Received: 10 January 2022 | Revised: 1 July 2022 | Accepted: 26 August 2022

DOI: 10.1111/exsy.13195

ORIGINAL ARTICLE

Expert Systems WILEY

Social network analytics and visualization: Dynamic topic-based influence analysis in evolving micro-blogs

Shazia Tabassum[1,2] | João Gama[1,2] | Paulo J. Azevedo[1,3] | Mario Cordeiro[1] | Carlos Martins[4] | Andre Martins[4]

[1]LIAAD, INESC TEC, Porto, Portugal

[2]Faculdade de Engenharia, Universidade do Porto, Porto, Portugal

[3]Departamento de Informática, Universidade do Minho, Braga, Portugal

[4]Skorr, Lisbon, Portugal

Correspondence
Shazia Tabassum, LIAAD, INESC TEC, Rua Dr. Roberto Frias, Porto, Portugal.
Email: shazia.tabassum@inesctec.pt

Funding information
Fundação para a Ciência e a Tecnologia, Grant/Award Number: UIDB/50014/2020

Abstract

Influence Analysis is one of the well-known areas of Social Network Analysis. However, discovering influencers from micro-blog networks based on topics has gained recent popularity due to its specificity. Besides, these data networks are massive, continuous and evolving. Therefore, to address the above challenges we propose a dynamic framework for topic modelling and identifying influencers in the same process. It incorporates dynamic sampling, community detection and network statistics over graph data stream from a social media activity management application. Further, we compare the graph measures against each other empirically and observe that there is no evidence of correlation between the sets of users having large number of friends and the users whose posts achieve high acceptance (i.e., highly liked, commented and shared posts). Therefore, we propose a novel approach that incorporates a user's reachability and also acceptability by other users. Consequently, we improve on graph metrics by including a dynamic acceptance score (integrating content quality with network structure) for ranking influencers in micro-blogs. Additionally, we analysed the topic clusters' structure and quality with empirical experiments and visualization.

KEYWORDS
dynamic topic modelling, micro-blogs, social network analysis, topic-specific influence analysis, visualization

1 | INTRODUCTION

Micro-blogging services such as Twitter, Instagram, Facebook have emerged as prominent platforms for interaction, information sharing and networking between users. The analysis of structural inter-dependencies in such network of users gives meaningful insights about their interests, preferences and behavioural characteristics (Tabassum et al., 2018). Influence analysis is one such task, which aims at the study of how the nodes in the network are influenced by each other. Precisely, the main purpose of it is to identify influencers for tracking spread or information dissemination. Therefore, it is critical for many applications such as, decision support and social recommender systems (Hajian & White, 2011), information diffusion (Bhattacharya & Sarkar, 2021), sentiment analysis (Chouchani & Abed, 2020), viral marketing (Kaple et al., 2017; Mostafa, 2021), politics and elections (Suau-Gomila et al., 2020), key players in vaccination discourses (Sanawi et al., 2017) using microblogging or social networking apps. Influencers are many a time also referred by other labels such as opinion leaders, dominant nodes, key players, topic experts, prestige/authority nodes or referrers.

As we have known the applicability of influence analysis in multifarious domains/topics from business to politics, it is obvious that the authoritative users in one topic need not be the leaders in all topics. As indicated in (Yu et al., 2016), one cannot have interest/expertise in all the topics.

Therefore, unlike most of the earlier researches that ignored the post content and discovered influencers not specific to topics, our method detects topics and influencers in them by applying various social network analysis (SNA) techniques. However, Influence computation is very costly, technically #P-hard under most influence models. Consequently, most existing studies have to compromise and consider it only on a static network (Yang et al., 2017). Therefore, we read the data as a stream of posts and compute influence score incrementally with a fading factor. This helps in tracking distinct influencers over time. Additionally, the process involves a choice of dynamic sampling strategies for scalability and capturing the changing trends in topic formation.

Inferring topics from unstructured data has been quite a challenging task. Typically, the data gathered from micro-blogs is in the form of posts generated by the users. Posts can be short texts, images, videos, messy data such as concatenated words, URLs, misspelled words, acronyms, slangs and more. Classification of posts into topics is a complex problem. While topic modelling algorithms such as Latent Semantic Analysis and Latent Dirichlet Allocation (LDA) are originally designed to derive topics from large documents such as articles, and books. They are often less efficient when applied to short text content like posts (Alash & Al-Sultany, 2020). Posts on the other hand are associated with rich user-generated hashtags to identify their content, to appear in search results and to enhance connectivity to the same topic. In (Wang et al., 2016) the authors state that hashtags provide a crowd sourcing way for tagging short texts, which is usually ignored by Bayesian statistics and Machine learning methods. Therefore, in this work, we propose to use these hashtags to derive topics using SNA methods, mainly community detection. Further, we analysed the topic clusters' structure and quality using empirical experiments. The results unveil latent semantic relations between hashtags and also show frequent hashtags in a cluster. Moreover, in this approach, the words in different languages are treated synonymously. Besides, we also observed top trending topics and correlated clusters.

Social network visualization techniques enhance the comprehensibility and explainability of analytics and results (Tabassum, 2020). Further, Sanawi et al. (2017) presented visualization of vaccination discourse network from twitter and reveal the identity characteristics of users with high SNA measures. Al-Shargabi and Selmi (2021) carried out explicit visualizations of SNA to portray dominant nodes. Similarly, we made use of visualization techniques to examine and illustrate results. Moreover to our knowledge, this work is first of a kind to use SNA and associated visualization layouts for topic modelling and illustrating topic clusters. This approach presents an overall view of network samples unlike the conventional way of using tables with limited capacity and missing structural patterns. Consequently, we tried to address the issues discussed above by proposing a framework with the contributions stated below:

1. We propose a scalable and incremental method using social network analytics for identifying influencers over time. Moreover, social network visualization techniques are applied to enhance comprehensibility.
2. Dynamic sampling mechanisms are employed to decrease space complexity and capture changing trends.
3. Multiple graph measures are compared, their correlations studied and new insights drawn. Additionally, we propose novel measures combining the content of post and structural relations.
4. Our topic model categorizes tags/words based on connectivity and modularity. In this way, the tags/words are grouped accurately even though they belong to different languages or if new hashtags appear.

Rest of the paper is organized as follows: In Section 2, we have presented a brief overview of the related works. Section 3 describes the data set and some statistics about it. The framework is described in Section 4. Methods are detailed in sections 5 and 6. The experiments and results are discussed in Section 7. Finally, Section 8 summarizes conclusions and some potential future works.

2 | RELATED WORK

With the prevalence of micro-blogging applications Influence Analysis has been an active area of research over the past decade. A recent and thorough survey on finding influencers is given by Ishfaq et al. (2022). To facilitate background understanding we classify below the literature according to their relevancy within this work.

