Applied Intelligence (2023) 53:5998–6024
https://doi.org/10.1007/s10489-022-03831-7

Topic-based influential user detection: a survey

Rrubaa Panchendrarajan[1] · Akrati Saxena[2]

Accepted: 30 May 2022 / Published online: 5 July 2022

Abstract

Online Social networks have become an easy means of communication for users to share their opinion on various topics, including breaking news, public events, and products. The content posted by a user can influence or affect other users, and the users who could influence or affect a high number of users are called influential users. Identifying such influential users has a wide range of applications in the field of marketing, including product advertisement, recommendation, and brand evaluation. However, the users' influence varies in different topics, and hence a tremendous interest has been shown towards identifying topic-based influential users over the past few years. Topic-level information in the content posted by the users can be used in various stages of the topic-based influential user detection (IUD) problem, including data gathering, construction of influence network, quantifying the influence between two users, and analyzing the impact of the detected influential user. This has opened up a wide range of opportunities to utilize the existing techniques to model and analyze the topic-level influence in online social networks. In this paper, we perform a comprehensive study of existing techniques used to infer the topic-based influential users in online social networks. We present a detailed review of these approaches in a taxonomy while highlighting the challenges and limitations associated with each technique. Moreover, we perform a detailed study of different evaluation techniques used in the literature to overcome the challenges that arise in evaluating topic-based IUD approaches. Furthermore, closely related research topics and open research questions in topic-based IUD are discussed to provide a deep understanding of the literature and future directions.

Keywords Topic-based influential user detection · Topic modeling · Influence mining · Online social network

1 Introduction

The advent of online social networks (OSN) has dramatically changed the way of communication between people, especially in producing, consuming, and sharing information. People tend to share real-time perceptions and interests on various topics, including events, products, and services. The shared information is quickly available to the connected users referred to as either *followers* or *friends* and they can further react to it in the form of like, comment, or share with their *followers* and *friends*. This enables rapid dissemination of information on the OSN, and users can influence or be influenced by other users. This aspect of analyzing the influence in OSN is called social influence analysis [26] and those who could influence many other users directly or indirectly in OSN are referred to as influential users [100].

In real-life, users' opinions often get influenced by their neighbors as they rely on the opinion of their neighbors due to the trust and bias they develop over time [90]. Therefore, identifying influential users in OSN has been a vital study among both, the research and business community, to analyze and predict the social trends, individual preferences, and community behaviors. This provides several opportunities for many real-life practical applications including maximize the spread of products [110], propagate the political agendas [34], spread scientific messages [57], news [42], and health awareness [71, 118].

The content posted by a user on OSN may discuss one or more topics of his or her interest ranging from political,

✉ Akrati Saxena
a.saxena@tue.nl

Rrubaa Panchendrarajan
rrubaa.p@sliit.lk

[1] Faculty of Computing, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka

[2] Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

finance, economy, or any sociological areas in real-time. Thus the influence made by the user through the content can be associated with the discussed topic, and users' influence usually varies in different topics due to their expertise and interest. Therefore, it is essential to analyze the user influence at the topic level and determine the topic-based influential users. This can be considered as a granular study of influential user detection problem, and a vast amount of interest has been shown towards inferring topic-based influential users in the literature [13, 26, 37, 38, 64, 69, 78, 98, 111, 125].

The research in topic-based influential users has been dependent on two following research topics, (i) topic detection techniques, and (ii) the existing influential user detection approaches. However, it adds another level of complexity to the problem due to how granular studies have been performed. Topics are extracted from the content posted by the users mainly using the medium of text. Whereas the user influence on this particular topic can result in various activities of influenced users in the OSN, such as following, reposting, liking, commenting, mentioning, publishing similar content, etc. Therefore, we require effective solutions to jointly analyze the content as well as the behaviors of the users in OSN to accurately infer the topic-based influential users.

In this survey, we perform a comprehensive review of existing topic-based influential user detection approaches with critical perspectives of challenges and limitations associated with these techniques. We present the taxonomy in topic-based influential user detection to understand the existing works and to identify the limitations and open research challenges. We also discuss various evaluation techniques used to assess the proposed approaches as well the used datasets. We further present a detailed study of a wide range of closely related research problems, and the open research problems to provide a depth insight into the existing techniques and the future trends.

To the best of our knowledge, this is the first work that presents a detailed study of existing topic-based influential users. In some of the existing surveys, researchers have reviewed related research topics. Riquelme et al. [87] presented a survey summarizing techniques that can be used to measure the activity, popularity and influence of users in OSN. In 2018, Jain et al. [45] presented characterization of opinion leaders whose opinion can influence or shape other's opinion in OSN. The authors further briefly discussed the techniques used to identify the opinion leaders. Later in 2019, a methodological review of detecting opinion leaders in social networks was presented by Bamakan et al. [10]. A similar study to determine influential users in OSN was presented by Ishfaq et al. [43] and Al-Garadi [4]. A recent review published in 2020 by Al-Yazidi et al. [5] presented techniques used to measure both the reputation and influence of users in OSN. The paper discusses features that are important to measure the reputation and influence, importance of measuring reputation and influence and finally the approaches that can be used to measure the reputation and influence. Different from these existing surveys, we focus on the study of identifying topic-based influential users and present a detailed review and comparison of the existing approaches.

The remainder of the paper is organized as follows. Section 2 briefly discusses the preliminaries to understand the problem domain following the presentation of taxonomy of topic-based influential users detection in Section 3. Sections 4 and 5 present a detailed study of techniques used in the literature to infer the influential users for a single topic and multiple topics, respectively. Section 6 talks about the correlation analysis carried out in the literature as a post-processing step to analyze the topic-based influential users. A detailed study of evaluation methods used for topic-based influential user detection is discussed in Section 7. Section 8 briefly introduces closely related research problems and Section 9 is devoted to future directions and open research problems. Finally, we conclude this review in Section 10.

2 Preliminaries

Social network mining has been a vast active area of research over the past two decades. In this section, we briefly discuss the background knowledge related to influence analysis and topic modeling in online social networks. We introduce the traditional IUD approaches, the propagation models, and random walk models that are widely used to rank nodes in a social network. Further, we briefly discuss the topic modeling and Latent Dirichlet Allocation (LDA), which is a traditional topic modeling approach used to infer topics from a collection of documents.

2.1 Social network mining

2.1.1 Influence in online social networks

In online social networks (OSNs), users post *microblog* that contain information in the form of text, images, audio, video, or multimedia. Once a user posts a microblog, it is available to all friends or followers of the user based on the working structure of the online social network. The friends and followers can further share this information with their friends and followers or react to the post in the form of a like or comments. Any of these actions carried out by friends or followers are being resulted due to the direct influence created by the user. When the friends and followers start sharing the information in OSN, it starts an information

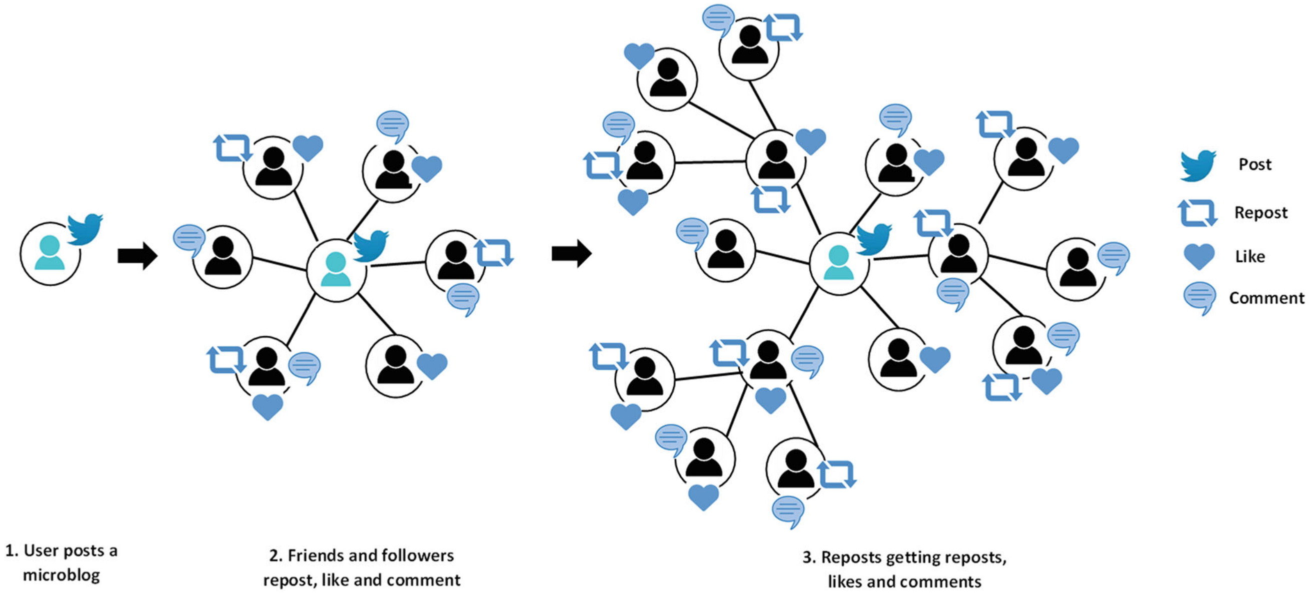

Post
Repost
Like
Comment
1. User posts a microblog
2. Friends and followers repost, like and comment
3. Reposts getting reposts, likes and comments

Fig. 1 Information and influence propagation in OSN

propagation on the network. The amount of propagation depicts the indirect influence the user has created in OSN. Both direct influence on the friends and followers and indirect influence in the OSN contribute to how much a user is influential. This process is explained in Fig. 1.

2.1.2 Influence network

Given an OSN represented as a graph G with nodes V and edges E, a subset of the network that is being influenced can be represented as an influence network $G'(V', E')$ and the analysis or inference of influence can be performed directly in the influence network. The influence network can be created by considering users as nodes involved in the considered information propagation. The follower-followee relationship or friendship relationship between the two users or any reaction to a post in the form of repost or like or comment can be modeled as edges. These edges are generally directed towards the users who follow/react to another user. Further, the influence network can be weighted networks where the edge-weight to an edge (u, v) will be assigned based on how many times user v reacted to user u's posts.

2.1.3 Influence propagation models

Quantifying the contribution of a user for the influence propagation in the influence network can reveal more influential users. Researchers have modeled the influence propagation on social networks and have proposed a variety of influence propagation models. Some of the well-known models are Independent Cascade Model (ICM) [49], Linear Threshold Model (LTM) [48], and compartment models, such as Susceptible-Infectious-Recovered (SIR) model [40], Susceptible-Infectious (SI) model [6]. While executing the propagation model, at a given timestamp, each node in the network is in either an active or inactive state. The active nodes will influence their neigbours based on the execution dynamic of the influence propagation model. During the information propagation, the probability of a node becoming active might increase with time as more of its neigbours will become active.

These propagation models are used to formalize the influence maximization problem, where the goal is to find k nodes in the network, such that by activating these k nodes, we can maximize the expected number of nodes that eventually get activated when the influence propagates. The identified top k nodes indicate the top k users who are more likely to produce large influence-driven cascades in the given influence network.

2.1.4 Random walk algorithms

Random walk algorithms in graphs exploit a random path over the nodes, where the walker starts from a random node and travels to its neigbour depending on the transition probability defined between any two nodes in the network. The walk will be continued until the stop criteria is satisfied. These algorithms are applied to many real-world problems, including influential user detection, where the chances of

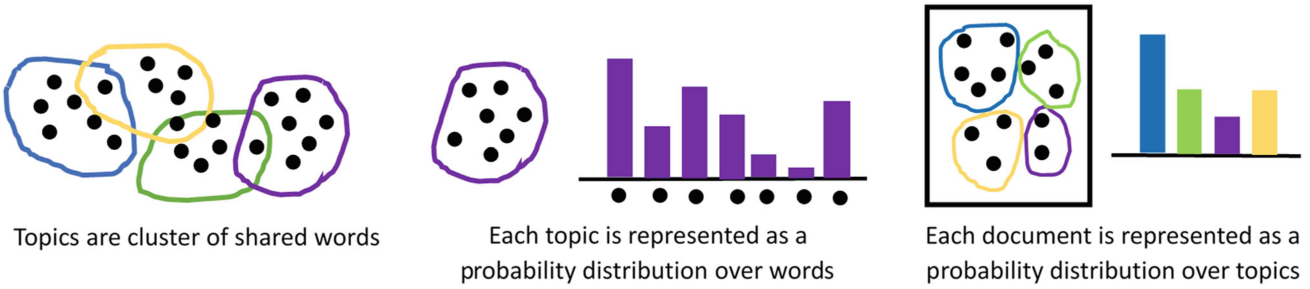

Topics are cluster of shared words
Each topic is represented as a probability distribution over words
Each document is represented as a probability distribution over topics

Fig. 2 Underlying idea of topic modeling

visiting a node over multiple random walks indicate how influential is the given node [114].

