Machine Learning (2019) 108:1831–1850
https://doi.org/10.1007/s10994-018-05776-w

Collaborative topic regression for predicting topic-based social influence

Asso Hamzehei[1,3] · **Raymond K. Wong**[1] · **Danai Koutra**[2] · **Fang Chen**[1,3]

Received: 20 April 2018 / Accepted: 24 December 2018 / Published online: 24 January 2019

Abstract

The rapid growth of social networks and their strong presence in our lives have attracted many researchers in social networks analysis. Users of social networks spread their opinions, get involved in discussions, and consequently, influence each other. However, the level of influence of different users is not the same. It varies not only among users, but also for one user across different topics. The structure of social networks and user-generated content can reveal immense information about users and their *topic-based* influence. Although many studies have considered measuring *global* user influence, measuring and estimating topic-based user influence has been under-explored. In this paper, we propose a collaborative topic-based social influence model that incorporates both network structure and user-generated content for topic-based influence measurement and prediction. We predict topic-based user influence on unobserved topics, based on observed topic-based user influence through their generated contents and activities in social networks. We perform experimental analysis on Twitter data, and show that our model outperforms state-of-the-art approaches on recall, accuracy, precision, and F-score for predicting topic-based user influence.

Keywords Social influence · Influence measurement · Topic-based influence prediction · Collaborative topic regression

Editors: Masashi Sugiyama and Yung-Kyun Noh.

✉ Asso Hamzehei
assoh@cse.unsw.edu.au

Raymond K. Wong
wong@cse.unsw.edu.au

Danai Koutra
dkoutra@umich.edu

Fang Chen
fang.chen@data61.csiro.au

[1] Computer Science and Engineering, University of New South Wales, Sydney, Australia

[2] Computer Science and Engineering, University of Michigan, Ann Arbor, MI, USA

[3] Data61, CSIRO, Sydney, Australia

1 Introduction

Social networks are ubiquitous nowadays and consist a prominent way of communication: people share their opinions and discuss a variety of topics that they are interested in with a wide range of online users. Users of social networks share information, ideas, and opinions that may trigger other users and, consequently, influence them. However, the level of influence of different users varies widely. For one user, her level of influence can significantly be different across topics. This rich property of social networks has attracted many researchers and businesses to focus on studying and identifying influential users (Herzig et al. 2014).

In terms of influence measurement and prediction, there are topic-based and global user influence prediction. One of the key differences of *topic-based influence* studies to the global user influence studies is that it takes the user-generated content into account in addition to network structure. On the other hand, global user influence is based on the network structure and dismiss the topics in the measurement. Also, there are influence maximization and propagation, which aim to find a group of users that have the highest power of information propagation. However, influence maximization is *not* the focus of this paper. To the best of our knowledge, there is no available study to predict topic-based user influence for a *new* topic, for which there is no observed data.

Topic modeling is a main phase in topic-based influence prediction. LDA as a standard topic model sees documents as a mixture of topics. and each topic is represented by set of words made by a distribution of words over the bag of words. These topics are drawn from the patterns of words co-occurrence within the document. Number of issues will raise if LDA is applied in short text corpus due to the low frequency of words and sparsity of co-occurrence of words. It is difficult for LDA to infer the document/word correlations in short text corpus and sense of words become more ambiguous (Cheng et al. 2014). As a result, specialized topic models are required to infer topics from short text corpus given those limitations.

We repeatedly use *observed topics* and *unobserved topics* terms in this paper. Whether a topic is observed or unobserved is defined as follows. The observed topics are the discovered topics in the historical data that our model is trained on and unobserved topics are those topics that our model is predicting their user influence. For example, if we train our model with user influence on US 2012 election topic, that topic is observed topic, while if we want to predict user influence on an upcoming topic like US 2016 election, it is unobserved topic.

Topic-based user influence prediction is an important challenge and the main focus of this paper. This task is significantly important for different applications such as marketing and election campaigns. In this work, we predict topic-based user influence on unobserved topics based on observed influence of users on the topics extracted from history of their posts on social network. We propose a Social Influence Collaborative Topic Regression (SICTR) approach to learn user, topic, and social factor latent spaces for estimating user social influence on an unobserved topic in social network. To infer the topics of social networks, SICTR, unlike CTR, is using a generative biterm topic model where words co-occurrence are explicitly modelled throughout the corpus.

Our approach represents users with their topic interests and their social influence on each observed topic. In more detail, our contributions are:

- We propose a novel topic-based influence prediction approach to integrate the user-topic relationships, topic content information, and social connections between users into the same principled model. Our model is inspired by the collaborative topic regression (CTR) model (Wang and Blei 2011), which has been successfully applied to article recommendation, and combines both user-topic and topic content information for influence prediction. SICTR improves the shortcoming of CTR in social network and influence prediction.

- Instead of considering user-to-user influence and global user influence, the proposed model considers individual users' influence and interests in a topic, which gives the capability of predicting one's influence in a new topic.
- The usefulness of considering content of topics and co-occurrence in the user-topic matrix are confirmed by our experiments on topic-based user influence prediction: topic-aware methods show better performance over other approaches that just consider a generic item and neglect its characteristics (content-based, e.g., LDA; and item-based, e.g., collaborative filtering methods).
- Finally, We have prepared and used a unique dataset from real-world social network to evaluate our proposed approach.

The remainder of this paper is organized as follows. We first discuss existing approaches for topic-based influence analysis in Sect. 2. Next, we define the research problem that we focus on, and then propose our approach and algorithms in Sect. 4. We describe our dataset and discuss the results in Sect. 5. Finally, we conclude the paper in Sect. 6.

2 Related work

One of the main approaches to study user influence in social networks has been through network structure as well as user's position and connectivity in the network. The traditional centrality measures such as closeness and betweenness are measured for users, to discover how well connected a user is to the rest of users in the network and whether a user is acting as a hub (Romero et al. 2011). The major adopted algorithms for network structure based influence measurement include PageRank (Haveliwala 2002) and HITS (Kleinberg 1999). Numerous works have applied PageRank algorithm variations on social network graph to rank user influence according to the network structure. An example of PageRank algorithm variations is the work by Kwak et al. (2010), in which they ranked users by applying PageRank on follower/following graph in Twitter (along with number of followers and number of retweets). The network structure is relatively static compared to the activities of users in social networks. Some studies have included the social network related meta data (in case of Twitter, the meta data are retweets, mentions, and likes) (Hajian and White 2011).