2.1 | Social network analytics for influence analysis

Analysing the patterns of interplay between connections or relationships is paramount for social networks. Therefore, we excluded the works ignoring social elements. However, we briefed them in the section below. Nevertheless, a number of studies have incorporated SNA methods in the context of identifying influential users. The popular one among them is PageRank (Page et al., 1999) and its variants. Hajian and White (2011) obtained an alternative called Influence rank based on PageRank, which is calculated from a combination of user's follow, like, comment, and retweet activities. Kaple et al. (2017) tried to find influencers from people of common interest by implementing community detection and then applying PageRank over a small subset of Facebook users. Mao and Zhang (2016) introduced PageRank for Micro-blogs (PG4MB), combined with

measures incorporating number of users, blogs and outbound users in a small twitter network. However, on comparison with PageRank the authors did not prove which one is better rather discussed the differences and reasons, as both the results were quite different. Some of their top influencers with PageRank were Bill Clinton, Bill Gates, Barack Obama, while the top influencers with PG4MB were news channels such as Huffington Post, BBC News, The New York Times. Therefore, the choice of algorithm is sometimes application specific. Other variant included signed-PageRank (Yin et al., 2021). However, other measures such as Degree, Closeness, Betweenness, Eigen Vector centralities have also been utilized extensively to rank influencers (Jianqiang et al., 2017; Mostafa, 2021; Sanawi et al., 2017). Recently, Corradini et al. (2021) implemented a social network model based on homophily and betweenness for finding negative influencers. Typically, the above works calculate the overall influence of nodes considering a global picture of the network. However, there is a small subset of studies focused on topic-oriented influence models considering graphs and social structures, which are listed below.

2.2 | Topical influence analysis

Recognizing the importance of topical models, recent researches in Online Social Networks have shifted their focus towards them. Similar to our approach, Xiao et al. (2014) used hashtags to define user communities. Otherwise, they did not detect topics from hashtags but they collected specific topic tweets and retrieved their hashtags. Likewise other researches based on twitter data, their work proposed two ranks called RetweetRank and MentionRank. These ranks assume a user to have a high influence if he gets retweeted or mentioned from other users and higher if those users are influential themselves. Two accessors were asked to judge the efficiency. However, the study was focused only on the news topic. Alp and Öğüdücü (2018) found topics from tweets using a tailored LDA version. Furthermore, they presented personalized PageRank (PPR) that uses information obtained from network topology and also user actions and activities for detecting topical influencers. The evaluation or information diffusion was estimated with a measure 'spread score' as they named it. Spread score was calculated by normalizing the number of retweets for a user over the number of tweets of that user. However, the credibility of the score was not established. Moreover, as the results of PPR lead to a good spread score, the more trivial spread score could replace the sophisticated PPR for finding influencers. Another interesting work (Jain & Sinha, 2020) applies a bunch of centralities over user-tweet graph and user-user graph for creating a normalized feature score matrix. These feature scores are multiplied by their relative impact weights to obtain Weighted Correlated Influence measure for each user in two specific topics. Most of the topic-aware models discussed above explore one or two pre-selected topics, unlike our method that discovers topic clusters and finds influencers in any topic. Rest of the topical models do not demonstrate scalability that is essential for micro-blogs data, which is typically very large. Interestingly enough Bi et al. (2014) demonstrated scalability on large clusters by presenting distributed Gibbs sampling algorithm for Followship-LDA (FLDA). FLDA is a Bernoulli-Multinomial mixture model that jointly models both content topic discovery and social influence analysis in the same generative process. Consequently, they proposed a general search framework where a user types a set of keywords and gets an ordered list of key influencers by their influence scores accordingly. Similarly, Yu et al. (2016) adopted distributed Gibbs sampling for scalability and Link-LDA (Erosheva et al., 2004) with users topic distribution and followee distribution jointly. Additionally, a similarity-based weight scheme was incorporated to eliminate bias. Furthermore, they demonstrated better results compared to the above FLDA over the data sets from Sina Weibo and Tencent Weibo. However, most of the topic-based models that are listed above, manually choose one or few topics to obtain data and run influence analysis. Besides, the models that pursued topic detection have considered influencers as a static set or bear high computational complexities. Besides, LDA-based models use a predefined set of topics, which contradicts the evolving nature of a real-time streaming microblog data.

2.3 | Topic modelling

Research works focusing on topic modelling are mostly based on inferring abstract topics from long text documents. Latent dirichlet allocation (Blei et al., 2003) is one of the most popular techniques used for topic modelling where the topic probabilities provide an explicit representation of a document. However, it assumes fixed number of topics that a document belongs to. Other well-known models include Latent semantic analysis (Deerwester et al., 1990), Correlated topic models (Blei & Lafferty, 2006), Probabilistic latent semantic indexing (Hofmann, 1999). Word2Vec (Mikolov et al., 2013) is another popular word representation techniques. This model outputs a vector for each word, so it is necessary to combine those vectors to retrieve only one representation per product, title or post, since there is the need to have the entire sentence representation and not only the values of each word. Word2Vec output dimensions can be configurable, and there is no ideal number for it since it can depend on the application and the tasks being performed. Moreover, these types of models are very common and can be expensive to train. However, traditional topic models also known as flat text models are incapable of modelling short texts or posts due to the severe sparseness, noise and unstructured data (Hong & Davison, 2010; Wang et al., 2016).

Recently, several researchers have focused on specifically hashtags clustering. In (Muntean et al., 2012) the authors clustered hashtags using K-means on map reduce to find the structure and meaning in Twitter hashtags. Their study was limited to understanding the top few hashtags

from three clusters. They found the top hashtags to be understandable as they are popular and while increasing the number of clusters the hashtags are dispersed into more specific topics. In another interesting work, multi-view clustering was used to analyse the temporal trends in Twitter hashtags during the Covid-19 pandemic (Cruickshank & Carley, 2020). The authors found that some topic clusters shift over the course of pandemic while others are persistent. Topic modelling was also applied on Instagram hashtags for annotating images (Argyrou et al., 2018). In (Bhakdisuparit & Fujino, 2018) the authors clustered twitter hashtags into several groups according to their word distributions. The model was expensive as Jensen-Shannon divergence was calculated between any two hashtags from the data. However, they considered a very small data set and calculated the probabilities for top 20 frequent hashtags while the structure and quality of clusters was not analysed.

While most of the models above were run on small-scale data sets crawled from one of the social media applications, we used a considerably large one which is composed of data from several micro-blogging applications and also visualized the quality and structure of our clusters.

