IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 27, 2018, accepted October 15, 2018, date of publication October 29, 2018, date of current version November 30, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2878494

Modeling Topic Evolution in Twitter: An Embedding-Based Approach

MUHAMMAD ABULAISH[1], (Senior Member, IEEE), AND MOHD FAZIL[2]
[1]Department of Computer Science, South Asian University, New Delhi 110021, India
[2]Department of Computer Science, Jamia Millia Islamia, New Delhi 110025, India

Corresponding author: Muhammad Abulaish (abulaish@sau.ac.in)

ABSTRACT In last two decades, online social networks have grown vertically as well as horizontally. Due to various users activities in these networks, huge amount of data, mainly textual, is being generated that can be analyzed at different levels of granularity for various purposes, including behavior analysis, sentiment analysis, and predictive modeling. In this paper, we propose a word embedding-based approach to analyze users-centric tweets to observe their behavior evolution in terms of the topics discussed by them over a period of time. We also present a word embedding-based proximity measure to monitor temporal transitions between the topics using five topic evolution events – *emergence*, *persistence*, *convergence*, *divergence*, and *extinction*. The proximity between a pair of topics is defined as a function of the content and contextual similarity between their word distributions, wherein the contextual similarity is calculated using word embedding. The proposed approach is evaluated over three `Twitter` datasets in line with the existing state-of-the-art approaches in literature and the experimental results are encouraging.

INDEX TERMS Social network analysis, twitter data analysis, temporal evolution, topic modeling, word embedding.

I. INTRODUCTION

Online social networks (OSNs) have become the primary source of communication and users engagement as almost one third of the world population is registered on atleast one OSN.[1] `Twitter`, a microblogging platform, is a special kind of OSN where users can register and share their views, thoughts, and information about any event, incident, or topic of interest using a concise message, limited to maximum 280 characters. Due to large user-base and various functionalities, OSNs are becoming the source of vast amount of data, and accordingly establishing new research dimensions, such as social computing, predictive modeling, and big data analytics. In the line, researchers have analyzed the evolution of structural and interactional properties of different OSNs and their users [1], [2]. Further, existing literatures have different approaches to track the evolving vocabulary, themes, and topics in OSNs. In this direction, Blei and Lafferty [3] presented a state space model-based approach utilizing the natural parameters of multinomial distribution to monitor various topical dynamics over a time-scale. They generated per-document topic distribution and per-topic word distribution for different time-intervals and observed the evolution of a topic over different time-intervals. However, they only observed topic generation over the time-intervals with fixed number of topics for each interval and did not track the topic evolutions.

In this paper, we propose an embedding-based approach to analyze users-specific tweets to observe their behavioral evolution in terms of the topics discussed by them over a period of time. The transition between topics is based on their proximity, which is defined as a function of the content and contextual proximity between their word distributions. The contextual proximity is based on the embeddings of unmatched words of the underlying topics. The proposed approach tracks topic transitions over different time-intervals using five topic evolution events – *emergence*, *persistence*, *convergence*, *divergence*, and *extinction*.

The rest of the paper is organized as follows. Section II presents a brief review of the existing literatures on topics evolution in large document corpus in general, and in OSN contents in particular. It also discusses how our proposed approach differs from existing approaches. Section III presents a brief description of the basic concepts that are used

[1]https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

as building blocks for the proposed approach. Section IV presents the functional details of the proposed embedding-based topic evolution modeling approach. Section V presents a brief description of the experimental dataset and topic extraction process. It also presents experimental details and evaluation results. Finally, section VI concludes the paper with future directions of research.

II. RELATED WORKS

Online social networks are the rich source of information comprising of structured, semi-structured and unstructured information, which are vital for various research fields. Since inception, social networks are the basis of various researches and studies, which did not exist or possible before. In this direction, Leskovec *et al.* [1] analyzed the connection forming behavior of individual OSN user to model the behavior of groups of users and processes describing the evolution of network structure. In other words, authors used users' microscopic behavior to model the network-level macroscopic processes. Similarly, Yang *et al.* [2] analyzed the temporal evolution of OSN user interactions to model interaction dynamics of the users. In addition, they also incorporated the users connection formation dynamics in their model. Researchers have also proposed different algorithms to track the community formation behavior of OSN users and devised several transition events to model their dynamic behavior over time [4], [5].

In existing literatures, there are several approaches to track the content, vocabulary, and topical evolution of document collections, databases, and OSN user timelines. In this direction, Spiliopoulou *et al.* [6] presented a cluster transition monitoring framework to track changes in the content of a data stream, in which change between clusters is based on overlapping words of the clusters. They also presented a method to capture inter-cluster transitions, such as *absorption*, *split*, and *survival*, and to track intra-cluster changes, such as change in cluster size like *shrink*, *expand* etc. In another similar approach, Mei and Zhai [7] performed temporal text mining to model theme evolution in a text stream. They partitioned text stream into different time-intervals and extracted latent themes from the underlying text corpuses using a probabilistic mixture model [8]. Thereafter, they modeled the extracted themes as theme evolution graph and applied KL-divergence [9] between the word distributions of the themes to observe their temporal behavior. Blei and Lafferty [3] proposed a probabilistic approach to model topics dynamic and evaluated it on a large collection of documents. They used a state space model based on multinomial distribution to extract fixed number of topics across each time-interval.

All approaches mentioned above extract topics over different time-intervals to find similarity between topics based on word distributions of the underlying topics pair, completely ignoring the contextual similarity between the unmatched words. Further, none of the discussed approaches have defined a complete set of evolutionary events, such as *appear*, *disappear*, *split*, *merge*, and so on to monitor different transition events. More significantly, all these approaches have been targeted and evaluated for topic evolution in text corpuses, rather than OSN data, which is generally informal, short, and ambiguous. However, Lauschke and Ntoutsi [10] proposed a change detection framework based on users profile constructed using the topics discussed by them over `Twitter`. They used bisecting k-means technique to extract topics from tweets and applied KL-divergence on the word distributions of the topics to observe the change between every pair of topics. They defined changes between topics using three evolution metrics – number of surviving topics, number of appearing topics, and number of disappearing topics. Similarly, Caro *et al.* [11] proposed a topic evolution tracking framework using Latent Dirichlet Allocation (`LDA`) model. They devised different events to monitor the change between the topics of different time-intervals and evaluated on a scientific papers repository.

To the best of our knowledge, there is no topic monitoring approach for OSN users that tracks the evolution between topics, incorporating both textual similarity of the matched words and contextual similarity of the unmatched words. Our proposed approach also tracks changes between the word distributions of the topics using different transition events, such as *emergence*, *persistence*, *convergence*, *divergence*, and *extinction*.