2.1 Topic-based Influence

Following the influence studies (overall user influence) on social networks, less studies have shed light on topic-based influence. More recently, topic-based influence studies have combined content of user posts with link-based metrics. Haveliwala (2002) proposed a topic-sensitive extension of PageRank to rank query results in regards to the query topics. The idea of topic-sensitive PageRank was later used and adjusted for social networks such as Twitter for ranking topic-based user influence. Topical authorities were also studied in Pal and Counts (2011). They proposed a Gaussian-based ranking to rank users efficiently. They used probabilistic clustering to filter feature space outliers and showed that mentions and topical signals are more important features in ranking authorities. Kong and Feng (2011) intended to identify and rank users that are posting quality tweets. They defined a topic-based high quality tweet with the author's topic-specific influence, topic related author's behavior. They applied their proposed metric on graph of following and retweets. Xiao et al. (2014) aimed at detecting topic related influential users by looking at hashtag user communities where hashtags are pre-identified from news keywords. They proposed RetweetRank and MentionRank as content-based and authority-based influential users. Similarly, Hu et al. (2013) worked on detecting topical authorities with the assumption that retweeting propagates topical authority. Sung et al. (2013) proposed another extension of PageRank, and unlike Weng et al. (2010), it does not need predefined topics for topic-based user influence. Cano et al. (2014) introduced a PageRank-based user influence rank algorithm that the user links have weights based on their topics of interest similarities. The topic-based influence framework of Liu et al. (2014), considers retweet frequency and link strength. The link strength is estimated by poisson regression-based latent variable model on user's frequency of retweeting each other. Welch et al. (2011) found out that topical relevance is better detectable through the retweet link rather than following links. They used two variations of PageRank algorithm to on retweet and following graphs for that purpose. Montangero and Furini (2015) also measured Twitter topic-based user influence where they identify topics by hashtags. Although hashtags can reveal the tweet's topic correctly, over 80% of tweets do not have hashtags. These results are neglecting the majority of tweets and can mislead a topic-based user influence, as 4 out of 5 of her tweets are not considered for measuring her influence. Cataldi and Aufaure (2015) and Bingol et al. (2016) estimated user influence for topics based on PageRank. For that purpose they build a topic information exchange graph to take the information diffusion and degree of information shared into account for user influence estimation. They manually considered seven topic categories and later assign each tweet to those categories through an n-gram model. However, their approach is unable to identify topics in the lower level of the main categories. For example, if someone is detected as influential in the sports category we do not know which sport the influence belongs to. Weng et al. (2010) offered Twitter-Rank, a PageRank extension, that measures user influence by calculating topical similarities of users and their network connections. For topic identification, they used the unsupervised text categorization technique, LDA, by aggregating all tweets of a user into a document. Although this approach is presented as topic-sensitive, this approach cannot discriminate the user influence for the topics. In a recent work by Katsimpras et al. (2015), they proposed a supervised random walk algorithm for topic sensitive user ranking. As it is obvious from the algorithm name, it needs labeled data which is not very practical in many cases specially with the volume of social networks. It is worth mentioning that similar works exist that are only consider the identification of global influencers instead of influencers for specific topics (Barbieri et al. 2012). An example of such works is where they extended the Linear Threshold Model and Independent Cascade Model to be topic-aware, the topics are still obtained based on the network structure, while totally ignoring the valuable content information.

2.2 Collaborative topic regression

Users and items information is considered valuable to improve recommendation performance. Wang and Blei (2011) proposed Collaborative Topic Regression (CTR) that utilizes user and item information into topic modeling based Collaborative Filtering (CF) models to further improve recommendation performance. CTR is further extended in some studies. For example, Purushotham et al. (2012) and Chen et al. (2014) proposed two models (i.e., CTR-SMF and CTR-SMF2) to incorporate user social network into CTR to further improve item recommendation performance. Wang et al. (2013) proposed a model to incorporate item social relationship into CTR to further improve tag recommendation performance in social tagging systems. More recently, HU et al. (2018) incorporated three source of information, rating, item review, and social connections, to predict rating via latent factors and hidden topics. To address the sparsity problem, Zhao et al. (2018) developed a heterogeneous social-aware movie recommendation system where they included movie poster image as an input as well as textual data, ratings, and social connections. Furthermore, Chen et al. (2017) have introduced a neural network-based recommender system and tried to efficiency issues as neural network-based algorithms are computationally expensive.

In contrast to other works, in this paper we take one step further in topic-based influence measurement, by proposing an approach that measures topic-based user influence and predicts user's influence on an unobserved topic.

3 Background

Next, we give preliminaries for Probabilistic Topic Modeling and Collaborative Topic Regression.

3.1 Biterm topic model

Probabilistic topic modeling , such as Latent Dirichlet Allocation (LDA), represents a low dimensional space of corpus by detecting a set of latent topics. The basic idea of Probabilistic Topic Modeling is having a t hidden variable for each word's co-occurrence in the collection of documents. t can range among k topics where each topic is a distribution over a fixed vocabulary. Given a corpus, a document may contain multiple topics and the words are assumed to be generated by those topics.

Given a set of documents denoted by $N_D = \{d_q\}_{q=1}^{D}$, topic modeling generates a set of $\mathcal{T}_K$ topics denoted by $\mathcal{T}_K = \{t_k\}_{k=1}^{K}$. The topics of documents can be represented by a multinomial distribution $\theta = \{\theta_k\}_{k=1}^{K}$ with $\theta_k = P(k)$ and $\sum_{k=1}^{K} \theta_k = 1$ and if consider t as topic variable and $\theta_k = P(t = k)$. Each topic is related to a weighted representation over m words denoted by $t_k = \{w_m\}_{i=m}^{W}$, where w_m is the weight representing the probability that word m belong to topic t_k (e.g., $P(w|t)$). The word distribution over all topics, $P(w|t)$, can be represented by matrix ϕ where row ϕ_k corresponds to topic k and the matrix entries are $\phi_{k,w} = P(w|t = k)$ and $\sum_{w=1}^{W} \phi_{k,w} = 1$. In BTM, documents are seen as a collection of biterms N_B where $B = \{b_m\}_{m=1}^{N_B}$ and $b_m = (w_{m,1}, w_{m,2})$ (Cheng et al. 2014; Li et al. 2016).

BTM can be generated over a process as follows:

1. Obtain θ, a distribution over topics to generate a document. This distribution is drawn from a Dirichlet distribution with a corpus-specific hyperparameter α.
2. For each topic to be generated
   (a) Draw ϕ_k from a Dirichlet distribution with hyperparameter β.
3. Then for each biterm to be generated;
   (a) Assign topics by drawing upon the distribution over topics
   (b) Finally, generate $w_{m,1}, w_{m,2}$ from distribution of topics over biterms in dictionary, which means biterms of each document come from a mixture of topics.

We aim to use probabilistic topic modeling to extract topics from the social networks and also as a part of SICTR process to create topic latent spaces.

3.2 Matrix factorization (MF) and collaborative topic regression (CTR)

CF analyzes the relationships between users and their associations with items by relying on historical user behavior (e.g., movies rating), without requirement of explicit user profiles. A basic approach of CF is neighborhood-based methods which analyze the relationship between items or users. For the item-based approach, the rating of a user on an item j is estimated based on her ratings on similar items, while user-based approach estimates item j's rating by looking at the rating behavior of other users with similar interests. Another CF approach is known as latent factor model, (e.g., matrix factorization Koren et al. 2009), which estimates a rating by utilizing both user and item patterns. It factorizes user-item matrix into a user-specific matrix and a item-specific matrix. The objective function of a matrix factorization model can be formulated as follows,

$$\pounds = \sum_{i,j} \left(U_i^T V_j - r_{i,j} \right)^2 + \lambda_U \, \|U\|_F + \lambda_V \, \|V\|_F \tag{1}$$

in which the first term is the difference between observed and prediction and the rest are regularization terms.

CTR (Wang and Blei 2011) is proposed on top of the matrix factorization and utilizes probabilistic topic modeling. It assumes that items (documents such as news and movie reviews) are generated by a topic model which is represented as topic latent vector v_j. In CTR, users are represented by topic interests. Similar to matrix factorization, CTR computes latent parameter of users u_i and items v_j. Latent variable ϵ_j captures the differences between topics for a user based on the users ratings on items. Equation 2 draw the ϵ_j as:

$$\epsilon_j \sim N(0, \lambda_v^{-1} I_k) \tag{2}$$

where N is probability density function of the Gaussian distribution with mean zero and variance equal to λ_v^{-1} and I is the indicator function that is equal to 1 if user i rated item j and equal to 0 otherwise.