The most typical random walk algorithm in the computer science era is the PageRank [77] algorithm. The PageRank algorithm was introduced to compute the importance of web pages by randomly walking on the world wide web network based on the hyperlinks between the web pages. Given a transition probability matrix M_{NXN}, where M_{ij} indicates the probability of transiting from node i to j, N indicates the number of nodes in the graph and M^T indicates the transpose matrix of M, then the PageRank scores of all nodes at time $t+1$, R_{t+1} is computed as,

$$R_{t+1} = M^T * R_t \tag{1}$$

Extension of the above model called personalized PageRank algorithm [46] was proposed with damping factor and it is being widely used as a traditional solution for IUD,

$$R_{t+1} = (1-\alpha)M^T * R_t + \alpha * p \tag{2}$$

where α is a damping factor, and p is a personalized vector that reflects the importance of each node in a graph for a specific user. Here the damping factor solves the problem of reaching nodes without any outgoing or incoming edges during the random walk by adding a probability of randomly choosing a node instead of visiting the neighbour. After a certain number of iterations, eventually, the PageRank score R_{t+1} converges, and the converged scores indicate the influential score of each node for the problem of IUD. PageRank algorithm has been widely used with various topic detection techniques in the literature to determine the topic-based influential users [7, 7, 14, 23–25, 37, 39, 67, 68, 75, 84, 88, 96, 108, 111, 115, 124].

2.2 Topic modeling

Topic modeling is a statistical modeling approach to uncover topics from a collection of documents. They automatically find the word clusters that represent topics based on the co-occurrence of words. Latent Dirichlet Allocation (LDA) is a traditional topic model, which assumes topics as a distribution over unique words found in the document collection. Moreover, the documents are considered as a mixture of topics, that enables to associate each word in the document with a topic. Figure 2 shows the underlying idea of topic modelling. Let, there are K number of topics and V number of unique words found in a document collection D. Then the generation of these documents are modeled in LDA as follows,

1. For each topic $k \in [1, K]$, draw a distribution over words $\phi_k \sim Dir(\beta)$
2. For each document $d \in [1, D]$,
   (a) Draw a topic distribution $\theta_d \sim Dir(\alpha)$
   (b) For i^{th} position in document d,
      i Draw a topic $z \sim Multi(\theta_d)$
      ii Draw a word $w_i \sim \phi_z$

where, $Dir(\lambda)$ and $Multi(\lambda)$ are Dirichlet and Multinomial distributions, respectively, defined using the parameter λ. Here, the topic assignment for each word in a document z is referred to as the latent variable; and Gibbs Sampling [33] approach is widely used to learn the latent variables. Once the latent variables are learned, the topic distribution of a document θ_d, word distribution of a topic ϕ_k, and the dominant topic discussed in a document are computed using the number of times a word is being assigned to a topic in each document.

3 Taxonomy of topic-based IUD

In this section, we introduce the taxonomy of topic-based influential user detection that is further followed in the paper to present state-of-the-art research.

Single vs. multiple topics Content posted by a user can be associated with single or multiple topics, and similarly, an influential user may influence other users in OSN on one or more topics. While identifying influential users specific to a topic enables more detailed study about the topic-based influence, identifying users who are influential across multiple topics benefits a larger business community. This has enabled the research community to focus on identifying influential users specific to a single topic as well multiple topics.

Existing techniques in single or multiple topic-based IUD generally first determine the topic and then infer the influential users using the posts and activities related to the topic on OSN. It is observed that while common influence detection techniques are used in both single and multiple topic-based IUD, the techniques and challenges vary at the topic detection level. Single topic-based IUD approaches generally come up with a manual representation of the topic of interest, such as a set of keywords, query, set of users, etc., whereas multiple topic-based IUD approaches attempt to infer topics from a collection of posts overtime automatically. Inferring influential users in multiple topics further enables us to determine the overall influential users across all the topics. We present a detailed discussion on single topic-based IUD techniques in Section 4 and multiple topic-based IUD in Section 5.

Topic detection Single topic-based IUD approaches aim to infer the influential user of a known topic. The known topic is represented using a set of keywords [16, 18, 24, 26, 29, 53, 63, 68, 75, 76, 81, 88, 93, 106, 107, 124, 128, 129], a query [2, 7, 99, 115] or a set of users [65], which enable them to filter the posts that contains keywords or query words, or posted by representative users, respectively. Finally, influential users are inferred from posts and activities related to the topic of interest. While most of these approaches use manually hand-crafted keywords to represent a topic, a topic can not be represented by a limited number of keywords; hence the influential users determined will be limited to the keywords defined. Further, these techniques can not be adapted to infer multiple topics detection as coming up with a manual representation of all the topics discussed in a huge volume of posts is a cumbersome process.

While very few approaches have used manually hand-crafted keywords to define multiple topics, [14, 131], various other techniques, such as topic modeling [7, 9, 13, 20, 23, 28, 37, 39, 47, 60, 64, 67, 82, 84, 96, 98, 104, 108, 109, 111, 115, 117, 117, 121, 125, 130, 132], machine learning [85, 108], and platform structures [50, 126] are used to infer multiple topics in the literature. The way content is organized on a platform can be considered as a representation of a topic based on platform structures; e.g., a board on the Pinterest platform discusses a single topic. However, not all social media platforms automatically organize or group the content; therefore, this technique is limited to specific platforms. It is observed that topic modeling is widely used to determine multiple topics from a collection of posts. This is mainly due to their nature of effectively inferring topics from a huge collection of posts without the need for any labeled data or external knowledge. Further, they are very powerful in modeling features related to the content, e.g., features in OSN such as different relationships between users and activities, which makes it more suitable for extracting topics for the topic-based IUD problem. Further, the machine learning techniques are too explored in recent research works as an alternative to topic models in the presence of labeled data for supervised topic detection. Subsections of Sections 4 and 5 discuss various topic detection techniques used for single and multiple topics respectively.

Influential user detection In the literature, different techniques, such as statistical measures [18, 26, 29, 53, 63, 81, 93, 99, 106, 107], random-walk approaches [2, 7, 7, 14, 16, 23, 24, 37, 39, 65, 67–69, 75, 76, 84, 88, 96, 108, 111, 115, 124, 126, 128], propagation algorithms [28, 85, 108, 115], machine learning [25, 50, 129–131], topic modeling [9, 13, 20, 64, 82, 98, 104, 109, 117, 121], similarity measures [47, 117, 125], and optimization algorithms [113] are used for topic-based influential user detection. The Pagerank algorithm, i.e. a random-walk based approach, has been used among majority of the research works due to its powerful nature of automatically ranking nodes in a large network. Further they enable to incorporate the topic based information on OSN into various stages such as constructing topic-specific influence network, determining the topic-specific damping factor and determining topic-specific transition probability.

Besides this, statistical measures, such as the number of reactions to a topic-related post, are also used for influential user detection as this measure directly reflects the influence of the post. However, these approaches fail to capture the indirect influence made in the influence network. Recent works on topic-based IUD approaches use machine learning techniques to infer the influential users. However, the adaption of machine learning algorithms is limited to the availability of labeled data to perform supervised influence detection. As an alternative to it, topic models are used as the influence detection technique by jointly modeling the topics and influence on OSN.

Figure 3 presents the summary of existing topic-based influential user detection approaches organized according to the taxonomy. The rest of the sections are structured in line with the taxonomy for a better understanding of the existing techniques, their limitations and to identify future trends.

4 IUD for a single topic

While a user can be influential on a varying number of topics, the influence of a user has to be analyzed at a topic level, as any action of the user that creates the influence

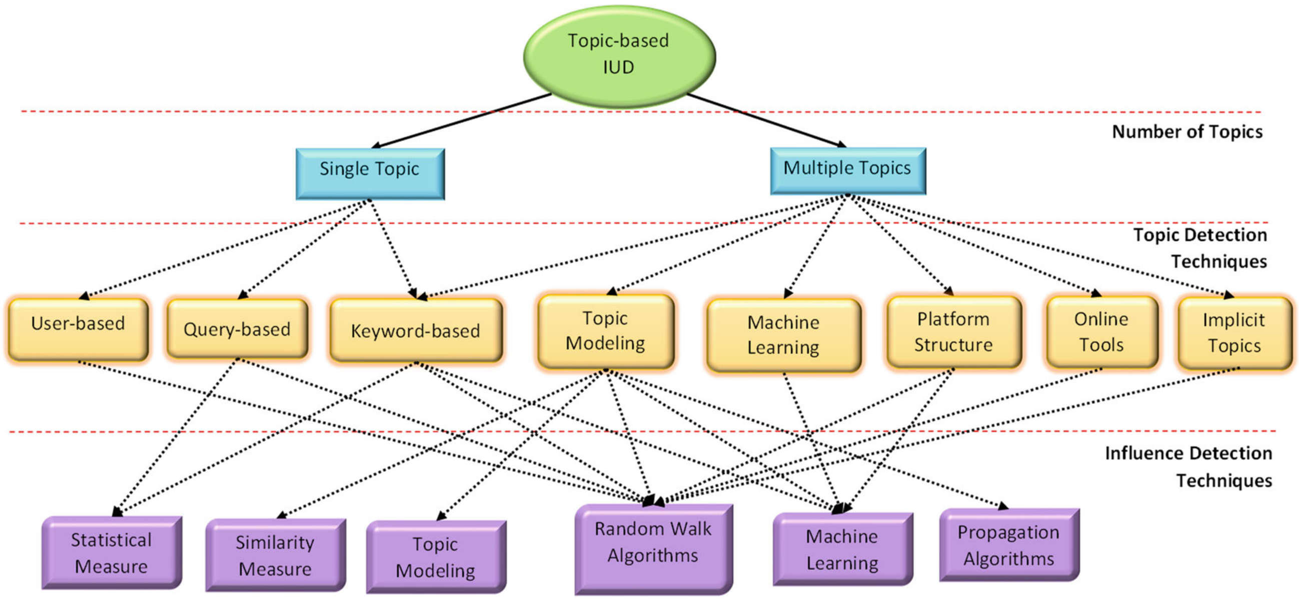


Fig. 3 Summary of topic-based influential user detection technique presented according to the taxonomy

in OSN can be associated with a topic. Such actions include publishing a post that discusses a specific topic and reposting, commenting, or liking a post that discusses that topic. This enables researchers to focus on IUD problems on a single topic that they are interested in instead of dealing with multiple topics. In this section, we discuss topic-based IUD approaches which infer influential users only for a single topic.

4.1 Keyword-based topic detection

Keywords are semantic representatives, often used to express a topic. While very few works [76, 88] assume that a single keyword is sufficient to represent a topic, many topic-based IUD approaches use a hand-crafted set of keywords including words [18, 24, 26, 29, 81, 124, 128] and/or hashtags [16, 53, 73, 73, 75, 75, 78, 78, 93, 130] (words stating with # symbol in the microblog) to denote a topic. Once the keywords are identified, posts that contain any of the keywords representing the topic are collected, and influential users are inferred from topic-related posts using various IUD techniques. The following subsections discuss such IUD techniques which use keyword-based approaches for topic detection.

4.1.1 Statistical measures for influence detection

Statistical measures observed in an OSN, such as the number of followers for a user who often discusses a topic, number of posts published about a topic, number of reactions to a post that discusses a topic, are good indicators of influence made on a specific topic. These statistical measures can be directly quantified [18, 26, 29, 53, 63, 81, 93, 106, 107] to define a topic-based influence score for a user.

Embar et al. [26] use a rank aggregation technique to infer the influential user from the collection of posts that discusses a topic denoted using hand-crafted keywords. They rank the users based on five different aspects, such as follower count, number of posts, number of reposts and comments, network centrality measure computed using PageRank algorithm, and time-taken by other users to react to the posts. Finally, they aggregate the ranks determined for each aspect to find the overall topic-based influence ranking of users. Apart from aggregating the ranks or scores on multiple aspects, [29, 81] compute topic-based influential scores independently based on various aspects including the number of posts, the number of reposts received, and the number of mentions received. They observe, the influential score obtained using the number of posts plays a prominent role in user influence and has a direct relationship with the level of influential measures.

Statistical measures in the influence network can be used to derive further meaningful scores. Wang et al. [106, 107] first quantify a user's strength on a topic as a fraction of posts published by the user about a topic as compared to other users. Similarly, a user u's impact on the neighbours on k^{th} topic is computed as follows,

$$f(u,k) = \sum_{v \in NB(u)} \frac{w_{uv}^{k}}{y_{v}^{k}} \tag{3}$$

where, $NB(u)$ indicates the neighbours of user u, w_{uv}^{k} indicates the number of reactions to user u's post by user v as repost and comments and y_v^k indicates the number of posts published by user v about topic k. Finally, the topic-based influence score for the user is computed as a multiplication of the user's strength and neighbour influence score $f(u, k)$.

Liengpradit et al. [63] focus on identifying topic-based influential users for different geographic regions. The authors construct a topic-specific influence network by considering users as nodes and friends who liked or commented on a user's posts as edges. Using the influence network, they compute the in-degree of the nodes as the number of friends and friends-of-friends who live in the same area or work/study in the same organizations, which is directly mapped to an influential score of a user. This enables them to determine the topic-based influential users with similar geographic features. In order to develop a more reliable set of keywords denoting a topic, Cha et al. [18] consults the news websites and informed individuals. They assume the three activities in the influence network, including number of followers, number of retweets, and number of mentions received by a user, that can represent the amount of influence he or she has made.

Hashtags are good indicators of emerging topics in social networks. Many works [53, 73, 78, 93] in the literature use single or multiple hashtags to denote a topic. Pal et al. [78] use a set of hashtags to denote a topic. They define several features for a user based on the activities in the network, including the number of keywords used and the self-similarity score that reflects how much a user borrows words from his/her previous post. Once the features are defined, Gaussian Mixture Model [86] is used to cluster the users into two groups, one representing the influential user and the other one representing the influenced users. Cluster with the highest scoring values for features are chosen as the influential user cluster, and then the users in this cluster are ranked according to the accumulated ranking based on all the defined features.

