 1(c)). Again, in our second

Fig. 1 Influential Users Ranking at Different Time Intervals for a Specific Query in a Social Graph



rank list at a time interval (I_{p_2}), the user (D) has achieved a lower rank than the user (A) because it has less connection strength, although user (D) has more number of tweets than the user (A).

All the parameters that are applied to stimulate our proposed framework are represented in [Appendix](#).

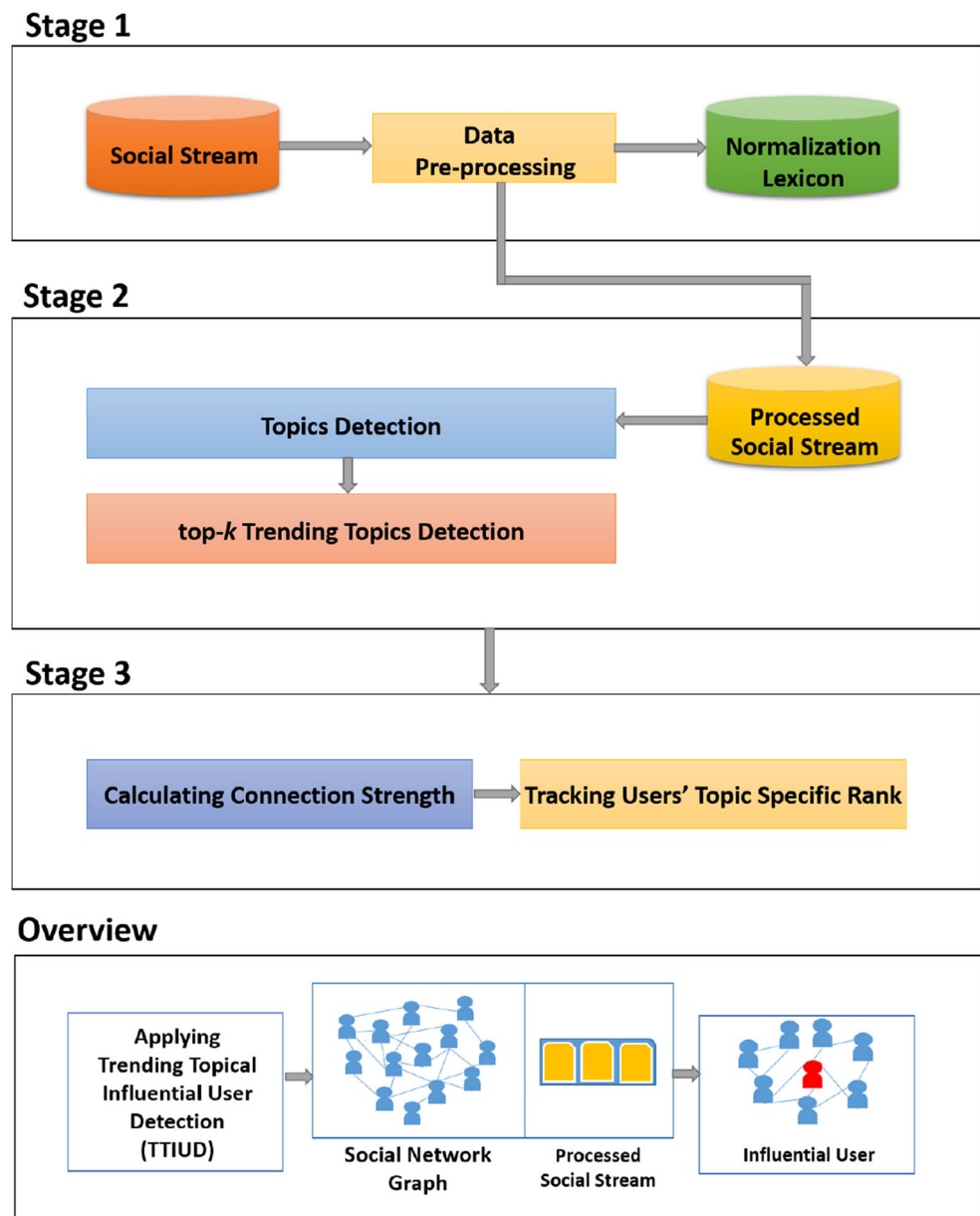
3.2 Influential user detection approach

The proposed approach of finding influential users for top- k trending topics can be summarized in two main steps eventually. The first one is to preprocess the unstructured tweets' stream and then identify the top trending topics. The next one is to carry out the user's influential score and rank

analysis of the head first trending case concerning Temporal TwitterRank and common neighbor's effect. So, we have three primary stages involved in the process presented in Fig. 2 in our proposed approach.

From Fig. 2, it is shown that data preprocessing is performed to eliminate extra information from the social stream. Then, we apply topic modeling or topic distillation method on the cleaned data to recognize the hidden topics. In this step, we have used Twitter LDA (T-LDA) to extract hidden topics. Next, we apply our model to find the top- k trending topics at each time. Finally, we develop algorithms and apply those algorithms on the processed social streams alongside the social network graph (G) to discover query-specific influential users. Our method determines the rank

Fig. 2 Workflow of Trending Topical Influential Users Detection (TTIUD) Framework



of influential users by calculating connection strength using the common neighbor's effect and no of tweets over a specific topic. The whole process is depicted briefly in the overview section in Fig. 2.

3.2.1 Data pre-processing and topic detection

On Twitter, tweets are typically created in a casual, easygoing manner and now and again contain syntactically incorrect sentence structures with an orthographic mistake and non-standard words (e.g., coook for cook, sooooo for so). Twitter users frequently use hashtags (for instance, #Obama, #Ronaldo, etc.) in tweets in text form to spread the information effectively. The utilization of hashtags is discretionary, and there are no standard guidelines for using a hashtag to indicate a topic. As a consequence, it is difficult for any framework to extricate topics from hashtags [11] successfully. Topic modeling is an unsupervised learning approach that creates information and analyzes words from reports and documents by connecting words with similar highlights and separating the uses of terms with different implications. There are a few unique methodologies such as Probabilistic Latent Semantic Analysis (PLSA), Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), etc.

Latent Semantic Analysis (LSA) is one of the former techniques of topic modeling. The core function assumes a document matrix and decomposes it into the document-topic matrix and a topic-term matrix. This model usually terms frequency-inverse document frequency score to replace row counts in document terms. Probabilistic Latent Semantic Analysis (PLSA) is considered a probabilistic method to analyze two-mode and co-occurrence data. Latent Dirichlet Allocation (LDA) is one of the most popular models in terms of topic modeling. It applies Dirichlet priors for the document-topic and word-topic distribution. It maps each document by determining a bag of words. It works better than PLSA because of its ability to generate new records more efficiently and also serves sample point for fast forward.

In our work, we have performed normalization of the tweets through direct replacement of lexical variations with their standard structures with a normalization lexicon proposed by Han et al. [20] to improve the standard of our tweet corpus. To retrieve the text-based content from a tweet, Twitter-LDA (T-LDA) [48], a compelling extension of LDA [12] is utilized here for topic distillation. It formats every tweet into two parts: main topic words and background words. This process works better in topic extraction because of semantic coherence compared to other methods.

In this procedure, every user's topical interest (ϕ_i) is represented by a distribution over topics (N). Every word

from a set of topics consists of background word distribution (θ_{bk}) and topic word distribution (θ_{kw}). The latent value (y) differs every time from the distribution process. If $y = 0$, it verifies that the word is from background word distribution θ_{bk} and if $y = 1$, it is from topic word distribution θ_{kw} . The latent value can be modified by π , which is a common factor indicating the rate of θ_{kw} and θ_{bk} is the same [44]. The distribution of topics has been governed through Symmetric Dirichlet Distribution. Because of this, Twitter-LDA can extract topics if there are any noisy data in tweets and capture the related background words. Twitter-LDA uses Gibbs sampling to obtain model inference. The generative process of Twitter-LDA has been described in Algorithm 1, and graphical representation of Twitter-LDA is shown in Fig. 3.

Algorithm 1 Twitter Latent Dirichlet Allocation Algorithm (Twitter-LDA).