3 | CASE STUDY

An anonymized data set is collected from a social media activity management application. The data set ranges from January to May 2020; comprises 1,002,440 posts with 124,615 hashtags posted by users on different social networking platforms (Twitter, Facebook, Instagram, Google). Besides hashtags, the given posts are also associated with number of likes, shares and comments. Additionally, they include 110,909 friendship links between 8451 users who posted (Figure 1). The degree distribution of Friendship network can be seen in Figure A1. The content of posts is not available but the posts are identified with posts IDs and the users are identified with anonymous user IDs. Figure 3 displays the distribution of hashtags vs posts. A few hashtags are used by large number of posts and many different hashtags are discussed by only some users. This satisfies a power law relation which is usually seen in most of the real world social networks (Tabassum et al., 2018). Each post can include one or more hashtags or none. The number of posts per day is given in Figure 2, which shows the seasonality of data. As one can observe there is decreased activity on weekends (Saturday and Sunday) compared to other days with the peaks on Fridays. The data in the last week of April had not been available which can be seen as an inconsistency in the curve with abnormally low activity close to zero. Figure 4 displays the top 10 trending hashtags in the given data set. This type of analysis with the help of topic modelling or trending hashtags can be used to detect events. In the figure, we can see the top two of frequent hashtags are relating to Covid19. What we need from our model is to cluster these hashtags and also the one is that are less frequent (such as Covid, Covid19, corona etc.) to be classified as one topic relating to Covid19. Similarly, with the other tags and their related posts. To achieve this we follow the methodology briefed below.

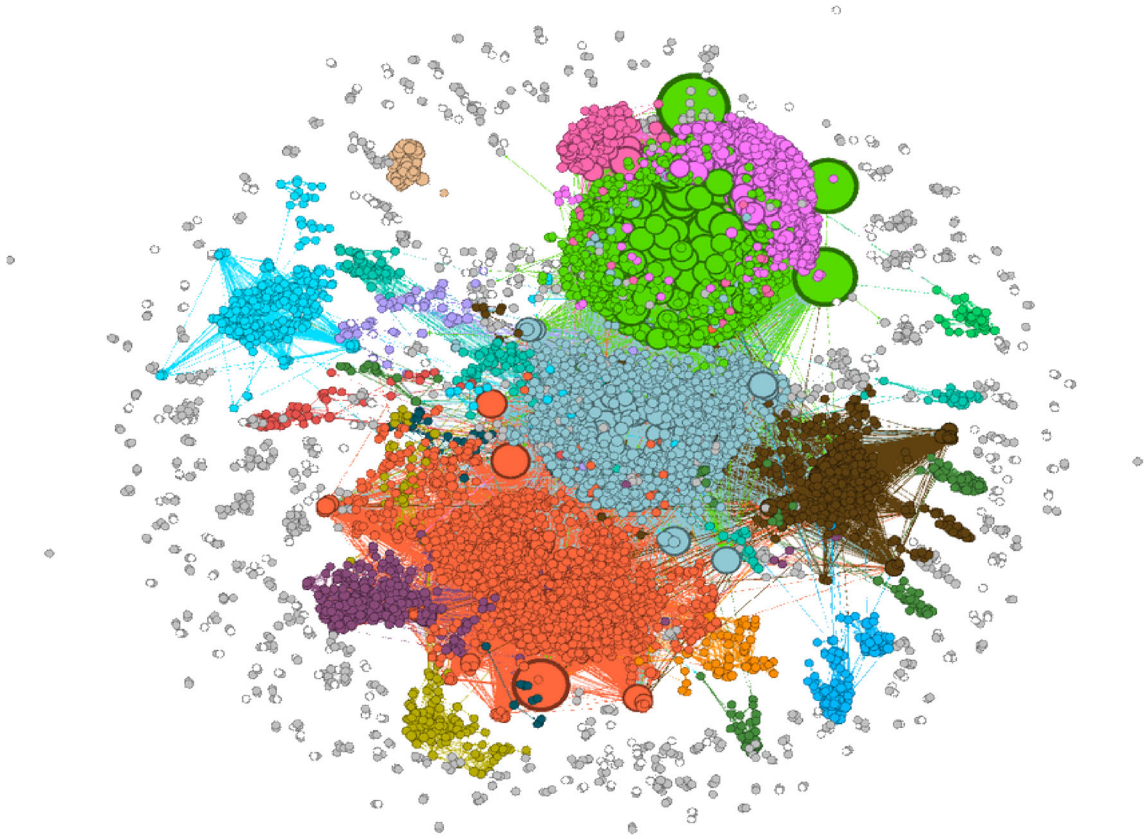


FIGURE 1 Representation of friendship links in the data. Colours represent communities

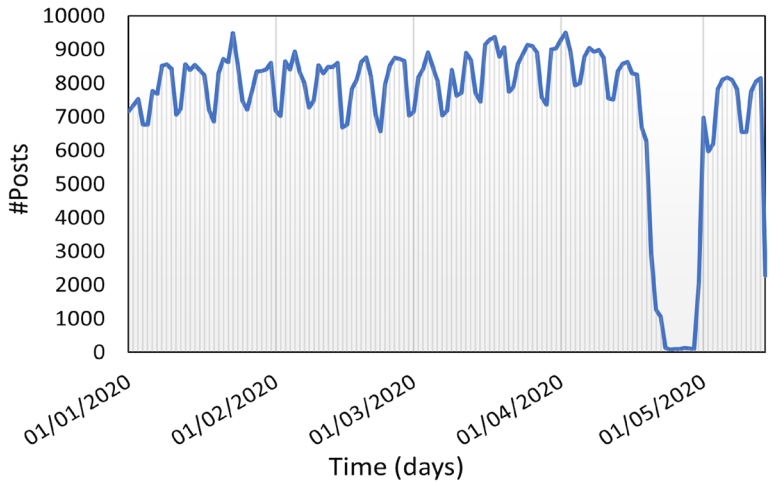


FIGURE 2 Temporal distribution of posts per day

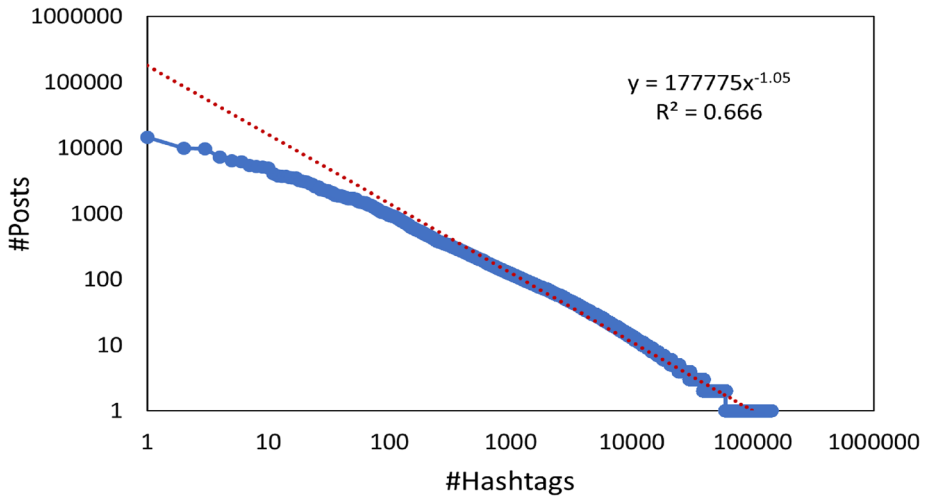


FIGURE 3 Posts versus hashtags distribution (blue line). Power curve following given function (red)

