Knowledge-Based Systems 163 (2019) 944–954

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Influence Factorization for identifying authorities in Twitter

Zeynep Zengin Alp [a,*], Şule Gündüz Öğüdücü [b]

[a] Institute of Science and Technology, Istanbul Technical University. Maslak, Istanbul, 34469, Turkey
[b] Department of Computer Engineering, Istanbul Technical University. Maslak, Istanbul, 34469, Turkey

ARTICLE INFO

Article history:
Received 3 April 2018
Received in revised form 10 October 2018
Accepted 12 October 2018
Available online 25 October 2018

Keywords:
Data mining
Influence analysis
Social media analysis
Matrix Factorization
Collaborative filtering
Influence prediction
Influence maximization

ABSTRACT

Prevalent usage of social media attracted companies and researchers to analyze its dynamics and effects on user behavior. One of the most intriguing aspects of social networks is to identify influencers who are experts on a specific topic. With the identification of these users within the network, many applications can be built for user recommendation, information diffusion modeling, viral marketing, user modeling and many more. In this paper, we aim to identify topic-based experts using a large dataset collected from Twitter. Our proposed approach has three phases: (1) identification of topics on social media posts (more specifically, tweets), (2) user modeling, based on a group of user specific features, and (3) Influence Factorization to identify topical influencers. The main advantage of the proposed method is to identify future influencers as well as current ones on Twitter. Moreover, it is an easy to implement algorithm using Spark MLlib, which can be easily extended to include other user specific features, and compare with other methodologies. The effectiveness of the proposed method is tested on a large dataset that contains tweets of 180K user over 70 day period. The experimental results show that our proposed method identifies influencers successfully when used with a hybrid user specific feature that contains follower count and authenticity information, and is a highly scalable and extensible algorithm.

1. Introduction

There have been many application areas based on social networks where researchers try to infer meaningful information. Some of them are: event detection [1,2] where the aim is to detect social gatherings or important sudden events like earthquakes, terrorist attacks; sentiment analysis [3–7] to infer positive or negative opinions about any topic within network; community detection [8,9] to identify structure of communities in large networks; user/topic recommendation [10] based on collaborative filtering on networks; and influence analysis to find out users who have great influence on other users.

Influence analysis can be used for different applications such as to identify authorities on networks so that those authorities can be used as a leverage for maximizing the spread of information. Another use of identification of the authoritative users is to recommend them to other users in the network. It also can be used to model information diffusion on the network, and finding these hub people can help the diffusion prediction.

Influence analysis concept has risen with sociology and psychology trying to understand human behavior and the affect of other's on human behavior [11–15]. Studying sociological and psychological aspects of influence helps us better understand how certain trends spread more and faster than the others and who are the key influencers on these trends. Influence analysis is gaining more and more attention parallel to broadening of its application areas, with any kind of network where peers interact with each other. With the increasing usage of social media in our lives, social networks became important medium that need to be analyzed to understand human behavior and interaction. Katz at al. [16] theorized that a few users called "influencers" can create a chain-reaction of influence that is based on word-of-mouth approach and result with reaching a broad audience. This idea is similar to viral marketing where companies try to reach as many customers as fast as possible over social media with a low cost.

In this work, we define social influence as the positive effect of users on others' behavior on social media that result with sharing information similar to definitions in [17–19]. Similarly, Aral et al. [20] recognize social influence as a key factor in the propagation of ideas, behaviors, and economic outcomes in society.

In an earlier research [21], where experiments were conducted on a large group of users in Twitter, it has been concluded that people tend to be authorities on specific topics such as sports, economy, politics rather than being global authorities. This result led us to explore more on topical authorities rather than global ones. Our proposed approach considers users' personal information and tries to identify and predict topical authorities. This personal information can be, for example, how focused a user is on a specific topic,

* Corresponding author.
E-mail addresses: zzalp@itu.edu.tr (Z.Z. Alp), sgunduz@itu.edu.tr (Ş.G. Öğüdücü).

https://doi.org/10.1016/j.knosys.2018.10.020
0950-7051/

how actively users are posting on a specific topic, or how authentic a user is on a specific topic.

One of the most important aspects of social media is that it is so dynamic and fast, nodes and edges appear and disappear very frequently. Hence, it is also important to make predictions about who may be influential in the future rather than only identifying current influencers. In this study, we identify influencers, who would spread the information more, on specific topics. We believe that these users tend to be topical authorities, meaning that they would be authorities for a small number of topics. They are regarded as experts on a few number of topics. Moreover, since social media is very dynamic and unpredictable, detecting future influencers becomes harder. In a previous study, influencers were identified using topological features of the social network and user features [22] extracted from the network, which are both very useful to determine influential users. In this work, we use those user features to predict latent factors of these features using Matrix Factorization and identify potential influencers as well as current influencers.

One of the most promising recommendation approaches for social media mining is Matrix Factorization, that can also be used making predictions on elements of a matrix. It has been proven to be superior to the classic nearest-neighbor algorithms with the Netflix Prize competition [23] for recommender systems. It works on a sparse user–item matrix and tries to identify latent factors on the user preferences. And finally, generates a new user–item matrix where we can also see predicted ratings of users on items which were not present in the initial matrix. In a traditional recommendation model, this approach helps one to predict ratings of items for each user and recommend them the high rated items as a prediction set.

Our approach, Influence Factorization (IF), is an efficient approach that utilizes user specific features to identify topical influencers. Two different sub-approaches are utilized within IF. First one uses a User–User Matrix where entries are normalized retweet rates. This approach is called User–User Influence Factorization (U-UIF). The second approach uses User–Topic matrices, where entries of these matrices are filled with user specific features, like focus rate, activeness, follower count or combinations of them. For each user feature, a different matrix is generated. This sub-approach of IF is called User–Topic Influence Factorization (U-TIF).

Different from previous studies, we focus on predicting the future influencers as well as current influencers of the network. For this aim, we will address to predict user features that are related to being influential for a given set of users and topics. The proposed approach can be run in a distributed manner which makes it efficient when working with large scale data sets. The method enables incorporation of user specific features and topic specific features which benefits the learning of latent factors for user specific features for each user and topic. We also introduced hybrid features that combines two different user specific feature. The proposed method is compared with similar methods in the literature. A relatively large data set is collected and used in the experiments compared to many recent works [18,24–28]. The results of various tests show that, the proposed method is superior to other methods in terms of predicting future influencers and current influencers.

The rest of the paper is organized as follows. First, a thorough related work will be explained with similarities and differences of proposed methodology. Secondly, baseline and comparative methods will be given that we compare performance of proposed method with. Then, Influence Factorization method and its variations will be explained in detail. Afterwards, results of experiments will be demonstrated and explained. Finally, conclusion of the research and future directions will be stated.

2. Related work

In this section, influence detection algorithms that are generally used for social media will be explained and analyzed.

There are several influence analysis approaches, many of which use relationship graphs to model user influence on other people. These graph-based approaches are able to model diverse types of information obtained from network topology. These methods can also be categorized into two sub-groups, such as diffusion models and influence models. Although these fields are two separate areas of research, they can be both used to identify influencers since we define influence as the maximized information diffusion and influencers as the users who contribute most of the diffusion of information. There are some classical information diffusion models such as Linear Threshold (LT) [29] and Independent Cascades (IC) [30] models that try to identify diffusion of information using the network topology. Both models classify nodes in the network as active and inactive depending on the information exposure. These are threshold based models such that when a user becomes active, it activates its followers with a probability. If this probability is above a certain threshold, the follower becomes active too. Assigning these probabilities and thresholds, which is an NP-Hard problem [31], is another challenge as addressed in many studies [17,32–34,24].