Input: $N = (\theta_{bk}, \theta_{kw}, \phi_i), \gamma, \pi$

Output: Extract meaningful topics

```

1:  $\theta_{bk} \in Dir(\lambda), \pi \in Dir(\gamma)$ 
2: for each topic  $n = 1, \dots, N$  do
3:    $\theta_{kw} \in Dir(\beta)$ 
4: for each user  $u = 1, \dots, U$  do
5:    $\phi_i \in Dir(\alpha)$ 
6:   for each tweet  $s = 1, \dots, N_u$  do
7:      $z = \text{Multi } \phi_i$ 
8:     for each word  $n = 1, \dots, N_{us}$  do
9:        $y = \text{Multi}(\pi)$  where  $y = (0, 1)$ 
10:      if  $y = 0$  then
11:         $w \in \text{Multi}(\theta_{bk})$ 
12:      else
13:         $w \in \text{Multi}(\theta_{kw})$ 
```

3.2.2 Topic sensitive temporal pagerank

Our perception is that social users' degree of topical interests changes over time, i.e., they have an alternate level of goods on different topics at different time intervals. As a result, our proposed framework has outranked TwitterRank [44] in the following aspects:

- Tracking top- k trending topics in the different time intervals
- Modelling users' temporal degree of topical interests by emphasizing the *freshness* or *newness* of users' recent activities
- Analyzing the effect of familiar neighbors (both direct and indirect) among the users while estimating user's influence
- Evaluating topic-sensitive user's rank considering connection strength rather than topic similarity

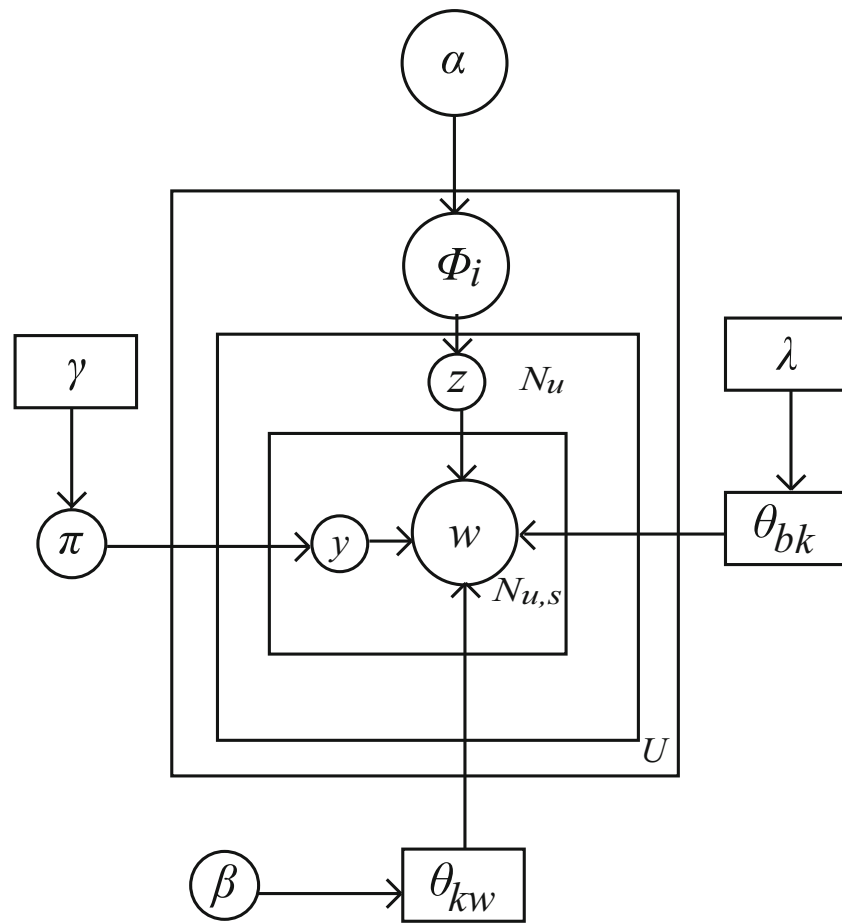


Fig. 3 Graphical Depiction of Twitter-LDA Model

- Input query (Q) in our proposed model includes multiple topics instead of a single or solitary topic

In our proposed model, we emphasize more on users' most recent activities as they are most likely to go with users' current topical interests by a measure called *recency score*, denoted by μ in (1).

$$\mu(u_i, \Psi_{u_i}, t_n) = \exp(-age(u_i, \Psi_{u_i}, t_n)) \quad (1)$$

In (1), $age(u_i, \Psi_{u_i}, t_n)$ denotes the quantity of time passed since the activity happened which has been done by user (u_i) at time point (t_n). For example, we can assume a specific time interval that starts from June 15 and ends on June 17. For estimating the recency score, the age of one tweet is calculated with the difference of its published time and “specific period ending time” which is June 17 in this case. Suppose the user (A) published a tweet on 16 June at 7:15:02 pm and user (B) published a tweet on 16 June at 10:20:12 pm. Considering 17 June and 11:59:59 as the specific interval ending time, we can achieve the recency score. User (B) will have the higher recency score in this example. Next, we compute trending score ($\sigma_{(T_j, I_p)}$) for

each topic T_j at time interval I_p according to (2).

$$\sigma_{(T_j, I_p)} = \frac{\sum_{i=1}^{|U|} \mu(u_i, T_j, t_n)}{NT_{I_p}} \quad (2)$$

Where $\mu(u_i, T_j, t_n)$ indicates the total number of activities related to topic T_j and NT_{I_p} denotes the total number of tweets in all issues at time interval I_p .

Our proposed approach changes TwitterRank [44] model by incorporating all the above factors and distinguishes itself from the PageRank algorithm [13] in the following features:

- It recognizes in the transition probability from one user to another in a particular topic where an arbitrary surfer keeps a *topic-specific* random walk.
- It builds a *topic-specific relationship network* among the users.
- It compares the *topic relevance* between users.

The transition matrix for topic T , denoted as P_T is defined as follows:

$$P_T(i, j) = \frac{|\Psi_j|}{\sum_{a: s_i \text{ follows } s_a} |\Psi_a|} \times sim_T(i, j) \quad (3)$$

$$\text{sim}_T(i, j) = 1 - |DT'_{it} - DT'_{jt}| \quad (4)$$

P_T = Transition probability matrix used for topic T from user s_i to follower s_j

$|\Psi_j|$ = Number of tweets issued by s_j on those topics

$\text{sim}_T(i, j)$ = Similarity between s_i and s_j in topic T

DT' = Row-normalized matrix in topic distillation

$\sum_{a: s_i \text{ follows } s_a} |\Psi_a|$ = Sums up the number of tweets published or issued by all of s_i 's friends

DT'_j = Contains the probability of Twitter user s_j 's interest in various topics

In this case, (4) represents two insights. Firstly, a twitterer (s_i) may follow multiple users, and every user can post several tweets. Then, the user (s_j) who posts a higher portion of tweets will have a greater influence on the user (s_i). Secondly, the influence is also correlated with topic similarity in the context of homophily [32].

The teleportation vector of the random surfer in topic T is clarified as:

$$E_T = DT'_T \quad (5)$$

We also estimate *connection trust* ($CT(i, j)$) between two users (i and j) as,

$$CT(i, j) = \frac{|N(i) \cap N(j)|}{|N(i) + N(j)|} \quad (6)$$

where the set $N(i)$ stipulates the neighbors of user (s_i). Next, we calculate the overall connection strength by contemplating both transition matrix ($P_T(i, j)$) and *connection trust* ($CT(i, j)$) using this:

$$W_e(i, j) = (\lambda \times P_T(i, j)) + (\beta \times CT(i, j)) \quad (7)$$

where, the controlling factor λ and β stabilizes the above two factors and $\lambda + \beta = 1$.

With the transition probability matrix and teleportation vector elucidated, the topic-specific temporal PageRank can be measured.

$$TR_T = \gamma W_e(i, j) \times TR_T + (1 - \gamma) E_T \quad (8)$$

where, TR_T is the user's topic-specific rank, E_T is the teleportation vector and γ is a parameter ($0 < \gamma \leq 1$) to authorizing the probability of teleportation.

3.2.3 Trending topical influential users detection algorithm and its overview

Algorithm 2 Trending Topical Influential Users Detection (TTIUD).