4 | METHODOLOGY: FRAMEWORK

In order to discover topics and topic-based influencers we build three networks derived from micro-blogs data incrementally. An illustration of framework is shown in Figure 5. From a given stream of posts $\{p_1, p_2, p_3...\}$ arriving in the order of time we start by building a co-occurrence network from the streaming hashtags incrementally. Explicitly, we create an edge e between the hashtags h_i and h_j that have been tagged together in a post. Therefore $e = (h_i, h_j, t)$ where $i, j \in \mathbb{N}$ and t is the time stamp when it occurred. Similarly, a Like, Share and Comment Network between users connected with weighted links based on likes, shares and comments is also derived. However, with the limited data we have it forms a one level disconnected graph, which limits us from applying centralities except degree centrality as in Section 6.

The method is space efficient as all the data from posts is not stored in the memory. It stores the results, that is, the influence score of users per topic T at any time t and increments it. The hashtags that need to stay in the network are decided based on the choice of the sampling algorithm in Section 5. Thereafter the communities are detected in the network as detailed in Section 5.1.

To handle the Friendship Network there are two alternatives; First, one is to store the centralities per user rather than the whole network and re-calculate. This is a viable option as the Friendship Network is a very slow evolving graph. Second is to store the network and calculate the centralities for users discussing same topic rather than entire network. In this case in our data, the nodes discussing same topic are often disconnected (direct link do not exist between them). Therefore, we consider the nodes discussing the topic and their neighbours to form a network for information dissemination.

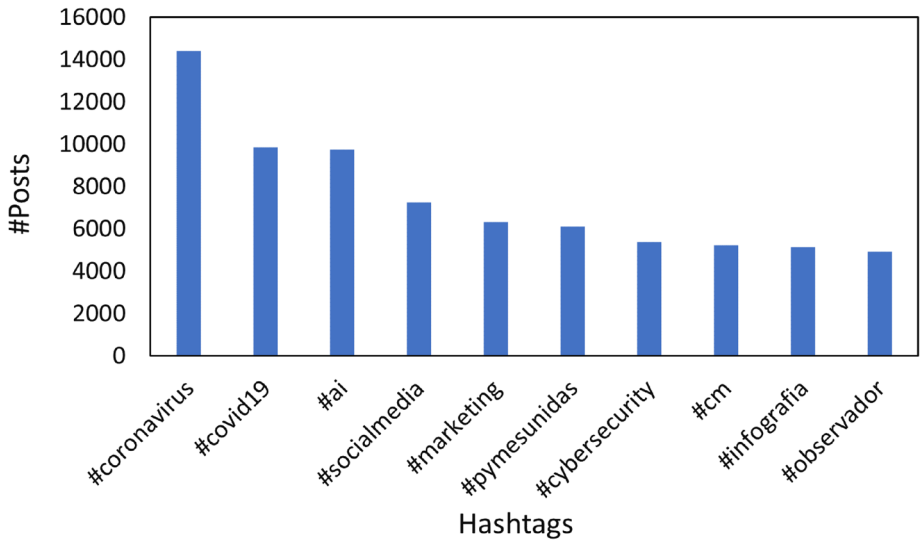


FIGURE 4 Top 10 trending hashtags distribution

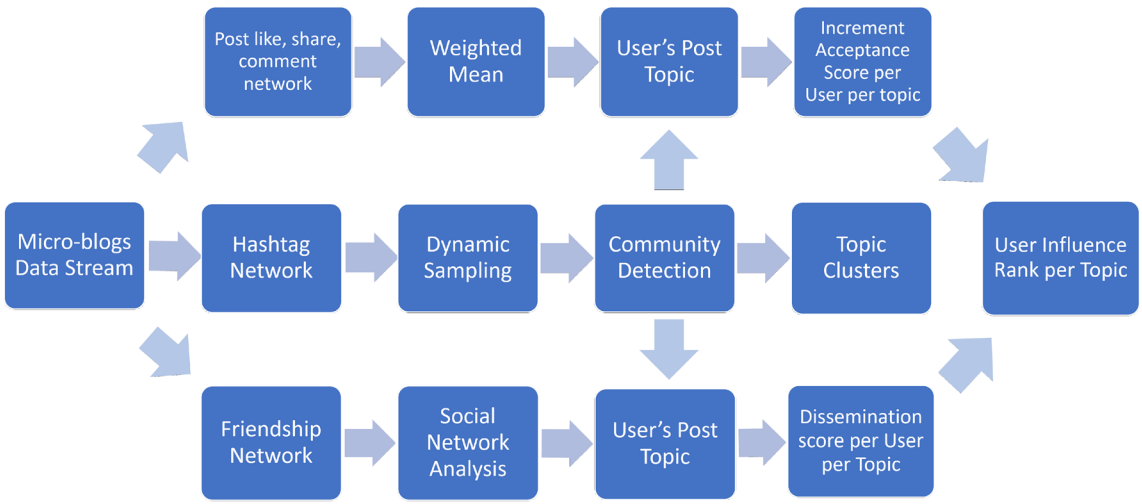


FIGURE 5 Framework for topic-based influence analysis

5 | STREAM SAMPLING

As posts are temporal in nature generating in every time instance, so are the hashtags. Also, there are new hashtags emerging over time. Moreover, the context for grouping hashtags may change over time. For example, hand sanitizers and face masks were not as closely related as with the onset of Covid19. Therefore, we employed the approach of exploiting the relation between hashtags based on the recent events or popular events by using the real-time dynamic sampling techniques below.

5.1 | Sliding windows

Sometimes applications need recent information and its value diminishes by time. In that case, sliding windows continuously maintain a window size of recent information (Gama, 2010). It is a common approach in data streams where an item at index i enters the window while another item at index $i - w$ exits it. Where w is the window size, which can be fixed or adaptive. The window size can be based on number of observations or length of time. In the later, case an edge (h_i, h_j, t) enters window while an edge $(h_i, h_j, t - w)$ exits.