Riquelme et al. [35] recently proposed a centrality measure based on Linear Threshold model and compared the spread of information between their proposed centrality measure and other ones.

Lagree et al. [28] recently proposed an improvement to diffusion models. Their principal argument was that offline calculation of activation probabilities in IC and LT models is not effective for online social networks. In their proposed iterative model, a learning agent picks the nodes among the candidates those from which a new diffusion process is initiated in the network. The agent gathers some feedback on the activations and adapts the subsequent steps of the campaign which makes the process online and learning on the go.

Another graph based approach for finding influencers in a social network is Google's well-known PageRank algorithm [36]. This algorithm was proposed to rank Web pages in Web search results by assigning an importance score to each Web page. Kwak et al. [37] used this algorithm to compare the performance of PageRank scores with retweet rate and follower count. They identified that PageRank scores are highly parallel to follower count of users. On the other hand, retweet rate and PageRank scores do not yield to similar influencer sets. Similarly, in our proposed approach we also will demonstrate that using follower count in combination with other user features improves influencer identification performance. However, follower count by itself does not yield to the best performance.

Similar to PageRank, Weng et al. [38] proposed TwitterRank algorithm. Their algorithm considers topical information of tweets and topical similarity between each pair of users. They proposed an algorithm which runs on topic specific networks and they used biased transition between users with respect to topical similarities between them. This algorithm will also be analyzed in detail in Section 4 as a baseline methodology.

Recently, PageRank like algorithm, PPR [22], was proposed. This methodology incorporates user features to PageRank algorithm to favor users in the network who are more active, authentic or focused. This algorithm was implemented on Spark which can run in parallel and can be scaled to as many nodes as available.

Another recent research used a bio-inspired approach for influence maximization problem. In their research Sankar et al. [25] proposed a new approach which works on a Twitter network that is formed as the re-tweet network of hashtag #KissofLove. They

inspired by waggle dance algorithm which is a communication process of honey bees. The proposed algorithm utilizes the global–local search capacity of the Artificial Bee Colony algorithm to solve the influence maximization problem. This is one of the few nature inspired algorithms that is used to solve NP-hard Influence Maximization problem.

Recently, there are studies [39,40] that address to sample the social network to make it easier to process while not losing valuable information. Tsugawa et al. [40] proposed a node sampling strategy in the identification of influencers and tried on several different networks. They demonstrated that %30 of the network is adequate for influencer identification. However, these studies focus on investigating the effects of different network sampling strategies on influence identification when using different influence measures. In this study, our proposed algorithm can run in parallel and can be scaled to as many nodes as available. This is one of the advantages of our method to handle big data sets without the need of sampling the original data set.

There are other approaches which do not consider network topology. Instead, these approaches usually focus on user behavior. For instance, Cha et al. [41] recently analyzed the effect of follower count, retweet count and mention count of users on being influential. They used a Twitter sub-network and identified that follower count is not a key indicator of influence, rather retweet and mentions are better indicators of being influencer. This means that influencers are users who increase information diffusion. This definition is also similar to our definition of being influential. Similarly, Pal et al. [42] identified several user metrics such as topic of interest or originality of tweets, and using these features they applied Gaussian clustering and ranking algorithms to identify authorities in Twitter.

Matrix factorization models which rely on using information between two groups of entities are also utilized for influencer identification in social networks. Cui et al. [43] proposed an item level matrix factorization methodology to rank users based on their social influence scores. They tried to identify "who should share what?" in order to maximize information diffusion for each user. Their approach used a user–item matrix formed where items of the matrix are posts of users. Hence, they needed to take only a small sample since matrix factorization is a complex approach in terms of computation. Our proposed approach uses user level or topic level matrices. This makes it more scalable than using item level matrices.

Zhao et al. [44] proposed a behavior factorization model in order to recommend topics to users. This matrix factorization based work uses Google+ dataset and they suggest that users behave differently on different topics. For example, one user posts on different topics while he/she comments, puts a plus one, or reshares on other topics. They identified topics of posts using *Google Knowledge Graph* and created different user–topic matrices for each user behavior. The output of their system is used to recommend topics to users for each behavior. This approach is highly feasible and effective for topic recommendation since they use user–topic matrices rather than user–item matrices which are too large since items are posts of users. Similar to this idea, we also apply our algorithm on user–topic matrices rather than using a large user–item matrix. In addition, we have incorporated other user specific features rather than using just count of user actions.

In this work, we propose a novel model, called IF, for finding influencers for different topics of a information network. This approach models the users with user specific features, considers topical expertise and leverages latent information among users, topics and users features. It is a highly personalized, efficient and effective approach. Although we previously utilized influence scores of a given network of users and their tweets using PPR algorithm [22], this approach was calculating influence over a known universe of users. The tweets were collected in a certain period of time and the network was constructed using a snapshot of entire Twitter network. However, by nature, social media evolves all the time and is never static. In this work, using the information we have collected, we are addressing to find latent factors of influence and predict future influencers as well as current ones. We experimented on Matrix Factorization with different input matrices and identified that U-TIF methodology is best performing approach among Influence Factorization and benchmark methodologies where entries are used as a hybrid feature using authenticity and follower count information.

3. Baseline and comparative methods

In this section, we will compare our proposed model with state of the art and new methodologies. The first method is the well-known PageRank [36] (denoted as PR) algorithm. Second one is TwitterRank [38] (denoted as TR) algorithm and the last one is newly proposed Personalized PageRank [22] (denoted as PPR) algorithm. All algorithms use network information as well as topical information. PPR algorithm also incorporates user features that improve the influence identification process.

PR is a well known algorithm to rank Web pages. However, now it is widely used to rank nodes in any type of network. It recursively assigns importance scores to nodes such that a node A contributes to its friends' scores with its own score divided by its out degree. Thus, a high scoring user (influential) contributes more to its friends especially when she has fewer friends.

Eq. (1) is the formulation that is used to calculate scores of Web pages with PR algorithm. The first line of the equation is the initialization of scores where the initial score of each node ($PR^{(0)}(v)$) is set to 1. In the second line of the equation, F_v denotes the set of the nodes pointing to v, and N_u denotes outgoing degree of node u. Here, new rank of node v is calculated using its previous rank and the weighted ranks of its incoming links. This calculation is computed for each node in each iteration. Page et al. [36] demonstrated that having 50 iterations is sufficient for convergence of scores in most of the cases.

$$\begin{aligned} PR^{(0)}(v) &= 1, \\ PR^{(i+1)}(v) &= \sum_{u \epsilon F_v} PR^{(i)}(u)/N_u \end{aligned} \tag{1}$$

A damping factor d is incorporated into this equation as shown in Eq. (2) to deal with sinks, cycles or disconnected components in the graph. Damping factor is used to model a random surfer who jumps to any other Web page rather than following links on current Web page. Eq. (2) models the random surfer model for PageRank calculation. The damping factor d is usually set to 0.85.

$$\begin{aligned} PR^{(0)}(v) &= 1, \\ PR^{(i+1)}(v) &= (1-d) + d \sum_{u \epsilon F_v} PR^{(i)}(u)/N_u \end{aligned} \tag{2}$$

In this research, PR is applied to topical networks rather than one global network. Topical networks only contain users who write on a specific topic as in [45]. Hence, topic specific version of PR is used as a baseline algorithm.

For the second baseline methodology, TR algorithm was implemented as in [38]. TR is similar to PR except the way of transition between nodes. While PR's random surfer jumps to a random node arbitrarily, TR favors users who are similar to each other. The similarities between users are defined based on the topics that users posted on. Thus, a user–user similarity matrix should be stored for each topic, which makes space complexity quadratic to user count.