Similarly, Mittal et al. [73] use a set of hashtags as keywords to represent a topic and collect tweets containing the hashtags. The authors create the influence network with nodes consisting of users, tweets, and hashtags, and the edges using relationships, such as follower-followee relationship, repost, like, comment and mention. From the influence network, they use different network-based centrality scores and popularity-based scores, including the number of reposts and likes, and aggregate them to compute the overall influence score of a user.

Koutrouli et al. [53] assume that a hashtag represents only a single topic, and they infer further hashtags representing a topic from a small set of annotated hashtags using the co-occurrences of two hashtags along with the same URL from posts of a user. Once the hashtags are inferred, they collect posts with at least one of the hashtags, then use a score-based mechanism to compute the topic-based influence score for a user as a sum of the influence score of each hashtag. This includes the number of reposts and likes for posts published by the user with a hashtag representing a particular topic. Whereas, Shalaby and Rafea [93] uses both hand-crafted keywords and hashtags associated with a topic to collect topic-related posts. The overall topic-based influence score for a user has been developed as the weighted sum of the number of followers, reposts, likes, and mentions.

It is observed the statistical measures to directly quantify the influence of a user have been widely explored in single topic-based IUD approaches as compared to multiple topic-based IUD methods. This could be mainly due to the simplification of the problem when you are dealing with a single topic, where the entire collection of documents discuss a single topic. Table 1 summarizes the statistical measures used to determine the influential users for a single topic. It is observed, the features, such as the number of followers, number of posts, number of reposts, and number of likes are widely used, whereas the time taken to react to the influential content and statistical measures that can be derived from the content are rarely used. Next, we discuss the random walk-based techniques used to infer the influential users once the topic is being defined using a set of keywords.

4.1.2 Random walk based influence detection

Random walk based approaches to determine an influential user require constructing an influence network prior to the execution of random-walk algorithms. In order to determine the topic-based influential users, the influence network specific to a topic is created as a subset of OSN from topic-specific activities in the OSN.

Once the posts discussing a topic are collected using a hand-crafted set of keywords, one can construct the follower-followee network to infer the influential users [24, 124]. Zhang et al. [124] applies the PageRank algorithm [77] on the follower-followee network to rank the topic-based influential users. Whereas, Dong et al. [24] further assumes that a follower is influenced by a followee if there is any post published by the follower after any of the posts published by the followee. This assumption has been modeled to compute the edge weight between two users u and v as $e^{-\frac{t_{uv}}{\delta}}$, if there is a post published by follower v after followee u, or zero otherwise. Here, t_{uv} is the minimum time gap between the posts by user u and v and δ is a controlling parameter.

Table 1 Summary of statistical measures used to determine influential users for single topic

Ref	Number of Followers	Number of Posts	Number of Reposts	Number of Comments	Number of Likes	Number of Mentions	PageRank Score	Time to React	Number of Keywords
[26]	✓	✓	✓	✓	–	–	✓	✓	–
[106, 107]	✓	✓	✓	✓	–	–	–	–	–
[78]	✓	✓	✓	–	–	✓	–	–	✓
[18]	✓	–	✓	–	–	✓	–	–	–
[73]	✓	–	✓	✓	✓	✓	✓	–	–
[93]	✓	–	✓	–	✓	✓	–	–	–
[63]	✓	–	–	–	✓	–	–	–	–
[99]	✓	–	–	–	–	–	–	–	✓
[29, 81]	–	✓	✓	–	–	✓	–	–	–
[53]	–	–	✓	–	✓	–	–	–	–

Similarly, Luiten et al. [68] use keywords inferred from Wikipedia articles to define a topic and collect posts that contain at least one of the keywords. Influence network specific to a topic is constructed using a follower-followee relationship, and then the PageRank algorithm is used to infer the topic-based influential user. The fraction of posts by the user that is related to a topic is used as a weight to modify the damping factor, and this allows to increase the probability of randomly choosing a user who has published relatively more posts about a topic.

Different from exploring a single influence network, Zhaoyun et al. [128] define four influence networks, each created with users as nodes and different aspects being modeled as edge and edge weights, such as,

1. repost network - reposts represent an edge, and fraction of reposted tweets defines the edge weight
2. reply network - replies represent an edge, and fraction of replied tweets defines the edge weight
3. reintroduce network - content similarity between posts defines the edge weight
4. read network - similarity between tweet posting rate defines the edge weight

Finally, the authors propose a random walk approach that can walk through the four influence networks to rank the users.

Santos et al. [88] assume that one keyword can be used to better represent a topic instead of a set of keywords, and crawl all the blog posts and comments in which the keyword appears. The influence network is constructed with users who post or comment on the topic as nodes and comments made by one user on another user's post as directed edge. Finally, they apply the PageRank algorithm on the influence network to determine the topic-based influential users. Oro et al. [76] follow a similar approach where a keyword or phrase (e.g., *TV series*) is used to collect posts associated with a topic. They generate a multi-layer network composed of three networks, a user network representing the repost behavior, an item network representing the subtopics as nodes and their similarity as edges, and a keyword network modeled using words as nodes and co-occurrence of words as edges. They apply a generalization of HITS [51] algorithm called TOPHITS to infer the influential user from the multi-layer network. While HITS introduces the concepts of hub and authority, and iteratively computes scores for these two dimensions for all the nodes in the network, TOPHITS adds the third dimension with a topic score for nodes.

As hashtags are good representatives of hot topics, Bogdanov et al. [16] use the set of hashtags to denote a topic. They assume that a hashtag represents only a single topic and manually annotate a set of hashtags representing a topic. They create an influence network specific to a topic as a subnetwork of the follower-followee network, where the edges are created if the followee has used a hashtag before the follower or, in other terms, if the follower has adopted any topic-specific hashtags from the followee. The number of hashtags adopted by the follower is used to determine the edge weight, and finally, they apply random walk in the topic-specific influence network to rank the influential users.

Similarly, in [75], a set of hashtags is used as keywords to represent a topic and then the tweets containing the hashtags are collected. They employ PageRank [77] algorithm on the influence network with user nodes and edges representing the repost, mention, and reply relationships to determine the influential users from topic-related posts. Different from other approaches, [25] modifies the Pagerank [77] algorithm by incorporating the following information from the influence network into a weight factor to control the ranking.

- User Trust: Measured using the number of friends, number of reposts, mentions, likes received for the posts of a user, and fraction of topic-specific keywords used by the user
- Influence Period: Measured using the friendship time between two users
- Similarity: Measured as the ratio of the number of topic-specific keywords used by two users

Here, the topic-specific keywords are defined manually. It is observed that different types of influence networks including follower-followee networks, repost networks, and other reaction networks are explored in the literature to infer topic-based influential users using keyword-based topic detection and Pagerank algorithm.

4.1.3 Machine learning based influence detection

Despite the adoption of Deep learning solutions to solve a wide variety of research problems due to its appreciable performance, a very little effort has been made to adopt them to infer influential users in OSN. Zheng et al. [129] attempted to infer HIV related influential users from tweets collected over two years using the keywords HIV and AIDS. The authors proposed a graph neural network model to infer influential users. The identified influential users were manually verified and the authors concluded that deep learning can infer the long-term influential users in OSN despite a larger volume of tweets. The promising results show that the deep/machine learning models can be further enhanced by incorporating additional information, such as the propagation network of the influence and users' attributes.

Even though determining a topic based on hand-crafted keywords or hashtags is widely used and simple, they do not have the luxury of utilizing all the words representing a topic, and hence the identified topic-based influential users are limited to the defined keywords.

4.2 Query-based topic detection

Given a query composed of a set of terms, the problem of IUD can be further narrowed down to find influential users specific to that query. The labels or words appearing in a search query are generally closely related and comprise a topic. Hence there are certain studies in the literature [2, 7, 99, 115] which attempt to model the problem of topic-based influential user detection as a query-based influential user detection. Once the posts related to the query are extracted, traditional IUD techniques can be used to determine the influential users for the given query.

4.2.1 Statistical measures for influence detection

Subbian et al. [99] allow to query for an influential user using the set of keywords and a time period. They assume that there is a keyword flow from the followee to the follower if a follower publishes a post with any of the queried keywords after the followee. This keyword flow from a followee to follower is quantified with a decaying time function, and the keyword flow between any two users in the influence network is quantified as a sum of keyword flow through the shortest path between the users. The overall influence between two users is computed as a sum of the influence of all the queried keywords.

4.2.2 Random-walk based influence detection

Once a query-specific influence network has been created, random walk algorithms can be applied to the influence network to infer influential users specific to the given query. Alp et al. [7] use the PageRank algorithm [77] to determine the query-based influential users from the influence network of a query. Here the influence network of a query is constructed by considering users as nodes, and user reactions such as comment, like, repost, and mention to posts with any of the keywords in the query as edges. This enables them to model the edge as a bag of words of posts to which the user reacted. Further, the similarity score between the bag-of-words of an edge and the query is used to define the edge weight. The authors explore three different similarity scores as follows,

- *label space* - measures the fraction of word matches between query and bag-of-words of an edge
- *people space* - measures the fraction of common users who reacted to the query and for a user's posts
- *Word2vec Space* - measures the word-embedding [72] vector similarity between query and bag-of-words of an edge

Agness et al. [2] follow a similar approach and use the Latent Semantic Analysis [41] to calculate the similarity between the query and bag-of-words of an edge.

Instead of considering only the posts with the search query, [115] proposes a co-occurrence based technique to find hashtags that are related to the query and then retrieve all the post that contains the top-n related hashtags. Co-occurrence based score computed for each hashtag considers the number of times a hashtag and search query appear together in a post compared to the number of times they appear individually. Using the retrieved posts specific to a query, they build the mention and retweet networks and infer the influential users in each network using PageRank algorithm [77].

It is observed that query-based approaches utilize the content from OSN to model the topic-based influential users in various forms, such as constructing query-specific influence network and representing the content similarity between two users as edge weights. While query-based approaches are more realistic as the influential users can be filtered out based on the query, similar to the cons of keyword-based approaches, a query may not include all the representative terms for a topic. Hence the identified influential users are limited to the terms appearing in the query. Section 5 discusses more advanced methods that overcome these problems.

4.3 User-based topic detection

Users who often discuss a topic can be considered as the representatives of those topics. Liu et al. [65] explore such an idea by manually identifying a set of users who discuss a topic most and extracting the followers and their interaction details to construct an influence network. Influence strength between any two users in the influence network is assumed as a Gaussian distribution to model the problem as an optimization problem solved using poison regression. They consider various features, including the number of friends, number of followers, number of reposts, number of mentions, the ratio of posts with URL, hashtags, and mentions as the features to build a similarity vector to input to poisson regression method. Given the values of a feature x for users u and v, the similarity score for the feature x is computed as follows,

$$sim(x_u, x_v) = -plog(p) - (1 - p)log(1 - p) \quad (4)$$

where $p = \frac{x_u}{x_u + x_v}$. The influence strength inferred between two users using the poison regression technique is used as the edge weight in the follower-followee network. Finally, the PageRank algorithm is used to infer the overall influence score of a user. However, this approach assumes that users who post more on a topic are not necessarily the most influential users on the same topic. Therefore, they attempted to represent a topic using a set of users and identify users who could influence others on a selected topic.

5 IUD for multiple topics

While dealing with a single topic denoted using a set of keywords or queries simplifies the IUD problem, still, the posts published in OSN discuss various co-occurring hot topics, and identifying a precise set of keywords denoting a topic is really challenging. This opens up the research problem of how to infer topics from a collection of documents that discusses various topics prior to the inference of influential users or inferring the topics and influence of users together. Moreover, this enables to identify overall influential users using influence score across multiple topics [111]. In this section, we will discuss research works that focus on identifying influential users on multiple topics structured according to the taxonomy.

5.1 Keyword-based topic detection

Manually hand-crafting precise keywords sets denoting multiple topics is very challenging and rarely explored [14, 131] as a multiple topic detection techniques in the literature. In this section, we discuss multiple topic-based IUD approaches that use a set of keywords to denote each topic.

5.1.1 PageRank based influence detection

Applying the PageRank algorithm for multiple topics requires the construction of influence networks at the topic level and executing the ranking algorithm for each topic-specific influence network independently. Bingöl et al. [14] use the follower-followee network constructed using topic-related posts as the influence network to infer influential users. They manually define a weight indicating the importance of each topic denoted using a set of keywords. For each user in the influence network, a word histogram is computed using all the posts published by the user. Then, the topic score for each user is computed by multiplying the number of times a word is being used and the weight of the topic to which the word belongs. This score has been further multiplied by the total number of reposts and likes a user received for posts related to a topic, and then, normalized to define the edge weight in the influence network. Finally, they use the weighted PageRank algorithm [116] to determine the topic-based influential users.

5.1.2 Machine learning based influence detection

Zheng et al. [131] use neural networks to formulate the problem of topic-based IUD as supervised learning. First, words in the posts published by a user are converted to a vector representation using Glove [80] word embedding, and then the vector representation is used as input to the proposed convolutional neural network to predict the influence score of the users. In order to generate the ground truth influential users, [131] proposes another neural architecture that can take the vector representation of users' posts and a set of keywords representing the topic, and predict the relatedness of the post with respect to the keywords. The user who posts content related to the keywords and has the highest number of followers are considered as ground truth influential users for training the convolutional neural network.