5.2 | Space saving

The Space Saving Algorithm (Metwally et al., 2005) is the most approximate and efficient algorithm for finding the top frequent items from a data stream. The algorithm maintains the partial interest of information as it monitors only a subset of items from the stream. It maintains counters for every item in the sample and increments its count when the item re-occurs in the stream. If a new item is encountered in the stream, it replaces the item with the least counter. The new item counting is assigned the value of the least counter plus 1.

5.3 | Biased random sampling

This algorithm (Tabassum & Gama, 2016) ensures every incoming item m in stream goes into the reservoir with probability 1. Any item n from the reservoir is chosen for replacement at random. Therefore, on every item insertion, the probability of removal for the items in the reservoir is $1/k$, where k is the size of reservoir. Hence, the item insertion is deterministic but deletion is probabilistic. The probability of n staying in the reservoir when m arrives is given by $(1-1/k)^{(t_m-t_n)}$. As the time of occurrence or index of m increases, the probability of item n from time t_n staying in reservoir decreases. Thus, the item staying for a long time in the reservoir has an exponentially greater probability of getting out than an item inserted recently. Consequently, the items in the reservoir are super linearly biased to the latest time. This is a notable property of this algorithm as it does not have to store the ordering or indexing information as in sliding windows. It is a simple algorithm with $O(1)$ computational complexity.

5.4 | Community detection

Community detection is very well-known problem in social networks. Communities can be defined as groups, modules or clusters of nodes, which are densely connected between themselves and sparsely connected to the rest of the network. The connections can be directed, undirected, weighted, and so on. Communities can be overlapping (where a node belongs to more than one community) or distinct. Community detection is in its essence a clustering problem. Thus, detecting communities reduces to a problem of clustering data points. It has a wide scope of applicability in real-world networks.

In this work, we applied the community detection algorithm proposed by Blondel et al. (2008) on every dynamic sample snapshot discretely as depicted in Figures 6, 7 and 8. The resultant graphs are analysed in Section 7 in detail. However, an incremental community detection algorithm can also be applied on every incoming edge. Nevertheless, the technique mentioned above is a heuristic based on modularity optimization. *Modularity* is a function that can be defined as the number of edges within communities minus the number of expected edges in the same at random (Newman, 2006) as computed below.

$$Q = \frac{1}{2m}\sum\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j), \tag{1}$$

where m is the number of edges, k_i and k_j represent, respectively, the degree of nodes i and j, A_{ij} is the entry of the adjacency matrix that gives the number of edges between nodes i and j, $\frac{k_i k_j}{2m}$ represents the expected number of edges falling between those nodes, c_i and c_j denote the groups to which nodes i and j belong, and $\delta(c_i, c_j)$ represents the Kronecker delta. Maximizing this function leads to communities with highly connected nodes between themselves than to the rest of the network. However, in very large networks the connections are very sparse and even a single edge between two clusters is regarded as strong correlation. Therefore, a resolution parameter is used to control high or low number of communities to be detected. Modularity is also used as a quality metric as shown in Table 1.

The above said algorithm has a fastest runtime of $O(n.log_2 n)$, where n is the number of nodes in the network. In our case, n is very small compared to the total number of nodes in the network, for instance n is equal to the number of hashtags in a sliding window.

6 | NETWORK ANALYSIS

The users posting about the topics are connected to each other via friendship links. Analysing the graph structure of Friendship Network could reveal the trust, information dissemination and influence patterns. The below graph metrics are implemented to rank the users based on their scores which represents the reachability, popularity and connectedness of users. It is explained in the subsection below as Dissemination scores.