Third baseline methodology is recently proposed Personalized PageRank (PPR) algorithm [22]. In PPR method, besides topical

information and network topology, user features like focus rate, activeness, authenticity, and speed of getting reaction are also incorporated to identify influencers in a topical network. A PageRank like algorithm has been also used, where transitions between users were favored with higher user features. Hence, if a user has many followers, especially with high scores, and is highly focused, active, authentic, and/or getting fast reaction, she ends up with higher influence scores. Eq. (3) gives formulation of PPR algorithm where w_u^t is a user feature specific to each user u on a topic t.

$$PPR^{(0)} = 1,$$
$$PPR^{(i+1)}(u) = (1 - w_u^t) + w_u^t \sum_{v \epsilon F_u} PPR(v)/N_v \quad (3)$$

Baseline models PR, TR, and PPR are topic specific same as the proposed methodology IF. Moreover, besides topical information, PR considers only network properties, TR also considers nodal features such as topical similarity, and PPR considers user specific features like focus rate, activeness, authenticity. Proposed IF model also leverages user specific features while it does not use network information. Lastly, IF model identifies current influencers as well as predicts future ones, while none of the baseline methodologies do prediction.

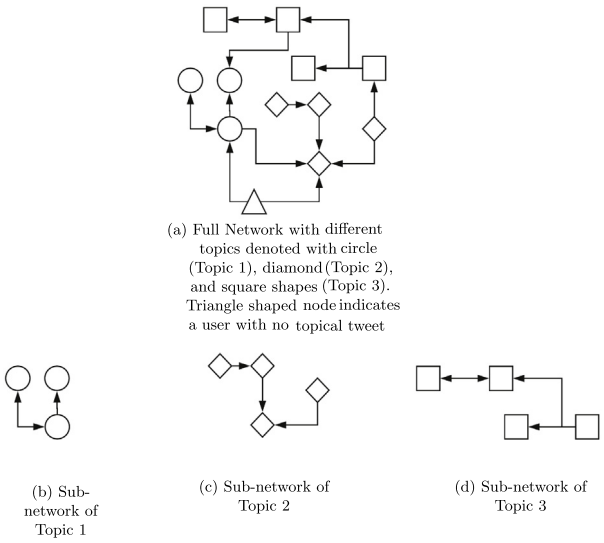

Fig. 1. Topical network construction example.

4. Influence factorization (IF)

In this section, we explain our data collection, preprocessing steps as well as the topic modeling, and influence analysis methodologies. Proposed methodology utilizes Matrix Factorization methodology on either User–Topic or User–User Matrices to identify highly influential users and are called User–Topic Influence Factorization (U-TIF) and User–User Influence Factorization (U-UIF) respectively.

4.1. Data collection and preprocessing

Prior to data collection, 20 Twitter users were manually identified that are known to be focusing on different topics such as politics, sports, TV, religion etc. Afterwards, Twitter user ids of friends and followers of these users are collected in a breadth first manner until sufficient number of users are obtained. Since Twitter APIs do not allow us to see information and tweets of protected profiles, protected users were eliminated from the initial set. Then, 20 tweets of remaining users were retrieved, and language code of these tweets were checked to identify the users whose tweets are in Turkish most of the time (80%). Users who do not post mostly in Turkish were also eliminated. At this point we had around 180K public user ids who post mostly in Turkish.

Using Twitter Streaming API, tweets of these users were collected between November, 4th 2015 and January 12, 2016. Afterwards, the data set was divided into training and test sets where tweets before December 30, 2015 are separated as the training set and the rest as the test set.

After data collection, we applied stemming to all of the tweets as the first preprocessing step of topic modeling. Afterwards, stopwords, punctuations, and mentions that have no effect for topic modeling were also eliminated.

4.1.1. Topic modeling

For topic modeling, Latent Dirichlet Allocation (LDA) was applied on stemmed and cleaned tweets with a tool called MAchine Learning for LanguagE Toolkit (MALLET) [46]. MALLET uses smoothed LDA for topic modeling as Blei et al. proposed in [47].

Since LDA is a bag-of-words approach, short text like tweets lowers the topic modeling performance. However, some pooling approaches [48–51] were proposed to enhance LDA performance on short text like Tweets. In a previous work [51], it has been shown that pooling tweets for each user and each day gives better performance than other pooling techniques. Hence the same approach was used in this research.

After running LDA on pooled tweets, topics were formed as explained in [51]. Specifically, three human experts analyzed the output of LDA to determine most coherent clusters of words that form a topic. Some of the clustered words was not semantically coherent enough to form a topic, thus, they were eliminated. Finally, experts labeled the coherent topics with appropriate topic titles.

After assigning topics to all of the tweets, topical networks were formed. For each user–topic pair percentage of topical tweets were calculated. Afterwards, topical labels were assigned to users according to topical contents of their tweets. Finally, for each topic, a sub-network is generated where nodes represent the users who post on the specific topic and edges represent the following relationship. A certain threshold is introduced in such a way that the user would be not added to the sub-network if the percentage of topical tweets of that user is below this threshold [51].

Fig. 1(a) demonstrates an example global network where each node is a different user. Shapes of the nodes represent the users who post on specific topic. Fig. 1(b), 1(c), and 1(d) are sub-networks constructed using topical information from global network. Notice that inter-topical edges are dropped, and users may not belong to any topical network if their tweets do not contain any topical information over the threshold. Moreover, a user can be in more than one network. This situation occurs when a user posts on different topics where each percentage of topical tweets are over the threshold.

4.2. User modeling

For topical influence analysis, after topic modeling and construction of topical networks, the next step of the IF is modeling users with user specific features. Several different user features, that are used in the literature, which might be correlated to being influential were identified for user modeling task. User features that will be employed for user modeling are: focus rate, activeness, and authenticity [22]. All of these features were calculated for each user and topic using simple parameters in Table 1 and are obtained using training set.

"Focus Rate" is a feature that shows how focused a user is given a topic. Intuitively the idea was that an authoritative user is usually focused on a single topic and would not write on many topics all the

Table 1
Parameters used for user modeling.

Param.	Explanation
p_u	tweets of user u
p_u^t	tweets of user u posted on topic t
$p_{u,rt}$	retweeted tweets of user u
$p_{u,rt}^t$	retweeted tweets of user u on topic t
rt_u	retweets of user u (retweeted by u)
rt_u^t	retweets of user u on topic t
d	total number of days
d_u^t	number of days user u posted on topic t
rt_{time}^p	duration passed for first retweet of post p
fc_u	follower count of user u

time. This feature is calculated as the percentage of users' tweets on topics as in Eq. (4), where p_u^t and p_u are used as in Table 1.

$$fr_u^t = \frac{|p_u^t|}{|p_u|} \tag{4}$$

"Activeness" is the feature to show how often does a user post tweets about a topic and calculated as in Eq. (5). The reason of using this feature as a user feature is that this feature might be a good indicator of whether users who frequently post on a specific topic are influential on that topic or not.

$$ac_u^t = \frac{d_u^t}{d} \tag{5}$$

"Authenticity" measures whether a user creates original content with her own words rather than re-tweeting most of the time. The intuition behind using this feature is that influential users tend to use their own words rather than transferring others' thoughts. It is calculated using Eq. (6):

$$au_u^t = \frac{|p_u^t| - |rt_u^t|}{|p_u^t|} \tag{6}$$

In this study, we also introduced hybrid features that combines two different user specific feature. In addition to the three aforementioned features, follower count is also used to calculate the hybrid features using Eq. (7). For each hybrid feature, exactly two different features are used. The weights of these features contributing to the hybrid feature are optimized using α and β coefficients. These coefficients were set using a validation set which is part of the training set consisting of chronologically last 20% tweets of each user. In the equation, um_u^t can be any of the three user features mentioned above in addition to follower count, fc_u. For hybrid feature calculations, all of the user specific features and follower count are normalized to 0–1 interval.

$$hy_u^t = \alpha \times (um_u^t)_1 + \beta \times (um_u^t)_2 \tag{7}$$

For instance, to combine follower count and authenticity features, $(um_u^t)_1$ corresponds to the follower count and $(um_u^t)_2$ corresponds to the authenticity in the above Equation.

