ISSN 2563-7568

RESEARCH ARTICLE

TTRank: A Temporal Model to Rank Online Twitter Users

Shifat Jahan Setu[1], Tahmina Islam[1], Md Al-Amin Bhuiyan[2], Md Musfique Anwar[1*]

[1]*Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh*
[2]*Department of Computer Engineering, King Faisal University, Eastern Region, Saudi Arabia*

Abstract

Twitter is an online social network or a news media where users can post their desirable topical interests in the form tweets. This is a networking model where each user can choose who can follow her and whom she wants to follow. We can find the users who are very active in the social networks and consider them as influential users. This research addresses on Temporal Twitter Ranking (TTRank) to rank the influential users in Twitter. We apply Twitter-LDA topic modeling method to find the users' topical interests. The time interval is an important factor as users' topical interest can change over time i.e. users' have different degree of topical interests at different time interval. So we give more emphasize on users' most recent tweets. Our proposed approach also considers the impact of "Follower Influence" and "Retweet Influence". The top influential users have been detected across different time intervals based on all the above mentioned factors and classified as "Highly Influential" and "Potential' users. Experiment results on a real Twitter dataset demonstrate the efficacy of the proposed system.

Key Words: *Online social network; Twitter-LDA; Active users; Highly-influential users; Potential users*

***Corresponding Author**: Md Musfique Anwar, Department of Computer Science and Engineering, Jahangirnagar University, Dhaka – 1342, Bangladesh, Tel: +88-01752-311590; E-mail: manwar@juniv.edu*

***Received Date**: July 09, 2020, **Accepted Date**: August 31, 2020, **Published Date**: October 30, 2020*

***Citation**: Setu SJ, Islam T, Bhuiyan MAA, et al. TTRank: A Temporal Model to Rank Online Twitter Users. Int J Auto AI Mach Learn. 2020;1(1):42-53.*

1. Introduction

Due to the great evaluation of the Internet and its availability, communication has become easy. Social media is a platform of websites and applications where people can exchange and share their ideas and thoughts such as they can share consumer-generated content and interact with each other through websites like Twitter, Facebook, blogs, YouTube, wikis, and so on [1]. People can express their opinion or experiences in online social media in the form of messages (or tweets), images, videos etc. Therefore, the social network establishes a sort of relationship with people, organizations, and other communal bodies. Hacking, cyber-bullying, fraud, and scams are some of the malicious issues that also appear on online networking sites.

Nowadays, Twitter is one of the popular social tools and a fusion of texting, blogging, and social media. It has been reported that the number of active users on Twitter is estimated to reach 275 million, monthly worldwide [2]. Twitter contains information on marketing, medical campaigns, election campaigns, e-governance, event monitoring, sentiment analysis etc [1]. Twitter users can tweet in 280 (earlier was 140) characters or less. The size limitation makes the tweets interesting and challenging. Twitter is a low-cost service for any type of person. In Twitter, who receive other people's tweets are called followers and the following someone means you will see Twitter updates in your timeline. Twitter employs a social-networking model called "following", in which each Twitter user is allowed to choose who she wants to follow without seeking any permission. Conversely, she may also be followed by others without granting permission first. In one instance of "following" relationship, the Twitter user whose updates are being followed is called the "friend", while the one who is following is called the "follower". As a consequence, those they are following in their timeline profile, their followers can see all tweets of the order of arrival [3].

In this work, we focus on finding influential users based on their topical degree of interests and follower-following relationships. Twitter users usually actively participate on their topics of interest. Existing approaches to find influential users in OSN's ranging from simply counting the immediate neighbors to complicated machine learning and message passing techniques [4,5]. Our observation is that users' have different degree of topical interests on different topics which vary widely over time i.e. they performed different actions and the number varies.

The main components that compromises Twitter are the users, followers, network model and also on the hashtags used in the tweets [6]. The number of followers and the number of following play vital roles for finding the influential users. Users and the followers being influential if they are active during a time interval. Hence, we consider follower influence (FI) and retweet influence (RI) as the two most important factors to rank the influential users. We also show the comparison in the rank list of influential users between the above two mentioned factors. Again, we categorize the top influential users highly influential and potential users based on their performances considering all the factors.

The remainder of the paper is organized as follows: Section 2 highlights the relevant works. Section 3 addresses the proposed methodology. Section 4 describes the experimental evaluation and Section 5 draws the overall conclusion.