III. PRELIMINARIES

In this section, we present a brief description of the basics concepts used in our proposed embedding-based approach for topic evolution modeling in `Twitter`. Starting with `LDA` model, the concepts of m-partite graph and word embedding are discussed in the following paragraphs.

a: LATENT DIRICHLET ALLOCATION

Since our proposed approach is primarily presented to model the topic transitions in `Twitter`'s users data, topics are extracted using `LDA` from a batch of tweets, rather than from each individual tweet, as tweets are generally short messages and it is difficult to get any meaningful topic from them. `LDA` is a generative probabilistic model that represents document as a mixture of topics and assigns every word of the document to be drawn from one of these topics [12]. The plate notation representation of `LDA` is shown in fig. 1, where α is the Dirichlet prior parameter that controls per-document topic distribution, β is another Dirichlet prior parameter that controls per-topic word distribution, θ represents document's topic distribution, ϕ is the per-topic word distribution, and z is the topic assignment to a particular word w, which is the only observed variable. In this paper, while extracting topics using `LDA`, we specified the value of hyper-parameters α and β as 0.1 and 0.01, respectively.

b: THE m-PARTITE GRAPH AND THRESHOLD

Considering the timeline of a user u which is divided into m partitions, topics are extracted using `LDA` from every partition

Algorithm 1 TopicEvolutionModeling(D, θ)

```
/* D is the set of tweets of user u                                   */
/* θ is the set of threshold values θ_ee, θ_cd, and θ_p               */
begin
    P ← dataPartitioning(D);   /* partition chronologically sorted tweets D into m parts
    and assign them to P */
    topicset_list ← [ ] ;                          /* initialize the topicset_list */
    M ← [ ][ ][ ] ;   /* initialize matrix M to hold the proximity values between every
    pair of topics of adjacent time-intervals */
    foreach parttion p in P do
        T_p ← preprocess(p)
        T_p ← topicExtraction(T_p) ;                /* extract k topics from T_p using LDA */
        topicset_list ← T_p;        /* append extracted topics from T_p into topicset_list */
    end
    foreach topicset in topicset_list do
        topicset_i ← topicset_list[i]
        topicset_{i+1} ← topicset_list[i + 1]
        foreach topic_j in topicset_i do
            foreach topic_k in topicset_{i+1} do
                P^E = explicitProximity(topic_j, topic_k) ;   /* calculate explicit proximity value */
                P^I = implicitProximity(topic_j, topic_k);    /* calculate implicit proximity value */
                P_x = P^E + P^I
                M[i][j][k] = P_x
            end
        end
    end
    call TopicEvolutionEvents(M, θ);
end
```

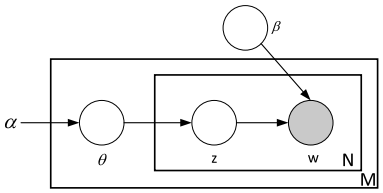

FIGURE 1. Plate notation of LDA model.

and represented as an *m-partite* graph, where nodes of the graph represent the topics. The similarity between the topics of adjacent time-intervals is calculated and edges are created between the topics if their proximity value is greater than a pre-defined threshold. In the proposed approach, we have defined three thresholds θ_p, θ_{cd}, and θ_{ee} for different transition events, where θ_p is used to determine persistent topic, θ_{cd} verifies the convergence and divergence events, and θ_{ee} is used to represent the emergence and extinction events. The values of these thresholds are chosen in such a way that $\theta_{ee} < \theta_{cd} < \theta_p$. The value of θ_p is set to 0.5 and kept highest as it is used to observe persistency between the topics which means that the topics are significantly similar. The value of θ_{cd} is kept as moderately high because in case of convergence coherently similar topics are merged into a single topic, whereas in case of divergence a topic is splitted into multiple coherent topics, but not as similar as persistent. Finally, the value of θ_{ee} is kept smallest because a topic to emerge at i^{th} interval implies that either the topic did not match (or insignificantly match) with all the topics of the $(i-1)^{th}$ interval, representing the emergence event. Similarly, a topic at $(i-1)^{th}$ interval to become dead implies that either the topic did not match (or insignificantly match) with all the topics of the i^{th} interval, representing the extinction event. Further details about all these events are presented in section IV-B.

c: WORD EMBEDDING

In word embedding, every word is mapped from an 1-dimensional space to a higher dimensional space by replacing it with a numeric vector that represents the position of the word in the vector space [13]. The numerical vector contains the contextual information of the word with respect to the co-occurring words. For example, if two words w_i, and w_j are contextually similar, then the distance between their embeddings, say e_i and e_j, will be low, showing that the two words are contextually similar and they are close to each other in the vector space. In the proposed approach, we have used GloVe word vector representation, which follows an

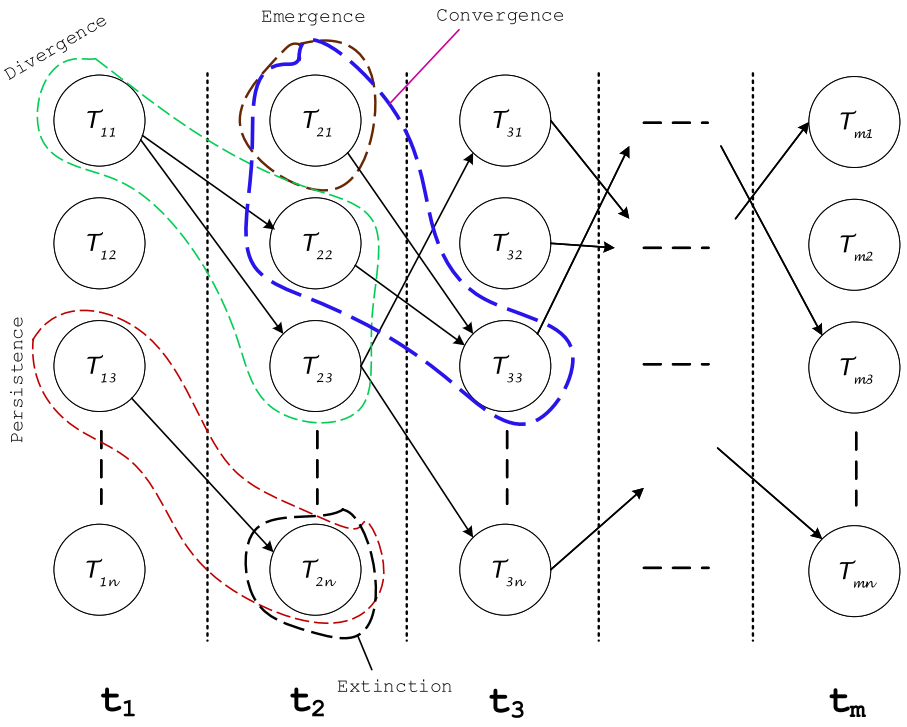

FIGURE 2. An *m*-partite graph representing relationships between the topics.

unsupervised approach to learn the vector representation of words [14]. GloVe uses the advantages of both *word2vec skip-gram* model and *matrix factorization* model.