CTR assumes that item latent vector v_j is close to topic proportion θ_j so that $v_j = (\theta_j + \epsilon_j)$ and draw it as:

$$v_j \sim N(\theta_j, \lambda_v^{-1} I_k) \tag{3}$$

The generative process of our SICTR model is inspired by CTR (Wang and Blei 2011).

4 Topic-based social influence prediction

4.1 Problem definition

Assume $G(\mathbb{V}, E)$ denotes a social network graph, where users are the vertex set $\mathbb{V}$ and users relationships are the edges set of E. Assume that users publish a set of texts $D = [d_1, d_2, \ldots, d_q]$, and talk about different topics $\mathcal{T} = [t_1, t_2, \ldots, t_j]$. Each user text (post) d_q holds one or more topics and receives engagement from other users by replying, liking, or re-publishing it. The engagement of other users in a post can reveal the influence of that particular post among its audience.

Quantifying the topic-based influence of each user based on social ties and other users' engagement in social network, we can identify the influence of user u_i on topic t_j, represented as F_{ij}. Then we have matrix $F = [F_{ij}]_{i \times j}$ that represents influence of all the users in all identified topics.

Given a list of users, topics, and social influence of each user on those topics, we are interested in predicting an unknown value in $F_{ij} = [F_{1j}, F_{2j}, \ldots, F_{ij}]^{\mathcal{T}}$; the social influence of user u_i on a new topic t_j where $\mathcal{T}$ is the set of topics. Specifically, we aim to estimate user influence on an unobserved topic, based on observed influence of users on the topics from their history of activities and generated contents.

We expect that an influential user in topic t can be influential on a similar topic t'. Assuming that there are patterns among users with similar topic-based influenc weights, the prediction can be performed with CF algorithms. However, the content-based methods use only the content information for recommendation. For example, if we want to predict influence for topic t_j, we can use the influence from the nearest neighbor in $\mathcal{T}$, the set of topics, based on the topics content similarity. We can also treat each topic as a label and use multi-label methods to train classifiers based on content information. Co-occurrence based methods use only the user-topic matrix F for prediction. For instance, if two topics t and t' co-occur for many users, and t is associated with u_i, we expect similar influence of t' to u_i. Both content-based methods and co-occurrence based approaches neglect useful information. As a result, they cannot achieve satisfactory performance in social influence prediction.

4.2 Our approach

To measure and predict social influence on unobserved topics in a social network, we propose SICTR (stands for Social Influence-based Collaborative Topic Regression), which predicts topic-based influence. In a nutshell, our model performs a two-part representation of the users: (i) latent feature representation of the users according to the social network, and their connections to other users, and (ii) latent feature representation of the users based on the topics they are active in. Our method adopts Collaborative Topic Regression, CTR, a well-known method that combines CF with topic modeling, to learn a model that uses the latent topic space to explain both the observed ratings and the observed words. We propose SICTR to predict topic-based user influence in social networks. SICTR computes the latent parameter of users U and topics $\mathcal{T}$.

4.2.1 Social network-based representation

We want to derive a k-dimensional feature U from the social network g to represent users. Let $U \in R^k$ be latent user metric with column vector U_i for user-specific latent feature vector. We have user and factor feature vectors after placing zero-mean spherical Gaussian prior on them as follows:

$$p(U \mid \sigma_U^2) = \prod_{i=1}^{m} N(U_i \mid 0, \sigma_U^2 I) \tag{4}$$

where $N(x \mid \mu, \sigma^2)$ is the probability density function of the Gaussian distribution with mean μ and variance σ^2, and I_k is a k-dimensional identity matrix (Table 1).

4.2.2 Topic-based representation

In our model, SICTR, items are the extracted topics from the corpus of user generated content in a social network, and users are represented by their topics of interests. SICTR predicts user influence on a topic according to similarity of items and other users' influence

Table 1 Key notations

Symbol	Description
$\mathcal{T}$	The set of topics extracted from Social Networks
$U, \|U\|_n$	The set and the number of users, respectively
V	The set of topics generated by SICTR
u	User latent vector
v	Topic latent vector
t	A topic
k	Number of latent dimensions
F_{it}	Influence of user i in t
$F_f(i,t)$	Follower strength influence measure for user i in t
$F_a(i,t)$	Activity influence measure for user i in t
$F_e(i,t)$	Engagement influence measure for user i in t
$F_c(i,t)$	Centrality influence of measure for user i in t
α	Offset term
β	Topic bias parameter
λ_u	Regularization parameter for u
λ_v	Regularization parameter for v
θ_t	k-dimensional topic distribution for t
ϕ_k	Word distribution for topic k
ω_{tn}	nth word of document in topic t
z_{tn}	The topic for the nth word in topic t
N_d	Number of words in document d
ϵ_t	Topic t's latent offset
c_{it}	Precision parameter for F_{it}
I_k	k-dimensional identity matrix

on similar topics. We identify the topics by applying probabilistic topic modeling on all the user-generated text. Each topic contains a set of social networks posts with all their related information and metadata, such as; content, replies, and republishing. For each tuple of $(user_i, topic_j)$, we measure the influence of user u_i on topic t_j in Sect. 4.5. That way, SICTR generates a topic latent space and a user latent space.

An important part of SICTR is generating topic latent vector $v_j = (\theta_j + \epsilon_j)$, where ϵ_j captures users interest in topic t_j and it assumes topic latent vector v_j is close to topic proportion θ_j. The expectation of F_{ij} is a linear function of θ_j, $E[F_{ij}|u_i, \theta_j, \epsilon_j] = u_i^T(\theta_j + \epsilon_j)$.

Moreover, from SICTR we know that item latent vector v_j is close to topic proportion θ_j and it generates item latent vector as $v_j = (\theta_j + \epsilon_j)$ where $\epsilon_j \sim N(0, \lambda_v^{-1} I_k)$ is equivalent to $v_j \sim N(\theta_j, \lambda_v^{-1} I_k)$ then

$$P(V \mid \sigma_v^2) \sim N(\theta_j, \lambda_v^{-1} I_k) \tag{5}$$

where V is the latent space for the set of topics $\mathcal{T}$ and $\lambda_v = \sigma_F^2/\sigma_V^2$.