4.3. Influence analysis

After topic modeling and user modeling steps, the next step is the Influence Analysis in order to identify topical authorities in the data set. Since social networks are highly dynamic, user's activity can change over time, new following relationships arise or disappear every second. Influence analysis studies are conducted usually on networks which are only snapshots of subsets of the Twitter network since the whole network is very huge to effectively analyze for influencers. An influential user might have been passive for any reason during the period of time that the snapshot is constructed, or a new user might have entered the network recently but would be highly influential later. Hence an effective method for influence identification needs to be developed to capture latent factors of being influential and be able to predict influence scores for later.

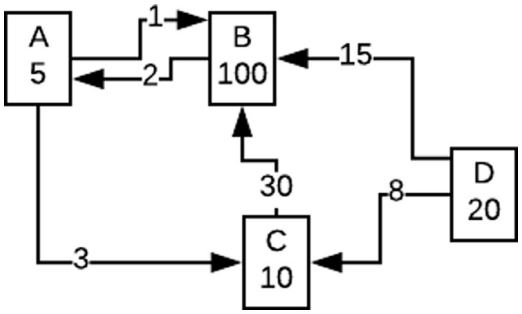

Fig. 2. Example retweet network.

We investigated on Matrix Factorization techniques with different matrices and identified the most efficient and best performing algorithms. Since Matrix Factorization gets a sparse user–item matrix and fills the empty entries with predicted user scores to those items, it can be used to predict user ratings for items for unseen observations. Using a similar idea, two separate set of experiments were developed using Matrix Factorization techniques, where we want to predict the unseen observations of those matrices. First one is User–User Influence Factorization (U-UIF), and the second is User–Topic Influence Factorization (U-TIF).

4.3.1. User–User influence factorization

Recall that our definition of being influential is positively correlated with the affect of spreading the word. To calculate the influence scores of users, normalized retweet rate of users were measured as spread score, which will be explained more in the Evaluation subsection. We want to predict who will get more retweets given a topic. With the knowledge of retweet rate of collected tweets, we can use Matrix Factorization and find out who would retweet whose tweets in the future.

User–User ($U-U$) matrices for each topic have been formed and filled with the normalized retweet rates of users. We have denoted the $U-U$ Matrix as $X \in \mathbb{R}^{M\times M}$ where M is the number of users. The entries of $U-U$ matrix are calculated as in Eq. (8).

$$X_{i,j} = \begin{cases} rt_{i,j}/p_j & if\ u_i\ retweeted\ u_j \\ 0 & otherwise \end{cases} \tag{8}$$

Fig. 2 is an example retweet network where arrows shows the retweet information of users. Values in the nodes represent the total number of tweets of the user (p_j) and the values of the links represent the retweet count per user pairs ($rt_{i,j}$). The U-U matrix of the example that will be used for Influence Factorization would be:

$$\begin{bmatrix} 0 & 0.01 & 0.30 & 0 \\ 0.40 & 0 & 0 & 0 \\ 0 & 0.30 & 0 & 0 \\ 0 & 0.15 & 0.80 & 0 \end{bmatrix}$$

For each topical network, different $U-U$ Matrices were generated where the smallest one has 3.5K users and the largest one has 55K users. Approximately, 0.01% of these matrix entries are filled, which shows that the data is highly sparse. The Matrix Factorization methodology would fill out the empty entries of the matrices with the predictions of retweets of users of other users' tweets. Since we only have the observed data that has been collected for 70 days, this idea of filling empty entries would increase the accuracy of influential user detection since social media is highly dynamic.

4.3.2. User–Topic influence factorization

Besides identifying unseen observations on retweet rate of users, another valuable information is to identify the unseen user specific features for each user and topic. For instance, a user might be highly focused and authentic on some topic, however, we might have captured only a small portion of her tweets due to some reason. For the purpose of user specific feature prediction, User–Topic ($U-T$) matrices were created for each user feature and also for hybrid features in order to predict unobserved values. Similar to previous approach, a Matrix Factorization algorithm were applied to those matrices.

We denote the $U-T$ Matrix where M is the number of users and T is the number of topics as $X \in \mathbb{R}^{M\times N}$. Entries of these matrices consist of user features (um) specific to each user feature for user i and topic j for each (i, j) pair as in Eq. (9).

$$X_{i,j} = \begin{cases} um_{i,j} & \text{if } u_i \text{ posted on topic } t_j \\ 0 & otherwise \end{cases} \tag{9}$$

For instance, suppose there are users such as A, B and C and four topics like T_1, T_2, T_3, T_4. The U-T matrix for authenticity user feature would be like below matrix where authenticity of A for T_4 is 100%, but 0 for other topics since she never posted a tweet on other topics. User B has 40% authenticity in T_1 and 50% authenticity in T_3 and user C has 30% authenticity in T_2.

$$\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0.40 & 0 & 0.50 & 0 \\ 0 & 0.30 & 0 & 0 \end{bmatrix}$$

There is a separate $U-T$ matrix for each user feature and for hybrid features. Approximately each $U-T$ matrix is 168K by 6 where about 360K entries are non-zero, that makes approximately 35% of the matrices are filled, which are much denser then the $U-U$ matrices.

For both IF approaches ($U-UIF$ and $U-TIF$), the influence prediction problem is transferred to the problem of predicting the unobserved entries in X.

We use Alternating Least Squares (ALS) algorithm that is an efficient Matrix Factorization methodology for collaborative filtering. Using Spark MLlib, we can apply ALS algorithm in a parallel manner which is highly efficient and gives good results [23]. ALS is used for Netflix competition to recommend movies to users [52].

To define ALS mathematically, let $X \in \mathbb{R}^{M\times N}$ be the user–topic matrix in an influence prediction system, where M and N are the number of users and topics respectively. The matrix factorization problem for this systems can be solved using Eq. (10), where Ω is the set of indexes for observed user features, with indices i and j; λ is the regularization parameter; $\|.\|_F$ denotes the Frobenius norm; w_i^T and h_j^T are the ith and the jth row vectors of the matrices W and H, respectively. The goal of problem is to approximate the incomplete matrix X by WH^T, where W and H are user and topic feature matrices respectively [53].

$$\min_{\substack{W\in\mathbb{R}^{m\times k}\\ H\in\mathbb{R}^{n\times k}}} \sum_{i,j\in\Omega} (X_{ij} - \boldsymbol{w}_i^T \boldsymbol{h}_j)^2 + \lambda(\|W\|_F^2 + \|H\|_F^2) \tag{10}$$

Eq. (10) is non-convex and NP-hard to optimize. However, when we fix W, then the objective function is a convex function of H, and vice versa.

Each user and each topic have a k-dimensional "feature vector" describing its characteristics, and each entry of the matrix (user's specific feature of a specific topic) is the dot product of user's and the topic's feature vector. Iterative process of ALS can be defined as in Algorithm 1.