5.2 Topic modeling based topic detection

As discussed in the preliminaries (Section 2.2), LDA [15] is a traditional topic-modeling approach to identify the topics automatically from a larger collection of documents, where each topic is defined as a probability distribution over the vocabulary of words from a document collection. This flexibility led to many topic-based IUD approaches in the literature using LDA [39, 67, 84, 96, 111] or extension of LDA [23] to first infer the topics, then grouping the posts based on the topic dominantly discussed in the post and inferring influential users from topic related posts using different IUD techniques.

5.2.1 PageRank-based influence detection

Hamzehei et al. [37] use the combination of LDA and PageRank algorithm to infer the topic-based influential users without any modification to the underlying models. Once the topics are learned using LDA, they use the interaction network (retweet and mention network) to execute the PageRank algorithm and develop an aggregation technique that includes PageRank score as one of the criteria along with the number of friends, number of posts, number of repost and mentions.

However, determining the topics from a collection of documents before the execution of the PageRank algorithm enables to incorporate the topic-specific information into the ranking algorithm. This includes defining topic-specific transition probability [23, 67, 84, 96, 108, 111], topic-specific damping factor [7] and even modifying the underlying PageRank algorithm specific to a topic [39]. Weng et al. [111] performed one of the earlier lines of research that modifies the PageRank algorithm by incorporating topic-based influence measures to infer topic-based influential users. Once the topic distribution of each tweet is learned using LDA, the authors compute a topic-specific similarity score between two users using the Jensen-Shannon Divergence [32] between the topic-specific word distribution of each user. This similarity score is used as the transition probability for the PageRank algorithm. Finally, they compute an overall influence score for a user as a weighted sum of all the topic-based influential scores. They propose the following two different weights to aggregate the topic-based influence score.

- General Influence - Calculated according to the number of words assigned to corresponding topics
- Perceived General Influence - Calculated according to the number of times the words in a user's posts have been assigned to corresponding topics

Similarly, Quan et al. [84] explore both topic-specific information and activities in the OSN to modify the transition probability of the PageRank algorithm. They define two metrics called *behavioral coefficient* and *topic influence coefficient*, and the transition probability is computed as a weighted combination of both. *Behavioural coefficient* is computed using the probability of an influential user publishing a post multiplied by the probability of influenced user forwarding it in a collection of time segments. *Topic influence coefficient* between two users u and v is computed as follows,

$$C(u, v) = \frac{1}{1 + exp(1 + |T^u - T^v|)} \tag{5}$$

where T^u is a topic profile vector of a user, where the k^{th} element indicate the probability of user u publishing a post about k^{th} topic, determined using the number of posts published by the user u assigned to topic k.

Dhali et al. [23] explore the similar idea of using topic profile vectors of two users to compute the transition probability. They use Twitter-LDA [127], an extension of LDA, i.e., proposed specifically for short text to first extract the topics and then compute the similarity between two users as follows using their topic profile T_u,

$$sim(u, v) = 1 - |T_u - T_v| \tag{6}$$

Once the topic-based transition probabilities are defined, the PageRank algorithm is used to rank the influential users.

Lu et al. [67] define two different topic-based influences between the users, called explicit and implicit influence, and their weighted combination determines the overall topic-based influence score between two users. The overall influence score is used to compute the transition probability. Once the topics are learned using LDA, the topic-based explicit influence score between two users is computed as the number of reposts by the influenced user, weighted by the topic distribution of the post. The implicit influence between two users is determined using the topic similarity computed as cosine similarity between the topic distributions and the cosine similarity between the activity distributions of two users. Here, the topic distribution is computed using the number of posts associated with a topic, and activity distribution is computed as the number of posts published in a time interval. Finally, the PageRank [77] algorithm is used to determine the global user influence on a topic.

Instead of using the statistical or similarity measures to quantity the transition probability, Shi et al. [96] use a supervised random walk approach to determine the transition probability between two users. A topic similarity score between two users is determined using the Jensen Shannon divergence [32] between the topic distribution of two users. This similarity score is used as the feature for the supervised random walk to determine the transition probability. Once the transition probabilities are determined, the PageRank algorithm is used to infer the influential users.

Similarly, Wang et al. [108] use machine learning techniques to predict the transition probabilities before the execution of the PageRank algorithm. Given a post published by a user and his/her follower, a logistic regression [112] model is trained to predict the response probability indicating the chances of reacting to the post by the follower as a repost or like or comment. The response probability for all the posts published by the user is aggregated to compute the transition probability between the user and the follower. Both user activity and followers activity are used as features for the logistic regression model, along with the topic similarity between the two users. Topics are determined using LDA, and similarity between two users u and v are computed as follows,

$$Similarity(u, v) = \sqrt{2 * D_{JS}(u, v)} \tag{7}$$

where, D_{JS} is the Jensen-Shannon divergence [32] between the topic distributions of users u and v.

Different from these approaches, Alp et al. [7] modify the damping factor of the Pagerank algorithm with topic-specific information learned using LDA. They explore the following four aspects as damping factors for the PageRank algorithm executed on a topic-specific follower-followee network.

- *Focus Rate* - Measured as the fraction of topics focused by the user. This assumes a user who focuses on fewer topics are more influential on those topics.
- *Activeness* - Measured using the number of active days, average tweets posted per day, and the average number of tweets on a topic.
- *Authenticity* - Measured as a fraction of tweets posted compared to the number of posts and reposts by the users.
- *Speed of getting reaction* - Measured using the time to receive the first repost.

Herzig et al. [39] further modify the PageRank algorithm as follows using two probability functions $\alpha(u)$, and $\beta(u, v)$,

$$R'_T(u) = d \cdot \alpha(u) + (1 - d) \sum_{v \in In(u)} R_T(v)\beta(u, v) \tag{8}$$

where $R'_T(u)$ and $R_T(u)$ indicate the current and previous PageRank score of the user u for topic T, respectively, d indicates the damping factor, and $In(u)$ indicates all the users who either liked or reposted or commented on user u's posts. $\alpha(u)$ indicates the probability of user u's post is being liked or reposted or commented by a random user and $\beta(u, v)$ indicates the probability that a random user v will like or repost or comment on u's post. $\alpha(u)$ is defined as the normalized cosine similarity between the word distribution of a topic and word distribution of a user u, and $\beta(u, v)$ is defined as the normalized cosine similarity between the word distributions of users u and v.

It is observed that the combination of the topic model and Pagerank algorithm is widely used in the literature to determine the topic-based influential users. This could be mainly due to the powerful nature of the Pagerank algorithm on ranking nodes in a large-scale network as well as the ability to incorporate the topic-specific information into the Pagerank algorithm by modifying the transition probability and/or the damping factor. Table 2 summarizes different user and network features used to model the topic-based IUD problem using PageRank algorithms. Topic-based influential users have been extracted using both user network and microblog networks by considering followership and various interaction relationships. Moreover, a range of statistical measures have been used at various stages of the PageRank algorithm to determining edge weights, transition probability and damping factor. We next discuss another traditional technique that focuses on propagation algorithms combined with topic models for topic-based IUD.

5.2.2 Propagation-based influence detection

Propagation algorithms model the possible or observed propagation of information over a given influence network and score the nodes in the influence network according to the amount of spread they might contribute or the observed contribution, respectively. The amount of spread made by a node in the influence network is an indication of the amount of influence created by the propagation of information in the propagation direction. Therefore, analyzing the propagation at a topic level can reveal the topic-based influential users [28, 101, 115].

Fang et al. [28] use the affinity propagation [31] algorithm to rank the users in the hypergraph generated from multimedia content. The method works as (i) collecting posts with images and text from Flickr, (ii) learning topics associated with each post from the text using LDA, and (iii) generating the hypergraph at the topic level. Both images and users are modeled as nodes in the hypergraph, and factors, such as image similarity, the similarity of words used with the images, and like or comment activity among the users are used to represent the edges in the hypergraph. Similarly, Tang et al. [101] use the affinity propagation in the co-author network to identify influential users. They use a topic model named Author-Conference-Topic (ACT) model [102] to identify the topics which was proposed to extract topics from scientific papers.

Table 2 Summary of information modelled in PageRank-based solutions for topic-based influential user detection

	Nodes		Edges					Other Statistical Measures											
Ref	Users	Microblogs	Reply	Repost	Mention	Likes	Followership	# Posts	# Friends	# Reply	# Repost	# Mention	# Likes	Friendship Time	Content Similarity	Topic Similarity	Topic Distribution	Active Days	# Topic Related Posts
[88]	✓	–	✓	–	–	–	-	–	–	–	–	–	–	–	–	–	–	–	–
[75]	✓	–	✓	✓	✓	✓	-	–	✓	✓	✓	✓	✓	✓	✓	–	–	–	–
[2, 7]	✓	–	✓	✓	✓	✓	-	–	–	–	–	–	–	–	✓	–	–	–	–
[115]	✓	–	–	✓	–	–	-	–	–	–	–	–	–	–	–	–	–	–	–
[65]	✓	–	–	–	–	–	✓	–	✓	✓	✓	✓	✓	–	✓	–	–	–	–
[14]	✓	–	–	–	–	–	✓	–	–	–	✓	–	✓	–	–	✓	–	–	–
[37]	✓	–	–	✓	✓	–	-	–	–	–	–	–	–	–	–	–	–	–	–
[23, 111]	✓	–	–	–	–	–	✓	–	–	–	–	–	–	–	–	✓	–	–	–
[84]	✓	–	–	–	–	–	✓	✓	–	–	✓	–	–	–	–	✓	–	–	–
[67]	✓	–	–	–	–	–	✓	✓	–	–	✓	–	–	–	–	✓	✓	–	–
[96]	✓	–	–	–	–	–	✓	–	–	–	✓	–	–	–	–	–	✓	–	–
[108]	✓	–	–	–	–	–	✓	–	✓	✓	✓	✓	✓	–	–	✓	–	–	–
[7]	✓	–	–	–	–	–	✓	✓	–	–	✓	–	–	–	–	–	–	✓	✓
[39]	✓	–	✓	✓	–	✓	-	–	–	✓	✓	–	✓	–	✓	✓	–	–	–
[126]	✓	–	–	–	–	–	✓	–	–	–	–	–	–	–	–	✓	–	–	✓
[17]	✓	–	–	✓	–	–	-	–	–	–	–	–	–	–	–	–	–	–	✓
[69]	✓	✓	✓	✓	✓	✓	-	–	–	✓	✓	✓	✓	–	✓	–	–	–	–

Xiao et al. [115] introduce a concept called *Gravity* to model the propagation between two users. They use LDA [15] to learn the topics and then compute the topic distribution P for each user, which is usually referred to as the *topic profile*. Using the topic profiles of two users, the *Gravity* between the users is computed as follows,

$$G_{uv} = G^s \frac{P_u * P_v}{D(u, v)^2} \tag{9}$$

where G^s is a constant controlling the gravity and $D(u, v)$ is the distance (minimum number of nodes to reach node v from node u) in the influence network. The average gravity of all the neigbours of a user is defined as an initial influence score of a user. Finally, they develop a propagation algorithm using the Markov model [30] to model the propagation of the influence to compute an overall influence score for the users.

5.2.3 Machine learning based influence detection

Machine learning algorithms are very powerful in performing classification and regression tasks. Eliacik et al. [25] model the influential user detection as a regression problem by attempting to predict the topic-based influence score of users. They consider both topic features and statistical features, such as the number of followers, and the number of friends, to train the prediction model. Topics from microblogs are extracted using LDA [15] and the overall topic distribution discussed in the microblogs collection is provided as a feature to the regression model. This enables the model to identify influential topics as well as to predict the high score for users who post content related to the influential topics. Different from the existing works that model the topic as a distribution over words, the authors generate phrases by combining words using their co-occurrence and then model the topics as a distribution over phrases. For the purpose of predicting the influence scores, three popular regression models, Decision Tree [66], Random Forest [62], and Gradient Boosting [123] are used in the work.

Similarly, [130] uses a neural network composed of a multi-layer neural network and LSTM [35] to learn the time and topic-based influential users. Two sets of matrices are generated as input to the neural network capturing the time-specific interaction details and time-specific topic influence details. The interaction matrix is composed of L types of interactions between any two users, and the topic influence matrix is composed of the number of times a user uses the top D words of each topic. To learn the topics and generate the topic influence matrix, they use seededLDA [44] which uses a set of words per topic referred to as *seed* to improve the topic learning.

Li et al. [60] aim to predict the influence of users in an upcoming social event using the influence on past events. A user's influence on his or her followers for an event is computed as the fraction of followers who attended the event. Using this proposed approach, a user-event matrix is computed, and the influence on upcoming events is detected using Matrix Factorization [52] approach. Hamzehei et al. [38] use a similar approach to predict the future influence using Matrix Factorization. Instead of using LDA to learn the topics, they use Biterm Topic Model (BTM) [119] which is specifically designed for short text, such as microblogs, to overcome the data sparsity issue.

5.2.4 Similarity measures for influence detection

The influence of a user on others in OSN may result in influenced user publishing posts related to the same topic with similar content to influence the user. Therefore, the content similarity between the two users could reveal the amount of influence between them.

Xu et al. [117] use the similar idea of quantifying the influence between two users by computing the similarity between the documents posted by them. They compute the influence between two documents posted by two different users on a topic using an exponentially decaying function which is further weighted by the similarity between the word distribution of two documents. The similarity scores are computed using KL-Divergence [103]. Finally, the maximum influence between any two documents posted by two different users on a topic is chosen as the topic-based influence score between the users. The topics are learned using the proposed topic model called *guided hierarchical LDA* which utilizes the knowledge of common representative words of popular topics to guide the topic extraction process.