Taking BTM into account, through Bayesian inference:

$$\begin{aligned} p(U, V \mid F, \sigma_F^2, \sigma_U^2, \sigma_V^2) \\ \propto p(F \mid U, V, \sigma_F^2) \\ \times p(U \mid \sigma_U^2) p(V \mid \sigma_V^2) \end{aligned} \tag{6}$$

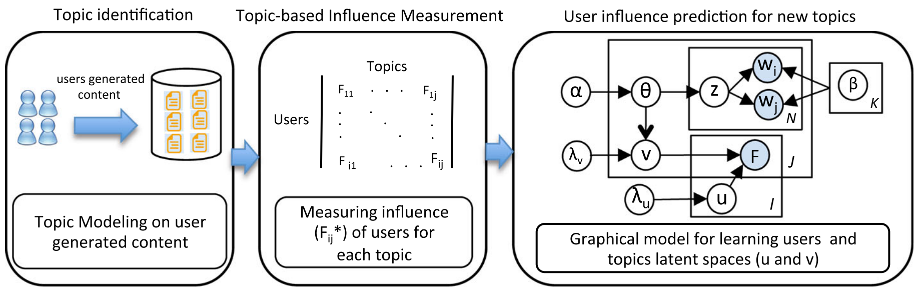

Topic identification
users generated content
Topic Modeling on user generated content
Topic-based Influence Measurement
Topics
Users
F11 . . . F1j
Fi1 . . . Fij
Measuring influence (Fij*) of users for each topic
User influence prediction for new topics
α
θ
z
wi
wj
N
β
K
λv
v
F
J
λu
u
I
Graphical model for learning users and topics latent spaces (u and v)

Fig. 1 The SICTR framework

Figure 1 shows the SICTR framework. It contains three main steps: collecting data and identifying topics, measuring user influence for each topic, and predicting user influence for a new topic. It shows the graphical model of SICTR for learning users and topics latent spaces. Also, Algorithm 1 describes the generative process of SICTR.

4.3 SICTR comparison with CTR

We compare SICTR with CTR to better understand how our model utilises topic modeling in collaborative topic regression and how it tackles short text corpus problem. Given a matrix of rating, CTR computes latent representation of items and users. CTR is using LDA to extract topics for explaining the users interest in items. It includes the latent topic space of the items into a model where observed ratings and learnt topics are combined in the recommendation systems. The topic modeling that is used in CTR, for each word of a document, it draws a topic k from topic distribution θ_d and then draw a word w from topic k. This means that the topic k of word w is dependent to the other words of the same document. This is not ideal for short text documents and social networks contents and affect poorly on estimation of k, θ_d, and topic-word distribution. This phenomenon, directly influence the CTR results when the items of the user/item matrix contain a bunch of short text documents to draw their topic latent spaces. However, to alleviate this shortcoming, SICTR uses a Biterm Topic Modeling (BTM) approach to learn the topic latent spaces of the items of the matrix. BTM unlike LDA, generate topic from biterms in the corpus rather than words in documents. In this way the data sparsity issue is alleviated as θ is learnt from a global topic distribution.

4.4 Learning the parameters of SICTR

We use an EM-style algorithm to learn the parameters (Purushotham et al. 2012). Maximization of the posterior is equivalent to maximizing the complete log-likelihood of U, V, $\theta_{1:t}$, F given U, V, and β.

$$L = \frac{-\lambda_u}{2}\sum_i u_i^T u_i - \frac{-\lambda_v}{2}\sum_t (v_t - \theta_t)^T (v_t - \theta_t) + \sum_t \sum_n \log\left(\sum_k \theta_{tk}\beta_{k,w_{tn}}\right) - \sum_{it} \frac{c_{it}}{2}\left(F_{it} - u_i^T v_t\right)^2 \tag{7}$$

where $\lambda_u = \sigma_F^2/\sigma_U^2, \lambda_v = \sigma_F^2/\sigma_V^2$ and Dirichlet prior (α) is set to 1. We optimize this function using gradient ascent by iteratively optimizing the CF on social network variables u_i, v_t and topic proportions θ_t. For u_i, v_t, maximization follows similar to matrix factorization. Given a current estimate of θ_t, taking the gradient of L with respect to u_i, v_t and setting it to zero helps to find u_i, v_t in terms of U, V, C,F, λ_v, λ_u. Solving the corresponding equations will lead to the following update equations:

$$u_i \leftarrow (VC_iV^T + \lambda_u I_k)^{-1}(VC_iF_i) \tag{8}$$

$$v_t \leftarrow (UC_tU^T + \lambda_v I_k)^{-1}(UC_tF_t + \lambda_v\theta_t) \tag{9}$$

where C_i is diagonal matrix with c_{it}; $t = 1 \cdots J$ as its diagonal elements and $F_i = (F_{it})_{t=1}$ for user i. For each topic t, C_t and F_t are similarly defined. Note that c_{it} is precision parameter for influence matrix F_{it}. Equation 9 shows how topic proportions θ_t affect the topic latent vector v_t, where λ_v balances this effect. Given U and V, we can learn the topic proportions θ_t. We define $q(z_{tn} = k) = \phi_{tnk}$ and then we separate the topics that contain θt and apply Jensen's inequality:

$$\begin{aligned} L(\theta_t) &\geq -\frac{\lambda_v}{2}(v_t - \theta)^T(v_t - \theta_t) \\ &\quad + \sum_n \sum_k \phi_{tnk}(\log \theta_{tk}\beta_{k,w_{tn}} - \log \phi_{tnk}) \\ &= L(\theta_t, \phi_t) \end{aligned} \tag{10}$$

The optimal ϕ_{ink} satisfies $\phi_{tnk} \propto \theta_{tk}\beta_{k,wtn}$. Note that we cannot optimize θ_t analytically, so we use projection gradient approaches to optimize $\theta_{1:t}$ and other parameters U, V, and $\phi_{1:t}$. After we estimate U, V, and ϕ, we can optimize for β,

$$\beta_{kw} \propto \sum_t \sum_n \phi_{tnk} 1\,[w_{tn} = w] \tag{11}$$

4.5 Influence measurement

We define social influence in a social network as importance of a user in the social network graph, user's activities, and involvement of others in the user's posts. Social influence can be analyzed through different modalities network structure and user's position in the network, scale of a user post's diffusion in the network, a user's activities and engagement in the social network, and message content that a user broadcast in the network (Embar et al. 2015).

From the network structure, we identify influence related attributes, such as user's friends and centrality of user in the social network. From the content of broadcasted text, we can identify topics, thus, the influence of user on broadcasted aspects. For instance, in Twitter, a tweet can contain user mentions, receive replies, and get retweeted by other users. All this information can reveal social influence of a user.

As a reminder of the Influence measures, let denote $D = \{D_t\}_{t=1}^{\mathcal{T}}$ as the set of collected texts, where D_t is texts related to topic t where there are $\mathcal{T}$ topics. Each text d_i contains a set of attributes as $(u_i, c_i, R_i, M_i, f_i)$ where u_i is the author of the text, c_i is the text, R_i is the list of users republished the text, M_i is the list of mentions for that text, and f_i is the number of followers of the text author.

We define four influence measures namely; follower scale F_f (the number of friends a user has in the network); Topic Activity F_a (topic-related activities of a user); Topic-based Attractiveness F_e (how other users are attracted to useru_i's post); and Network centrality F_c (centrality of a user in the graph of active users in topic t_j). The four influence measures described above F_f, F_a, F_e, F_c will be aggregated to form a single influence score F^* for user u_i in topic t_j.

For more details on the Influence measure please refer to Hamzehei et al. (2016) and Hamzehei et al. (2017).

Algorithm 1 SICTR generative process

```
Measure users influence for each identified topic from the corpus
For each user $u_i$ from the user's collection of posts of topic $t_j$, draw user latent vector of her topics of interest $u_i \sim N(0, \lambda_u^{-1} I_k)$
**for** topic (each topic extracted from the corpus and represented as a bag of words) in $j$ **do**
    Draw topic proportions $\theta_j \sim Dirichlet(\alpha)$
    Draw topic latent offset $\epsilon_j \sim N(0, \lambda_v^{-1} I_k)$ , set the topic latent vector as $v_j = \epsilon_j + \theta_j(d)$
    **for** word in $w_{jn}$ **do**
        Draw topic assignment $t_{jn} \sim Multinomial(\theta)$
        Draw word $w_{jn} \sim Multinomial(\beta_{t_{jn}})$
For each user-topic pair $(u_i, t_j)$, draw, $F_{ij}$, the rate of social influence of user $u_i$ for topic $t_j$
```

4.6 User influence prediction on a new topic

Having the parameters learned, our SICTR model can be used for in-matrix and out-matrix prediction. In-matrix prediction refers to predicting a user influence on a topic where influence rates of some other users are available. Out-matrix prediction refers to predicting user influence on a topic that no influence data is available (totally new topic).