Algorithm 1 ALS algorithm

Input: X, λ, *number_of_iterations*
1: Initialize W and H
2: **for** $l \leftarrow 1$ to *number_of_iterations* **do**
3: **for** $i \leftarrow 1$ to N **do**
4: optimize H given W is fixed (using Eq. (10))
5: **end for**
6: **for** $j \leftarrow 1$ to M **do**
7: optimize W given H is fixed (using Eq. (10))
8: **end for**
9: **end for**

4.4. Complexity analysis

Eq. (10) is a non-convex problem. However, when optimizing either W or H, it becomes a quadratic problem with a globally optimal solution. ALS monotonically decreases the objective function value in Eq. (10) until convergence [54]. Updating each element in H costs $O(n_i k^2 + k^3)$ where n_i is the number of topics that user u tweeted about (for U-T matrix), and similarly updating each element in W costs $O(n_j k^2 + k^3)$ where n_j is the number of users that have posted in topic j. Hence, complexity of Matrix Factorization is cubic with respect to the size of feature vector. However, it is proven that it can be parallelized and run much faster [52]. There are several ALS libraries on different platforms like Spark, Hadoop etc. that run in a distributed manner.

Fig. 3 shows the major steps of overall approach. Tweet Collection, Preprocessing, Topic Modeling and User Modeling steps are performed once before Matrix Generation and Influence Analysis steps. Tweet Collection and Preprocessing steps are linear with respect to the total tweet count. Topic modeling with LDA has time complexity proportional to $O(NVK)$ where N is the number of documents, V is the number of words and K is the number of topics. In our experiments, the number of documents corresponds to the number of pooled tweets by user and day. Thus, for each user, the number of documents is the number of days she posted at least one tweet. Moreover, number of words in our experiment is the number of unique words in all of the tweets. After topic identification, complexity of User Modeling is also linear to user count. The Matrix Generation and Influence Analysis steps constitute the actual proposed methodology that can be used with any other user model, not necessarily the one implemented in this research.

4.5. Evaluation

Evaluation of influence identification results is another hard task. Since there is no objective ground truth it is hard to calculate accuracy, precision, recall like metrics. Since the objective is to maximize the spread, there is several ways of evaluating the influence set. One of them is calculating retweet or mention rate of identified influencers [41,55], and another one is calculating node activation based on IC and LT models [17,24,31–34]. Spread score is a newly proposed measure for evaluation of influencer identification methods which is similar to the former approach [22].

In this paper, we used node activation and spread score to evaluate the performance of our proposed method. As in IC model, node activation is the total number of users that adopt the behavior of users who were initially active. In this research, we have calculated the number of users that would retweet a post that is tweeted by the users in the influencer set. To calculate the activated nodes using an initial activated seed set, edge probabilities and a threshold for activation should be calculated. In this research, edge probabilities are calculated as retweet rate between each pair of users, i.e. if a user u_i retweets 6 of user u_j's 10 topical

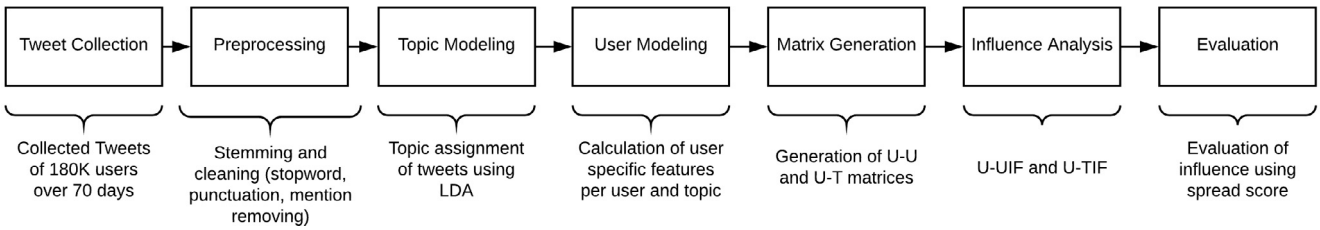

Fig. 3. Overall approach of influence factorization.

Table 2
Statistics about data sets.

	#users	#tweets	#links/follows
Training set	186K	31M	16M
Test set	181K	7M	15.6M

tweets, the edge probability for u_j to activate u_i is set to 0.6. In an earlier research [22], different threshold values were evaluated and seen that almost all threshold values yield to similar performance ordering of different techniques. Hence, a single threshold value is selected, which is 0.5 for this research.

Spread score is the total normalized retweet rate of influencer user set. In order to calculate spread score, for each experiment (baseline and proposed), we identified top 25 influencers of each topic t as the output of experiments. Normalized retweet rate, called spread score, for each user is calculated and the results are summed to calculate the overall information diffusion estimate of top 25 influencers identified as the influencer set of topic t (inf_t).

Eq. (11) demonstrates how potential information diffusion is calculated.

$$spread(t) = \sum_{u \epsilon\ inf_t} \frac{\left|p^t_{u,rt}\right|}{\left|p^t_u\right|} \sum_{p \epsilon p^t_{u,rt}} \left|retweets_p\right| \tag{11}$$

In the equation, $p^t_{u,rt}$ and p^t_u parameters are retweeted tweets and tweets of user u on topic t respectively, and inf_t is the set of users in the influencer set for topic t. For each retweeted tweet p of $p^t_{u,rt}$, we summed up the number of retweets and normalized this sum with the retweet rate of the user. This formula gives us the potential spread of posts of influential users. All of the variables in Eq. (11) are calculated using test data set.

5. Experimental results

This section demonstrates first data collection and preprocessing statistics. Then, results of topic modeling, user modeling and IF will be given respectively.

The friends and followers of a set of users which identified manually have been obtained in a breath first manner. Then, tweets of these users were collected for a 70 day period. Afterwards, several preprocessing steps are conducted such as stemming, and removal of stopwords, punctuations, mentions.

Table 2 demonstrates the statistics about the resulting data set. 38M tweets of 186K users were divided into training and test sets as 31M and 7M tweets respectively. As seen from the table, 181K of 186K users also appeared in the test set. The very few number of users who appeared only in test set were eliminated to get rid of cold start problem. Cold start problem should be tackled in another research and is kept out of scope for this work. Moreover, as seen from the table, not all users in the training set appear in the test set. Those users might have also been eliminated but have been kept in order to achieve a more real life scenario.

For topic modeling, an LDA tool called MALLET [46] has been used on the dataset where each document is represented with all tweets that a user has been tweeted on one day. The output of MALLET tool is clusters of words which indicates semantic/topical similarity between words within a cluster. In this research, three human experts identified six coherent topics and labeled them with a topical keyword. After this process, there were six topics formed and named as: "Politics/Breaking News", "Religion", "Spiritual", "Social Responsibility", "Soccer/Sports", and "TV/TV Shows". There was a major political election in Turkey in the middle of the tweet collection period. Hence, "Politics" and "Breaking News" topical words were highly overlapping, thus a single topic is formed for both topics. Table 3 demonstrates selected topics and top words in the topics.[1]

After topic modeling is completed, topical networks are constructed (recall Fig. 1) by forming graphs where nodes represent users who post on the specific topic and edges represent the following relationships where edge direction is from follower to friend. Users whose topical tweet ratio is less than 25% are discarded from topical networks in order to avoid users who rarely post on the topic as a noise cancellation process. After topical network construction, a user might end up in zero, one or many topical networks.

Table 4 demonstrates statistical information of each topical network.[2] As shown in the table, although the entire network is very large and more dense than the sub-networks, average number of tweets and retweets are not respectively large. This shows that our sub-networks are filtering out users that are not active and not passing information a lot. This makes sub-networks more valuable and computationally feasible. Table 4 also demonstrates that "TV/Shows" topic has larger average tweet rate than the other topics; however, its retweet rate is smaller. Thus, people are less willing to pass information on this topic. Table 4 also shows that retweet rate is very large for "Soccer/Sports" and "Politics/Breaking News" topics. Hence, people want to pass this type of information to others more than the other topics.