2. Related Work

Recently, enough attention has been paid to find the influential users in online social networks. Ma et al. proposed PageRank based temporal influence ranking TIR model [7] that performs a temporal analysis on the patterns of the users' activities find the dominant users. Some existing works applied topic modeling approach such as LDA (Latent Dirichlet Allocation) [8,9] to find topic specific influential users. Katsimpras et al. [9] applied random walk model to identify topic senstitive influential users. Other research works considered the sentiments of users that mentioned in the tweets on the trending topics to rank the top user influential users [5,10-12].

Cha et al. [13] accumulated a large amount of data from Twitter and compared the three measures of influence: in degree, retweets, and mentions. Similarly, Kwak et al. [14] used three measures: number of followers, PageRank algorithm-based scores on the following /follower network, and the number of retweets. Authors in [15] proposed FLDA method which integrates both content topic discovery and social influence analysis in the same generative process. Sendi et al. considered temporal topics of interest derivation based on belief-function and aging theories to discover influential users [16]. Zhao et al. TwitterRank estimates the dominance considering both the similar topic among Twitter users and the structure of the users' account links [17].

Gupta et al. proposed a model WTF [18], which means who to follow to provide recommendation services and serves that one can search for people who already follow. In [19] and [20], the authors address the problem of temporal interaction biased community detection combining with influence propagation model and a time interval parameter. Fang et al. [21] proposed spatial-aware community search. Li et al. [22] studied the problem of personalized influential topic search in social networks. Their goal is to find how important topics and influential users might be better factors to meet a specific user's information. Li et al. [23] defined a novel problem of maximum geographic spanning regions over location-aware social networks, which takes a query region, a budget k of seed selection, and a locally minimal covering ratio ρ as parameters.

Du et al. [24] studied on microblog posts to extract the hot topics using MF-LDA method considering the large number of shares, comments, and likes. They did consider only hot topics, though the non-hot topic is also important for research work. Patil et al. [25] proposed a topic modeling approach known as Fuzzy LDA to find topic-based sentiment analysis on social media. Some recent works focused on finding topic oriented active users and group those active users into different clusters [26,27].

3. Proposed Framework

This research proposes the TTRank model which is designed to categorize the most influential users within Twitter. Follower influence and retweet influence are two indicators that have considered to rank the influential users. TTRank model shows the user influence in two categories-highly influential users and potential users. The working flow diagram of TTRank model is depicted in Figure 1.

3.1. Social Stream

Social users performs actions (such as posting tweets in Twitter, publishing research papers in coauthor network) at different time points. All these actions collectively known as social stream which is a continuous and temporal sequence of users' activities at different time intervals.

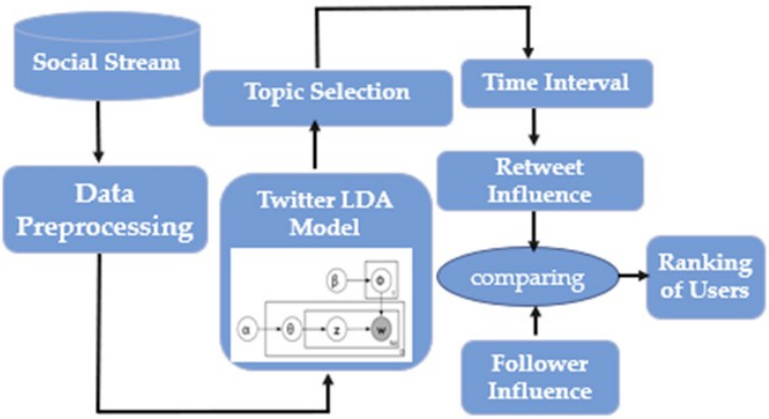


Figure 1: *Working flow diagram of the proposed methodology.*

3.2. Data Preprocessing for Topic Detection

In general, tweets are informally written and often contain grammatically incorrect sentence structures with misspellings and non-standard words. Tweets containing many non-standard forms (e.g., toook for took, goooood for good), informal abbreviations (e.g., tmrw for tomorrow, wknd for weekend), phonetic substitutions (e.g., 4eva for forever, 2day for today) etc. Table 1 shows example of Twitter slang words. In order to improve the quality of our tweet corpus and the performance of the subsequent steps, we performed normalization of the tweets through direct substitution of lexical variants with their standard forms with a normalization lexicon proposed by Han et al. [28].