For a given observed set of documents, D, prediction of influence of a user on a topic can be predicted as the expected value of:

$$E\left[F_{ij}|D\right] \approx E\left[u_i|D\right]^T \left(E\left[\theta_j|D\right] + E\left[\epsilon_j\right]\right) \tag{12}$$

In-matrix can be predicted by using point estimate of u_i, θ_j and ϵ_j to approximate their expectations (recall $v_j = \theta_j + \epsilon_j$),

$$F_{ij}^* \approx (u_i^*)^T(\theta_j^* + \epsilon_j^*) = (u_i^*)^T(v_j^*) \tag{13}$$

For out-matrix prediction, as a topic is new and no historical influence is available from users then $E\left[\epsilon_j\right] = 0$. As a result, we can predict a user's influence on a new unobserved topic as:

$$F_{ij}^* \approx (u_i^*)^T \theta_j^* \tag{14}$$

Putting everything together, we obtain our proposed SICTR method (Algorithm 1).

5 Results and experiments

In this section, we discuss the the details of conducted experiments. It includes the data, the influence measurement, and topic-based user influence prediction by SICTR.

5.1 Dataset

In order to conduct our experiments, we collected a dataset from Twitter and Google Scholar. The details of each dataset is is as following:

- *Twitter dataset* To validate our proposed method, we collected a unique dataset from Twitter using the Twitter Search API. We targeted Machine Learning community and identified 500 users that have mentioned machine learning or data science as a keyword in their profile description. To choose the users, we selected a set of well-known machine learning users as seeds and crawled among their friends and friends of friends for other machine learning-related users. For the prepared list of users, we gathered their timeline tweets which for majority of the users it covers their last 5 years tweets. For each tweet, we also, collected the related meta-data such as the list of users who have replied to each tweet (mention list) and the list of users who have retweeted each tweet (retweet list). The final dataset contains 101,363 tweets with their related metadata, mention lists, and retweet lists. The network that is built on retweet list contains 301,870 users. The dataset was collected on 4 May 2016.
- *Google Scholar dataset* If we consider number of citations of a research paper as a measure of influence of that paper and author in that research topic, then it is safe to claim that a user with a high number of citations in a topic is very influential in that topic. So, if the same author is an active Twitter user and tweets on the same topic, we expect that her influence will be visible in Twitter as well.

For validating our influence measurement results we prepared this Google Scholar dataset on the same users that are in the Twitter dataset described above. Specifically, we spotted 254 of our Twitter users in Google Scholar and extracted their top 20 papers with the highest number of citations. Each extracted paper contains author names, title, abstract, publication date, number of citations, Google Scholar link, and PDF link. The dataset was collected on 14 November 2016.

5.2 Evaluation

Our experiments contain two main evaluation tasks. We first measure topic-based user influence on the corpus. Second, we predict user's influence for unobserved topics.

5.2.1 User influence measurement on the identified topics from the tweet corpus

User influence is an abstract concept and evaluating user influence is very subjective, and thus challenging. That means, for the task of validation of the influence results, there is no ground truth. Moreover, there is no standard method in the literature to validate the algorithm output. Montangero and Furini (2015) compared their method to the number of followers of the users of interest. Cataldi and Aufaure (2015) used opinions from experts to evaluate their proposed influence measure. Lastly, Katsimpras et al. (2015) used PageRank scores for as their evaluation measure.

We evaluate the measured user influence results through expert opinion and a secondary dataset prepared for cross-validation. This dataset is prepared from Google Scholar. For the same Twitter users included in our study, we collected information about their research publications, as described in the previous subsection. We identify a user's influence for each topic based on her citations. For example, for the topic of 'Social Network Analysis', we measure user influence separately from Twitter and scientific publications (Google Scholar). Then, we cross validate the results from both datasets.

We have chosen the machine learning and data science research community on Twitter as our community of study for two main reasons. First, there is wide availability of experts in the domain, which allows us to verify the identified influential users through our algorithm. Second, as these users have research publications, we can measure their influence on the topics they have published by considering their number of citations per topic. This is based on the assumption that the measured user influence is correlated between the two datasets.

5.2.2 Prediction of user influence on unobserved topics through our proposed SICTR

We report recall, precision, accuracy and F-score. Due to the uncertainty of the meaning of zero influence, recall will be the main performance measure while we still report precision, accuracy, and F-score. Zero influence mean either the user has not been influential in a topic or her activities are not represented in our dataset. For experimental analysis, we split the dataset into 80% train and 20% test datasets. If we present each user by topics that are estimated as influential from SICTR, recall corresponds to the number of topics that the user u_i is predicted as influential over the total number of topics that the user u_i recognized as influential:

$$recall = \frac{\#\ of\ topics\ the\ user\ is\ predicted\ as\ influential}{\#\ of\ topics\ the\ user\ is\ influential\ in} \tag{15}$$

5.3 Topic-based influence measurement

Next, we proceed with identifying topics from the collection of all tweets and then measuring user influence for each topic. Number of topics is one of the inputs for topic modeling. As our aim in this paper is to predict topic-based user influence, we choose a number of topic for topic modeling that gives the best performance to SICTR.

In our proposed approach, we perform probabilistic topic modeling in two different rounds. The first round is for identifying the topics in the tweets dataset and the second round is to create topics latent space to take topics similarity into account for influence prediction.

The user tweets gathered from their timelines, belong to the identified topics with a probability. We set the probability threshold to 0.1 to consider whether a tweet belongs to a topic. Each tweet is mapped to at least one topic. Now that for each topic we have a collection of related tweets with their mention and retweet lists, we can measure user influence for them. In Sect. 4.5, we defined influence based on 4 measures; follower strength, activity, engagement, and network centrality. Follower strength will be taken from the number of users follow the user u_i on Twitter. Activity represents the number of tweets user u_i has in topic t_j. Engagement is the sum of number of mentions and retweets for all of user u_i's tweets in topic t_j. For measuring network centrality, we build the retweet graph for each topic separately from the corresponding retweet list and measure centrality of that user node through PageRank algorithm.

5.4 SICTR prediction results and comparisons

Here, we present the SICTR, in-matrix prediction results and compare them with WNTM-based approach, LDA-based approach, CF, a content-based model, and a random baseline. In WNTM-based approach (Zuo et al. 2016), CTR is using WNTM for generating topic latent vectors. In LDA-based approach, CTR is using LDA for topic vector generation. For the CF model, per-user and per-topic latent vectors are fixed with the influence values in the matrix. While, content-based model behaves as a content-only model. For content-based model, the per-user latent vector is generated by LDA and is fixed to influence entries in the matrix and per-topic latent vector is only based on the words of the topics. The random baseline is a random model, where a user randomly is predicted as influential for a topic. The random model predicts a user as influential using the probability of appearance of influential users in the user-topic influence matrix. As we mentioned before, in-matrix prediction considers influence prediction for the topics that already exist in the data and influence of at least one user is available for them.