Input: $G = (U, E, A), \mathcal{I}, Q$

Output: top- r influential users

```

1: for each  $I_p \in \mathcal{I}$  do
2:    $Q \leftarrow \text{TOP\_k\_Trending\_TOPICS}(S, I_p, \alpha)$ 
3:   for each  $u_i \in U$  do
4:     compute  $\mu_{(u,v,\Psi_{uv})}$ 
5:     select users that follow user  $u_i$ , denote the set as
        $s_z$ 
6:     select users with common neighbors with  $u_i$ ,
       denote this set as  $s_{n_z}$ 
7:     for each  $u_x \in s_z$  do
8:       Compute  $P_T(u_i, u_x)$ 
9:     for each  $u_x \in s_{n_z}$  do
10:      Compute  $CT(u_i, u_x)$ 
11:      Compute  $W_e(u_i, u_x)$ 
12:   execute the Topic Specific temporal PageRank
       method
13:   Output the top- $r$  influential users at each time interval
        $I_p$ 
14: Procedure TOP_k_Trending_TOPICS( $S, I_p, \alpha$ )
15:    $L \leftarrow \text{PriorityList}(k)$ 
16:   for each  $T_j \in \mathcal{T}$  do
17:     compute  $\sigma_{(T_j, I_p)}$ 
18:      $L.add(\eta_{(T_j, I_p)})$ 
19: return top- $k$  results from  $L$ 

```

The proposed algorithm, called TTIUD, first identifies top- k trending topics at each time interval I_p through a procedure TOP_k_Trending_TOPICS (line 14-19). It computes and adds the trending score $\sigma_{(T_j, I_p)}$ for each topic T_j to a priority list of size k (line 16-18). It then returns the top- k topics based on their trending scores. Next, the algorithm computes the recency scores of each activity executed by the users U in G (line 3-4). It then calculates the transition probability $P_T(u_i, u_x)$, connection trust $CT(u_i, u_x)$, weight of the edge (connection) $W_e(u_i, u_x)$ (line 5-11). Finally, it makes the lists of the top- r influential users related to Q by performing the modified TwitterRank algorithm (line 12-13).

We analyze the time complexity of Algorithm 2. Both operations of finding common neighbors as well as set of users that follow user u_i take constant time i.e. $O(1)$ as our proposed model will record these information during the pre-processing of the dataset. The computations of

$P_T(u_i, u_x)$ and $CT(u_i, u_x)$ take $O(s_z)$ -time and $O(s_{n_z})$ -time, respectively. These steps will take $O(s_z + s_{n_z})$ -time for each user u_i and so, it will take $O(|U| * (s_z + s_{n_z}))$ -time for set of all users U (here, $|U|$ is the number of social users). Again, the time complexity to execute topic-specific temporal PageRank algorithm will take $O(pg * |U|)$ -time, where pg is the number of iterations. The procedure of finding trending topics i.e. $TOP_k_Trending_TOPICS(S, I_p, \alpha)$ takes $O(\mathcal{T})$ -time. As a result, the total time complexity at each time interval I_p is $O(\mathcal{T} + (|U| * (s_z + s_{n_z})) + (pg * |U|))$. So, the overall time complexity of Algorithm 2 is $O(\mathcal{I} * (\mathcal{T} + (|U| * (s_z + s_{n_z})) + (pg * |U|)))$.

4 Experimental evaluation

All the experiments have been performed on an Intel(R) Core(TM) i7-6500U CPU 2.50 GHz - 2.60 GHz, Windows 10 PC with 12 GB RAM and 240 GB SSD.

4.1 Dataset

Twitter Dataset: We have selected a Twitter Dataset from Stanford large network dataset (SNAP), which includes a collection of different social networks' data with thousands of nodes and edges, web graphs, communities, etc. [27]. To carry out our research, we casually choose 4,00,000 users and contemplate their tweets from June 1, 2009, to July 9, 2009.

DBLP Dataset: An academic Co-author network dataset named Database Systems and Logic Programming (DBLP) includes information about significant computer science publications [40]. Here, we specify the research papers those are published from 2003 to 2014. Table 1 shows the information of our experimental datasets.

4.2 Analyzing trending topics

Our proposed method identifies the trend in Twitter by analyzing recent tweets delivered by online users on Twitter. Identifying the dynamic trend of topics involves users' concerns at a specific time and topic. Table 2 represents the top five trending topics from June 15 to June 23 in 2009 with three days intervals. This table indicates that these

Table 1 Applied Datasets on Suggested Framework

Dataset	# of Nodes	# of Edges	# of Activities
SNAP	400,000	5,357,560	573,832
Co-author	331	1,213	850

Table 2 Overall Trending Topics in SNAP Dataset from June 15 to 23

#	Trending Topics in Different Time Interval		
	June (15-17)	June (18-20)	June (21-23)
1.	Iran Election	Iran Election	Iran Election
2.	Microsoft	Swine Flu	Economy
3.	Tornado	Microsoft	Swine Flu
4.	Swine Flu	Gay Marriage	Barack Obama
5.	iPhone	Tornado	Gaza

topics are the most popular on Twitter according to the public interest. About 81,599 tweets are posted regarding Iran Election in just nine days. From 15 June to 17 June, there are 36,905 tweets, 34,880 tweets and 12,213 tweets about *Iran Election*, *Microsoft* and *Tornado* respectively which are the top topics. In the second time interval, topic *Swine flu*'s trendy score has increased, and it shifts up to the second position. We can see that topic *Microsoft* dominates in the first two time intervals, but it's not included in the last time interval. There are also different topics (for example *iPhone*, *Barack Obama*, *Gay Marriage* etc.) included into our trending topics list. The words used for symbolizing trending topics are illustrated in Figs. 4, 5 and 6.

Table 3 represents the top three trending topics from 2003 to 2014 with four years interval in the Co-author dataset. According to research interest, this table indicates that these topics are the most popular among authors within these years.

4.3 Ranking of influential twitterers in trending topics

We detect users' of different topics during a specific time interval using both **TwitterRank** and **TTIUD** approach. Table 4 represents the top ten users from June 15 to June 17 in 2009. Here, the Query set is $Q = \{\text{Iran Election, Microsoft, Tornado}\}$ which are the Top-3 trending topics. We performed our experiment on those users who posted at least 20 tweets from the entire query set. From June 15 to June 17, more than ninety-nine thousand tweets have been published.

In Table 4, the top five influential twitterers are "Bob", "Ayran", "Lewis", "Ramini", and "Copperhead" identified in TTIUD approach. Here, "Bob" has secured the first position in both approaches. He mainly tweeted about "Iran election" in those three days. He has the highest number of mutual connections in the table. "Ramini" is in the second rank in TwitterRank. Still, it shifts down to the fourth position in the TTIUD approach because "Ramini" has fewer common neighbors with these users in the list than "Ayran" and "Lewis". Figure 7 represents the influential