Afterwards, user modeling is performed based on the user specific features. Up to this point, each tweet has an owner and zero, one, or more than one topic assigned to it. With this information we can calculate how focused a user is, how active she is, or even how authentic she is as we also know if those tweets are retweets or her own words. These calculations are done for per user and per topic. Hence, we have many information about users which can be used to distinguish them. Moreover, follower count of each user is also calculated for using in hybrid user features.

For hybrid feature calculations, we also experimented on selection of different α values. For the experiments conducted on U-TIF (au&follower) experiment, where authenticity and follower count features are used to calculate the hybrid feature, we realized that different α values can be better for different topics as demonstrated in Fig. 4. These results show that while setting α to 0.5 is best for Politics/Breaking News topic, 0.8 is better for Religion topic. These α values vary for each hybrid specific feature as well as the topic.

[1] Topic words are translated into English for better representation in the paper.
[2] Data set is available upon request with limitations of Twitter privacy policy.

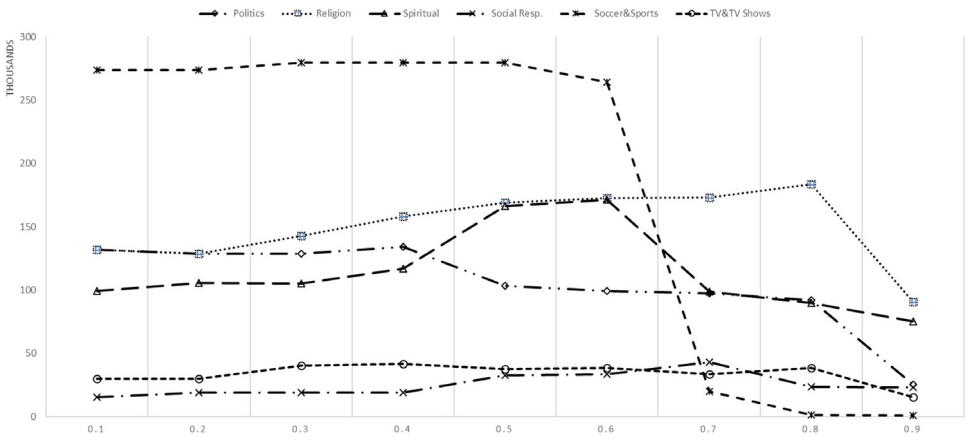

Fig. 4. Spread scores of different topics for U-TIF(au&follower) experiment using different alpha values.

Table 3
6 Topics generated by topic modeling using LDA.

Politics/News	Spiritual	Religion	Social Responsibility	Soccer/Sports	TV/TV Shows
akp[a]	love	allah	need for	galatasaray[b]	watch
hdp[a]	beautiful	muhammed	blood	soccergame	show
chp[a]	#loveis	Rabbi	rh	goal	scene
syria	happy	pray	#blood	team	trailer
#brkngnews	hearth	hz	#urgent	beşiktaş[b]	#kirazmevsimi[c]
martyr	peace	muslim	#thrombocyte	fenerbahçe[b]	#kiralıkaşk[c]
paris	poem	religion	#collectingbooks	soccer	#güneşinkızları[c]
terror	lover	belief	#3decemberdisabledday	league	#poyrazkarayel[c]
france	#whatislove	heaven	awareness	fan	#pantenealtınkelebek[c]
#lastminute	#loveisactually	islam	#helptheoneinneed	champion	#muhteşemygüzyılkösem[c]

[a]A Turkish political party name.
[b]A Turkish soccer team.
[c]A Turkish TV Show.

After all preprocessing and modeling steps, influence analysis is conducted. As discussed in Section 4, we used ALS algorithm that is implemented for Spark MLlib and can run in a distributed manner.

First, we generated U-U Matrix where elements of the matrix contain normalized retweet rate of users. We ran ALS on this U-U matrix for U-UIF experiment. The output of this process gives us prediction of users who will get most retweets. When we get the top 25 of these users and calculate their spread scores in the test set, the outcome of the average spread of those users were lower than the benchmark algorithms for most of the topics (See Table 5). For TV/TV shows network there is a huge performance increase when we applied U-UIF methodology. The reason of this can be since this network is very dynamic with higher tweet rate and being less dense with edges, latent factors can identify future influencers better. For this specific network, the mostly influential people in the test dataset were not the most influential in the training set and U-UIF methodology captured well these people. Hence, to optimize the alpha value different values should be tested with a validation set for each experiment.

Secondly, U-T Influence Factorization experiments were conducted. First set of these experiments were applied with single features such as focus rate (fr), activeness (ac), authenticity (auth), an follower count (fol). As demonstrated in Table 5 single user features do not provide better influencer detection results than benchmark methodologies. Among single feature experiments, only follower count feature yielded better spread score than some of the benchmark methodologies such as PPR, and for some topics PR, TR. Other single feature U-TIF experiments underperformed the baseline algorithms. This can be explained by that all benchmark methodologies already use the follower information while using network topology. Hence, follower count information is an important aspect of being influential, but not necessarily the only important information. For instance, for Politics/Breaking News topic, U-TIF (fr) algorithm provides better than TR algorithm, or for Spiritual topic it outperforms both TPR, and TR algorithm but under-performs PR, and all PPR based algorithms. Like U-UIF, this U-TIF (fol) based method is also better than all the benchmark methodologies for "TV/TV Shows" topic.

Table 4
Statistics about networks.

	nodes	edges	avg tweets	avg retweets
Entire Network	185,898	16,5M	168	106
Politics/brkgNews	33,462	2,5M	394	456
Spiritual	57,877	2,3M	309	320
Religion	15,092	286K	312	350
Social Resp.	1292	6K	383	270
Soccer/Sports	14,335	266K	276	560
TV/shows	3516	42K	445	208

Second set of U-TIF experiments was conducted on hybrid features. Note that, the spread scores for each hybrid feature in Table 5 is calculated with its best α value. First of all, every possible combination of features were created to find best hybrid features but only some of them are demonstrated in Table 5, for the sake of clarity and space.

We clearly see that scores for U-TIF is better when hybrid features are used. Especially when follower count is incorporated with other user features. Although we proved earlier that having the most followers does not necessarily mean being the most influencer [22,56,57,41], follower count information increases identification of influencers performance when used with other features. As seen from the table, using only follower count as single feature, U-TIF (fol), did not outperform the benchmark methodologies on

Table 5
Spread scores of benchmark and proposed methodologies.

Method group	Method name	Politics/BrkngNews	Spiritual	Religion	Social Resp.	Soccer/Sports	TV/TV shows
Benchmark experiments	PR	107,639	129,901	101,650	9100	224,765	967
	TR	69,803	69,259	52,089	6091	33,346	792
	PPRfr	114,624	114,433	162,128	39,105	274,717	1822
	PPRac	94,708	156,482	71,928	39,091	257,377	3,862
	PPRauth	**150,163**	162,816	173,546	39,128	275,663	2.949
U-U Inf. Fact.	U-UIF	20,134	120,844	61,998	14,350	27,289	37,733
Single feature U-T Inf. Fact.	U-TIF(fr)	19,726	5623	9732	3632	11,165	359
	U-TIF(ac)	18,470	6786	5790	3214	12,125	213
	U-TIF(auth)	35,090	11,567	8,898	4785	18,650	914
	U-TIF(fol)	85,652	106,896	148,361	13,127	237,303	23,190
Multi-feature U-T Inf. Fact.	U-TIF(ac&auth)	53,167	42,540	56,768	13,127	122,450	3134
	U-TIF(fr&fol)	78,467	156,346	165,754	21,274	212,546	32,913
	U-TIF(ac&fol)	117,470	161,776	154,980	38,287	192,134	41.716
	U-TIF(auth&fol)	134,400	**183,642**	**174,277**	**43,190**	**279,827**	**41,879**

Table 6
Activation scores of benchmark and proposed methodologies.