Twitter users often publish spam/noisy tweets which are unrelated to their interest. Noisy tweets must be filtered from the datasets because they can lead to false analysis. We check the structure of the tweets, and filter out tweets that have more than 2 user mentions or more than 2 hashtags, or less than 4 text tokens. The idea behind this structure-based filtering is that tweets that have many user mentions or hashtags, but lack enough clean text features, do not carry any topic-like content, or are generally very noisy. This step filters many noisy tweets. Before applying topic modeling approach, all the words from the abstracts of the research papers are converted into a seed word (stemming word) for example: plays, playing, etc.-> play by using Lucene 4.9.0 Java API.

3.3. Topic Modeling with Twitter-LDA Method

Twitter users often use hashtags (for example, #Obama, #Ronaldo etc.) to indicate topics of the tweets. Use of hashtag is optional and there are no specific standard rules of using hashtag to mention a topic. As a result, it is difficult for any system to correctly extract topics from hashtags.

To retrieve the textual content from a tweet, Twitter-LDA (T-LDA) [29,30] an effective extension of LDA is used here for topic distillation. Twitter-LDA is better in topic semantic coherence by presuming the ratio between topic and background words is indifferent for each user's tweets. The graphical representation of T-LDA is shown in Figure 2.

Table 1:
Example of slang words in Twitter.

English words	Slang words
today	todaaayyy, toodaay, twoday, tody, todaayyy, todaaay, today
network	netwrk, ntwork, networ, netwerk, netwrks, network
good	gooda, goodi, ggoood, gud, gooodd, goooddd, ggod
crazy	crazyyy, craaazy, craaazzzyyy, crrrazy, crayzay, craazyy
search	serach, seach, serch, searc, searh, srch

T-LDA is based on the assumptions as follows: Each tweet is indicated by a topic number in T-LDA and contains a set of related words. Every individual user's topical interest φu is represented by a distribution over k topics. Each word in the tweet assigned by topic k is generated from a topic word distribution θk. The latent value y determines whether the word is a background word or a topic word.

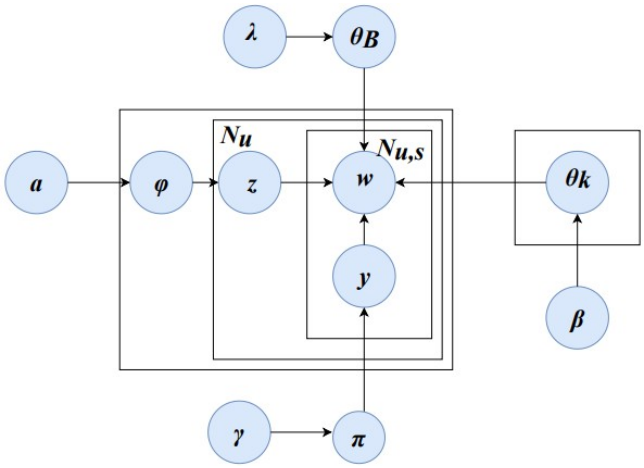


Figure 2: *Graphical Representation of Twitter-LDA Model.*

Generally, not all the past activities of a user are equally important. We assign greater importance to user's most recent activities by a measure called recency score. We divide each time interval into two equal halves. For each Twitter user, we consider 50% of her tweets posted in the first half of a time interval and all the tweets posted in the second half of the time interval. For example, if a user X posted 10 tweets in first half and 5 tweets in other half in a particular time interval, then the number of tweets that will be considered is 10 (5+5) in our proposed model. Again, if another user Y posted 4 tweets in first half and 9 tweets in other half in a particular time interval, then the number of tweets that will be considered for Y will 11 (2+9). Although user Y has posted less number of tweets (13) compared with user X (15), but Y has more number of recent tweets and hence, Y will be considered more active than X.

3.4. Impact of Follower Numbers and Retweet Factor

We consider two different factors such as the follower influence and retweet influence which are described briefly in below.