A similar idea has been used by Kefato et al. [47]. They use LDA to learn the topics and determine the topic distribution of each user using the number of times a topic is being assigned to a word posted by the user. This distribution is referred to as *topic profile* of the user and used to compute the topic similarity between two users as the absolute value of the sum differences in the probabilities. The topic similarity is further weighted by different popularity aspects of an influential user, such as the number of reposts, likes, and followers. Finally, they compute the overall topic-based influence of a user recursively as a weighted sum of his or her followers' influence score for a topic.

Instead of computing the topic-level influential score for a user, Zhao et al. [125] compute an overall influence score using collective credit allocation (CCA) approach [95]. The CCA approach was proposed to allocate credits for authors of a paper using the information available in the citation network. The authors emphasize that the importance or weight of a topic should be considered while computing the overall influential scores for the users. Hence, they build a topic popularity matrix representing the topic distribution over all the users in the network. Then a link strength vector specific to a target user is constructed using the in-degree of nodes in a sub-network of the target user. Here a sub-network represents the follower-followee network of a target user. Following CCA, a user-specific topic weight allocation vector is computed as the product of the topic popularity matrix and link strength vector. Finally, the influential score of a user is computed as a similarity between the overall topic distribution and the topic weight allocation vector of a user. KL Divergence [103] is used to compute the similarity between two topic distributions.

5.2.5 Other approaches for influence detection

A reliable user or content could create more influence in the OSN, hence it can be used to model the influence of the user. Li et al. [61] use a similar idea, where they model the reliability of a document on the web and the topic-based influence of a user as inter-related aspects. The reliability of a document is defined as the sum of the influential score of users who reacted to the document (forward, view and comment), and the topic based influential score of a user is defined as the sum of reliability of all the documents posted by the user on that particular topic. LDA [15] is used to infer the topics associated with the documents on the web. The reliability of the documents and the influential score of the user are updated alternatively until both converge.

5.3 Topic modeling based topic and influence detection

Topic models are very powerful generative models, which can learn unknown parameters automatically as far the problem is defined as a generative process. By extending the traditional topic model LDA [15], many real-world generative problems have been modeled as generative processes, and the unknown parameters of the models are estimated using Gibbs Sampling algorithm [33]. This idea has been explored for the problem of identifying topic-based IUD as well, by modeling the influence between two users in a topic model as a generative problem that infers both topics and influence together from an unlabelled collection of data [9, 13, 20, 64, 82, 98, 104, 109, 117, 121]. Different from the previous approaches for topic-based IUD, which are modeled as a two-step solution to extract the topics first and then infer the influential users, we discuss topic modeling based solutions in this section which infers both topics and influence together.

Tang et al. [64] presented one of the earlier research work that extends LDA to learn topics and influence together. While LDA assumes each document is a distribution over topics, [64] further models that each document is associated with a user, and each user is represented as a distribution over topics. Additionally, they assume that the behavior of an influenced user can be generated in two ways, either depending on his/her interest or influenced by one of his/her friends. Therefore, the extended topic model includes an additional parameter to model influence type for each word in a document. During the inference, if the influence type for a word is determined as an influence by one of his/her friends, then the model further chooses an influential user for each word. This enables us to infer an influence matrix that quantifies the influence between any two users, based on the number of times an influential user is being assigned to the words in documents posted by influenced users. Using the direct influence learned using the proposed topic model, they further propose a propagation algorithm developed using matrix multiplication to infer the indirect influence, and the global influence score for each user is computed as a sum of direct and indirect influence.

Extending LDA, [9, 13] propose a topic model, called *Followership-LDA*, that models both post generation and influence network generation in a single model. In *Followership-LDA*, the generative process of the user's content is the same as LDA, where each word in the post is associated with a topic, hence each post is represented as a distribution over topics. In order to model the influence network, they assume that the users follow other users either because of the content-independent popularity or due to the topic-based influence. While modeling the generation of influence network, for each follower-followee link, *Followership-LDA* samples a binary variable to indicate whether the reason for the follower-followee link is due to the followee's content or not. This enables to obtain two levels of scores for each user, topic-based influence score and global popularity score. Both scores are computed based on the number of times a followee is being followed by a follower due to their topic-based influence or content-independent popularity, respectively.

Similar to the *Followership-LDA*, [82, 121] propose a new topic model which models the generation of posts and follower-followee relationship. The generation of posts is modeled same as LDA, and for each follower-followee relationship, they sample a latent variable indicating the topic associated with the relationship. This enables to find the topic-based influential users who have more followers for a particular topic. Moreover, while computing the probability of assigning a topic to a follower-followee relationship, they introduce a weight factor computed using the similarity between the follower and followee to control the probability. Both content similarity and structural similarity are computed and summed to find the overall similarity between the users. The content similarity is computed using the cosine similarity between TF-IDF [3] vector representation of all the content posted by the user and structural similarity is computed using the fraction of overlapping followers between the two users.

Chen et al. [20] further assume that users can be grouped into communities and each community is associated with a specific topic interest. They further assume that each community creates influence on other users based on the community's topic interest. They propose a topic model that incorporates these assumptions by representing each community as a topic distribution and then introduce additional latent variables for each word to determine which community influenced the topic associated with that word. Moreover, for each follower-followee link, the model introduces a latent variable to determine the community that the link belongs. This enables to determine which groups of users are influencing more at a topic level based on the number of times a community is being assigned to the word and the number of times an influential user (followee here) is being assigned to a community.

Similar to this, Vega et al. [104] assume that the users can be grouped into communities, and each community is associated with a specific topic interest. Therefore, for each user, they sample a latent vector indicating the community membership as a probability distribution over C number of communities. Then, for each post published by the user, they sample a community assignment along with the topic assignment, where the topic is sampled from the corresponding community's topic distribution. Furthermore, for each follower-followee relationship, they sample a latent variable indicating whether the relationship is due to the content influence or popularity. This enables to identify influencing communities based on the number of times a community is being assigned to posts, as well as to identify influencing users based on the number of times a follower-followee relationship is resulted due to the content influence.

A time-based topic-influence inference has been carried out by [98, 109]. They highlight that influence on a user changes over time, and propose a new topic model called *TIT*, a topic-level influence over time that jointly models, topic, the link between users, and time. Different from other approaches, which consider only the follower-followee link [13, 82, 104, 121], *TIT* also considers interactions, such as repost and mention activities while modeling the links between two users. Similar to *Followership-LDA*, *TIT* samples a binary variable to indicate whether the time-specific link has been created based on a topic or not. This results in a topic and time-based influential score for each user. The authors propose an aggregation mechanism with decaying weight to sum the temporal influence of a user on a specific topic.

Table 3 lists the summary of user and network features used in topic modeling approaches to jointly learn topics

Table 3 Summary of OSN information modelled in topic modeling approaches to jointly infer topics and topic-based influential users

Ref	Source of influence			Community	Time
	Textual content in OSN	Followership	Interaction		
[64]	Personal/Friends	–	–	–	–
[9, 13]	–	Content Popularity or Topic		-	
[82, 121]	–	Always a Topic	–	–	–
[20]	–	Community	–	Associated with a distribution of topics	–
[104]	Community	–	–	Associated with a single topic	–
[98, 109]	–	–	Content Popularity or Topic	–	Change in Influence over Time

and influential users. It is very evident that topic models are widely used to model source of topic-based influence on choosing words from posts, on followership, and on user interactions. Further, they model various aspects of OSN, such as user communities with different topic interests and changes in influence over time.

5.4 Machine learning-based topic detection

A cluster or group of documents that discuss a topic can be determined directly instead of explicitly identifying the topics. Hence, existing clustering algorithms can be used to group the documents based on the dominantly discussed topic. Ramya et al. [85] propose a clustering approach named as weighted partition around medoids (WPAM) to cluster the tweets, where each cluster represents a topic. WPAM is shown to be more stable and less sensitive to outliers. They convert the tweets into bag-of-words to be used as features for the clustering algorithms. This includes the prepossessing steps of stopword, URL, numbers, and special character removal, and stemming and lemmatization which brings the words into their root form (e.g., eating to eat and Ate to eat). Once the clusters representing the topics are formed, influential users from each topic cluster are identified using the Maximum Likelihood Estimation [74] technique.

Wang et al. [108] use a supervised approach to determine the topics. They use Naive Bayes Classifier [19] to train a classification model which can classify a given text to the predefined set of topics. They define six topics, health, work, life, law, policy, and encouragement, and set of terms as keywords of the topics to train the classifier for predicting the presence of a topic accordingly. Once the topics are detected, a topic-specific user-index matrix is created, where each row of the matrix indicates a user and each column indicates different OSN information, such as the number of posts, reposts, likes, followers, friends, etc. for a particular topic. Finally, Singular Value Decomposition [105] technique is used to transform the user-index matrix into a user-influence score.

Transformer-based architectures are proved to be very effective in many text classification tasks. Especially the Bidirectional Encoder Representations from Transformers (BERT) models caused the stir in the natural language processing community by presenting state-of-the-art results in a wide range of tasks including topic detection [1, 8, 36]. Utilizing pre-trained BERT word embedding along with other text features, such as TF-IDF and Named Entity Recognition, is shown to generate easily interpretable topics. This encourages adapting BERT models for topic-based IUD tasks in the future.

5.5 Platform structure-based topic detection

The way the posts are organized on a platform may reveal the topics associated with the posts [50, 126]. This eliminates the need to infer the topics, and the platform-specific solutions can be developed to infer the topic-based influential users. Kim et al. [50] focus on inferring topic-based influential users on the Pinterest platform, where a board in Pinterest is considered to be associated with a topic. They develop a hierarchical influence graph that includes users and boards as nodes, and following either a user or board is modeled as edges. They develop five node features and two edge features in the influence graph based on the number of followers and the number of items on a board. Finally, a continuous conditional random field (CCRM) [83] model is used to infer the influential score for each user given the node and edge features.

Zhao et al. [126] explore a similar idea for a question-answering platform, where the questions are associated with a hierarchy of topics. Topics are divided into two levels, and the influential users are identified at the first level. They use the number of answers made by the user on the main topic as the damping factor and similarity between probability distribution obtained using the number of questions or answers posted on each subtopic by any two users as the transition probability. Pagerank algorithm is executed with the topic-specific damping factor and transition probability to identify the influential users for each main topic.

5.6 Online tools based topic detection

Cano et al. [17] use an online tool, named OpenCalais[1], to tag the topics associated with each microblog, and then build a retweet network per topic to infer the influential users. In the retweet network, the normalized value of the retweet count of all the users who retweeted the posts of a user has been used to define the transition probability between two users. Moreover, 1/*Number of users posted on a topic* is used as the topic-specific damping factor. Finally, they use the Pagerank algorithm to determine the topic-based influential users.

5.7 Implicit topic detection

While all the topic-based IUD techniques discussed so far explicitly identify a topic to infer the influential users, a few works [69, 97, 113] in the literature utilize the posts published on the OSN to infer the overall topic-based influential users without explicitly inferring the

[1] http://www.opencalais.com/

topics. They assume that the content similarity captured in their proposed approaches implicitly models the topic-level influence between two users. Such influential users are assumed to be influential across all the topics discussed in the microblogs collection.

Ma et al. [69] construct two influence networks, a user behavior network and user content networks, and use them to infer the influence using PageRank [77] and HITS [51] algorithms. The user behavior network is created with users as nodes and activities, such as repost, comment, and like as edges between the nodes. The number of activities between two users is used to define the edge weight. For user content networks, microblogs are considered as nodes, and co-occurrence of words and the number of co-occurring words between two microblogs are used to define the edge and edge weight, respectively. This enables to model both behavioral influence and content or topic influence on the other users.

Wu et al. [113] model the influential user detection problem as an optimization problem where the objective of the optimization problem is defined as finding users having higher influence with a minimum propagation time. The influence network input to the optimization algorithm is constructed by considering users as nodes and defining influence probability and time propagation as edge weights. They use content similarity between two users to define the two measures, *influence probability* and *propagation time*. If a post p_1 has a similarity above a threshold with another post p_2 published after p_1, then p_1 is considered to have a direct influence on p_2 and the *influence probability* of p_1 on p_2 is computed as $1/k$, where k is a constant. Here, the Word2Vec model [72] has been used to convert the text to a vector representation, and the cosine similarity between the vector representation of two posts is used to compute the similarity scores. Finally, the *influence probability* between two users is computed as a multiplication of *influence probability* between any two posts published by them. Influence *propagation time* between any two users is computed as a time gap between a post published by the influential user and a post published by the influenced user with the largest similarity. These two pieces of information are used to solve the optimization problem of identifying a user with the highest influence in a shorter propagation time.

Table 4 lists the summary of techniques used to infer the topic-based influential users.

6 Correlation analysis in topic-based influence

So far, we have discussed different approaches to identify topics-based influential users who influence others on a specific topic. In this section, we discuss different correlation analyses carried out as a post-processing step to infer further insights on topic-based influence.

Cha et al. [18] analyzed the consistency of influence across different topics by inferring the influential users across diverse topics and determining the correlation between their rankings. The topics were identified using hand-crafted keywords sets, and tweets were collected using the keywords associated with each topic for 60 days. The number of followers, number of reposts, and number of mentions received by a user are used to determine the influential users on a specific topic. The correlation analysis performed on influence ranks of users across multiple topics reveals that the ranking of top 1% of influential users is highly correlated across different topics. This shows that the users who are influential on a certain topic are highly likely to be influential on other topics discussed during the same time period.