IV. PROPOSED TOPIC EVOLUTION MODELING

This section presents a detailed description of the proposed approach for modeling topics evolutions over different time-intervals. The proposed approach is formally represented in algorithm 1, which consists of two main components – (i) *topical proximity*, and (ii) *topic transition events* are described in the following sub-sections.

A. TOPICAL PROXIMITY

This section presents a detailed description of the proposed topical proximity measure to find similarity between every pair of the topics of adjacent time-intervals. The proximity measure is mathematically defined in (1), where $\mathcal{T}_{ki}$ and $\mathcal{T}_{(k+1)j}$ are the i^{th} and j^{th} topics of the k^{th} and $(k+1)^{th}$ time-intervals, respectively. The topics of m different time-intervals are organized as an m-partite graph as shown in fig. 2, where edge between the topics $\mathcal{T}_{ki} \in \mathcal{T}_k$ and $\mathcal{T}_{(k+1)j} \in \mathcal{T}_{k+1}$ of the adjacent partitions t_k and t_{k+1} is created only when the proximity value is greater than one of the thresholds and satisfies the corresponding transition conditions. We have not considered the intra-partition topics proximity as the proposed approach monitors topics evolution over different time-intervals, rather than within a time-interval. The proximity between every topics pair is calculated as given in (1), which is the sum of the two types of explicit and implicit proximities. Further details about these proximities are given in the following sub-sections.

$$\mathcal{P}_x(\mathcal{T}_{ki}, \mathcal{T}_{(k+1)j}) = \mathcal{P}^{\mathcal{E}}(\mathcal{T}_{ki}, \mathcal{T}_{(k+1)j}) + \mathcal{P}^{\mathcal{I}}(\mathcal{T}_{ki}, \mathcal{T}_{(k+1)j}) \quad (1)$$

1) EXPLICIT PROXIMITY

The explicit proximity between a pair of topics determines the proximity value based on the matching of their word distributions. Two topics with similar word distributions are likely to describe similar discussion topics. In explicit proximity, we match the occurrence of words between every pair of topics of the adjacent time-intervals. The explicit proximity represented by $\mathcal{P}^{\mathcal{E}}$ between two topics $\mathcal{T}_{ki}$, and $\mathcal{T}_{(k+1)j}$ is calculated as the ratio of the overlapping words occurring in the word distributions of both topics to the higher-length topic. Mathematically, it is defined in 2.

$$\mathcal{P}^{\mathcal{E}}(\mathcal{T}_{ki}, \mathcal{T}_{(k+1)j}) = \frac{|(\mathcal{T}_{ki} \cap \mathcal{T}_{(k+1)j})|}{max(|\mathcal{T}_{ki}|, |\mathcal{T}_{(k+1)j}|)} \quad (2)$$

2) IMPLICIT PROXIMITY

What about the words between the word distributions of topics pairs that do not match? Are they really different or just two different words like *Modi* and *BJP* that are contextually coherent. We compute implicit proximity between a pair of topics to find the contextual proximity of the unmatched words. The contextual proximity between two different words is calculated using their contextual word vector. In order to observe the contextual proximity, each word is represented as a d-dimensional contextual vector of real numbers. In existing literatures, there are various models like word2vec [15],

and GloVe [14] that already have learned word vectors from different large corpus to represent each word as a contextual vector of the real numbers. In the proposed approach, we have used GloVe model to replace every unmatched word with their corresponding 25-dimensional embedding. If unmatched word distributions of the topics $\mathcal{T}_{ki}$ and $\mathcal{T}_{(k+1)j}$ are $\mathcal{T}'_{ki}$ and $\mathcal{T}'_{(k+1)j}$, respectively, then the implicit proximity between them is calculated as shown in 3, where em_r and em_s are the embeddings of the unmatched words r and s, and p and q are the number of unmatched words in these topics.

$$\mathcal{P}^{\mathcal{I}}(\mathcal{T}_{ki}, \mathcal{T}_{(k+1)j}) = \frac{\sum_{r \in \mathcal{T}'_{ki}} \sum_{s \in \mathcal{T}'_{(k+1)j}} cosine(em_r, em_s)}{p \times q} \tag{3}$$

B. TOPIC TRANSITION EVENTS

This section presents a detailed description of different transition events to track topics evolution in users-centric tweets over different time-intervals. Considering topic $\mathcal{T}_{ki} \in \mathcal{T}_k$ as the i^{th} topic of the k^{th} partition and $\mathcal{T}_k$ as the underlying topic set, every topic $\mathcal{T}_{ki}$ is matched with every topic of the $(k+1)^{th}$ time-interval to monitor the transition of the topics of k^{th} time-interval to the topics of the $(k+1)^{th}$ time-interval, based on topics proximity defined in section IV-A. Different transition events between the topics of adjacent time-interval are based on two parameters – topics proximity value and the number of matching topics. Corresponding to the topics of all m partitions an m-partite graph is constructed and proximity is calculated between every pair of topics of the adjacent partitions. Finally, edge between pair of topics is created if the proximity value is greater than the respective threshold and it satisfies the underlying transition condition. On the basis of threshold value and number of matching topics of the k^{th} time-interval with the topics of the $(k+1)^{th}$ time-interval, we have defined five topic transition events, which are described in the following paragraphs.

a: EMERGENCE

If word distributions of a topic $\mathcal{T}_{(k+1)j}$ of $(k+1)^{th}$ time-interval insignificantly match with the word distributions of all the topics of preceding k^{th} time-interval, such that the proximity value between every topics pairs is less than the threshold θ_{ee}, then it represents the fact that topic $\mathcal{T}_{(k+1)j}$ was not in discussion during k^{th} time-interval. The topic $\mathcal{T}_{21}$ of t_2 time-interval shown in fig. 2 shows an exemplary topic emerging in the 3^{rd} time-interval which was absent during the second time-interval. Emergence of topics between any two time-intervals is represented by $\mathcal{E}$ which is a *none-to-one* relationship, as represented by the relation in the first row of table 1.

b: PERSISTENCE

Whenever a topic $\mathcal{T}_{ki}$ of k^{th} time-interval matches to only one topic $\mathcal{T}_{(k+1)j}$ of $(k+1)^{th}$ time-interval such that the proximity value between the topics is greater than the persistent threshold θ_p, it is said that the topic $\mathcal{T}_{ki}$ has sustained over the time-period $t_k \rightarrow t_{k+1}$. This observation is shown in fig. 2 using the topic transition $\mathcal{T}_{13} \rightarrow \mathcal{T}_{2n}$. The persistency of a topic is equivalent to an *one-to-one* relationship and represents long-term discussion of a theme, showing the interest and intention of a user towards joining the Twitter. Mathematically, persistence event, represented by $\mathcal{P}$, between a pair of topics $\mathcal{T}_{ki}$ and $\mathcal{T}_{(k+1)j}$ of the adjacent time-intervals k and $(k+1)$ is represented in the second row of table 1.