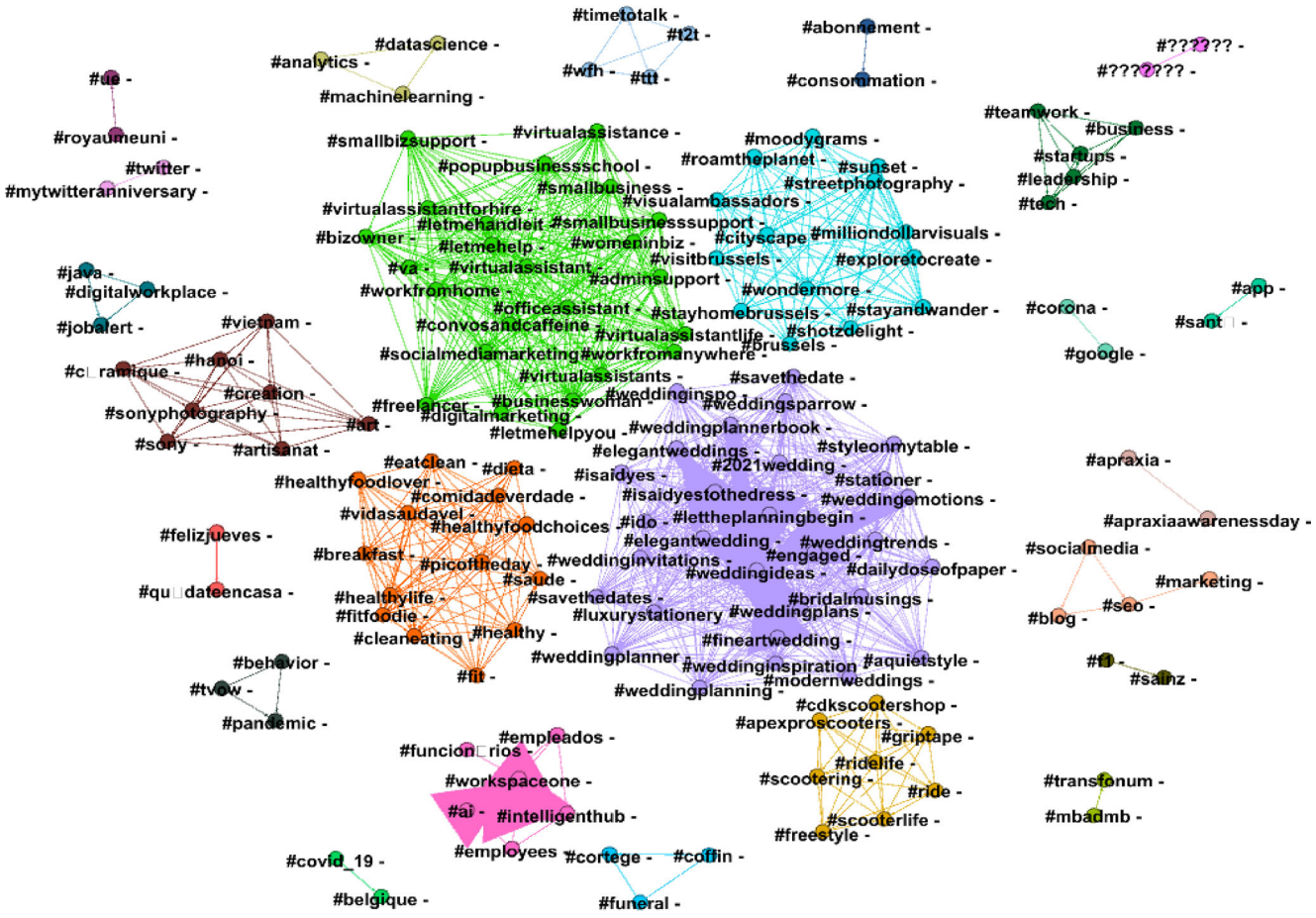


FIGURE 6 Sliding window

6.1 | Dissemination scores

6.1.1 | Degree centrality

The degree centrality of a node is the number of nodes directly connected to it or the number of edges incident on it. In the case of recurring links between two nodes the number of unique edges are considered. In the case of digraphs, it can be categorized into two types based on the direction of edges. It is also regarded as a central and important measure in network analysis. In social networks, it portrays the importance of nodes with number of friends, posts or followers/followees. The direction of the links also distinguishes the importance. It is usually denoted as $deg(v)$ or $C_D(v)$ for a degree centrality of a vertex v. The space complexity is $O(n)$, where n is the number of nodes or users being observed. The time complexity is $O(n)$ and in the streaming incremental algorithm $O(1)$ per edge for the addition operation.

6.1.2 | Betweenness centrality

The betweenness centrality (C_B) of a node is the number of shortest paths that pass through it. Where a path is a sequence of connected edges between any two nodes in the network. It can be expressed as in Equation (2).

$$C_B(v) = \sum_{i,j \in V(G) \setminus i} \frac{\sigma_{ij}(v)}{\sigma_{ij}} \tag{2}$$

where σ_{ij} denotes the number of shortest paths between vertices i and j (usually $\sigma_{ij} = 1$) and $\sigma_{ij}(v)$ expresses the number of shortest paths passing through node v. High betweenness nodes act as a gateway to pass or block information between different communities in the network. Hinz et al. (2011) found high betweenness and degree centrality nodes yielding better results in viral marketing when chosen as seed points.

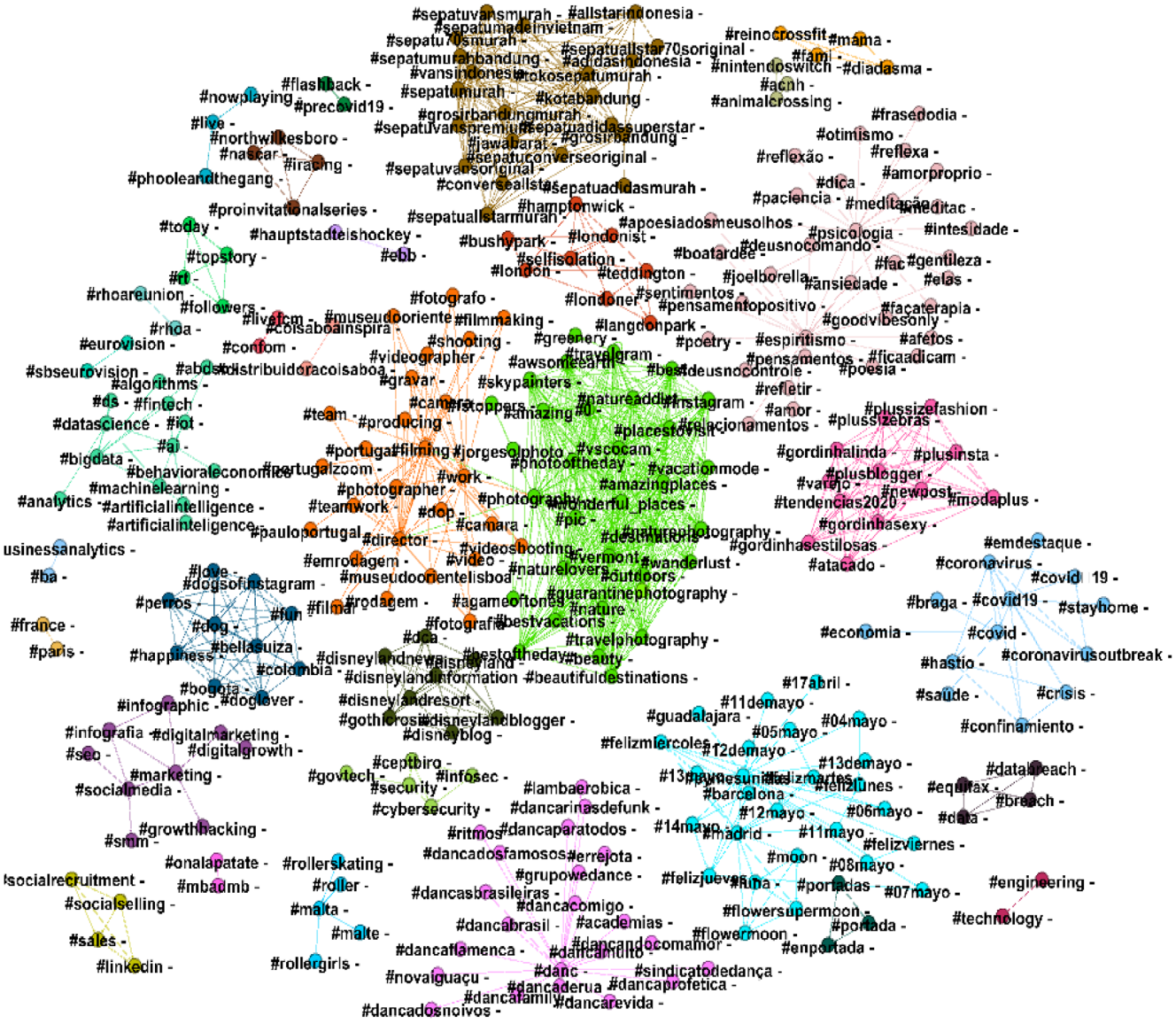


FIGURE 7 Space saving

Betweenness requires computing shortest paths between all the nodes, which increases the space complexity to $O(n + m)$ with Brandes algorithm (Brandes, 2001) and $O(nm)$ time. Where n and m denote the number of nodes and edges.

6.1.3 | Closeness centrality

The reciprocal of the average length of shortest paths between a node and all the other nodes in the graph is said to be its closeness centrality (C_C). It signifies how close a focal node is to all the other nodes in the network. The nodes with high closeness centrality are able to reach all the nodes in the network easily relative to the other nodes in the network. In the networks with a number of components, the closeness can be computed on the nodes within the components. Since the closeness of nodes across components will be ∞. The equation is given below:

$$C_C(v) = \frac{n-1}{\sum_{u \in V(G)\setminus v} d(u, v)}. \quad (3)$$

where n is the number of nodes in the graphs. Bond III et al. (2004) found closeness centrality to be related with reputational effectiveness in an organization. As closeness centrality is also a path-based algorithm it follows the same complexities as the Betweenness above.