The relationship between the use of language in posts published by a user and the influence made by the same user on a specific topic has been analyzed in [92, 94]. They chose a selected number of influential users on a particular topic and see whether there is a correlation between the linguistic feature of the posts of the user and the other activities of the user, such as number of posts, number of reposts, follow count, etc. A quantitative measure of the linguistic feature for correlation analysis is obtained by training a language model using highly reposted posts, and then computing the perplexity [22] measure of any post using the trained language model. Perplexity indicates the likelihood of a specific language model generating that piece of text; the lower perplexity indicates that the language of the text is closer to the model. The authors show that there is an inverse correlation exists between the perplexity and the number of reposts of an influential user post about a topic, which indicates that there is a correlation between the use of language and the influence.

7 Evaluation

Evaluating topic-based influential user detection is challenging as there is no generic dataset available. Developing a ground truth dataset is also a tedious process due to the higher number of users and microblogs to be analyzed manually, and the ranking of topic-based influential users can be a very subjective decision. In this section, we perform a detailed analysis of the available datasets as well as different evaluation techniques used to overcome the challenges.

7.1 Platforms

Twitter is one of the widely used OSN Platforms for social network mining. While the majority of the topic-based IUD

Table 4 Summary of topic-based influential user detection techniques

Number of Topics	Topic detection approaches	Influence detection approaches	Ref
Single	Keywords	Statistical measures	[18, 26, 29, 53, 63, 81, 93, 106, 107]
		Random-walk approach	PageRank [24, 68, 75, 88, 124]
			TOPHITS [76]
			Other [16, 128]
		Machine learning	Graph neural network [129]
	Query	Statistical measures	[99]
		Random-walk approach	PageRank [2, 7, 115]
	User	Random-walk approach	PageRank [65]
Multiple	Keywords	Random-walk approach	PageRank [14]
		Machine learning	Neural network [131]
	Topic modeling	Random-walk approach	LDA (T) & PageRank (I) [7, 23, 37, 39, 67, 84, 96, 108, 111]
		Propagation algorithms	LDA (T) & Affinity propagation (I) [28]
			LDA (T) & Markov Model (I) [115]
		Machine learning	LDA (T), Decision tree (I), Random forest (I) & Gradient boosting (I) [25]
			seededLDA & Neural network (I) [130]
			Biterm topic model & Matrix factorization (I) [60]
		Similarity measures	LDA (T) [47, 125]
			Guided hierarchical LDA [117]
	Machine learning	Probabilistic approach	Weighted partition around Medoids (T) & Maximum likelihood estimation (I) [85]
			Naive Bayes classifier (T) & Singular value decomposition (I) [108]
	Platform structure	Probablisitic approach	Conditional random field (I) [50]
		Random-walk approach	Pagerank [126]
	Implicit topics	Random-walk approach	PageRank and HITS [69]
		Optimization problem	[113]
	Topic modeling		[9, 13, 20, 64, 82, 98, 104, 109, 117, 121]

Technique labelled as (T) indicates a topic detection approach and technique labelled as (I) indicates an influential user detection approach approaches use Twitter for evaluating their solutions, the following Platforms are also explored in the literature.

1. Sina Weibo [13, 47, 61, 67, 69, 84, 98, 106, 107, 109, 125, 132]
2. Instragam [73, 79]
3. Flickr [28]
4. Facebook [63]
5. Reddit [55]
6. Yelp [76]
7. Webblog [113]
8. Tencent Weibo [121]
9. Pinterest [50]
10. Scoop.it [50]
11. Google Scholar [38]
12. DBLP [101]
13. ACM Digital Library [82]
14. DoubanEvent [60]

7.2 Datasets

Table 5 lists the datasets used in the literature to evaluate topic-based influential users. It is observed that some of the works attempt to manually annotate the influential users [13, 24, 93, 130, 131]; however, various other techniques also have been used to develop the ground truth dataset. This includes using the ranking provided by the platform [50, 98, 109, 132], using platform-specific information to develop ranking [38, 60, 121], and using search query responses to identify influential users [99].

7.3 Evaluation techniques

With the availability of ground truth data, the performance of topic-based IUD techniques can be compared using the following metrics defined in the literature.

1. **Accuracy**, **Precision**, and **Hit Count** [24, 38, 50, 75, 79, 94, 96, 98, 99, 117, 128, 130, 131] - Fraction of correctly identified influential users among the top k influential users inferred.
2. **Recall** [24, 38, 75, 79, 96, 99, 117, 128] - Fraction of ground truth top k influential users inferred correctly.
3. **F1-Score** [38, 75, 128, 131] - Given *Precision* and *Recall*, *F1-Score* is computed as follows,

$$\frac{2 * Precision * Recall}{Precision + Recall} \tag{10}$$

4. **Mean Average Precision** (MAP) [13, 98, 109, 131] - On average, the fraction of correctly identified influential users among the top k influential users inferred across all the topics.
5. **Mean Absolute Error** (MAE) [29, 60] - Given the ground truth influential score for top k users inf^G and

Table 5 Dataset used to evaluate topic-based IUD

Ref	Domain of data	# Users	Platform	Size	Annotation approach
[131]	HIV, Suicide	550-640	Twitter	340K-470K Tweets	Manual Annotation
[130]	Time range	1,000- 3M	Twitter, Reddit	126K-1.8M Posts	
[24]	-	120,130	Twitter	1.4M Tweets	
[93]	Time range	1,221	Twitter	8168 Tweets	
[98],[109]	Time range	0.4M	Sina Weibo	207M words	Platform Ranking
[132]	Time range	1.26M	Sina Weibo	114M Posts	
[50]	-	8,624	Pinterest	4.1M Posts	
[60]	Time range	100K	DoubanEvent	356K User-Event Pairs	RSVP Responses are used to develop the ground truth data
[38]	-	254	Scholar	101K articles	Users with a higher number of citations are considered as influential users
[99]	Time range	293K	Twitter	1.91M tweets	Search queries are used to develop the ground truth data
[13, 121]	KDD Cup 2012	1.78M	Tencent Weibo	492M Words	VIP users are considered as influential users

the corresponding influential score determined by a topic-based IUD approach inf, MAE is computed as follows,

$$\frac{1}{k}\sum_{i=1}^{k}|inf^{G}{}_{i} - inf_i| \tag{11}$$

6. **Root Mean Square Error** (RMSE) [14, 60, 132] - Given the ground truth influential score for top k users inf^G and the corresponding influential score determined by a topic-based IUD approach inf, RMSE is computed as follows,

$$\sqrt{\frac{1}{k}\sum_{i=1}^{k}(inf^{G}{}_{i} - inf_i)^2} \tag{12}$$

7. **Area Under Curve (AUC)** [131] - Area under the curve that plots fraction of correctly identified top k influential users against fraction of wrongly identified top k influential users.
8. **Normalized Discounted Cumulative Gain** (NDCG) [131] - Discounted Cumulative Gain is a widely used approach to measure the ranking quality. Given the ground truth relevance score rel for the top k users U_k inferred by a topic-based IUD approach and the corresponding sorted list of users UI_k according to the relevance score, then NDCG is computed as follows,

$$NDCG_k = \frac{DGC_k}{IDGC_k} \tag{13}$$

$$DGC_k = \sum_{i=1}^{k}\frac{2^{rel_{U_i}} - 1}{log_2(i+1)} \tag{14}$$

$$IDGC_k = \sum_{i=1}^{k}\frac{2^{rel_{UI_i}} - 1}{log_2(i+1)} \tag{15}$$

9. **Correlation Coefficient** [61] - Given a ground truth ranking of topic-based influential users and ranking determined by a topic-based IUD approach, the correlation between two rankings can be used as a metric to compare various topic-based IUD approaches. Spearman correlation coefficient [122] is widely used in the literature to quantify the correlation between two set of observations.

However, as we discussed earlier, there is no commonly used ground truth dataset for topic-based IUD problems, and the generation of ground truth dataset is quite challenging. Therefore, different techniques have been used in the literature to compare the topic-based IUD techniques. Here, we list commonly used approaches in the literature to compare the topic-based IUD techniques without a ground truth dataset.

1. **Visualization** [23, 26, 37, 53, 56, 76, 78, 113] - Top k influential users inferred using multiple topic-based IUD approaches can be visualized for manual verification.
2. **User Study** [2, 54, 65, 78, 98, 115] - The top k inferred influential users can be presented to a set of users for rating the ranking given by various topic-based IUD approaches [2, 54, 78, 98, 115]. Liu et al. [65] request a set of users to make a binary decision of whether an influential user from top k influential users is relevant or not and the fraction of relevant identified influential users can be used as a metric to compare the topic-based IUD approaches.
3. **Correlation Analysis** [16, 17, 25, 67, 68, 73, 84, 88, 93, 94, 96, 97, 111] - Correlation analyses are performed on various intuitions or assumptions that the ranking of topic-based influential users is correlated with different observable aspects. Many works [16, 17, 67, 68, 84, 88, 111] in the literature have performed a correlation analysis on ranking among various existing techniques to observe the similarity of the proposed approach in ranking influential users with that of the existing approaches, especially with the traditional techniques like PageRank algorithms. Following are other aspects explored in the literature to perform a correlation study.

   – Correlation among features of the proposed solution [93, 97]
   – Correlation with sentiment expressed by the user [25]
   – Correlation in ranking across different topics [96]
   – Correlation in ranking across different time intervals [73]

   While correlation analysis could give meaningful insights related to the problem, it does not help in directly comparing the better-performing topic-based IUD approaches.
4. **Prediction Analysis** - Inferred topic-based influential users can be used as features to various other OSN mining problems. Hence the performance of a model in a new problem can be used as a comparison technique as it will be indirectly affected by the performance of topic-based IUD approaches. Following are some sample problems explored in the literature as a comparison technique.

   – *Link Prediction* [20, 55, 56, 82] - Predict whether a user will follow another user or not. As in OSNs, it is observed that if a user is more influential, the chances of that user being *followed by others* will be high.
   – *User Recommendation* [28, 70, 79, 125] - *Recommend a user to be followed*. Here, the more the

user is influential, the chances of he or she is being recommended to others will be high.

- *User Group Recommendation* [106, 107] - Recommend a group of users with similar interests. Here, the group with more influential users will have a higher chance of getting recommended to others.
- *Content Recommendation* [28] - Recommend a content to users in OSN. Here, the more the user is influential, the chance of her/his content is getting recommended to others will be high.
- *Retweet Prediction* [20, 64] - Predict whether a friend or follower will repost the tweet published by a user. Here, the more the user is influential, the chances of his or her followers reposting the tweet will be high.

5. **Quantifying Aspects of Topic-based Influential Users** - Similar to the correlation analysis, different aspects of a topic-based influential user that are correlated with the user's influence can be directly quantified to compare different topic-based IUD approaches. Following is a list of such aspects used in the literature as a comparison technique.

   - *Influence Spread* [39, 47] - This indicates the capability of a user in spreading the influence in the influence network. Using the top k influential users inferred, the influence propagation is initiated using independent cascade model [49] and the fraction of activated users are quantified as the influence spread of top k influential users.
   - *Semantic Coverage* [69] - It is computed as a fraction of content covered by top k users through their posts, compared to the content published by all the N number of users in the influence network. Let Θ_i represents the semantic information of a user i, then the Semantic Coverage of top k influential users is computed as follows,

$$COV(k) = \frac{\sum_{i=1}^{k} \Theta_i}{\sum_{j=1}^{N} \Theta_j} \tag{16}$$

   - *Semantic Contribution* [69] - It is computed as a fraction of unique content covered by top k influential users through their posts compare to the content published by non-influential users among the N users. Let Θ_i represents the semantic information of a user i and $k \bigcap l = \emptyset$, then the Semantic Contribution of top k influential users is computed as follows,

$$CON(k) = 1 - \frac{\sum_{i=1}^{l} \Theta_i}{\sum_{j=1}^{N} \Theta_j} \tag{17}$$

6. **Case Study** [64, 99, 104] - A case study can be performed at selected topic(s) level to infer meaningful insights about a topic-based IUD approach.

Among the techniques used for comparison without a ground truth dataset, correlation analysis is being widely used to derive meaningful insights. However, these techniques still require at least a small set of ground truth influential users to determine the approach that correctly identifies the topmost topic-based influential users.

8 Closely related research topics

In this section, we discuss closely related research topics of topic-based IUD, which has overlapping objectives, methodologies, and techniques.

Topic-based influential post detection Similar to the inference of influential users, influential posts can also be determined for various marking purposes [12]. This can be considered a fine-grained study of IUD, where modeling posts as nodes in the influence network instead of users would enable the reuse of existing solutions to the IUD problem to determine the most influential posts. Moreover, the solutions to determine the influential posts can also be extended for IUD problems by aggregating the influence score of all the posts published by a user to determine an overall influential score for the user.

Topic-aware influence maximization Influence maximization is the problem of identifying top-k nodes in the OSN that maximizes the number of influenced nodes or influence propagation in the network. This is generally achieved using exploring the network structure of OSN, and the identified nodes indicate the potential influential users of the network. The problem has been further studied on topic-specific influence maximization [11, 21, 27] since users are more likely to be influenced on specific topics based on their interest. Existing Topic-aware Influence Maximization problems generally use the traditional topic models, such as LDA [15] and LSA [41], to infer the topics, and the traditional propagation models, such as the Independent Cascade model, are used to infer the top-k nodes that can maximize the topic-level influence. Topic-based IUD and topic-aware influence maximization problems are very interrelated as the influential users can potentially maximize the influence propagation in OSN. Therefore, the insights inferred from one problem can be used as an initial direction to the other problem.