Fig.4 Word Cloud of Topics from June 15 to June 17



Fig. 5 Word Cloud of Topics from June 18 to June 20

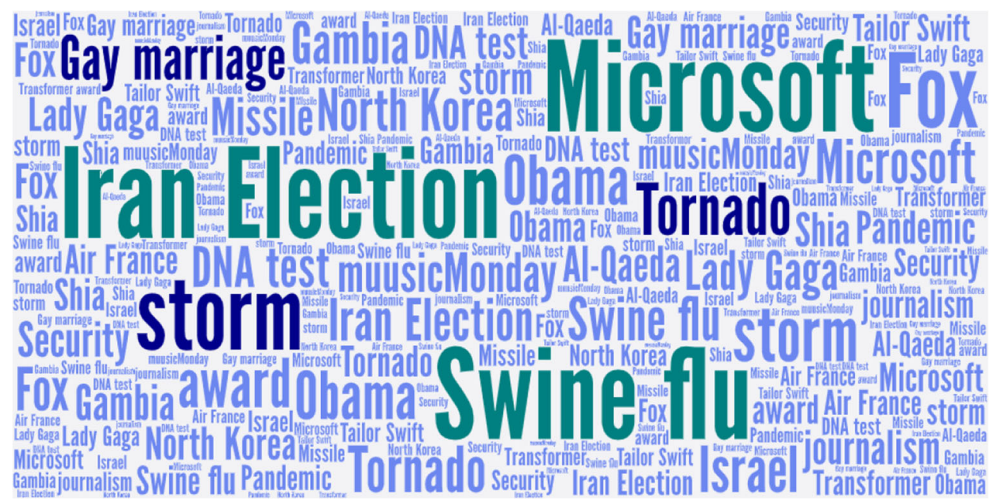


Fig. 6 Word Cloud of Topics from June 21 to June 23

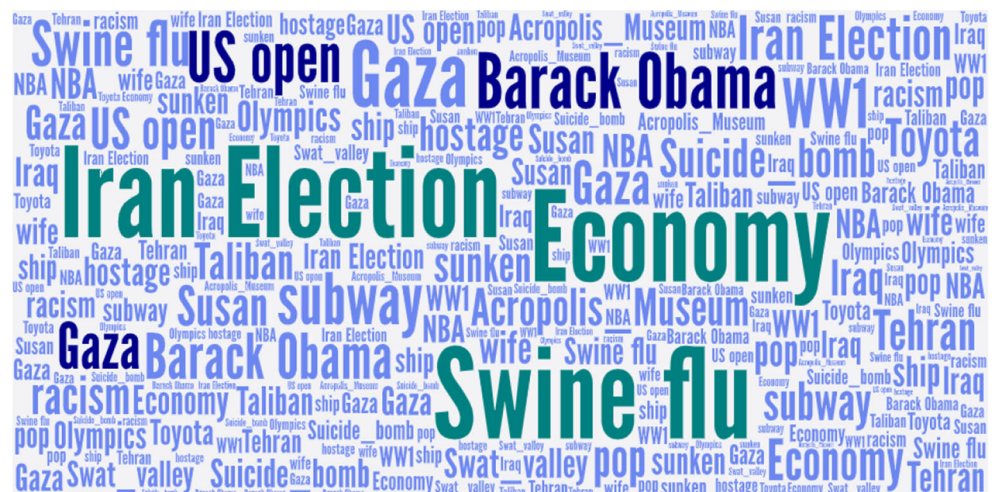


Table 3 Overall Trending Topics from 2003 to 2014 in Co-author Dataset

#	Trending Research Topics in Different Time interval		
	Year (2003-2006)	Year (2007-2010)	Year (2011-2014)
1.	Semantic Web	Machine Learning	Machine Learning
2.	Computer Graphics and Multimedia	Semantic Web	Social Network Analysis
3.	Machine Learning	Social Network Analysis	Semantic Web

Table 4 Top Influential Twitterers in Top-3 Trending Topics for the First Time Interval

Time Span	#	TwitterRank	TTIUD
15 June - 17 June	1	Bob	Bob
	2	Ramini	Ayran ↑
	3	Ayran	Lewis ↑
	4	Copperhead	Ramini ↓
	5	Lewis	Copperhead ↓
	6	Jack	Chris Williams1 ↑
	7	Alexandre	Jack ↓
	8	Chris Williams1	Ham ↑
	9	Jethro	Alexandre ↓
	10	Ham	Jeanpaul ↑

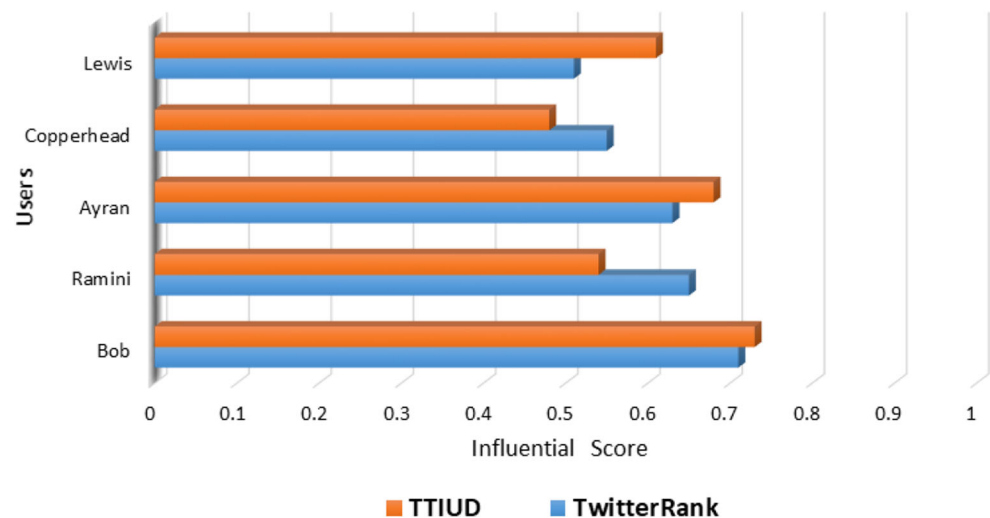
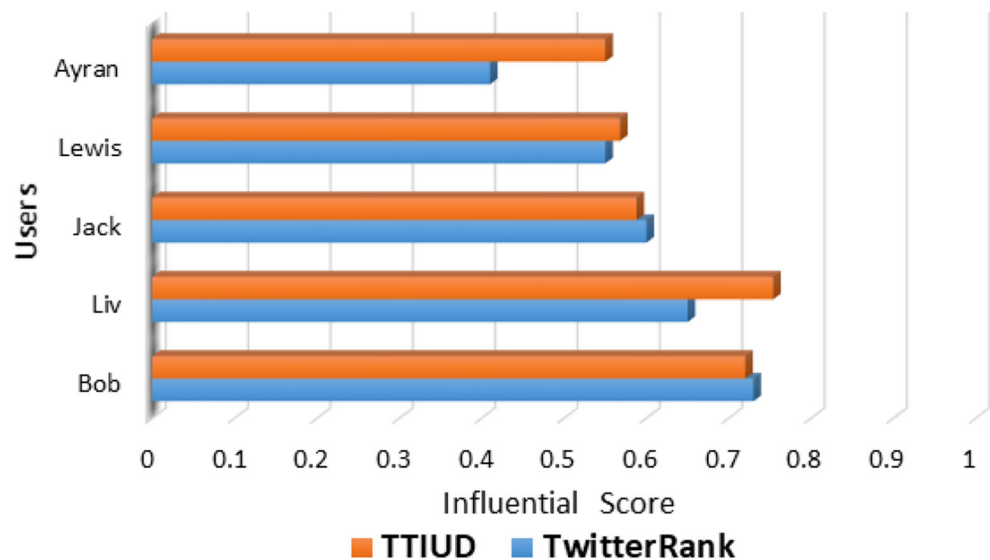
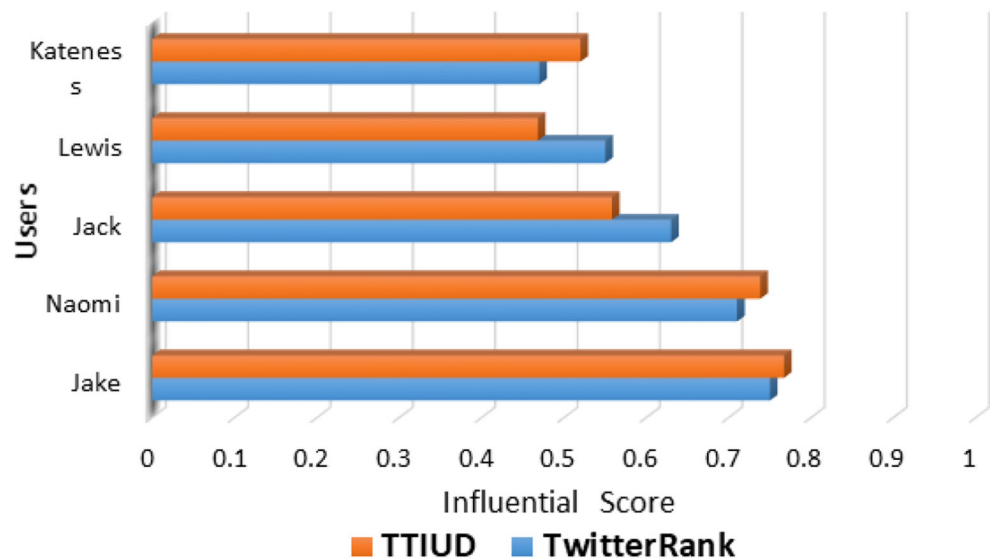
Fig. 7 Influential Score of Users for the First Time Interval

Table 5 Top Influential Twitterers in Top-3 Trending topics for Second Time Interval

Time Span	#	TwitterRank	TTIUD
18 June - 20 June	1	Bob	Liv ↑
	2	Liv	Mike ↑
	3	Jack	Bob ↓
	4	Lewis	Jack ↓
	5	Ayran	Lewis ↓
	6	Chris Williams1	Ayran ↓
	7	Jethro	Jethro
	8	Greg	Greg
	9	Mike	Chris Williams1 ↓
	10	Copperhead	Alexandre *

Fig. 8 Influential Scores of Users in Second Time Interval**Table 6** Top Influential Twitterers in Top-3 Trending Topics for Third Time Interval

Time Span	#	TwitterRank	TTIUD
21 June - 23 June	1	Jake	Jake
	2	Naomi	Naomi
	3	Jack	Jack
	4	Lewis	Kateness ↑
	5	Kateness	Lewis ↓
	6	Ramini	Ramini
	7	Jason Asbahr1	Jason Asbahr1
	8	Bob	Bob
	9	hazeee777	Daryn *
	10	Chris Williams1	hazeee777 ↓

Fig. 9 Influential Scores of Users in Third Time Interval

score of some top users from 15 June 2009 to 17 June 2009 using both TwitterRank and TTIUD.