3.4.1. Follower Influence (FI)

Follower means a platform that increased connectivity through the actual user (active in a time interval). We consider a user as influential when she has more followers because a tweet posted by her will be broadcasted to a large number of her followers. For example: (User A has a higher follower than B but B has a large number of following. So, based on our consideration B has more influence than A user). People who follow one, there has a possibility that they may not see his or her tweet (called inactive followers) whereas following a user has a smaller chance to see all the tweets he/she follows. So, we observed that there are two elements to direct Follower Influence (FI):

- The number of followers
- Follower to following ratio

3.4.2. Retweet Influence (RI)

A retweet is another action that users often perform on Twitter which is a re-posting of a tweet. Twitter's Retweet feature helps users to quickly share that a tweet with all of their followers. So, retweeting is a broadcast way that can make a user famous. For example, we consider X, Y, Z are three Twitter users. User X tweets "HappyWomen Day2019" and user Y sees the tweets and retweets it. Then it will appear to user Y's followers as "RT@X HappyWomen Day2019" and again if one of Y followers Z retweets it then it appears as "RT@Y@X HappyWomen Day2019". From this scenario, the influence of X is 1 or 2 which is depending on the Y (Number of followers and the total number of retweets of the user). Retweet is a factor that plays a role to prejudice such tweets since it has been witnessed that the users usually retweet their messages that they follow. This kind of relationship is also known as "homophily", which means a user retweets just to do a favor and not because of the tweet content [3], [2]. We consider the following things to measure the follower influence and retweet influence:

i. Calculate follower influence according to specific topic.
ii. Calculate retweet influence.
iii. Record the three parameters FI, RT, and time.
iv. Identify the top scoring users.

We compute FI_U and RI_U of the users in accordance with their activities and from these two factors, we can select the influential users independently.

Table 2:
Notations used in this work paper.

Notation	Description
U	User who tweets T
T	User U tweets at time t_i
${}_aF_U$	Number of followers that follow user U
${}_afw$	Number of following ${}_afw_i$ that follow of user
FI_U	Follower influence indicator of user U
RI_U	Retweet influence indicator of tweet T

3.4.3. Follower Influence Indicator (FI_U)

Our proposed approach takes into account the following parameters to measure the impact of follower influence FI_U:

- Let $u_1,.....,u_n$ is the number of followers of users U
- Let aFw_i be the number of a follower for $i = 1,.....,n$, where w_i is the following numbers of user U.
- HFu is the highest number of followers. We assume that w_i have the possibility of reading a tweet coming from U which is the inverse with aFw_i.
- uHw_i is the highest number of following w_i

$$FI_U = \frac{\sum aFw_i}{HF_U}\alpha + \frac{\sum w_i}{uHw_i}(1-\alpha) \quad (1)$$

Equation 1 indicates the follower influence $FI_U$where a is a weighting parameter that balances two factors such as the number of followers and follower to following ratio.

3.4.4. Retweet Influence Indicator (RI_U)

Retweet indicator RI_U is used to find the users who shared (retweeted) the tweet T posted by user U as shown in Equaiton 2.

$$RI_U = \frac{RI_T}{aF_U} \quad (2)$$

Where, FI_U is the followers of U who retweeted tweet T.

4. Experimental Result

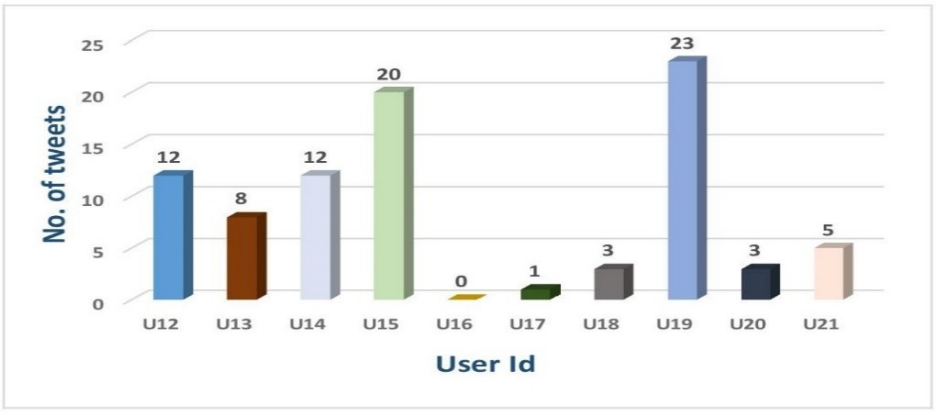
This section presents the results which are obtained by our proposed method. We use a Twitter dataset named SNAP [31]. SNAP contains more than 400 million Twitter posts from the month of June, 2009 to the month of December, 2009. We randomly choose 7,288 users and consider their tweets from June 01, 2009 to June 30, 2009. We first remove the noisy information (such as users' activities that indicate an invalid topic) from the dataset. Next, we divide the above period into two-

time intervals such as (1-15 i.e. June 01, 2009 to June 15, 2009) days as the first interval and (16-30 i.e. June 16, 2009 to June 30, 2009) days as the second interval.