Topic-aware influence minimization Minimizing negative influence in OSNs is required to avoid the spreading of malicious rumors and disinformation, especially on certain topics [91]. Therefore, the problem of topic-aware influence minimization has also received a good deal of attention recently [120]. Similar to the topic-aware influence maximization problem, a combination of topic models and the propagation models can be used to infer the top *k* nodes that minimize the influence at the topic level. Further, this can be related to topic-based IUD on negative topics, where blocking the most influential users can minimize the negative influence in an OSN, and the top-*k* nodes that minimize the influence in the OSN can be a potential influencer for a negative topic.

Topic-level opinion influence modeling Similar to the influence of the topic of interest among the users in OSN, the user's opinion may also be influenced by other users, especially by the neighbours. In real-life, an opinion is expressed towards a topic, and this opens up research for modeling the topic-level opinion influence of neighbours [58, 59]. Once the topic-level opinion influence has been modeled, this information can be used to predict the opinion (positive, negative, or neutral) of a new post published by the user using the opinion of his or her neighbours. They generally use the traditional topic models such as LDA [15] to infer the topics and lexicon-based word comparison approaches to infer the opinion of a post. This enables to determine the probability of expressing a certain opinion towards a topic and the probability of agreeing or disagreeing with the neighbour. Finally, probabilistic models are developed to predict the opinion of new published posts [58, 59]. This problem can be seen as another fine-grained study of the topic-level influence that focuses specifically on the opinion towards a topic which could be positive, negative, or neutral.

9 Further directions/open research questions

In this section, we discuss possible future directions based on our comprehensive analysis of existing topic-based IUD approaches. Following are some interesting questions that can be looked at as future direction to the topic-based IUD problem.

Multi-platform topic-based IUD An influential user can be an active influencer across multiple social networking platforms. This enables to gather cross-platform information and model generic solutions to identify influential users across multiple platforms instead of focusing on platform-specific solutions.

Multi-media content analysis for topic-based IUD Content published in OSN is generally multi-media with various forms, such as text, image, audio, video, emojis, and memes. Therefore, relying on a single medium of content will not accurately reflect influential users in OSN as the information expressed towards the other medium will be lost. It is therefore important to develop multi-modal solutions to determine the topic-based influential users precisely.

Identifying influential users of trending topics On OSNs, people share a lot of content on different topics; however, only a few topics manage to attract enough attention to become viral and temporally trend among users [89]. Therefore, it is evident that determining the trending topics and corresponding influential users is important for a meaningful inference. While existing works use topic-modeling techniques to infer topics from a collection of documents before the inference of influential users, they do not distinguish between the trending topics. Therefore, identifying influential users of trending topics remains an open problem.

10 Conclusion

Over the past decade, a substantial amount of interest has been shown towards analyzing the topic-level influence in OSN. The purpose of this survey is to draw attention to this area with a comprehensive study of the topic-based influential user detection (IUD) problem. To the best of our knowledge, this is the first comprehensive study of this type. Our survey summarizes various topic detection techniques used in the literature for single and multiple topic-based IUD problems. Further, we present a wide range of influential user detection techniques combined with different topic detection approaches. It is observed that the combination of traditional topic-modeling approaches, such as LDA [15] and the traditional random-walk approaches, such as PageRank algorithms [77] are widely used to infer the topic-based influential users. Moreover, it is evident that topic models are very powerful in automatically modeling the topic and topic-level influence together along with different features of OSN, including follower-followee relationship, user communities, and time-based analysis, without the need for any external knowledge.

We conclude that the PageRank algorithm has been widely used to infer the influential users mainly due to its powerful nature of effectively ranking nodes in large networks. It is evident that the existing works have modeled the various relationships between users and microblogs as well a range of statistical measures to infer the influential users. At the same time, we also highlighted that very little amount of work has been done to adapt machine learning solutions in this area. This could be mainly due to the requirement of having labeled data according to the problem domain to perform supervised learning.

Furthermore, we present a detailed analysis of evaluation techniques used in the literature to overcome the challenges that arise due to the unavailability of the public dataset. Apart from this, we have also discussed various metrics used to compare the topic-based IUD approaches using a ground truth list of influential users. Moreover, our brief discussions on similar research topics and open research problems in the domain of topic-based IUD would help to get a deep understanding of the problem. We hope our survey will help researchers in the domain of online social network mining to understand the topic-based IUD problem thoroughly and work towards interesting future directions.

References

1. Abuzayed A, Al-Khalifa H (2021) Bert for arabic topic modeling: an experimental study on bertopic technique. Procedia Comput Sci 189:191–194
2. Agness JA, Raj R JR (2018) An integrated approach for identifying topical experts in microblogs. In: Proceedings of 3rd international conference on internet of things and connected technologies (ICIoTCT), pp 26–27
3. Aizawa A (2003) An information-theoretic perspective of tf–idf measures. Inf Process Manag 39(1):45–65
4. Al-Garadi MA, Varathan KD, Ravana SD, Ahmed E, Mujtaba G, Khan MUS, Khan SU (2018) Analysis of online social network connections for identification of influential users: survey and open research issues. ACM Comput Surv (CSUR) 51(1):1–37
5. Al-Yazidi S, Berri J, Al-Qurishi M, Al-Alrubaian M (2020) Measuring reputation and influence in online social networks: a systematic literature review. IEEE Access 8:105824–105851
6. Allen LindaJS (1994) Some discrete-time si, sir, and sis epidemic models. Math Biosci 124(1):83–105
7. Alp ZZ, Öğüdücü SG (2018) Identifying topical influencers on twitter based on user behavior and network topology. Knowl-Based Syst 141:211–221
8. Asgari-Chenaghlu M, Feizi-Derakhshi M-R, Balafar M-A, Motamed C et al (2021) Topicbert: a cognitive approach for topic detection from multimodal post stream using bert and memory–graph. Chaos, Solitons & Fractals 151:111274
9. Balmin AL, Bi B, Sismanis J, Tian Y (2016) Identifying influencers for topics in social media. Google Patents. US Patent 9,449,096
10. Bamakan SMH, Nurgaliev I, Qu Q (2019) Opinion leader detection: a methodological review. Expert Syst Appl 115:200–222
11. Barbieri N, Bonchi F, Manco G (2013) Topic-aware social influence propagation models. Knowl Inform Syst 37(3):555–584
12. Bashari B, Fazl-Ersi E (2020) Influential post identification on instagram through caption and hashtag analysis. Measur Control 53(3-4):409–415
13. Bi B, Tian Y, Sismanis Y, Balmin A, Cho J (2014) Scalable topic-specific influence analysis on microblogs. In: Proceedings of the 7th ACM international conference on web search and data mining, pp 513–522
14. Bingöl K, Eravcı B, Etemoğlu CO, Ferhatosmanoğlu H, Gedik B (2016) Topic-based influence computation in social networks under resource constraints. IEEE Trans Serv Comput 12(6):970–986
15. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
16. Bogdanov P, Busch M, Moehlis J, Singh AK, Szymanski BK (2014) Modeling individual topic-specific behavior and influence backbone networks in social media. Soc Netw Anal Min 4(1):204
17. Cano AE, Mazumdar S, Ciravegna F (2014) Social influence analysis in microblogging platforms–a topic-sensitive based approach. Semant Web 5(5):357–372
18. Cha M, Haddadi H, Benevenuto F, Gummadi K (2010) Measuring user influence in twitter: the million follower fallacy. In: Proceedings of the international AAAI conference on web and social media, vol 4
19. Chakrabarti S, Roy S, Soundalgekar MV (2003) Fast and accurate text classification via multiple linear discriminant projections. VLDB J 12(2):170–185
20. Chen L, Prakash BA (2019) Joint post and link-level influence modeling on social media. In: Proceedings of the 2019 SIAM international conference on data mining. SIAM, pp 262–270
21. Chen S, Fan J, Li G, Feng J, Tan K, Tang J (2015) Online topic-aware influence maximization. Proce VLDB Endowm 8(6):666–677
22. Chen SF, Beeferman D, Rosenfeld R (1998) Evaluation metrics for language models
23. Dhali A, Gomasta SS, Anwar MM, Sarker IH (2020) Attribute-driven topical influential users detection in online social networks. In: 2020 IEEE Asia-Pacific conference on computer science and data engineering (CSDE). IEEE, pp 1–5
24. Dong G, Li B, Wei X, Qin T (2019) Mining key users of microblog topics based on trust model. Int J Performability Eng 15(11):3024
25. Eliacik AB, Erdogan N (2018) Influential user weighted sentiment analysis on topic based microblogging community. Expert Syst Appl 92:403–418
26. Embar VR, Bhattacharya I, Pandit V, Vaculin R (2015) Online topic-based social influence analysis for the wimbledon championships. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1759–1768
27. Fan J, Qiu J, Li Y, Meng Q, Zhang D, Li G, Tan K-L, Du X (2018) Octopus: An online topic-aware influence analysis system for social networks. In: 2018 IEEE 34th international conference on data engineering (ICDE). IEEE, pp 1569–1572
28. Fang Q, Sang J, Xu C, Rui Y (2014) Topic-sensitive influencer mining in interest-based social media networks via hypergraph learning. IEEE Trans Multimed 16(3):796–812

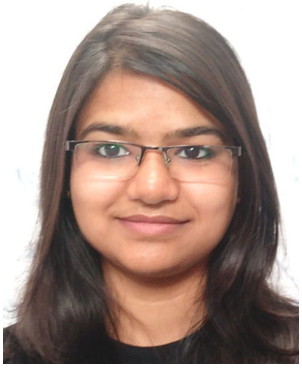
29. Farahani HS, Bagheri A, Saraf EHKM (2017) Characterizing behavior of topical authorities in twitter. In: 2017 International conference on innovative mechanisms for industry applications (ICIMIA). IEEE, pp 581–586
30. Fei H (2006) Application of markov model in stock market forecast [j]. Friend of Science Amateurs 6:.
31. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Science 315(5814):972–976
32. Fuglede B, Topsoe F (2004) Jensen-shannon divergence and hilbert space embedding. In: International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings. IEEE, p 31
33. Gelfand AE (2000) Gibbs sampling. J Amer Stat Assoc 95(452):1300–1304
34. González-Bailón S, Borge-Holthoefer J, Rivero A, Moreno Y (2011) The dynamics of protest recruitment through an online network. Sci Rep 1(1):1–7
35. Graves A (2013) Generating sequences with recurrent neural networks. arXiv:1308.0850
36. Grootendorst M (2020) Bertopic: leveraging bert and c-tf-idf to create easily interpretable topics. 4381785 https://doi.org/10.5281/zenodo
37. Hamzehei A, Jiang S, Koutra D, Wong R, Chen F, et al. (2017) Topic-based social influence measurement for social networks. Australas J Inf Syst 21
38. Hamzehei A, Wong RK, Koutra D, Chen F (2019) Collaborative topic regression for predicting topic-based social influence. Mach Learn 108(10):1831–1850
39. Herzig J, Mass Y, Roitman H (2014) An author-reader influence model for detecting topic-based influencers in social media. In: Proceedings of the 25th ACM conference on hypertext and social media, pp 46–55
40. Hethcote HW (2000) The mathematics of infectious diseases. SIAM Rev 42(4):599–653
41. Hofmann T (2013) Probabilistic latent semantic analysis. arXiv:1301.6705
42. Hu M, Liu S, Wei F, Wu Y, Stasko J, Ma K-L (2012) Breaking news on twitter. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 2751–2754
43. Ishfaq U, Khan HU, Iqbal S, Alghobiri M (2021) Finding influential users in microblogs: state-of-the-art methods and open research challenges. Behaviour & Information Technology, 1–44
44. Jagarlamudi J, Daumé III H, Udupa R (2012) Incorporating lexical priors into topic models. In: Proceedings of the 13th conference of the European chapter of the association for computational linguistics, pp 204–213
45. Jain L, Katarya R (2018) A systematic survey of opinion leader in online social network. In: 2018 International conference on soft-computing and network security (ICSNS). IEEE, pp 1–5
46. Jeh G, Widom J (2003) Scaling personalized web search. In: Proceedings of the 12th international conference on world wide web, pp 271–279
47. Kefato ZT, Montresor A Personalized influencer detection: topic and exposure-conformity aware
48. Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining, pp 137–146
49. Kempe D, Kleinberg J, Tardos E (2005) Influential nodes in a diffusion model for social networks. In: International colloquium on automata, languages, and programming. Springer, pp 1127–1138
50. Kim D, Lee J-G, Lee BS (2016) Topical influence modeling via topic-level interests and interactions on social curation services. In: 2016 IEEE 32nd international conference on data engineering (ICDE). IEEE, pp 13–24
51. Kleinberg JM, Newman M, Barabási A-L, Watts DJ (2011) Authoritative sources in a hyperlinked environment. Princeton University Press
52. Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. Computer 42(8):30–37
53. Koutrouli E, Daskalakis C, Tsalgatidou A (2018) Finding topic-specific trends and influential users in social networks. In: International conference on discovery science. Springer, pp 405–420
54. Lahoti P, De Francisci Morales G, Gionis A (2017) Finding topical experts in twitter via query-dependent personalized pagerank. In: Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017, pp 155–162
55. Lee RK-W, Hoang T-A, Lim E-P (2018) Discovering hidden topical hubs and authorities in online social networks. In: Proceedings of the 2018 SIAM international conference on data mining. SIAM, pp 378–386
56. Lee RK-W, Hoang T-A, Lim E-P (2019) Discovering hidden topical hubs and authorities across multiple online social networks. IEEE Trans Knowl Data Eng 33(1):70–84
57. Letierce J, Passant A, Breslin J, Decker S (2010) Understanding how twitter is used to spread scientific messages
58. Li D, Shuai X, Sun G, Tang J, Ding Y, Luo Z (2012) Mining topic-level opinion influence in microblog. In: Proceedings of the 21st ACM international conference on Information and knowledge management, pp 1562–1566
59. Li D, Tang J, Ding Y, Shuai X, Chambers T, Sun G, Luo Z, Zhang J (2015) Topic-level opinion influence model (toim): an investigation using tencent microblogging. J Assoc Inform Sci Technol 66(12):2657–2673
60. Li X, Cheng X, Su S, Li S, Yang J (2017) A hybrid collaborative filtering model for social influence prediction in event-based social networks. Neurocomputing 230:197–209
61. Li Y, Ma S, Huang R (2015) Social context analysis for topic-specific expert finding in online learning communities. In: Smart Learning Environments. Springer, pp 57–74
62. Liaw A, Wiener M et al (2002) Classification and regression by randomforest. R News 2(3):18–22
63. Liengpradit P, Sinthupinyo S, Anuntavoranich P (2014) A conceptual framework for identify specific influencer on social network. Int J Comput Internet Manag 22(2):33–40
64. Liu L, Tang J, Han J, Jiang M, Yang S (2010) Mining topic-level influence in heterogeneous networks. In: Proceedings of the 19th ACM international conference on information and knowledge management, pp 199–208
65. Liu X, Shen H, Ma F, Liang W (2014) Topical influential user analysis with relationship strength estimation in twitter. In: 2014 IEEE International conference on data mining workshop. IEEE, pp 1012–1019
66. Loh W-Y (2014) Classification and regression tree methods. Wiley StatsRef: Statistics Reference Online
67. Lu M, Wang Z, Ye D (2019) Topic influence analysis based on user intimacy and social circle difference. IEEE Access 7:101665–101680
68. Luiten M, Kosters WA, Takes FW (2012) Topical influence on twitter: a feature construction approach. In: Proceedings of 24th Benelux conference on artificial intelligence (BNAIC 2012), pp 139–146
69. Ma Q, Luo X, Zhuge H (2019) Finding influential users of web event in social media. Concurr Comput Pract Exp 31(3):e5029
70. Ma X, Li C, Bailey J, Wijewickrema S (2017) Finding influentials in twitter: a temporal influence ranking model. arXiv:1703.01468