We split the data into train and test datasets. For measuring performance, we use 5-fold-cross-validation. We make sure every topic appears in all the folds so that each topic appears both in the train and test data.

5.4.1 SICTR experimental settings

SICTR contains parameters that need to be tuned to evaluate the proposed approach with the benchmarks. In the models, the a and b are tuning parameters for the parameters c_{ij} and d_{ij} The parameter λ_v is the precision parameter to balance the diverging of topic latent vector from the topic proportion θ_j. Similar to CTR, we increase λ_v to penalty v_j diverging from θ_j. If we set $v_j = \theta_j$ for per-topic latent vector, SICTR behaves as a content-based method and the topic vector θ_j will be just based on the words in the topic j. To find the best value of the parameters, we use grid search. For comparisons of the three models, the parameters are set to $k = \frac{number\ of\ topics}{4}$, $\lambda_v = 0.01$, $\lambda_u = 0.01$, $a = 1$, $b = 0.01$. Moreover, we analyze the SICTR performance for different values of λ_v, number of topics, and influence threshold.

5.4.2 Effect of sparsity on influence prediction

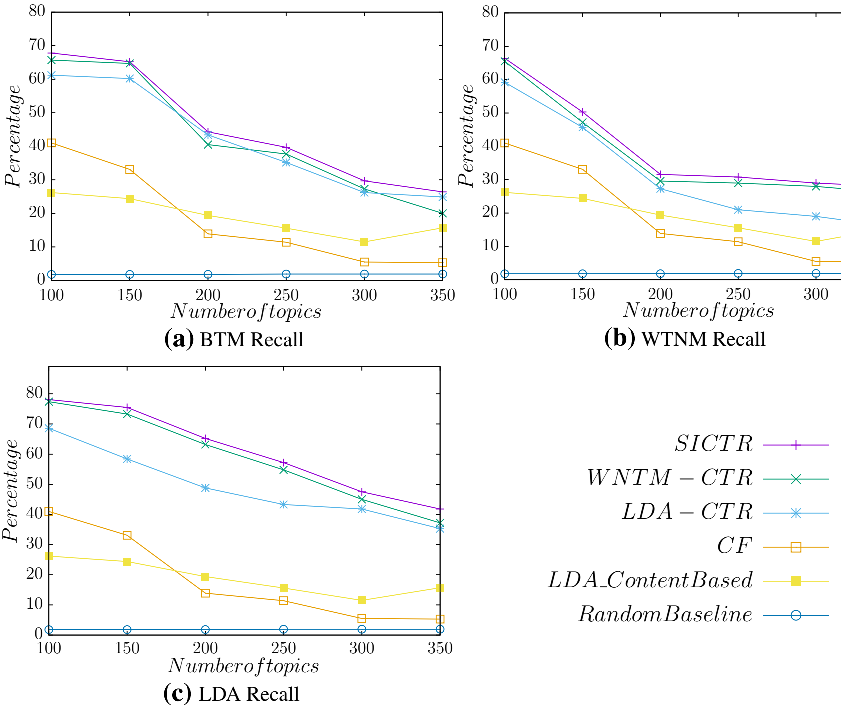
The matrix of topic-based user influence is a sparse matrix. The matrix sparsity can be increased with higher influence threshold. By increasing the influence threshold, fewer users will be identified as influential and as a result, the influence matrix becomes more sparse. Also, by increasing the number of topics, the topic-based influence matrix becomes more sparse. Figure 2 shows the effect of number of topics on recall. When the matrix becomes more sparse, recall decreases. Accordingly, it is shown in Figs. 3 and 4 that precision and accuracy for SICTR also decrease when the matrix becomes more sparse.

5.4.3 Effect of influence threshold on SICTR performance

The topic-based user influence is in the range of zero to one. Zero indicates no influence over that topic and one means the user has the highest influence for all four influence measures in that particular topic. Higher influence threshold results in less number of users being identified as influential, and, consequently, makes the influence matrix more sparse. In the experiments, when influence threshold is increased, our model predicts influential users more precisely. However, due to sparsity, the recall measure decreases.

5.4.4 Effect of Lambda

For this experiment, we fix the number of topics to 300, the influence threshold to 0.3, the number of factors to 50, and the remaining parameters as explained in experiment settings. We vary $\lambda_v \in [0.01, 0.1, 1, 10, 100, 1000]$ to evaluate the SICTR performance for different λ_v values. As λ_v is the precision parameter to balance the divergence of the topic latent vector from topic proportion θ_j and is specific for SICTR, its change does not affect CF and content-based model performance. The results show that SICTR outperforms the CF by 29% in the best case scenario with $\lambda_v = 100$ and beats the CF in all experiments. This suggests



(a) BTM Recall **(b)** WTNM Recall

(c) LDA Recall

Fig. 2 Recall comparisons between SICTR, WNTM-base model, LDA-base model, CF, Content-based model, and Random baseline. The x-axis shows experiments for different number of topics. The influence threshold is set to 0.3

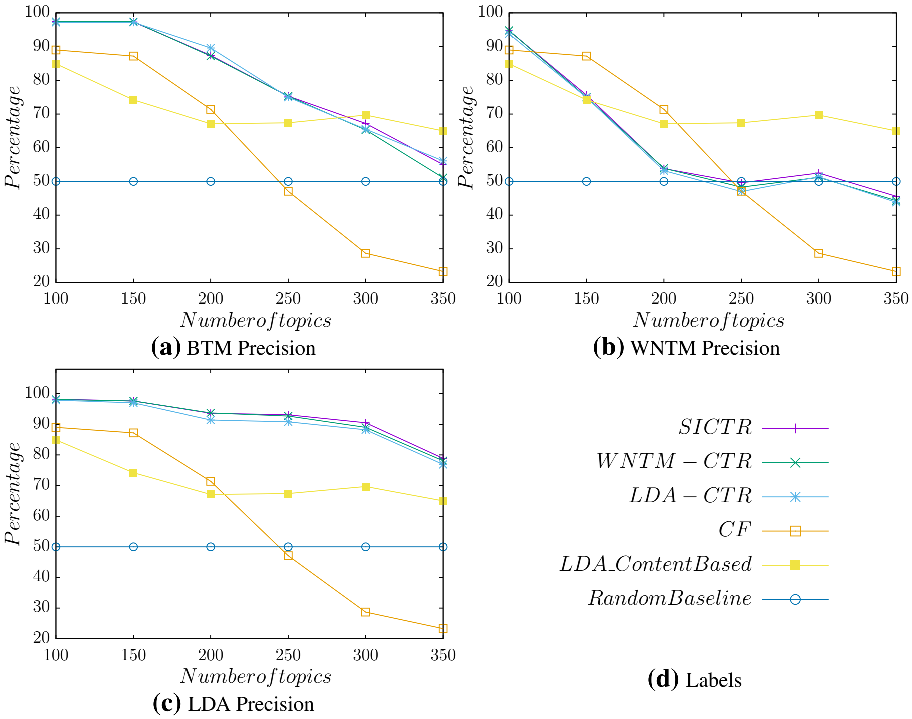
that taking topic features into account improves the influence prediction performance. Also, it indicates that a user's influence on a new unobserved topic can be estimated by similarity of the new topic with the previously observed topic-based user influences.