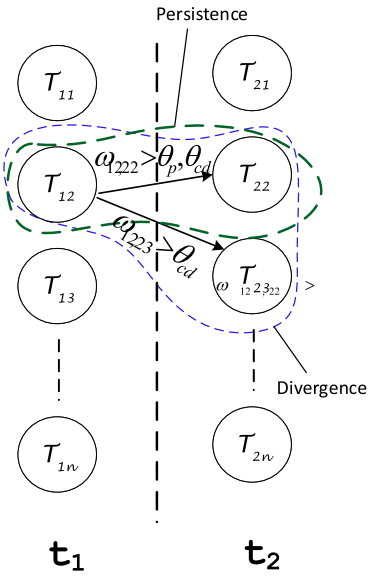

FIGURE 3. An *m*-partite graph representing the co-occurrence of *divergence* and *persistence* events.

c: DIVERGENCE

Suppose a topic $\mathcal{T}_{ki}$ discussed by user u at k^{th} time-interval matches with a subset of topics $\mathcal{T}^s_{(k+1)j} \subset \mathcal{T}_{k+1}$ of $(k+1)^{th}$ time-interval such that the proximity $\mathcal{P}_x(\mathcal{T}_{ki}, \tau) > \theta_{cd}\ \forall \tau \in \mathcal{T}^s_{(k+1)j}$ with respect to the divergence threshold θ_{cd}. The value of threshold θ_{cd} for both the convergence and divergence scenario is chosen such that $\theta_{cd} < \theta_p$ as divergence splits a topic into multiple topics, and thereby decreasing the proximity between the original and splitted topics in order to distribute the proximity over multiple topics. Fig. 2 shows an exemplary divergence transition event, represented by $\mathcal{D}$, using the topic transitions $\{\mathcal{T}_{13} \rightarrow \{\mathcal{T}_{22}, \mathcal{T}_{23}\}$. In addition, a diverging set of topics from k^{th} to $(k+1)^{th}$ time-interval can also include persisting topics, as represented by an exemplary persistent topic transition $\mathcal{T}_{12} \rightarrow \mathcal{T}_{22}$ in the diverging topic transition $\mathcal{T}_{12} \rightarrow \{\mathcal{T}_{22}, \mathcal{T}_{23}\}$, which is shown in fig. 3.

d: CONVERGENCE

Sometimes, users discuss multiple topics and later converge to a single topic, which is coherent with multiple previously discussed topics. In order to monitor such transitions, we have proposed the concept of topic convergence. If a user u has discussed $\mathcal{T}_k$ set of topics during the k^{th} time-interval and a subset $\mathcal{T}^s_k \subset \mathcal{T}_k$ of those topics match with a topic $\mathcal{T}_{(k+1)j}$ of the $(k+1)^{th}$ time-interval such that their topic proximity

TABLE 1. Mathematical relations representing different topic transition events.

Mathematical relation	Topic transition event
$\mathcal{E}=\left\{(\mathcal{T}_k, \mathcal{T}_{(k+1)j}) : \forall \mathcal{T}_{ki} \in \mathcal{T}_k, \mathcal{P}_x(\mathcal{T}_{ki}, \mathcal{T}_{(k+1)j}) < \theta_{ee}\right\}$	Emergence
$\mathcal{P}=\left\{(\mathcal{T}_{ki}, \mathcal{T}_{(k+1)j}) : \mathcal{P}_x(\mathcal{T}_{ki}, \mathcal{T}_{(k+1)j}) > \theta_p\right\}$	Persistence
$\mathcal{D}=\left\{(\mathcal{T}_{ki} \times \mathcal{T}^s_{k+1}) : \mathcal{T}_{ki} \in \mathcal{T}_k, \mathcal{T}^s_{k+1} \subset \mathcal{T}_{k+1}, \text{and } \forall \tau \in \mathcal{T}^s_{k+1}, \mathcal{P}_x(\mathcal{T}_{ki}, \tau) > \theta_{cd}\right\}$	Divergence
$\mathcal{C}=\left\{(\mathcal{T}^s_k \times \mathcal{T}_{(k+1)j}) : \mathcal{T}^s_k \subset \mathcal{T}_k, \mathcal{T}_{(k+1)j} \in \mathcal{T}_{k+1}, \text{and } \forall \tau \in \mathcal{T}^s_k \mathcal{P}_x(\tau, \mathcal{T}_{k+1}) > \theta_{cd}\right\}$	Convergence
$\mathcal{E}_x=\left\{(\mathcal{T}_{ki} \times \mathcal{T}_{k+1}) : \forall \mathcal{T}_{(k+1)j} \in \mathcal{T}_{k+1}, \mathcal{P}_x(\mathcal{T}_{ki}, \mathcal{T}_{(k+1)j}) < \theta_{ee}\right\}$	Extinction

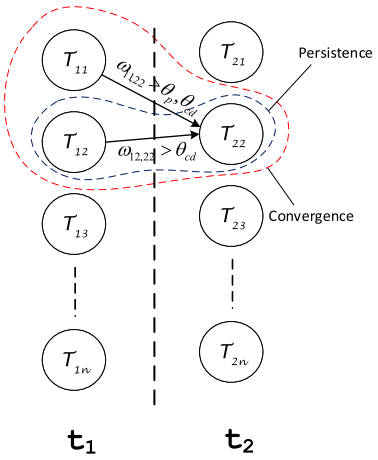

FIGURE 4. An *m*-partite graph representing the co-occurrence of *convergence* and *persistence* events.

is greater than the threshold θ_{cd}, then it is said that $\mathcal{T}^s_k$ has converged to a single topic. A convergence event, represented as $\mathcal{C}$, is shown using the topic transition $\{\mathcal{T}_{21}, \mathcal{T}_{22} \rightarrow \mathcal{T}_{33}\}$ in fig. 2. Similarly, a convergence event including persistence may also occur and it is shown in fig. 4. The convergence relation $\mathcal{C}(\mathcal{T}^s_k, \mathcal{T}_{(k+1)j})$ is shown in the fourth row of table 1. Mathematically, it is equivalent to a *many-to-one* relation.