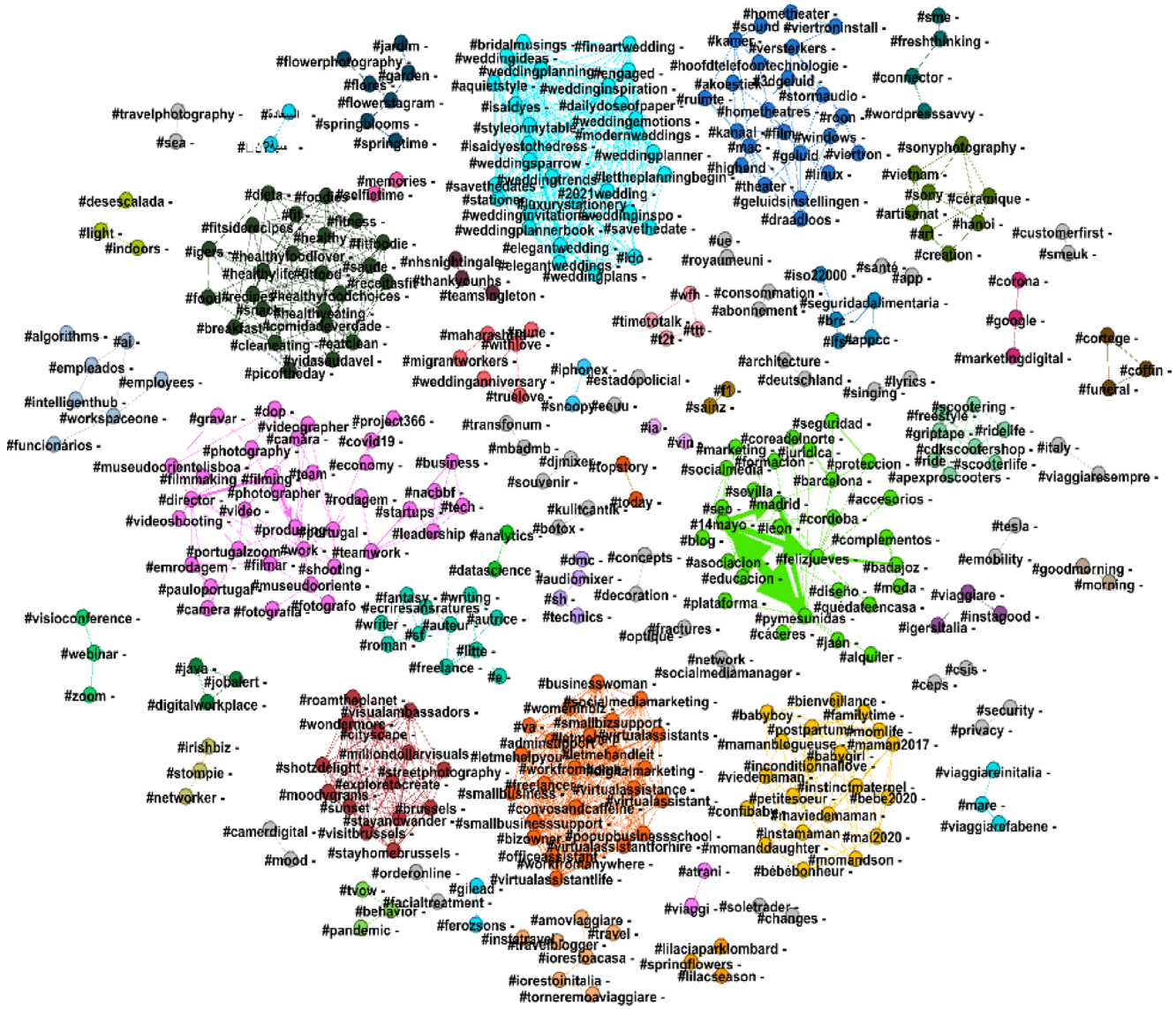


FIGURE 8 Biased random sampling

6.1.4 | PageRank

PageRank (Page et al., 1999) is a link analysis algorithm used by Google to rank webpages. The idea is to count the number of links received from high quality nodes. The more links received from high quality nodes, the higher is its PageRank. The quality of nodes depend on the number of links they receive. It requires to store the whole graph with complexity $O(n^2)$ and time $O(mnk)$, where k is the number of iterations.

6.1.5 | Laplacian centrality

This measure is a two hop degree centrality. In other words, Laplacian centrality of a node is its degree plus the degree of its neighbours (adjacent nodes). To estimate the centrality of a vertex, it considers not only the first order connections but also the importance of their neighbours. The results and related calculations describe the so-called Laplacian energy of the network. Therefore, this strategy is known as the Laplacian centrality.

The reader can find a detailed demonstration of Laplcian centrality in Cordeiro et al. (2018). It requires a space of $O(n^2)$ and time of $O(nd^2)$, where d is the maximum degree of a node in graph.

6.1.6 | Eigenvector centrality

It is considered as a measure of the influence of a node in a network. Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores. It is similar to pageRank which is a left hand eigenvector and also the number of iterations in pageRank depends on a damping factor. Its space complexity is $O(n^2)$ and time $O(n^3)$ using matrices.

6.1.7 | Clustering coefficient

Clustering coefficient of a node is the extent to which its neighbours are connected. It is the proportion of links between a node's neighbours to the total number of links that can exist between them. Average clustering coefficient is a global measure for the whole network given by averaging the clustering coefficients of all the nodes in a network.

Watts and Strogatz (1998) proposed a local version of the clustering coefficient, denoted c_i where $i = (1, ..., n)$. In this context, transitivity is a local property of a node's neighbourhood that indicates the level of cohesion between the neighbours of a node. This coefficient is, therefore, given by the fraction of pairs of nodes, which are neighbours of a given node that are connected to each other by edges (see Equation 4).

$$C_i = \frac{2\,|\,e_{jk}\,|}{k_i(k_i-1)} : v_j, v_k \in N_i, e_{jk} \in E \tag{4}$$

where N_i is the neighbourhood of node v_i, e_{jk} represents the edge that connects node v_j to node v_k, k_i is the degree of node v_i, and $|e_{jk|}$ indicates the proportion of links between the nodes within the neighbourhood of node v_i. The space required for computing clustering coefficient of all the nodes in a network is $O(n^2)$ and time complexity is $O(nm)$.