5.4.5 Effect of number of topics on SICTR, CF, and LDA

We evaluate SICTR overall performance against WNTM-baed model, LDA-based model, CF, Content-based model, and random baseline. Figures 2, 3, 4 and 5 show the performance of the SICTR and benchmarks in terms of recall, precision, accuracy and F-score. Each of the figures contain three subfigures. The difference between the three subfigures is in the topic modeling approach that is used for identifying the corpus' topics. For example, in Fig. 2, the subfigure a shows the results for recall measure where the corpus topics (items) in SICTR, WNTM-CTR, and LDA-CTR are identified through BTM.

In Fig. 2, it is clear that SICTR is outperforming the benchmarks on recall. The highest recall rate is achieved on 100 topics. The recall rate decreases as the number of topics grows due to increase in sparsity of the influence matrix. The second best performer in Recall measure is WNTM-CTR. Both SICTR (BTM-based) and WNTM are using topic modeling approaches that are developed for social networks. So, it is not out of expectations that they perform better than others. CF and Content-based model are the worst performers after the random baseline.

It is worth mentioning that the prediction results is highly dependent on the quality of topics generated in the first round of topic modeling. In Fig. 2, each subfigure is the experiment



(a) BTM Precision

(b) WNTM Precision

(c) LDA Precision

(d) Labels

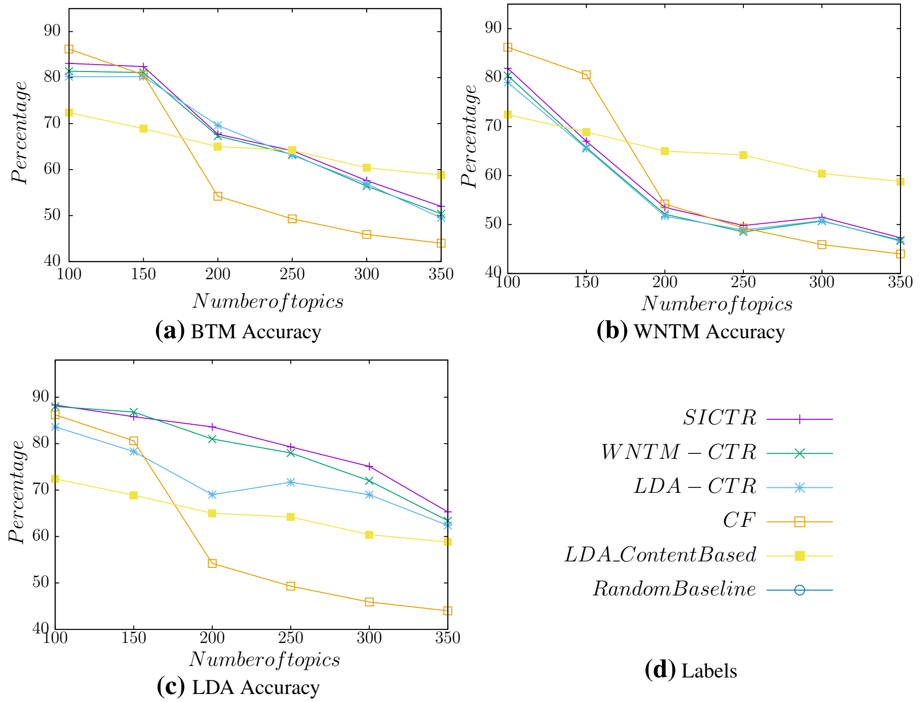
Fig. 3 Precision comparisons between SICTR, CF, LDA, and Random baseline. The x-axis shows experiments for different number of topics. The influence threshold is set to 0.3

result on a different topic modeling approach used in the first round to extract topics from the corpus. In subfigure b, the poor topics generated by WNTM impacted the prediction results. As SICTR and CTR related models in the benchmark are highly dependent on the first round of topic modeling, the content-based and CF outperformed the CTR based models in the experiment performed on WNTM topics.

Figure 3 shows precision metric results of SICTR and the benchmarks on the three sets of topics generated from BTM, WNTM, and LDA. The highest achieved precision is observed for results on 100 topics. The precision figures explain the quality of topic modeling as well. It can be seen that topics generated by BTM and LDA are leading to better precision by the models. The highlight of this subsection is that SICTR is slightly outperforming the benchmarks in terms of precision.

The results of accuracy metric can be found in Fig. 4. The accuracy of predicting influential users decreases when the number of topics increases, similar to recall and precision. Regardless of the topic modeling approach used in the pre-modeling stage, the SICTR has been more accurate against the benchmarks. WNTM-based and LDA-based models are second and third best performers. For instance, in Fig. 4c when number of topics is set to 150, accuracy measure is 88% for SICTR while it is 86% for WNTM-based model and 72% for LDA-based model respectively.

Furthermore, we report the F-score measure for this experiment. Figure 5 presents F-score and it shows the outperformance of SICTR in this evaluation as well. However, Fig. 5b shows otherwise when number of topics is less than 170. Similarly, the same trend was observed in Fig. 2b where CF was performing better on low number of topics when the topic modelling



(a) BTM Accuracy

(b) WNTM Accuracy

(c) LDA Accuracy

(d) Labels

Fig. 4 Accuracy comparisons between SICTR, CF, LDA, and Random baseline. The x-axis shows experiments for different number of topics. The influence threshold is set to 0.3

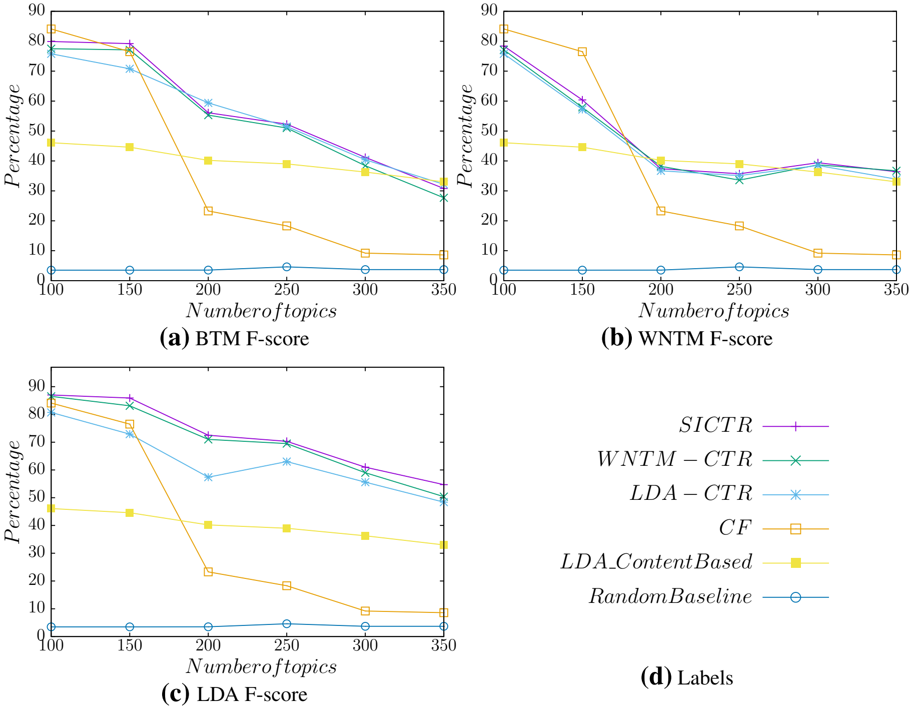
from the first round has been poor. According to Figs. 2b and 5b, CF's precision, accuracy, and F-score is higher than competitors when pre-modelling topics are in poor quality. Under the same circumstances, CF's outcome deteriorates when the number of topics increases.