e: EXTINCTION

In OSNs, users show very random behavior; they are sometimes very active and sometimes become inactive for a longer period of time. Similar is the case while discussing topics in social media. Although experts always discuss similar topics that are related to their specialization, but at the lower level of the topics representation they may show incoherent behavior. For example, a political analyst generally talks about politics, but during a topic discussion he/she discusses about specific election, political scenario of a particular region, government policies, etc. The proposed *extinction* event is to observe the topic-leaving behavior of the users. If a user discusses a topic $\mathcal{T}_{ki}$ in k^{th} time-interval and stops talking about it in the next time-interval, then the topic $\mathcal{T}_{ki}$ is called a extincted topic at $(k+1)^{th}$ time-interval. In order to track extincted topics, every topic $\mathcal{T}_{ki} \in \mathcal{T}_k$ of a time-interval is matched with every topics of the $(k+1)^{th}$ period subject to the threshold θ_{ee}. If $\mathcal{P}_x(\mathcal{T}_{ki}, \tau) < \theta_{ee}\ \forall \tau \in \mathcal{T}_{(k+1)j}$, then the topic $\mathcal{T}_{ki}$ is said to be extinct. However, it is possible that an extincted topic may resume in some later time-interval.

V. EXPERIMENTAL RESULTS

In this section, we present experimental results on three `Twitter` datasets to establish the efficacy of the proposed embedding-based topics evolution modeling approach. The following sub-sections present a detailed descriptions of datasets curation and processing, topics mapping, and impact analysis of different parameters on the proposed approach.

A. DATA CURATION AND PROCESSING

The dataset D_u of a user u is curated from `Twitter` using REST API, and it contains a set of n tweets collected from the timeline of u during a time-period T, i.e., $D_u = \{\tau_1, \tau_2, \ldots, \tau_n\}$. The crawled tweets are time-stamped and chronologically sorted and partitioned into m groups $P_1, P_2, \ldots, P_m$ containing equal number of tweets, and time-interval $t_1, t_2, \ldots, t_m$. Although the number of tweets are same in each group their time-intervals are no necessarily same due to random behavior of the OSN users; sometimes they are active and frequently post messages, whereas sometimes they are inactive for a longer period of time, even for months. Therefore, in order to avoid partitions without tweets or with very few tweets, equal-depth binning was applied to map tweets to different partitions. Thereafter, pre-processing steps were applied over the tweets to filter out non-English tweets and to remove stop-words using *nltk*[2] toolkit, a `python` suite of libraries for natural language processing. Similarly, each words was lemmatized to get its base form using `spaCy`,[3] another open-source `python` library for natural language processing. Thereafter, topics were extracted from the pre-processed tweets of every time-interval using freely distributed open-source implementation of `LDA` [16]. It takes text corpus as an input and provides topics in the form of word distributions, where each word is word is assigned a relevance score. Equation 4 presents an exemplary i^{th} topic of the first partition ($\mathcal{T}_{1i}$) extracted by `LDA`, where $r(w_n)$ represents the relevance score of the

[2]https://github.com/nltk/nltk
[3]https://spacy.io/

Algorithm 2 TopicEvolutionEvents(M, θ)

```
/* M is a 3-d matrix containing proximity value between the topics pairs of the
adjacent time-intervals */ /* θ is the set of threshold values θ_ee, θ_cd, and θ_p */
begin
    foreach M[i][.][k] in M do
        if none of the M[i][.][k] values is greater than θ_ee then
            event ← emergence ;      /* a topic has emerged in (i+1)^th interval which was not
            present in i^th interval */
        if M[i][.][k] has m values greater than θ_cd then
            event ← convergence ;    /* m topics from i^th interval have converged to a single
            topic of (i+1)^th interval */
    end
    foreach M[i][j][k] in M do
        if M[i][j][k] is greater than θ_p then
            event ← persistence ;    /* j^th topic of i^th time-interval persisted as k^th in (i+1)^th
            interval */
    end
    foreach M[i][j][.] in M do
        if M[i][j][.] has m values greater than θ_cd then
            event ← divergence ;     /* j^th topic of i^th time-interval diverged into m topics in
            (i+1)^th interval */
        if none of the M[i][j][.] values is greater than θ_ee then
            event ← extinction ;  /* j^th topic of i^th time-interval extincted in (i+1)^th interval
            */
    end
end
```

word w_n.

$$\mathcal{T}_{1i} = \{(w_1, r(w_1)), (w_2, r(w_2)), \ldots, (w_n, r(w_n))\} \quad (4)$$

B. TOPIC EVOLUTION MAPPING

In this section, we present the topics evolution modeling on three Twitter datasets, out of which two datasets are curated from the timelines of two most influential personalities (Barack Obama and Donald Trump), whereas the third dataset is curated from the timeline of a Twitter socialbot. Further details and experimental results are given in the following sub-sections.

1) BARACK OBAMA

In this case, we crawled 3200 tweets from the timeline of the former American President, Barack Obama, which includes tweets from 29th August 2014 to 15th February 2018. The tweets were chronologically sorted on the basis of the associated time-stamp. Thereafter, equal-depth binning was applied on the tweets to divide them into 10 groups (partitions), each one containing 320 tweets and an associated time-interval. Thereafter, LDA was applied on each partition for topics extraction. In the topic extraction process, we manually observed the efficacy of LDA for different values of the *number of topics*, α, and β parameters. On analysis, we found that adjusting *number of topics* parameter at 10 provides the best result. Therefore, we extracted 10 topics from each partition and observed different topic transition events over all time-intervals and plotted the occurrence of each topic transition event, which is shown using blue line in fig. 5. Further analysis reveals that Obama has continuously tweeted about the same set of topics during fourth to sixth time-intervals before switching to some new topics during the seventh time-interval. Accordingly, highest number of new topics emerged between sixth and seventh time-intervals, as shown in fig. 5(a). On analysis, we found that it was at the beginning of the last year (late 2015) of his presidentship when he started talking about the achievements made during the last seven years of his presidency and assurance to keep on working in future. A word-cloud representation of the topics and transition events between three time-intervals t_1, t_2 and t_3 is shown in fig. 6. The word-cloud is a visualization technique to plot word distributions where font-size of a word is proportional to its relative relevance score. The word-cloud shown in fig. 6 depicts all transition events between the topics of all three intervals, except one convergence event including topics $\mathcal{T}_{24}$, $\mathcal{T}_{25}$, and $\mathcal{T}_{29}$ of time-interval t_2 that converge into 8th topic $\mathcal{T}_{38}$ of t_3. It can be observed that the last topic of time-interval t_3 emerges as a new topic because theme of the discussion is unclear and ambiguous and does not overlap or show contextual similarity with any topics of t_2. Similarly, 8th topic of t_1 where former president was asking for public support towards many social welfare movements disappeared during the time-interval t_2.