Table 5 and Fig. 8 are denoted for the second time interval, which represents the ranking and scores of users from June 18 to June 20 in 2009, respectively. For this interval, the Query set is $Q = \{\text{Iran Election, Swine Flu, Microsoft}\}$. Here, “Liv” obtains the first rank in the TTIUD approach. He has forty-two mutual connections with “Mike” who ranks the second position in the list. He is also followed by “Bob”, “Jack” and “Lewis”. “Liv” has tweeted about politics and social events, for example (# Iran Election, # Gay rights). On the contrary, the number of tweets is lesser about Tech or entertainment-related news. The user named “Chris Williams1” ranks 6th position in the TwitterRank approach, but he shifts down to 9th position in the TTIUD method. Despite having followers followed by other influential users, he has tweeted less about the top topics (# Microsoft, # Swine Flu).

Similarly, Table 6 and Fig. 9 represent those users from 21 June to June 23 in 2009 where Query Set is $Q = \{\text{Iran Election, Economy, Swine Flu}\}$. Here, top-3 users are the same for both TwitterRank and TTIUD, but there is a new user named “Daryn” ranks the ninth position in TTIUD. “Daryn” mostly tweets about financial issues, economic management. He is followed by some influential twitterers as well. For example, “Daryn” is followed by “Kateness”

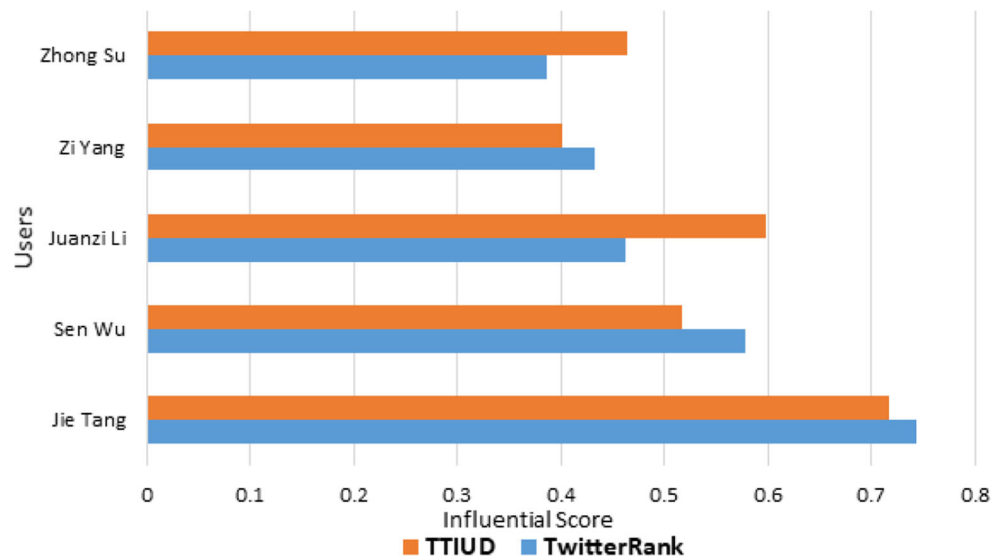
and they share 83 mutual connections. Again, “Kateness” is followed by “Jake”; who is followed by “Lewis” and “Naomi”.

Table 7 represents the top five users from 2003 to 2006. Here, the Query set is $Q = \{\text{Semantic Web, Computer Graphics and Multimedia, Machine Learning}\}$ which are the Top-3 trending topics. Here, “Sen Wu” is in the second rank in TwitterRank, but it shifts down to the third position in the TTIUD approach because “Sen Wu” has less common neighbors with these users in the list than “Juanzi Li”. The author named “Zhong Su” has thirteen familiar neighbors, where “Zi Yang” has only seven mutual neighbors. So, the position of “Zhong Su” goes up in the TTIUD approach because of having more common neighbors. Figure 10 represents the influential score of some top Co-authors from 2003 to 2006 using both TwitterRank and TTIUD.

Table 8 and Fig. 11 are denoted for the second time interval of the Co-author dataset, which represents the rank and score of authors from 2007 to 2010, respectively. For this interval, the Query set is $Q = \{\text{Machine Learning, Semantic Web, Social Network Analysis}\}$. “Wenbin Tang” has published more than six papers and has nine mutual connections. But, “Jibin Gong” has more common links with “Juanzi Li”; who has secured the second rank in TwitterRank and TTIUD method. So, “Wenbin Tang” gets the fourth position in TTIUD, but he is in the third position

Table 7 Top Influential Co-author in Top-3 Trending Topics from 2003 to 2006

Time Span	#	TwitterRank	TTIUD
2003 - 2006	1	Jie Tang	Jie Tang
	2	Sen Wu	Juanzi Li ↑
	3	Juanzi Li	Sen Wu ↓
	4	Zi Yang	Zhong Su ↑
	5	Zhong Su	Zi Yang ↓

Fig. 10 Influential Scores of Co-authors from 2003 to 2006**Table 8** Top Influential Co-author in Top-3 Trending Topics from 2007 to 2010

Time Span	#	TwitterRank	TTIUD
2007 - 2010	1	Jie Tang	Jie Tang
	2	Juanzi Li	Juanzi Li
	3	Wenbin Tang	Jibin Gong ↑
	4	Sen Wu	Wenbin Tang ↓
	5	Jibin Gong	Sen Wu ↓

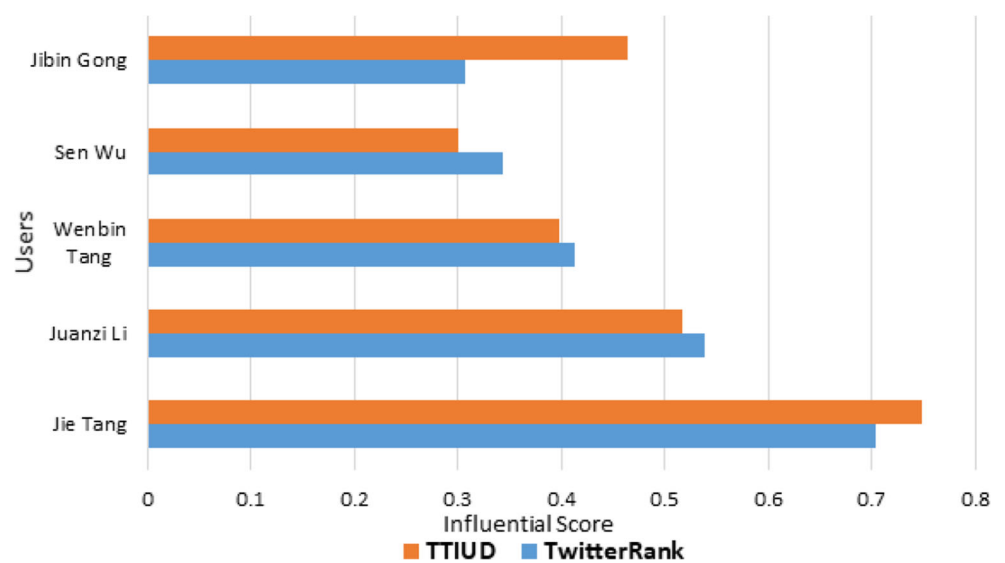
Fig. 11 Influential Scores of Co-authors from 2007 to 2010

Table 9 Top Influential Co-author in Top-3 Trending Topics from 2011 to 2014

Time Span	#	TwitterRank	TTIUD
2011 - 2014	1	Jie Tang	Jie Tang
	2	Juanzi Li	Juanzi Li
	3	Jibin Gong	Bo Gao ↑
	4	Sen Wu	Zhong Su ↑
	5	Bo Gao	Wenbin Tang ↑

in TwitterRank. We can also observe that the part of “Sen Wu” has gone down in TTIUD as it has fewer standard connections with other members of the dataset.