5.5 Implications and applications

This section describes the real-world implications and applications of our model. Identifying topic-based influential users is similar to the problem of finding experts and authorities. Spotting the elite group of users for topics can inform and improve available systems, such as search engines. The query result for both contents and users can be returned and ranked using the score provided by our system.

One of the main applications of this work is in Marketing. Marketing campaigns benefit from targeting the influential users in the related topic, as this often leads to a more productive and cost effective campaign. Influential users act as hubs in the network and have a central position in the network in terms of information diffusion, and also attract and engage more users into their conversations. The influence prediction power of SICTR brings the ability to form a stronger campaign by predicting users influence for the new product.

Unlike the work in the literature, our model is able to detect *new* and *surprising* topics. This makes our work capable of working in the real-world and detecting new topics and related influential users. As a result, it eliminates the need for manually defining the topics, and at the same time it does *not* miss recent and new topics (which could have been missed via a manual process). SICTR can also be applied to a time-evolving setting, where it can



(a) BTM F-score

(b) WNTM F-score

(c) LDA F-score

(d) Labels

Fig. 5 F-score comparisons between SICTR, CF, LDA, and Random baseline. The x-axis shows experiments for different number of topics. The influence threshold is set to 0.3

detect when topics get viral and which users are influential in those topics at different periods of time.

6 Conclusions

In this study, we presented a collaborative topic regression model, SICTR, to predict topic-based user influence in social networks. We identified topics from user posts on social networks, measured each user's influence on each topic, based on the influence definition we proposed, and used SICTR to predict user influence for unobserved topics. The main contributions our work include:

- proposing a topic-based user influence prediction approach in social network that incorporates both network structure and user-generated content for topic-based influence measurement and prediction,
- opening a new discussion for user influence prediction in social networks that has not been explored in the literature.

We tested our topic-based influence measurement system and prediction model using a unique dataset that we collected from Twitter and Google Scholar. Our experiments show that SICTR is effective for predicting influential users for a new topic. For instance, SICTR can be valuable for marketing campaigns by selecting the most influential people for for a new product's marketing campaigns. In future work, we are interested in studying how topic-based user influence changes over time and improves the topic-based influence prediction. We will also investigate alternative methods for combining various influence measures.

References

Barbieri, N., Bonchi, F., & Manco, G. (2012). Topic-aware social influence propagation models. In *ICDM* (pp. 81–90).

Bingol, K., Eravci, B., Etemoglu, C. O., Ferhatosmanoglu, H., & Gedik, B. (2016). Topic-based influence computation in social networks under resource constraints. *IEEE Transactions on Services Computing*, *PP*(99), 1–1.

Cano, A. E., Mazumdar, S., & Ciravegna, F. (2014). Social influence analysis in microblogging platforms: A topic-sensitive based approach. *SWJ*, *5*(5), 357–372.

Cataldi, M., & Aufaure, M. A. (2015). The 10 million follower fallacy: Audience size does not prove domain-influence on twitter. *KAIS*, *44*(3), 559–580.

Chen, C., Zheng, X., Wang, Y., Hong, F., & Lin, Z. (2014). Context-aware collaborative topic regression with social matrix factorization for recommender systems. In *AAAI*.

Chen, T., Sun, Y., Shi, Y., & Hong, L. (2017). *On sampling strategies for neural network-based collaborative filtering, KDD '17* (pp. 767–776). New York: ACM.

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, *26*(12), 2928–2941.

Embar, V. R., Bhattacharya, I., Pandit, V., & Vaculin, R. (2015). Online topic-based social influence analysis for the wimbledon championships. In *KDD* (pp. 1759–1768).

Hajian, B., & White, T. (2011). Modelling influence in a social network: Metrics and evaluation. In *PASSAT* (pp. 497–500).

Hamzehei, A., Jiang, S., Koutra, D., Wong, R. K., & Chen, F. (2016). TSIM: Topic-based social influence measurement for social networks. In *Proceedings of The 14th Australasian data mining conference*.

Hamzehei, A., Jiang, S., Koutra, D., Wong, R., & Chen, F. (2017). Topic-based social influence measurement for social networks. *Australasian Journal of Information Systems*, *21*, 61.

Haveliwala, T. H. (2002). Topic-sensitive pagerank. In *WWW* (pp. 517–526).

Herzig, J., Mass, Y., & Roitman, H. (2014). *An author–reader influence model for detecting topic-based influencers in social media, HT '14* (pp. 46–55). New York: ACM.

Hu, G. N., Dai, X. Y., Qiu, F. Y., Xia, R., Li, T., Huang, S. J., et al. (2018). Collaborative filtering with topic and social latent factors incorporating implicit feedback. *ACM Transactions on Knowledge Discovery from Data*, *12*(2), 23:1–23:30.

Hu, J., Fang, Y., & Godavarthy, A. (2013). Topical authority propagation on microblogs. In *CIKM* (pp. 1901–1904).

Katsimpras, G., Vogiatzis, D., & Paliouras, G. (2015). Determining influential users with supervised random walks. In *WWW* (pp. 787–792).

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *JACM*, *46*(5), 604–632.

Kong, S., & Feng, L. (2011). A tweet-centric approach for topic-specific author ranking in micro-blog. In *ADMA* (pp. 138–151).

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 30–37.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter: A social network or a news media? In *WWW* (pp. 591–600).

Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016). Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval* (pp. 165–174). ACM.

Liu, X., Shen, H., Ma, F., & Liang, W. (2014). Topical influential user analysis with relationship strength estimation in twitter. In *ICDM* (pp. 1012–1019).

Montangero, M., & Furini, M. (2015). Trank: Ranking twitter users according to specific topics. In *CCNC* (pp. 767–772).

Pal, A., & Counts, S. (2011). Identifying topical authorities in microblogs. In *WSDM* (pp. 45–54).

Purushotham, S., Liu, Y., & Kuo, C. C. J. (2012). Collaborative topic regression with social matrix factorization for recommendation systems. In *ICML*.

Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011). Influence and passivity in social media. In *PKDD* (pp. 18–33).

Sung, J., Moon, S., & Lee, J. G. (2013). The influence in Twitter: Are they really influenced? (pp. 95–105).

Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *KDD* (pp. 448–456).

Wang, H., Chen, B., & Li, W. J. (2013). Collaborative topic regression with social regularization for tag recommendation. In *IJCAI*.

Welch, M. J., Schonfeld, U., He, D., & Cho, J. (2011). Topical semantics of twitter links. In *WSDM* (pp. 327–336).

Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010). Twitterrank: Finding topic-sensitive influential Twitterers. In *WSDM* (pp. 261–270).

Xiao, F., Noro, T., & Tokuda, T. (2014). Finding news-topic oriented influential twitter users based on topic related hashtag community detection. *JWE*, *13*(5–6), 405–429.

Zhao, Z., Yang, Q., Lu, H., Weninger, T., Cai, D., He, X., et al. (2018). Social-aware movie recommendation via multimodal network learning. *IEEE Transactions on Multimedia*, *20*(2), 430–440.

Zuo, Y., Zhao, J., & Xu, K. (2016). Word network topic model: A simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, *48*(2), 379–398.

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