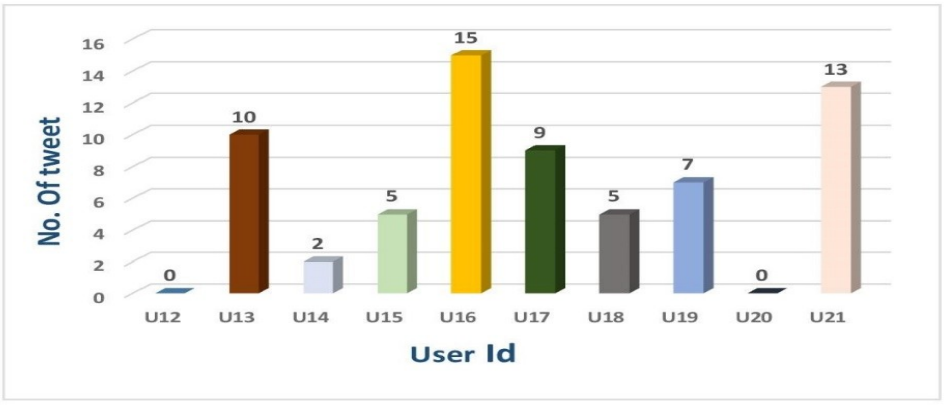
Figure 3 illustrates the most active users at different time intervals such as I1 and I2. User U19 has the most number of tweets 23 at I1 whereas U16 has the highest number of tweets of 15 at I2. User U16 has no activity at I1and in I2, user U12 and U20 didn't post any tweet. So, we can say that users have different degree of activess at different time intervals.

Table 3 lists the values of the parameters such as $aFw_i, HF_U, W_i, UH_{wi}$ related to follower influence FI_U.

The ranking of the influential users based on retweet influence factor RI_U is shown in Table 4.



(a) First Time Interval I1(1-15)



(b) Second Time Interval I2 (16-30)

Figure 3: *Most active users at different time intervals.*

4.1. Ranking Comparison between FI_U and RI_U

Follower influence FI_U and retweet influence RI_U are the two indicators that are used to find and rank the influential users. Table 5 shows the comparison result between these two indicators.

In the Table 5, a is a weighting variable whose value can be $0 < a \leq 1$. We choose three different values of a such as 0.3, 0.5 and 0.7. Bold values indicate the top 3 influential users in each result

columns. We avoid the original users' names for privacy issues and instead we assume some random user IDs.

Table 3:
Follower Influence FI_U table that is used for ranking influential users.

User ID	aFw_i	HF_U	W_i	UH_{wi}
12	1206	4560	355	3435
13	689	4560	408	3435
14	885	4560	106	3435
15	50	4560	7	3435
16	211	4560	425	3435
17	3440	4560	1535	3435
18	567	4560	135	3435
19	2212	4560	637	3435
20	4560	4560	3435	3435
21	4060	4560	356	3435

Table 4:
Ranking of active or non-active users based on Retweet Influence RI_U.

Time Interval	User Id	RI_T	aF_U	RI_U	Rank List
(1-15)	12	15	1206	0.012	5
	13	0	689	0	-
	14	52	885	0.059	3
	15	10	50	0.2	1
	16	39	211	0.185	2
	17	53	3440	0.015	4
(16-30)	12	23	1206	0.019	3
	13	37	689	0.054	1
	14	9	885	0.010	5
	15	0	50	0	-
	16	5	211	0.024	2
	18	0	3440	0	-