Method group	Method name	Politics/BrkngNews	Spiritual	Religion	Social Resp.	Soccer/Sports	TV/TV shows
Benchmark experiments	PR	134	44	74	31	35	27
	TR	139	51	65	30	27	29
	PPRfr	146	53	87	35	35	44
	PPRac	106	58	84	27	34	45
	PPRauth	153	69	125	32	37	44
U-U Inf. Fact.	U-UIF	71	35	74	32	24	84
Single feature U-T Inf. Fact.	U-TIF(fr)	89	36	65	33	11	22
	U-TIF(ac)	92	48	66	27	25	21
	U-TIF(auth)	96	46	75	25	24	27
	U-TIF(fol)	102	55	112	36	38	67
Multi-feature U-T Inf. Fact.	U-TIF(ac&auth)	87	48	98	31	29	64
	U-TIF(fr&fol)	106	69	130	29	35	75
	U-TIF(ac&fol)	136	71	129	36	33	84
	U-TIF(auth&fol)	**155**	**83**	**132**	**38**	**41**	**86**

most of the topics, but it increased performance drastically when used with other features, especially with authenticity feature, as in U-TIF (au&fol). However, for topic "Politics/Breaking News", none of the IF Method outperformed the previously proposed PPRauth methodology. The interpretation for this result can be that this network is mature and not very dynamic, so that simpler approaches that can only capture current influencers are well enough to detect real influencers.

Follower count has not been used in benchmark methodologies that are PageRank variations. That is because PageRank indirectly uses follower count by adding up scores of followers to calculate the scores of users. Hence, users with more followers get more contributions, but not necessarily larger contributions. High follower count also has a positive effect on PageRank like algorithms. In PPR, user features also incorporated to PageRank and authenticity was the best feature for most of the topics. Similarly, within IF methodologies using U-T matrix that has the hybrid feature calculated using both authenticity and follower count outperforms other IF based algorithms.

Node activation is also calculated to evaluate the proposed algorithm. Defining activation threshold is another hard problem to solve. If the activation threshold of the network is below this value, when *B* is activated, it also activates *A*. Table 6 shows performance of activation on different experiments on all topics. Table 6 also depicts that performance of experiments with spread scores in Table 5 are parallel to each other. Since spread score is easier to calculate, it can be used for influence evaluation on social networks.

Some insights from these results can be summarized as follows:

- Prediction of retweet for each user–user pair does not yield good results for large matrices. U-UIF methodology underperforms U-TIF for our datasets.
- Using hybrid features obtained by follower count incorporated with other user features increases the performance of IF methodologies significantly. The parameters of the hybrid feature can be tuned using a validation set.
- Nature of the network such as how dynamic it is, frequency of the edge and node changes in the network, is very important for methodology selection.
- Authenticity is a good user specific feature that can be used with other features together to promote the results.
- The experiments have been performed on a large data set. The results show the scalability of matrix factorization based approaches.

6. Conclusion and future work

In this paper, we propose a novel method called Influence Factorization and its variations. The proposed method is based on the idea that users may have different expertise and/or interests in various topics which lead them to be topic-specific influencers. The proposed model also leverages the latent factor identification of Matrix Factorization methodology to predict future influencers as well as current ones.

Several nodal features; *focus rate*, *activeness*, and *authenticity*, and hybrid features that have been calculated using these features in addition to *follower count* were used. The proposed method also uses ALS algorithm on Spark MLlib implemented in a distributed manner to increase the speed of calculations. Experimental results on a large data-set collected from Twitter show that using both user specific features and network features is more effective in identifying topical influencers.

As a future work, PPR and U-TIF methodologies can be combined. For example, U-TIF algorithm can be used to calculate user

specific features with latent factors, and those features can be used as the input to the PPR algorithm. Another future work can be using the *IF* algorithm for other purposes than detecting the topical influencers. For instance, we are going to investigate if *IF* algorithm can be used for identification of "troll" accounts. We believe that some demographic information and authenticity feature can be used with *IF* to detect trolling in social media. We may also compare the performance of *PPR* and *IF* algorithms on troll detection.

Acknowledgment

This research is partially funded by 2211 - TUBITAK BIDEB Ph.D. Scholarship Fund.