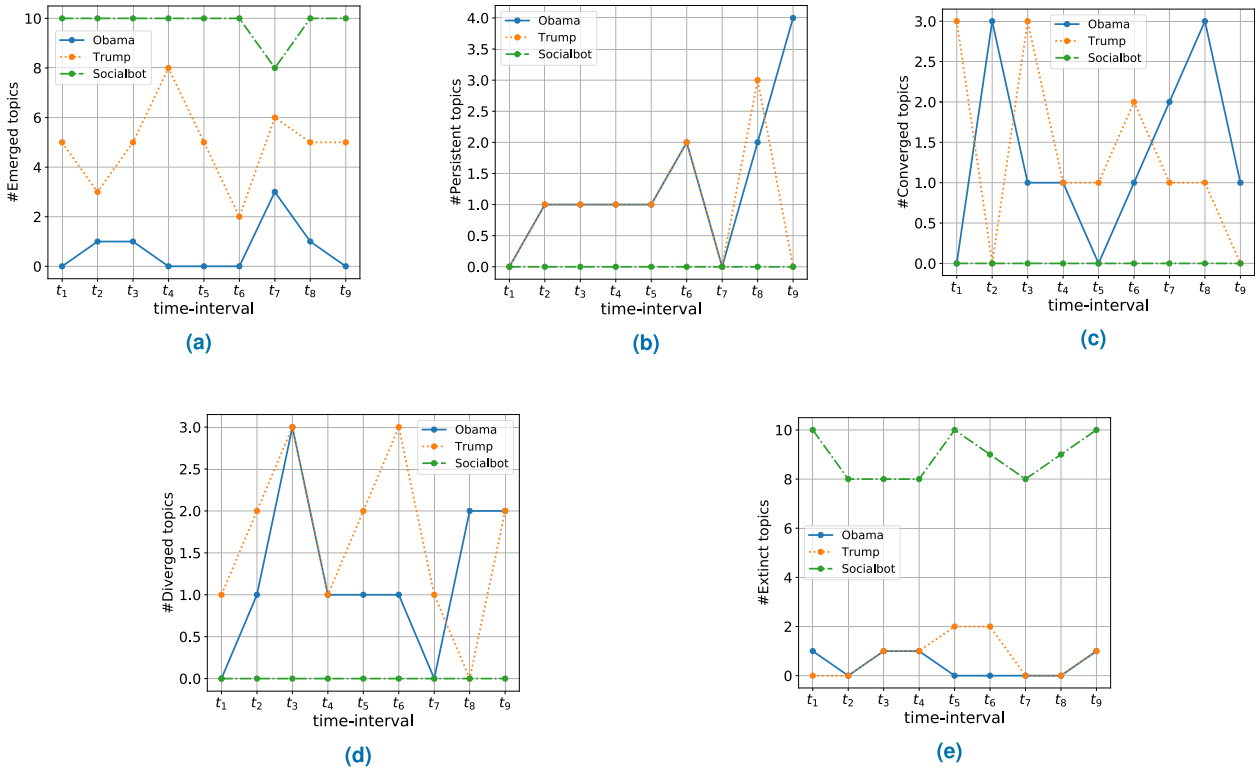

FIGURE 5. Number of topic transition events observed from the `Twitter` timelines of Barack Obama, Donald Trump, and a socialbot.

Similarly, other transition events can be observed in this figure. All these transition events in this figure are realized using threshold values for θ_p, θ_{cd} and θ_{ee} as 0.5, 0.425, and 0.35, respectively. Further, it can be observed that a number of topics across all three time-intervals do not have any edge, which does not mean that such topics are semantically independent, rather their semantic similarity is not strong enough with respect to the chosen threshold values.

2) DONALD TRUMP

In this case too, we curated total 3200 tweets from the `Twitter` timeline of the present American president Donald Trump posted during 30^{th} January 2017 to 10^{th} May 2018, which is just 10 days after he assumed office as an American president. Like previous case, crawled tweets were partitioned, pre-processed, and topics were extracted from them using `LDA`. On analysis, it is observed that Trump discussed a number of topics and a number of them persisted over two time-intervals, but none of the topics persisted over longer period of time. It is due to the fact that Trump continuously talked about current affairs, such as economy reform, *taking a knee* movement, building Mexican wall, and so on. We also observed the topic transitions over different time-intervals and plotted the occurrence of each type of transition events using a dotted yellow line, as shown in fig. 5. On analysis, we found that president Trump is very vocal on `Twitter` and keeps on discussing new topics which can be observed from fig. 5(a) where number of emerging topics over the time-interval is significantly higher in comparison to his predecessor Obama. However in terms of topics *persistence* and *extinction*, both show nearly identical behavior which is obvious in figs. 5(b) and 5(e). On the other hand, both show extremely contrasting behavior in terms of *divergence* and *convergence* as shown in figs. 5(d) and 5(c). On analysis, it is found that during initial days of his presidency, Trump mainly talked about the topics that were semantically similar and later converged. It is due to the fact that he was posting on Obama policies and how he will change them to frame better policies that were semantically coherent topics. Over the time-intervals, a number of converging topics started decreasing and between the none of the topics converged in last two time-intervals. It is because, as time passed, he started tweeting about diverse topics depending on the current events that were generally incoherent, resulting in decreasing number of converged topics over these time-intervals, as shown using the fig. 5(c). The topic transition events between the topics of first three time-intervals t_1, t_2 and t_3 are visualized using word-cloud and shown in fig. 7. Over the three time-intervals, none of the topics extincted, however 8 new topics emerged during this period, as shown in the fig. 7.

3) TWITTER SOCIALBOT

This section presents an analysis of `Twitter` socialbot profile having "@DharmeshhPawar" handle, which has been suspended. The profile was injected and used by a political party to favor and diffuse the government policies in an