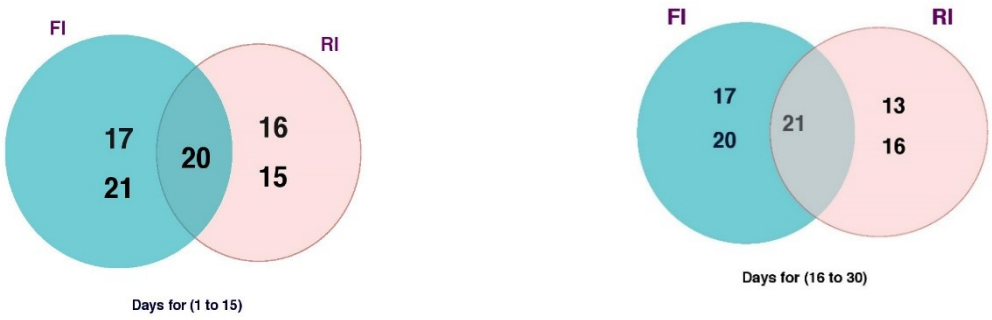
4.2. Highly Influential and Potential Users

In the proposed approach, we also categorized users into highly influential and potential users. If a user is in the top-3 rank for both indicators (i.e. considering follower influence and retweet influence), we mark that user as a highly influential user and others as potential users.

Table 5:
Comparison of Top-ranking users between two indicators (FI_U and RI_U).

User ID	Time Interval I1 (Days 1 - 15)				Time Interval I2 (Days 16 - 30)			
	FI_U $(\alpha=0.3)$	FI_U $(\alpha=0.5)$	FI_U $(\alpha=0.7)$	RI_{Ui_1}	FI_U $(\alpha=0.3)$	FI_U $(\alpha=0.5)$	FI_U $(\alpha=0.7)$	RI_{Ui_2}
12	0.137	0.173	0.216	0.014	0.132	0.145	0.186	0.021
13	0.119	0.143	0.162	0.0	0.118	0.132	0.149	**0.049**
14	0.089	0.089	0.187	0.061	0.057	0.097	0.156	0.012
15	0.008	0.012	0.009	**0.21**	0.011	0.017	0.008	0.0
16	0.110	0.097	0.081	**0.183**	0.127	0.077	0.078	**0.026**
17	**0.547**	**0.711**	**0.745**	0.017	**0.389**	**0.428**	**0.481**	0.0
18	0.068	0.081	0.119	0.0	0.098	0.085	0.097	0.013
19	0.305	0.367	0.461	0.002	0.165	0.254	0.307	0.011
20	**0.837**	**0.869**	**0.904**	**0.173**	**0.782**	**0.832**	**0.833**	0.005
21	**0.423**	**0.542**	**0.781**	0.023	**0.389**	**0.422**	**0.562**	**0.043**

In Figure 4 (a), we see that user ID 20 is listed in top-3 influential users considering both FI_U and RI_U at time interval I1 and hence, she is the highly influential user and other users such as users' ID of 15, 16, 17 and 21 are considered as potential users. Similarly, at time interval I2, user ID 21 is considered as highly influential user and others such as users' ID of 13, 16, 17 and 20 are the potential users as shown in Figure 4 (b).



(a) Days (1-15) *(b) Days (16-30)*

Figure 4: *Ranking of top-3 influential users considering FI_U (a = 0.3) and RI_U.*

6. Conclusion

In this research work, we proposed a model to rank the influential users in micro-blogging services. This is a networking model where each user can choose who can follow her and whom she wants to follow. Our observation is that users' influences depend on their followers and following numbers. We also categorized influential users as highly influential and potential users on the basis of two factors such as follower influence and retweet influence. In future work, we like to include the users' topical degree of interests in our proposed model to rank topic specific active influential users at different time intervals.

Conflict of interests: The authors declare that there is no conflict of interest.