Here, Table 9 and Fig. 12 refers top five authors from 2011 to 2014 with their influential score. The Query set is $Q = \{\text{Machine learning Analysis, Social Network Analysis, Semantic Web}\}$. Here, the author “Bo Gao” has lifted his position from the fifth position in TwitterRank to the third position in the TTIUD approach.

4.4 Performance analysis

This section evaluates the performance of proposed algorithms according to “*recommendation*” task and also a correlation of rank lists between TTIUD and other approaches. Here, we have chosen three methods: TwitterRank, PageRank, and Topic-specific PageRank.

- **TwitterRank** which is proposed by Weng et al. to measure the influence of Twitterers considering the topic similarity and link structure of users connection [44].
- **PageRank** is applied to measure users’ influence based on the link structure of the social network stream [13].

- **Topic-specific PageRank** is quite similar to PageRank but it considers topical influence based on transition probability matrix [21].

4.4.1 Correlation

We use Kendall’s τ to measure the correlation between the rank lists obtained by different algorithms. It calculates the strength of their relationship. It ranges from -1 to 1. $\tau = 1$ represents that those lists are the same and have the most vital relationship. Reversely, $\tau = -1$ represents the weakest relationship between the two rank lists. If the lists are pretty independent of each other τ will be approximately 0. Table 10 indicates Kendall’s τ value of representing dependencies between various algorithms. We experimented on SNAP [27], our used Twitter dataset, and the time interval was considered from June 15 to June 17. We found that the three top trendings (sequentially) of that time is *Iran Election, Microsoft and Tornado*.

Here, TwitterRank is indicated as TR, PageRank is indicated as PR, Topic-specific PageRank is indicated as TSPR. This table and the figure show that TR and TSPR have higher concurrence with TTIUD than with PR in 3 different lists of trending topics.

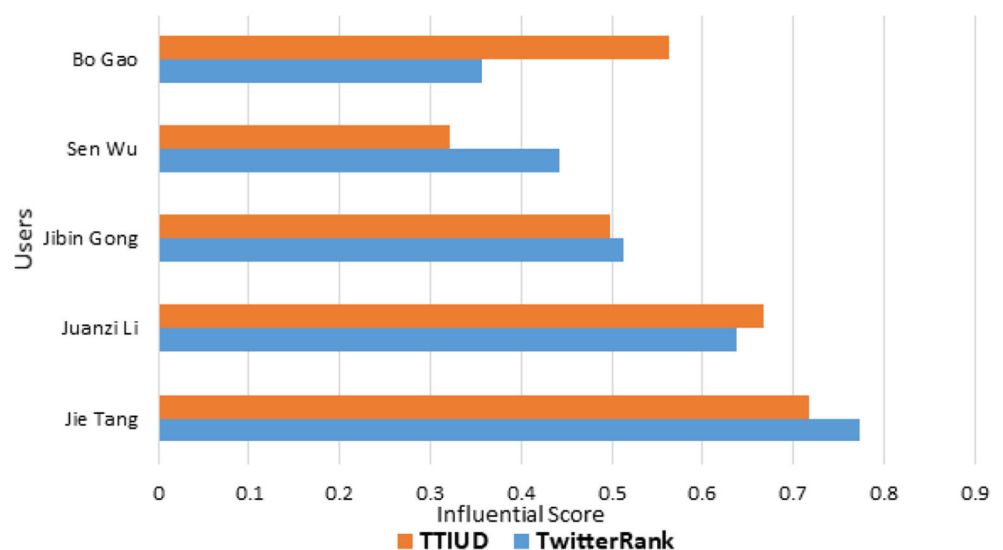
Fig. 12 Influential Scores of Co-authors from 2011 to 2014

Table 10 Correlation Between Rank Lists by Four Algorithms

Algorithms	Kendall's τ Value		
	Trending Topic #1 (Iran Election)	Trending Topic #2 (Microsoft)	Trending Topic #3 (Tornado)
TTIUD vs. TR	0.676	0.651	0.684
TTIUD vs. TSPR	0.556	0.592	0.538
TTIUD vs. PR	0.621	0.637	0.672

4.4.2 Performance in recommendation task

The assessment of Trending topical Influential Users Detection (TTIUD) is performed based on the recommendation task. L_t is the set of “following” or “common neighbor” relationship formed among twitterers and $l_o \in L_t$ where l_o is a rank list recommended by any of algorithms. The quality of recommendation denotes by $R(A)$.

$$R(A) = |\{t_i | t_i \in T_s, \text{ and } l_o(t_i) < l_o(t_f)\}| \quad (9)$$

In this (9), t_i is a twitterer who belongs to a new social network (T_s) which doesn't contain any common neighbors or any following relationship, $l_o(t_i)$ and $l_o(t_f)$ represents the rank of t_i in l_o and the rank of user (t_f) who belongs to “following” or “common neighbors” social network. The performance of any algorithm is inversely proportional to the value of $R(A)$.

There are four standard cases to generate “ L_t ” in anticipation of analyzing the proposed algorithm's performance.

- (i) **Case 1:** Based on t_f 's number of *followers*, two “ L_t ” lists can be denoted: P_{FH} and P_{FL} . P_{FH} and P_{FL} indicate t_f 's high follower count and low follower count respectively.

- (ii) **Case 2:** Depending on t_f 's number of tweets, two lists P_{TH} and P_{TL} are formed. Here, TH and TL are the thresholds for a high number of tweets and a low number of tweets, respectively, in a specific time interval.
- (iii) **Case 3:** Considering topical similarities, two lists P_{DS} and P_{DL} are generated in which DS and DL represent the high topical similarities and low topical similarities between users and followers.
- (iv) **Case 4:** Two lists P_{CH} and P_{CL} can be initiated considering the impact of common neighbors. CH and CL represent the high number and less number of common neighbors of the user.

Figure 13 points out the effectiveness of four algorithms in recommendation tasks. Our method Trending Topical Influential Users Detection (TTIUD), has outranked almost all cases except P_{TH} and P_{TL} . In the case of P_{TL} , TwitterRank (TR) has the worst performance because of the lower similarity of topics extracted by the LDA method. Our approach (TTIUD) has performed significantly better except PageRank (PR) because of the recency score and more advanced topic distillation method (T-LDA). In the context of P_{TH} , Topic-specific PageRank (TSPR) achieves better results due to using an identical transition probability

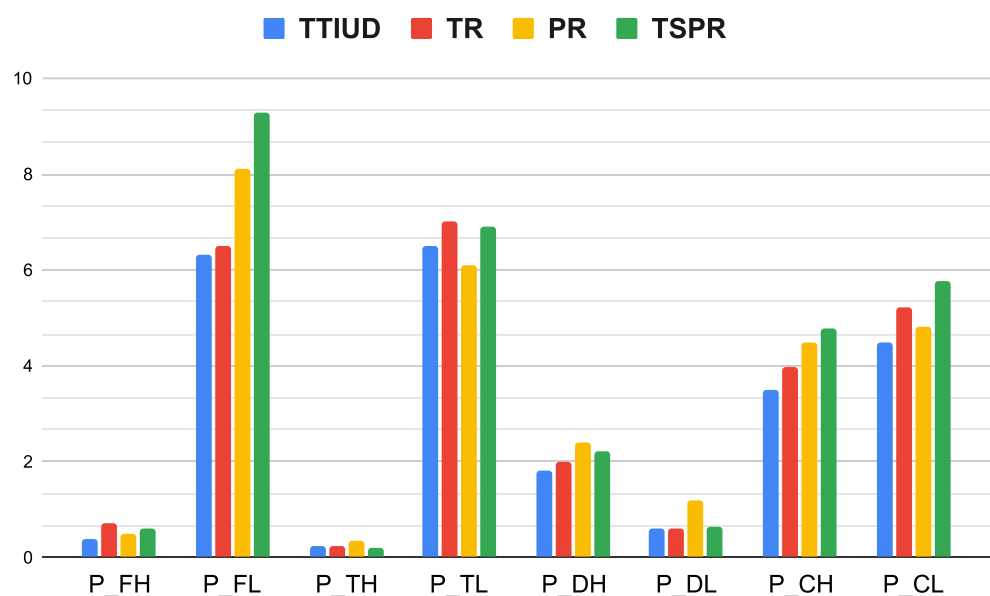
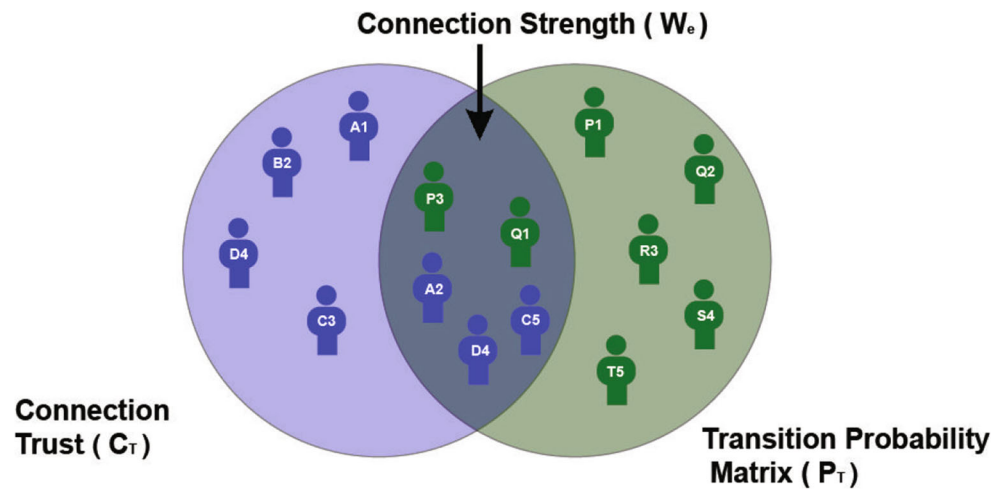
Fig. 13 Comparison of Performance (measured by $R(A)$) in Recommendation Task