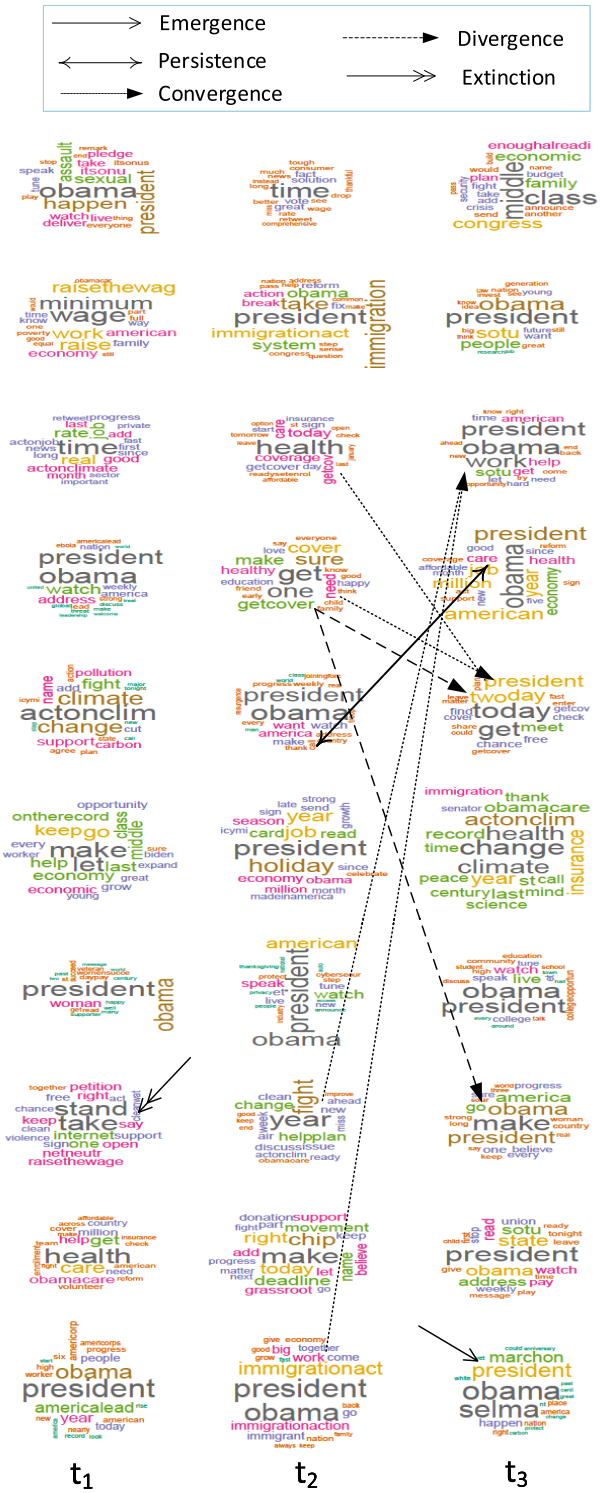

FIGURE 6. Exemplary topic transition events observed from the Twitter timeline of Barack Obama over three time-intervals t_1, t_2, and t_3.

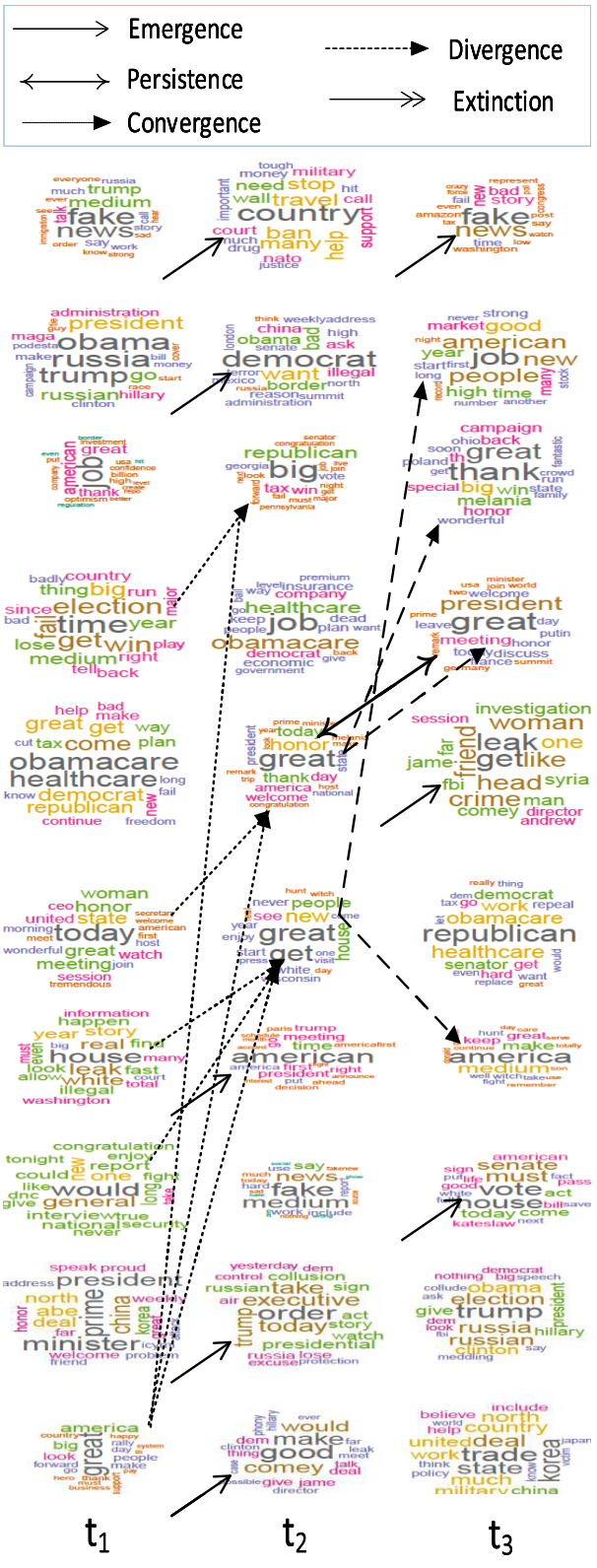

FIGURE 7. Exemplary topic transition events observed from the Twitter timeline of the American president Donald Trump over three time-intervals t_1, t_2, and t_3.

Indian state. We crawled all tweets of the socialbot profile and filtered all non-English tweets, as this profile had posted a number of tweets containing Hindi words. Finally, 1700 tweets were retained for further analysis. Since this profile belongs to a Hindi-speaking country, a number of Hindi words, such as *lathicharge*, *madhya*, etc. were contained in the tweets. Similarly, there were some words that may have different region-specific contexts. Therefore, for

a significant number of words embeddings are either not available in the GloVe or contextually not effective. In order to overcome this problem, we learned word embeddings from the tweets corpus of the whole network of the socialbot which includes approximately 2 million tweets. The embeddings were learned using the word2vec model implemented in gensim[4] toolkit. Thereafter, we divided the 1700 tweets into 10 partitions of chronological order and extracted 10 topics from each partition using LDA. On analysis of the topic transition events over the time-intervals of this profile, it is observed that the similarity between the topics is very low and very few transition events occur over the time-intervals. This is mainly due to the fact that the tweets of the socialbot have very incoherent topics, depending on the events related to the government. Further, all topics in each interval were either emerging or dying in consecutive intervals which can be observed from the transition counts shown using green line in fig. 5 where number of emerging and extinct topics are approximately equal to the number of topics in the interval. On the other hand, persistence, convergence and divergence values are zero over the time-intervals, reflecting the very incoherent and abnormal behavior of the socialbot. All these analytical results and discussions prove that socialbots are injected to defend and propagate the government and their masters' agendas, and accordingly they post as per the requirement. We found that the socialbot profile was continuously posting tweets to praise the government on different issues and events, and to defend the government as and when needed.