References

[1] D.T. Nguyen, J.E. Jung, Real-time event detection for online behavioral analysis of big social data, Future Gener. Comput. Syst. 66 (2017) 137–145.
[2] G. Burel, H. Saif, M. Fernandez, H. Alani, On semantics and deep learning for event detection in crisis situations, in: Workshop on Semantic Deep Learning (SemDeep), at ESWC 2017, 2017.
[3] Z. Xiaomei, Y. Jing, Z. Jianpei, H. Hongyu, Microblog sentiment analysis with weak dependency connections, Knowl.-Based Syst. 142 (2018) 170–180.
[4] J. Bernabé-Moreno, A. Tejeda-Lorente, C. Porcel, H. Fujita, E. Herrera-Viedma, Quantifying the emotional impact of events on locations with social media, Knowl.-Based Syst. 146 (2018) 44–57.
[5] S. Rosenthal, N. Farra, P. Nakov, SemEval-2017 task 4: Sentiment analysis in Twitter, in: Proceedings of the 11th International Workshop on Semantic Evaluation, 2017, pp. 502–518.
[6] E. Cambria, D. Das, S. Bandyopadhyay, A. Feraco, Affective computing and sentiment analysis, in: A Practical Guide to Sentiment Analysis, Springer, 2017, pp. 1–10.
[7] G. Petz, M. Karpowicz, H. Fürschuß, A. Auinger, V. Stříteskỳ, A. Holzinger, Reprint of: Computational approaches for mining user's opinions on the Web 2.0, Inf. Process. Manag. 51 (4) (2015) 510–519.
[8] N. İlhan, c.G. Öğüdücü, Feature identification for predicting community evolution in dynamic social networks, Eng. Appl. Artif. Intell. 55 (C) (2016) 202–218.
[9] S. Cavallari, V.W. Zheng, H. Cai, K.C.-C. Chang, E. Cambria, Learning community embedding with community detection and node embedding on graphs, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17, ACM, New York, NY, USA, 2017, pp. 377–386.
[10] M. Rawashdeh, M. Shorfuzzaman, A.M. Artoli, M.S. Hossain, A. Ghoneim, Mining tag-clouds to improve social media recommendation, Multimedia Tools Appl. 76 (20) (2017) 21157–21170.
[11] M. Deutsch, H.B. Gerard, A study of normative and informational social influences upon individual judgment, J. Abnorm. Soc. Psychol. 51 (3) (1955) 629–636.
[12] R. Lippitt, N. Polansky, S. Rosen, The dynamics of power; a field study of social influence in groups of children, Hum. Relat. 5 (1952) 37–64.
[13] A.R. Cohen, Some implications of self-esteem for social influence, Personal. Persuas. (1959) 102–120.
[14] J. Holdershaw, P. Gendall, Understanding and predicting human behaviour, 2008.
[15] N. Majumder, S. Poria, A. Gelbukh, E. Cambria, Deep learning-based document modeling for personality detection from text, IEEE Intell. Syst. 32 (2) (2017) 74–79.
[16] E. Katz, P.F. Lazarsfeld, Personal influence, the part played by people in the flow of mass communications, Can. J. Econ. Polit. Sci. 21 (6) (1957).
[17] A. Goyal, F. Bonchi, L.V.S. Lakshmanan, A data-based approach to social influence maximization, Proc. VLDB Endowment 5 (1) (2011) 73–84.
[18] S. Jendoubi, A. Martin, L. Liétard, H.B. Hadji, B.B. Yaghlane, Two evidential data based models for influence maximization in twitter, Knowl.-Based Syst. 121 (2017) 58–70.
[19] A. Goyal, Social Influence and its Applications (Ph.D. thesis), The School of the Thesis, The University of British Columbia, Vancouver, 2013.
[20] S. Aral, D. Walker, Tie strength, embeddedness, and social influence: A large-scale networked experiment, Manage. Sci. 60 (6) (2014) 1352–1370.
[21] Z.Z. Alp, Ş.G. Öğüdücü, Influential user detection on Twitter: Analyzing effect of focus rate, in: Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on, IEEE, 2016, pp. 1321–1328.
[22] Z.Z. Alp, Ş.G. Öğüdücü, Identifying topical influencers on twitter based on user behavior and network topology, Knowl.-Based Syst. 141 (2018) 211–221.
[23] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer 42 (8) (2009).
[24] Y. Mei, W. Zhao, J. Yang, Influence maximization on twitter: A mechanism for effective marketing campaign, in: Communications (ICC), 2017 IEEE International Conference on, IEEE, 2017, pp. 1–6.
[25] C.P. Sankar, S. Asharaf, K.S. Kumar, Learning from bees: An approach for influence maximization on viral campaigns, PLoS One 11 (12) (2016) e0168125.
[26] W.-X. Lu, C. Zhou, J. Wu, Big social network influence maximization via recursively estimating influence spread, Knowl.-Based Syst. 113 (2016) 143–154.
[27] W. Liu, K. Yue, H. Wu, J. Li, D. Liu, D. Tang, Containment of competitive influence spread in social networks, Knowl.-Based Syst. 109 (2016) 266–275.
[28] P. Lagrée, O. Cappé, B. Cautis, S. Maniu, Effective large-scale online influence maximization, in: 2017 IEEE International Conference on Data Mining (ICDM), 2017, pp. 937–942.
[29] M. Granovetter, Threshold models of collective behavior, Am. J. Sociol. 83 (6) (1978) 1420–1443.
[30] J. Goldenberg, B. Libai, E. Muller, Talk of the network: A complex systems look at the underlying process of word-of-mouth, Mark. Lett. 12 (3) (2001) 211–223.
[31] D. Kempe, J. Kleinberg, E. Tardos, Maximizing the spread of influence through a social network, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03, ACM, New York, NY, USA, 2003, pp. 137–146.
[32] A. Guille, H. Hacid, A predictive model for the temporal dynamics of information diffusion in online social networks, in: Proceedings of the 21st International Conference on World Wide Web, WWW '12 Companion, ACM, New York, NY, USA, 2012, pp. 1145–1152.
[33] K. Saito, K. Ohara, Y. Yamagishi, M. Kimura, H. Motoda, in: M. Kryszkiewicz, H. Rybinski, A. Skowron, Z.W. Raś (Eds.), Foundations of Intelligent Systems: 19th International Symposium, ISMIS 2011, Warsaw, Poland, June 28-30, 2011. Proceedings, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 153–162.
[34] C. Long, R.C.-W. Wong, Viral marketing for dedicated customers, Inf. Syst. 46 (2014) 1–23.
[35] F. Riquelme, P. Gonzalez-Cantergiani, X. Molinero, M. Serna, Centrality measure in social networks based on linear threshold model, Knowl.-Based Syst. 140 (2018) 92–102.
[36] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, Technical Report 1999–66, Stanford InfoLab, 1999.
[37] H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a social network or a news media?, in: WWW '10: Proceedings of the 19th international conference on World wide web, ACM, New York, NY, USA, 2010, pp. 591–600.
[38] J. Weng, E.P. Lim, J. Jiang, Q. He, TwitterRank: Finding topic-sensitive influential Twitterers, in: Proceedings of the Third ACM International Conference on Web Search and Data Mining, in: WSDM '10, ACM, New York, NY, USA, 2010, pp. 261–270.
[39] K. Kimura, S. Tsugawa, Estimating influence of social media users from sampled social networks, in: Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on, IEEE, 2016, pp. 1302–1308.
[40] S. Tsugawa, K. Kimura, Identifying influencers from sampled social networks, Physica A 507 (2018) 294–303.
[41] M. Cha, H. Haddadi, F. Benevenuto, P.K. Gummadi, Measuring user influence in Twitter: The million follower fallacy, ICWSM 10 (10–17) (2010) 30.
[42] A. Pal, S. Counts, Identifying topical authorities in microblogs, in: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM '11, ACM, New York, NY, USA, 2011, pp. 45–54.
[43] P. Cui, F. Wang, S. Liu, M. Ou, S. Yang, L. Sun, Who should share what?: Item-level social influence prediction for users and posts ranking, in: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '11, ACM, New York, NY, USA, 2011, pp. 185–194.
[44] Z. Zhao, Z. Cheng, L. Hong, E.H. Chi, Improving user topic interest profiles by behavior factorization, in: Proceedings of the 24th International Conference on World Wide Web, WWW '15, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2015, pp. 1406–1416.
[45] T.H. Haveliwala, Topic-sensitive pagerank, in: Proceedings of the 11th International Conference on World Wide Web, WWW '02, ACM, New York, NY, USA, 2002, pp. 517–526.
[46] A.K. McCallum, MALLET: A machine learning for language toolkit, 2002, http://mallet.cs.umass.edu.
[47] M. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, Mach. Learn. Res. 3 (2003) 993–1022.
[48] L. Hong, B.D. Davison, Empirical study of topic modeling in Twitter, in: Proceedings of the First Workshop on Social Media Analytics, SOMA '10, ACM, New York, NY, USA, 2010, pp. 80–88.
[49] W.X. Zhao, J. Jiang, J. Weng, J. He, E.P. Lim, H. Yan, X. Li, Comparing twitter and traditional media using topic models, in: Proceedings of the 33rd European Conference on Advances in Information Retrieval, ECIR'11, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 338–349.

[50] R. Mehrotra, S. Sanner, W. Buntine, L. Xie, Improving lda topic models for microblogs via tweet pooling and automatic labeling, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13, ACM, New York, NY, USA, 2013, pp. 889–892.
[51] Z.Z. Alp, S.G. Ögüdücü, Extracting topical information of tweets using hashtags, in: 14th IEEE International Conference on Machine Learning and Applications, ICMLA 2015, Miami, FL, USA, December 9-11, 2015, pp. 644–648.
[52] Y. Zhou, D. Wilkinson, R. Schreiber, R. Pan, Large-scale parallel collaborative filtering for the netflix prize, in: Proceedings of the 4th International Conference on Algorithmic Aspects in Information and Management, AAIM '08, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 337–348.
[53] M. Zaharia, M. Chowdhury, M.J. Franklin, S. Shenker, I. Stoica, Spark: Cluster computing with working sets, HotCloud 10 (10–10) (2010) 95.
[54] H.F. Yu, C.J. Hsieh, S. Si, I.S. Dhillon, Parallel matrix factorization for recommender systems, Knowl. Inf. Syst. 41 (3) (2014) 793–819.
[55] I. Anger, C. Kittl, Measuring influence on twitter, in: Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, i-KNOW '11, ACM, New York, NY, USA, 2011, pp. 31:1–31:4.
[56] L. Gallos, S. Havlin, M. Kitsak, F. Liljeros, H. Makse, L. Muchnik, H. Stanley, Identification of influential spreaders in complex networks, Nat. Phys. 6 (11) (2010) 888–893.
[57] D.M. Romero, W. Galuba, S. Asur, B.A. Huberman, Influence and passivity in social media, in: Proceedings of the 20th International Conference Companion on World Wide Web, WWW '11, ACM, New York, NY, USA, 2011, pp. 113–114.