Fig. 14 Selected Influential Users through TTIUD Approach



matrix when calculating topic-specific ranks. In scenarios of P_{CH} and P_{CL} , TTIUD outperforms all other algorithms because all the methods ignore the effect of familiar neighbors while measuring influence score.

5 Discussion

The majority of our trending topic-related queries are extracted from large amounts of textual data over time. The change of trending questions is reasonably expected because of dynamic public reaction on a specific topic. Queries that achieve lower rank have less favored hashtags or are challenging to combine with particular expressions. After identifying the most trendy issues from tweets based on trending scores, we have computed which Twitterer is influential. We have observed temporal degree of topical interests, effects of familiar neighbors emphasizing connection trust, and transition matrix.

Our framework displays a positive outcome where indicators based on the number of recent tweets and followers proved not consistently accurate. The novelty of our proposed framework is that the parameters we have computed are not solely dependent on Twitter signals (that depict both the user and the user's tweets). Here, some users who have most of the recent tweets show variability in the case of connection trust ($CT(i, j)$). Ranking according to connection strength ($W_e(i, j)$) is rationally different from ranking solely based on absolute numbers of followers. Generally, these two indicators display a different set of influential users over a particular topic in a specific time interval. After correlating these two parameters, our proposed approach (TTIUD) has proved that it can detect the most efficient influential Twitterers at a specific time.

In Fig. 14, the results have obtained when searching for influential users of a specific trending topic. Here, the result shows that ranking users only based on the

number of followers represented by transition probability matrix ($PT(i, j)$), five users have been selected. User P1 is the most influential one; user Q2 is the second, and T5 ranks last. Again, ranking users based on connection trust ($CT(i, j)$) represents different influential Twitterers. At this point, A1 is the topmost influential Twitterer. The parameter, connection strength (W_e) from our proposed framework (TTIUD), considers the user's direct-indirect connections and the transition probability matrix. Observing (W_e) influential score, the ranking is quite different. According to the algorithm, the significant users are Q1, A2, P3, D4, and C5. A, C, and D are present in the set where the user's connection is considered, but we can see that users P and Q are missing there, and they belong to the other circle. In our approach, Q1 dominates all other users, and D4 has rank alike in connection trust ($CT(i, j)$). Again, C3 has ranked the third position considering one particular index ($CT(i, j)$), but, in our approach, it goes down to the fifth position represented by C5. It suggests more reasonable and accurate results than the one obtained by procedure dependent on the sole indicator of data. Our approach has successfully managed to point out more accurate results capturing different aspects of influence on people.

6 Conclusion and future work

Social Network entirely depends on the users and the activities where users are the core content generators. Detecting social influence accurately provides enormous benefits for the functionality of viral advertisement or product design. Through this research, we find the most influential users on a specific topic on any social network. We have suggested a strategy for identifying query-oriented topical significant social users. We not only focus on users' influence on their direct connections but also consider familiar neighbors. Moreover, we also propose a way to

detect the top-k trending topics. We also consider the impact of familiar neighbors. From the outcome, we obtain that direct or indirect relations between users and followers can change the rank drastically. In the meantime, we observe how the users' topical interest vary on online social network over time and different query topics. Our method's performance is measured by Kendall's τ which calculates the strength of multiple compared algorithms. We have also analysed the effectiveness of our algorithms' via a recommendation task and drawn comparison lists with another related algorithm. The potency of our proposed method has been exemplified by extensive experiments on two real benchmark datasets. Although the outcomes are entirely satisfactory, there are some limitations to our proposed method. We did not observe the misinformation or confusing hashtags in our data. So, the effect of misinformation can encounter a significant change in the identification process.

In the future, we plan to work on domain-sensitive ranking. We also target improving the correlation between communication recurrence and other micro-interactions (mentions/shares) between users. We also intend to work on dynamic network graphs considering geographical locations to detect influential real-time users and their communities. For predicting rating, we want to use linearity and non-linearity in user interaction. How geographical location can impact the influence of the user is also in our plan.

Appendix

In our proposed framework, we have used multiple mathematical parameters to calculate equations. All the parameters are listed in Table 11.

Table 11 Parameters Used in Proposed Method

Parameters Name	Symbol
Top trending topics	k
Input query	Q
Recency Score	μ
Individual User	u_i
Number of tweets on a topic	Ψ_{u_i}
Selective time point's set	t_n
Quantity of time after an activity	$age(u_i, \Psi_{u_i}, t_n)$
Topic trending score	$\sigma(T_j, I_p)$
Time Interval	I_p
Individual topic	T_j
Total number of activities related to a specific topic	$\mu(u_i, T_j, t_n)$
Total number of tweets in a time interval	NT_{I_p}
Transition matrix	P_T
Similarity between two users	$sim_T(i, j)$

Table 11 (continued)

Parameters Name	Symbol
Row-normalized matrix	DT'
Probability matrix of followers topical interest	DT'_{jt}
Teleporation vector	E_T
Normalized matrix of DT'	DT'_T
Connection trust	$CT(i, j)$
Neighbors set	N
Connection strength	$W_e(i, j)$
Controlling factor	λ and β
User's Topic-specific temporal Rank	TR_T
Authorizing factor	γ

Availability of data and material The data that support the findings of this study are openly available in [27] and [40].

Code Availability The code implemented in this paper is described in Sections 3.2.3 and 3.2.3. Detailed code can be found in <https://github.com/ron352/Trending-TIUD>.

References

1. Abu-Salih B, Chan KY, Al-Kadi O, Al-Tawil M, Wongthongtham P, Issa T, Saadeh H, Al-Hassan M, Bremie B, Albahlal A (2020) Time-aware domain-based social influence prediction. *J Big Data* 7(1):10
2. Al-garadi MA, Varathan KD, Ravana SD (2017) Identification of influential spreaders in online social networks using interaction weighted k-core decomposition method. *Physica A: Statistical Mechanics and its Applications* 468:278–288
3. Alp ZZ, Ögüdücü ŞG (2018) Identifying topical influencers on twitter based on user behavior and network topology. *Knowl-Based Syst* 141:211–221
4. Alshahrani M, Fuxi Z, Sameh A, Mekouar S, Huang S (2020) Efficient algorithms based on centrality measures for identification of top-k influential users in social networks. *Inf Sci* 527:88–107
5. Antonakaki D, Fragopoulou P, Ioannidis S (2021) A survey of twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Syst Appl* 164:114006
6. Anuar NIM, Mohamad SR, Zulkiffli WFW, Hashim NAAN, Abdullah AR, Rasdi ALM, Hasan H, Abdullah T, Deraman SNS, Zainuddin SA et al (2020) Impact of social media influencer on instagram user purchase intention towards the fashion products: The perspectives of students. *European Journal of Molecular & Clinical Medicine* 7(8):2589–2598
7. Anwar MM (2020) Query-oriented temporal active intimate community search. In: *Australasian database conference*. Springer, pp 206–215
8. Anwar MM, Liu C, Li J (2019) Discovering and tracking query oriented active online social groups in dynamic information network. *World Wide Web* 22(4):1819–1854
9. Aurpa TT, Khan F, Anwar MM (2020) Discovering and tracking query oriented topical clusters in online social networks. In: *2020 IEEE Region 10 symposium (TENSYP)*. IEEE, pp 1054–1057
10. Statistics and twitter (2020) <https://www.b2bmarketingzone.com/statistics/twitter/>