C. IMPACT OF PARAMETERS ON TOPIC EVOLUTION

In our proposed approach, efficacy of topic transition mainly depends on two parameters – *proximity threshold* and *word embedding dimension*. A detailed experimental evaluation of the impact of these two parameters on topic transition events is presented in the following sub-sections.

1) IMPACT OF THRESHOLDS

In this section, we analyze the effect of threshold values on transition events. All experimental results presented in section V-B were obtained at the threshold values θ_p, θ_{cd} and θ_{ee} as 0.50, 0.425, and 0.35, respectively. On analysis, we observed that increasing persistency threshold θ_p to 0.6 resulted in the survival of only one topic from Obama's tweets from 8^{th} to 9^{th} time-intervals where he was talking about newly appointed supreme court judges. In contrast, on lowering the persistency threshold to 0.4, a large number of topics started surviving. It may be due to the fact that most of the topics somehow revolve around Obama and his policies and they are moderately coherent. Therefore, choosing θ_p to 0.5 seems empirically logical and valid. On the other hand, when emergence and extinction threshold θ_{ee} was increased to 0.4, the number of emerging and vanishing topics increased rapidly, whereas on decreasing it to 0.3, no topic either emerged or vanished. Therefore, increasing or decreasing the values of θ_{ee} threshold has alarming impact on the transition events. Similarly, we analyzed the impact of convergence and divergence threshold on transition events and found that this threshold has little impact on convergence as compared to divergence.

[4]https://radimrehurek.com/gensim/tutorial.html

2) IMPACT OF EMBEDDING DIMENSION

In this section, we discuss the impact of embedding dimension on the topic proximity and resulting impact on transition events. The implicit proximity calculated in section 3 used 25-dimensional embeddings from GloVe. On analysis, we found that implicit proximity decreases on increasing embedding dimension. For example, cosine similarity between *Obama* and *Republic* using 25-dimensional embedding is 0.569, and it decreases to 0.411 when calculated using 100-dimensional embeddings. It is due to the fact that cosine similarity uses vector dot product in numerator, and scalar product of the square roots of the respective sums of the squared dimensional magnitudes of the vectors in the denominator. Therefore, increase in dimension increases the denominator at larger extent than numerator, resulting in smaller cosine similarity. Although, relative change in proximity value for all topics pairs will be same, the values will be lower and accordingly the proximity thresholds need to be adjusted.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we have proposed a word embedding-based topics evolution modeling approach to analyze user-centric tweets to observe their topical evolution over a period of time. The proposed approach models topics evolution in terms of five transition events – *emergence*, *persistence*, *convergence*, *divergence*, and *extinction*. The transition between topics over different time-intervals is based on proximity between their word distributions, which is determined as a function of explicit and implicit proximities. The explicit proximity is based on the number of words occurring in the word distributions of the topics, whereas implicit proximity finds contextual similarity between the unmatched words of the topics based on their embeddings. The proposed approach models topic transitions between different time-intervals as an m-partite graph in the edges between the partitions represents different types of topic transitions. Extension of the proposed approach to track the evolutionary behavior of different user groups over times seems one of the promising area of research, which could be useful to determine the evolving interests of the users of a particular product, open-source intelligence, and spambots detection.

ACKNOWLEDGMENT

The authors would like to thank the South Asian University (SAU), Delhi, India for the financial support under the start-up research grant provided to the first author of this article.

REFERENCES

[1] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, "Microscopic evolution of social networks," in *Proc. KDD,* Las Vegas, NV, USA, 2008, pp. 462–470.

[2] Z. Yang, J. Xue, C. Wilson, B. Y. Zhao, and Y. Dai, "Process-driven analysis of dynamics in online social interactions," in *Proc. COSN*, Palo Alto, CA, USA, 2015, pp. 139–149.

[3] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. ICML*, Pittsburgh, PA, USA, 2006, pp. 113–120.

[4] S. Y. Bhat and M. Abulaish, "HOCTracker: Tracking the evolution of hierarchical and overlapping communities in dynamic social networks," *IEEE Trans. Know. Data Eng.*, vol. 27, no. 4, pp. 1013–1019, Apr. 2014.

[5] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, "Finding statistically significant communities in networks," *PLoS ONE*, vol. 6, no. 4, pp. 1–18, Apr. 2011.

[6] M. Spiliopoulou, I. Ntoutsi, Y. Theodoridis, and R. Schult, "MONIC: Modeling and monitoring cluster transitions," in *Proc. KDD*, Philadelphia, PA, USA, 2006, pp. 706–711.

[7] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: An exploration of temporal text mining," in *Proc. KDD,* Chicago, IL, USA, 2005, pp. 198–207.

[8] C. Zhai, A. Velivelli, and B. Yu, "A cross-collection mixture model for comparative text mining," in *Proc. KDD*, Seattle, WA, USA, 2004, pp. 743–748.

[9] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[10] C. Lauschke and E. Ntoutsi, "Monitoring user evolution in Twitter," in *Proc. ASONAM,* Istanbul, Turkey, Aug. 2012, pp. 1972–1977.

[11] L. D. Caro, M. Guerzoni, M. Nuccio, and G. Siragusa, "A bimodal network approach to model topic dynamic," in *Proc. STI*, Leiden, The Netherlands, Sep. 2018, pp. 486–491.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. NIPS*, Stateline, NV, USA, 2013, pp. 3111–3119.

[14] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors forword representation," in *Proc. EMNLP*, Doha, Qatar, 2014, pp. 1532–1543.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean. (2013). "Efficient estimation of word representations in vector space." [Online]. Available: https://arxiv.org/abs/1301.3781

[16] D. Q. Nguyen. (2018). "jLDADMM: A Java package for the LDA and DMM topic models." [Online]. Available: https://arxiv.org/abs/1808.03835

MUHAMMAD ABULAISH (SM'12) received the Ph.D. degree in computer science from the Indian Institute of Technology Delhi in 2007. He is currently an Associate Professor with the Department of Computer Science, South Asian University, New Delhi. His research interests span over the areas of data analytics and mining, social computing, and data-driven cyber security. He has published over 84 research papers in reputed journals and conference proceedings, including three papers in the IEEE Transactions. He is a Senior Member of the ACM and CSI.

MOHD FAZIL received the master's degree in computer science and application from Aligarh Muslim University, Aligarh, India, in 2013. He is currently pursuing the Ph.D. degree in computer science with Jamia Millia Islamia (A Central University), New Delhi. He has qualified the UGC-JRF Exam in 2013. His research interests include data mining, data-driven cyber security, social network analysis, and machine learning. He is currently a recipient of the UGC Senior Research Fellowship.