References

1. Abulaish M, Kamal A, Zaki MJ. A survey of figurative language and its computational detection in online social networks. Acm T Web. 2020;14:1-52.
2. Weng J, Lim EP, Jiang J, et al. Twitterrank: Finding topic-sensitive influential. Proceedings of the Third ACM International Conference on Web Search & Data Mining, New York. 2010.
3. Montangero M, Furini M. Trank: Ranking twitter users according to specific topics. 2015 12th annual IEEE consumer communications and networking conference (CCNC), Las Vegas, NV, USA. 2015.
4. https://arxiv.org/abs/1608.02519
5. Lee C, Kwak H, Park H, et al. Finding influentials based on the temporal order of information adoption in twitter. Proceedings of the 19th international conference on World wide web, Raleigh, North Carolina, USA. 2010.
6. Garton L, Haythornthwaite C, Wellman B. Studying online social networks. J Comput Mediat Commun. 1997;3.
7. https://www.researchgate.net/publication/314263010_Finding_Influentials_in_Twitter_A_Temporal_Influence_Ranking_Model
8. Du Y, Yi YT, Li XY, et al. Extracting and tracking hot topics of micro-blogs based on improved Latent Dirichlet Allocation. Eng Appl Artif Intel. 2020;87:103279.
9. Katsimpras G, Vogiatzis D, Paliouras G. Determining influential users with supervised random walks. Proceedings of the 24th International Conference on World Wide Web, Florence Italy. 2015.
10. Furini M, Montangero M. TSentiment: On gamifying twitter sentiment analysis. 2016 IEEE Symposium on Computers and Communication (ISCC), Messina, Italy. 2016.
11. https://www.aclweb.org/anthology/S16-1001/
12. https://www.aclweb.org/anthology/S17-2088/
13. Cha M, Haddadi H, Benevenuto F, et al. Measuring user influence in twitter: The million-follower fallacy. fourth international AAAI conference on weblogs and social media, Washington. 2010.
14. Kwak H, Lee C, Park H. What is Twitter, a social network or a news media?. Proceedings of the 19th international conference on World wide web, Raleigh North Carolina USA. 2010.
15. Bi B, Tian Y, Sismanis Y, et al. Scalable topic-specific influence analysis on microblogs. Proceedings of the 7th ACM international conference on Web search and data mining, New York USA. 2014.
16. Sendi M, Omri MN, Abed M. Discovery and tracking of temporal topics of interest based on belief-function and aging theories. J Amb Intel Hum Comp. 2019;10:3409-425.
17. Zhao WX, Jiang J, Weng J, et al. Comparing twitter and traditional media using topic models. European conference on information retrieval, Dublin, Ireland. 2011.
18. Gupta P, Goel A, Lin J, et al. Wtf: The who to follow service at twitter. Proceedings of the 22nd international conference on World Wide Web, Rio de Janeiro Brazil. 2013.

19. Gupta A, Lamba H, Kumaraguru P, et al. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. Proceedings of the 22nd international conference on World Wide Web, Rio de Janeiro Brazil. 2013.

20. Alduaiji N, Datta Amitava, Li J. Influence propagation model for clique-based community detection in social networks. IEEE Trans Comput Soc Syst. 2018;5:563-75.

21. Fang Y, Cheng R, Li X, et al. Effective community search over large spatial graphs. Proc Vldb Endow. 2017;10:709-20.

22. Li J, Liu C, Yu JX, et al. Personalized influential topic search via social network summarization. IEEE T Knowl Data En. 2016;28:1820-34.

23. Li J, Sellis T, Culpepper JS, et al. Geo-social influence spanning maximization. IEEE T Knowl Data En. 2017;29:1653-66.

24. Du Y, Yi YT, Li XY, et al. Extracting and tracking hot topics of micro-blogs based on improved Latent Dirichlet Allocation. Eng Appl Artif Intel. 2020;87:103279.

25. Patil H, Palwe S. Fuzzy LDA for Topic Modeling: An Overview. CLIO An Annu Interdiscip J Hist. 2020;6:213-21.

26. Anwar M, Liu C, Li J. Discovering and tracking query oriented active online social groups in dynamic information network. WWWJ. 2018;22:1-36.

27. Anwar M, Liu C, Li J. Uncovering attribute-driven active intimate communities. 29th Australasian Database Conference, Australia. 2018.

28. Han B, Cook P, Baldwin T. Lexical normalization for social media text. ACM Trans Intell Syst Technol. 2013;4.

29. Dedeoglu BB, Taheri B, Okumus F, et al. Understanding the importance that consumers attach to social media sharing (ISMS): Scale development and validation. Tour Manag. 2020;76:103954.

30. Zhang H, Wheldon C, Dunn AG, et al. Mining Twitter to assess the determinants of health behavior toward human papillomavirus vaccination in the United States. J Am Med Inform Assoc. 2020;27:225-35.

31. Leskovec J, Krevl A. Snap datasets: Stanford large network dataset collection. 2014.