11. Belhadi A, Djenouri Y, Lin JCW, Cano A (2020) A data-driven approach for twitter hashtag recommendation. *IEEE Access* 8:79182–79191. <https://doi.org/10.1109/ACCESS.2020.2990799>
12. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *The Journal of machine Learning research* 3:993–1022
13. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems* 30(1–7):107–117
14. Choudhery D, Leung CK (2017) Social media mining: prediction of box office revenue. In: *Proceedings of the 21st international database engineering & applications symposium*, pp 20–29
15. Das BC, Anwar MM, Bhuiyan MAA, Sarker IH, Alyami SA, Moni MA (2021) Attribute driven temporal active online community search. *IEEE Access* 9:93976–93989
16. Dencik L, Hintz A, Carey Z (2018) Prediction, pre-emption and limits to dissent: Social media and big data uses for policing protests in the united kingdom. *New Media & Society* 20(4):1433–1450
17. Dewan A (2021) Detecting organic audience involvement on social media platforms for better influencer marketing and trust-based e-commerce experience. In: *Data analytics and management*, pp. 661–673. Springer
18. Dhali A, Gomasta SS, Mohanta S, Anwar MM (2020) Identification of query-oriented influential users in online social platform. In: *2020 IEEE Region 10 symposium (TENSymp)*. IEEE, pp 973–976
19. Feng Y, Chen H, Kong Q (2020) An expert with whom i can identify: the role of narratives in influencer marketing. *Int J Advert*, 1–22
20. Han B, Cook P, Baldwin T (2013) Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology (TIST)* 4(1):1–27
21. Haveliwala TH (2002) Topic-sensitive pagera7frfdddikznk, 2002. In: *Proceedings of the 11th association for computing machinery international conference on world wide web (ACM)*, pp 517–526
22. 36 twitter statistics all marketers should know in 2021 (2021). <https://blog.hootsuite.com/twitter-statistics/>
23. Javed Awan M, Mohd Rahim MS, Nobanee H, Munawar A, Yasin A, Zain AM (2021) Social media and stock market prediction: a big data approach
24. Jianqiang Z, Xiaolin G, Feng T (2017) A new method of identifying influential users in the micro-blog networks. *IEEE Access* 5:3008–3015. <https://doi.org/10.1109/ACCESS.2017.2672680>
25. Kemp S (2021) Digital 2020: July global statshot - datareportal – global digital insights. <https://datareportal.com/reports/digital-2020-july-global-statshot>
26. Kumar S, Singhla L, Jindal K, Grover K, Panda B (2021) Im-elpr: Influence maximization in social networks using label propagation based community structure. *Appl Intell*, 1–19
27. Leskovec J, Krevl A (2014) Snap datasets: Stanford large network dataset collection
28. Li H, Zhang R, Zhao Z, Liu X, Yuan Y (2021) Identification of top-k influential nodes based on discrete crow search algorithm optimization for influence maximization. *Appl Intell*, 1–17
29. Li Y, Li R, Xiong X, Gu X, Liang T, Xu M, Yuan Y (2021) Multi-topical authority sensitive influence maximization with authority based graph pruning and three-stage heuristic optimization. *Appl Intell*, 1–19
30. Lin Y (2021) 10 twitter statistics every marketer should know in 2021 [infographic]. <https://www.oberlo.com/blog/twitter-statistics>
31. Mandal S, Maiti A (2021) Deep collaborative filtering with social promoter score-based user-item interaction: a new perspective in recommendation. *Appl Intell*, 1–26
32. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27(1):415–444
33. Mittal D, Suthar P, Patil M, Pranaya P, Rana DP, Tidke B (2020) Social network influencer rank recommender using diverse features from topical graph. *Procedia Computer Science* 167:1861–1871
34. Othman J, Vanathas C (2017) Topic familiarity and its influence on listening comprehension. *The English Teacher*, 14
35. Pick M (2020) Psychological ownership in social media influencer marketing. *European Business Review*
36. Sasaki K, Yoshikawa T, Furuhashi T (2014) Online topic model for twitter considering dynamics of user interests and topic trends. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp 1977–1985
37. Sekulić I, Strube M (2020) Adapting deep learning methods for mental health prediction on social media. *arXiv:2003.07634*
38. Shi T, Ding S, Xu X, Ding L (2021) A community detection algorithm based on quasi-laplacian centrality peaks clustering. *Appl Intell*, 1–16
39. Subroto A, Apriyana A (2019) Cyber risk prediction through social media big data analytics and statistical machine learning. *J Big Data* 6(1):1–19
40. Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z (2008) Arnetminer: extraction and mining of academic social networks. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp 990–998
41. Tankovska H (2021) Twitter: monthly active users worldwide. <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>
42. Wang F, Li J, Jiang W, Wang G (2017) Temporal topic-based multi-dimensional social influence evaluation in online social networks. *Wirel Pers Commun* 95(3):2143–2171
43. Wang N, Luo F, Shvrtare Y, Badal VD, Subbalakshmi K, Chandramouli R, Lee E (2021) Learning models for suicide prediction from social media posts. *arXiv:2105.03315*
44. Weng J, Lim EP, Jiang J, He Q (2010) Twitterrank: finding topic-sensitive influential twitterers. In: *Proceedings of the third ACM international conference on Web search and data mining*, pp 261–270
45. 40 twitter statistics marketers need to know in (2020). <https://www.wordstream.com/blog/ws/2020/04/14/twitter-statistics>
46. Zareie A, Sheikahmadi A, Jalili M (2019) Identification of influential users in social networks based on users' interest. *Inf Sci* 493:217–231
47. Zhang J, Tang J, Li J, Liu Y, Xing C (2015) Who influenced you? predicting retweet via social influence locality. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 9(3):1–26
48. Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H, Li X (2011) Comparing twitter and traditional media using topic models. In: *European conference on information retrieval*. Springer, pp 338–349
49. Zheng C, Zhang Q, Young S, Wang W (2020) On-demand influencer discovery on social media. In: *Proceedings of the 29th ACM international conference on information & knowledge management*, pp 2337–2340

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Sarmistha Sarna Gomasta graduated from Jahangirnagar University in Bangladesh in 2022 with a B.Sc. degree in Computer Science and Engineering. She is currently pursuing Master of Science (M.Sc.) in Computer Science and Engineering at Jahangirnagar University, where she also works as a research assistant at the JU DM RG lab. She has produced or coauthored several research publications that have been published in prestigious IEEE confer-

ences and journals. Her research interests include data mining, deep learning, blockchain technology, and bioinformatics.



Md Musfique Anwar received the Ph.D. degree from the Swinburne University of Technology, Australia, in 2018. He is currently an Associate Professor at Jahangirnagar University, Bangladesh. His research interests include data mining, social network analysis, natural language processing, and software engineering.



Aditi Dhali earned a B.Sc. in Computer Science and Engineering from Jahangirnagar University in Bangladesh. She is presently working as a part-time research assistant in a R&D based Software Company named Pioneer Alpha and an active member of the JU DM RG lab while completing her M.Sc. in Computer Science and Engineering. She has coauthored multiple research papers that have appeared in renowned IEEE journals and international conferences.

Data mining, natural language processing, deep learning, and blockchain technology are among her research interests.



Iqbal H. Sarker received his Ph.D. in Computer Science and Software Engineering from Swinburne University of Technology, Melbourne, Australia in 2018. His professional and research interests include Data Science, Machine Learning & AI, Data-Driven Cybersecurity and Threat Intelligence, Smart Cities, Systems, and Security. He has published over 100 research papers including TOP-ranked Journals and Conferences with reputed sci-

entific publishers like Elsevier, Springer Nature, IEEE, ACM, Oxford University Press, etc. Moreover, he is an author of a research monograph book titled “Context-Aware Machine Learning and Mobile Data Analytics: Automated Rule-based Services with Intelligent Decision-Making”, published by Springer Nature, Switzerland, 2021. Based on his research interests, Dr. Sarker established “Sarker DataLAB”, a research platform of Advanced Technologies. Dr. Sarker also received the honor of the world’s TOP 2% research scientist listed by Elsevier and Stanford University, USA, 2021. He is a member of ACM and IEEE.