71. Makita M, Mas-Bleda A, Morris S, Thelwall M (2021) Mental health discourses on twitter during mental health awareness week. Issues Ment Health Nurs 42(5):437–450
72. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781
73. Mittal D, Suthar P, Patil M, Pranaya PGS, Rana DP, Tidke B (2020) Social network influencer rank recommender using diverse features from topical graph. Procedia Comput Sci 167:1861–1871
74. Myung IJ (2003) Tutorial on maximum likelihood estimation. J Math Psychol 47(1):90–100
75. Oo MM, Lwin MT Detecting influential users in a trending topic community using link analysis approach
76. Oro E, Pizzuti C, Procopio N, Ruffolo M (2017) Detecting topic authoritative social media users: a multilayer network approach. IEEE Trans Multimed 20(5):1195–1208
77. Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab
78. Pal A, Counts S (2011) Identifying topical authorities in microblogs. In: Proceedings of the fourth ACM international conference on Web search and data mining, pp 45–54
79. Pal A, Herdagdelen A, Chatterji S, Taank S, Chakrabarti D (2016) Discovery of topical authorities in instagram. In: Proceedings of the 25th international conference on world wide web, pp 1203–1213
80. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
81. Petre A-N, Florea EAM, Ismail EA-A (2014) Searching for topical authorities on twitter
82. Qian Y, Liu Y, Jiang Y, Liu X (2020) Detecting topic-level influencers in large-scale scientific networks. World Wide Web 23(2):831–851
83. Qin T, Liu T-Y, Zhang X-D, Wang D-S, Li H (2008) Global ranking using continuous conditional random fields. Adv Neural Inform Process Syst 21:1281–1288
84. Quan Y, Song Y, Deng L, Jia Y, Zhou B, Han W (2019) Identify influentials based on user behavior across different topics. In: International conference on artificial intelligence and security. Springer, pp 476–487
85. Ramya GR, Sivakumar PB (2021) An incremental learning temporal influence model for identifying topical influencers on twitter dataset. Soc Netw Anal Min 11(1):1–16
86. Reynolds DA (2009) Gaussian mixture models. Encycl Biom 741:659–663
87. Riquelme F, González-Cantergiani P (2016) Measuring user influence on twitter: a survey. Inf Process Manag 52(5):949–975
88. Santos HenriqueDP, Wives LK Popular topical authors in brazilian blogosphere using comments as relationships
89. Saquib S, Ali R (2017) Understanding dynamics of trending topics in twitter. In: 2017 International conference on computing, communication and automation (ICCCA). IEEE, pp 98–103
90. Saxena A, Hsu W, Lee ML, Leong Chieu H, Ng L, Teow LN (2020) Mitigating misinformation in online social network with top-k debunkers and evolving user opinions. In: Companion proceedings of the web conference 2020, pp 363–370
91. Saxena A, Saxena P, Reddy H (2022) Fake news propagation and mitigation techniques: a survey. In: Principles of social networking. Springer, pp 355–386
92. Shalaby M (2014) Identifying the topic-specific influential users in twitter
93. Shalaby M, Rafea A (2013) Identifying the topic-specific influential users and opinion leaders in twitter. Acta Press 793:16–24
94. Shalaby M, Rafea A (2015) Identifying the topic-specific influential users using slm. In: 2015 First international conference on arabic computational linguistics (ACLing). IEEE, pp 118–123
95. Shen H-W, Barabási A-L (2014) Collective credit allocation in science. Proc Natl Acad Sci 111(34):12325–12330
96. Shi L-L, Liu L, Wu Y, Jiang L, Panneerselvam J, Crole R (2019) A social sensing model for event detection and user influence discovering in social media data streams. IEEE Trans Comput Soc Syst 7(1):141–150
97. Shinde M, Girase S (2016) Identification of topic-specific opinion leader using spear algorithm in online knowledge communities. In: 2016 International conference on computing, analytics and security trends (CAST). IEEE, pp 144–149
98. Su S, Wang Y, Zhang Z, Chang C, Zia MA (2018) Identifying and tracking topic-level influencers in the microblog streams. Mach Learn 107(3):551–578
99. Subbian K, Aggarwal CC, Srivastava J (2016) Querying and tracking influencers in social streams. In: Proceedings of the ninth ACM international conference on Web search and data mining, pp 493–502
100. Tang J, Sun J, Wang C, Yang Z (2009) Social influence analysis in large-scale networks. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, pp 807–816
101. Tang J, Wu S, Gao B, Wan Y (2011) Topic-level social network search. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, pp 769–772
102. Tang J, Zhang J, Yao L, Li J (2008) Extraction and mining of an academic social network. In: Proceedings of the 17th international conference on World Wide Web, pp 1193–1194
103. Van Erven T, Harremos P (2014) Rényi divergence and Kullback-Leibler divergence. IEEE Trans Inf Theory 60(7):3797–3820
104. Vega L, Mendez-Vazquez A, López-Cuevas A (2021) Probabilistic reasoning system for social influence analysis in online social networks. Soc Netw Anal Min 11(1):1–20
105. Wall ME, Rechtsteiner A, Rocha LM (2003) Singular value decomposition and principal component analysis. In: A practical approach to microarray data analysis. Springer, pp 91–109
106. Wang J, Liu Z, Zhao H (2015) Topic oriented user influence analysis in social networks. In: 2015 IEEE/WIC/ACM International conference on web intelligence and intelligent agent technology (WI-IAT), vol 1. IEEE, pp 123–126
107. Wang J, Zhao H, Liu Z (2017) Exploring user influence for topical group recommendation. Chin J Electron 26(1):106–111
108. Wang L, Lu T, Gu H, Ding X, Gu N (2015) Influential user recommendation through svd based topic diversification. In: 2015 IEEE 19th International conference on computer supported cooperative work in design (CSCWD). IEEE, pp 176–181
109. Wang Y, Zhang Z, Su S, Chang C, Zia MA (2016) Topic-level influencers identification in the microblog sphere. In: Proceedings of the twenty-second european conference on artificial intelligence, pp 1559–1560
110. Weinberg T (2009) The new community rules: marketing on the social web. O'Reilly Sebastopol, CA
111. Weng J, Lim E-P, 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. WSDM '10. Association for Computing Machinery, New York, pp 261–270. https://doi.org/10.1145/1718487.1718520
112. Wright RE (1995) Logistic regression
113. Wu J, Sha Y, Li R, Liang Q, Jiang B, Tan J, Wang B (2017) Identification of influential users based on topic-behavior influence tree in social networks. In: National CCF conference on

natural language processing and Chinese computing. Springer, pp 477–489
114. Xia F, Liu J, Nie H, Fu Y, Wan L, Kong X (2019) Random walks: a review of algorithms and applications. IEEE Trans Emerg Top Comput Intell 4(2):95–107
115. Xiao F, Noro T, Tokuda T (2014) Finding news-topic oriented influential twitter users based on topic related hashtag community detection. J Web Eng 13(5&6):405–429
116. Xing W, Ghorbani A (2004) Weighted pagerank algorithm. In: Proceedings second annual conference on communication networks and services research, 2004. IEEE, pp 305–314
117. Xu E, Hsu W, Lee ML, Patel D (2014) Inferring topic-level influence from network data. In: International conference on database and expert systems applications. Springer, pp 131–146
118. Xu S, Markson C, Costello KL, Xing CY, Demissie K, Llanos AdanaAM (2016) Leveraging social media to promote public health knowledge: example of cancer awareness via twitter. JMIR Public Health Surveill 2(1):e17
119. Yan X, Guo J, Lan Y, Cheng X (2013) A biterm topic model for short texts. In: Proceedings of the 22nd international conference on world wide web, pp 1445–1456
120. Yao Q, Shi R, Zhou C, Wang P, Guo L (2015) Topic-aware social influence minimization. In: Proceedings of the 24th international conference on world wide web, pp 139–140
121. Yu Y, Mo L, Wang J (2016) Identifying topic-specific experts on microblog. KSII Trans Internet Inform Syst (TIIS) 10(6):2627–2647
122. Zar JH (2005) Spearman rank correlation. Encycl Biostat 7
123. Zemel RS, Pitassi T (2001) A gradient-based boosting algorithm for regression problems. Advances in neural information processing systems pp 696–702
124. Zhang S, Zhang S, Yen NY, Zhu G (2017) The recommendation system of micro-blog topic based on user clustering. Mob Netw Applic 22(2):228–239
125. Zhao Q, Yang J, Wang S, Li M, Zhang W (2019) High-value user identification based on topic weight. IEEE Access 7:175917–175928
126. Zhao T, Huang H, Fu X (2018) Identifying topical opinion leaders in social community question answering. In: International conference on database systems for advanced applications. Springer, pp 372–387
127. Zhao WX, Jiang J, Weng J, He J, Lim E-P, Yan H, Li X (2011) Comparing twitter and traditional media using topic models. In: European conference on information retrieval. Springer, pp 338–349
128. Zhaoyun D, Yan J, Bin Z, Yi H (2013) Mining topical influencers based on the multi-relational network in micro-blogging sites. China Commun 10(1):93–104
129. Zheng C, Wang W, Young SD (2021) Identifying hiv-related digital social influencers using an iterative deep learning approach
130. Zheng C, Zhang Q, Long G, Zhang C, Young SD, Wang W (2020) Measuring time-sensitive and topic-specific influence in social networks with lstm and self-attention, vol 8. IEEE, pp 82481–82492
131. 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
132. Zhou J, Wu G, Tu M, Wang B, Zhang Y, Yan Y (2017) Predicting user influence under the environment of big data. In: 2017 IEEE 2nd international conference on cloud computing and big data analysis (ICCCBDA). IEEE, pp 133–138

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

Rrubaa Panchendrarajan received the Bachelor of Science degree in computer science and engineering from the University of Moratuwa, Sri Lanka, in 2016, and the Master's degree in computer science from the National University of Singapore, in 2021. She is currently a Lecturer at the Sri Lanka Institute of Information Technology. Her research interests include natural language processing, information retrieval, and social media analysis.



Dr. Akrati Saxena is a Research Fellow at the Department of Mathematics and Computer Science, Eindhoven University of Technology (TU/e), The Netherlands. She received her Ph.D. in Computer Science and Engineering from IIT Ropar, India and then worked as a Research Fellow at the National University of Singapore, Singapore. Her research interests cover Social Network Analysis, Complex Networks, Computational Social Science, Data Science, and Fairness. Her current research is focussed on designing fairness-aware solutions for social problems using Network Science and Data Science techniques. She has written several book chapters on social network analysis and social media data analytics. She also co-edited the Deep Learning for Social Media Data Analytics book that is published at Studies in Big Data, Springer book series.