6.2 | Acceptance score

The user may share the information/posts with other users as simulated by the metrics above but it may or may not be accepted, endorsed or forwarded by them. Considering friend links alone does not reflect the user's information sharing ability. Therefore, we propose an acceptance score that indicates the number of times a user's post is accepted by other users in the form of likes, shares and comments.

In other words, the Acceptance score of a user per topic is given as a weighted average of the number of users who like, share and comment on his/her posts on a particular topic. The weights are determined by the scaling factor. As the range of values for the given three variables is dissimilar, the values are multiplied with a weight, which is the scaling factor of that variable. The scaling factor is determined by simply dividing the mean of variables with the highest mean among them.

Over time over a number of posts this value gets decremented with a fading factor. Which makes it capable of capturing time varying influence. The results given (Section 7) are achieved in the end of file.

$$Q_t = \mu_t + (\alpha - 1)Q_{t-1} \tag{5}$$

where Q_t is the Acceptance score of a user u at t. $\mu_t = \sum_{i=1}^{z} w_i v_i^u$ is the weighted mean of the values v_i from variables #likes, #shares, #comments or more per u at time t in a specific topic T and z is the number of variables.

It requires $O(n)$ space, which is equal to the output size and a time complexity of $O(n)$. In a streaming scenario $O(1)$ per edge for the addition operation.

6.3 | Influence rank

We get the influence score by combining the scores from measures that represent information dissemination and the measure that reflects the acceptance or recognition of spread (Section 6). Precisely, it is an average of Dissemination and Acceptance scores (detailed above) of a user per topic. In other words, each of the dissemination scores given above are upgraded with the acceptance score. Therefore, we call them as improved scores and present as Degree+, Betweenness+, Closeness+, PageRank+, Laplacian+, EigenVector+ and Clustering Coefficient+ .

7 | EXPERIMENTAL EVALUATION

The experiments are conducted to evaluate the topic model and identify topic-specific Influencers from the data detailed in Section 3. Figure 9 presents the topics and influential nodes from them as an example illustration. Similarly, Figures 10 and 11 demonstrate an instance of results using Degree centrality. The graphical representations of influential nodes and their dissemination/spread (one hop) in other topics is displayed in Figure B2. To facilitate visual evaluation and demonstration, the size of samples is fixed to be 1000 edges. In the case of sliding windows, the

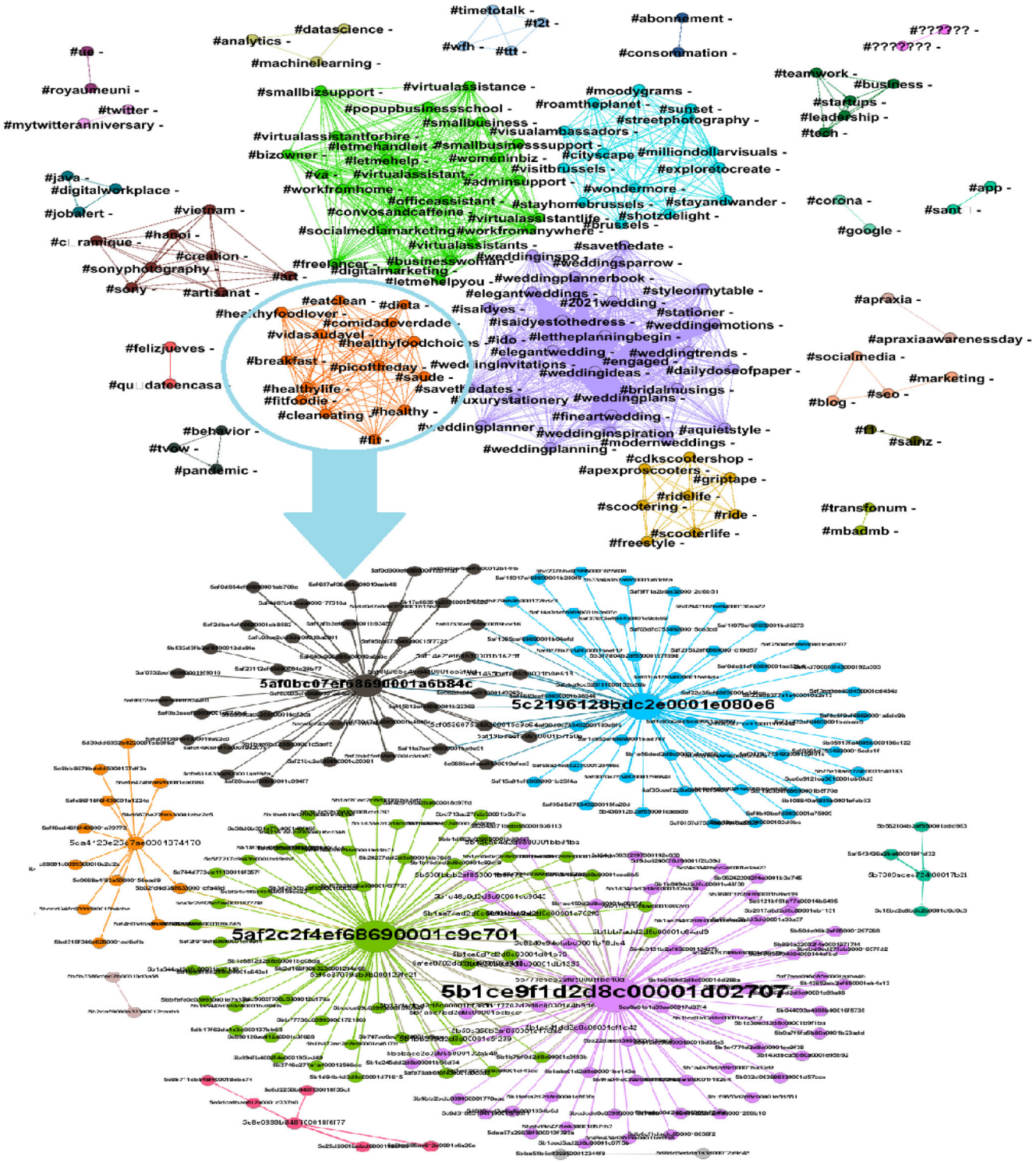


FIGURE 9 Illustration of topic modelling and finding influential nodes in a topic cluster using a sliding window

	A	B
1	Hashtag	Community ID
2	#dieta	0
3	#comidadeverdade	0
4	#fitfoodie	0
5	#picoftheday	0
6	#healthyfoodchoices	0
7	#healthylife	0
8	#vidasaudavel	0
9	#cleaneating	0
10	#breakfast	0
11	#healthy	0
12	#fit	0
13	#healthyfoodlover	0
14	#saude	0
15	#eatclean	0

FIGURE 10 Demonstration of results with the list of hashtags in a community

	A	B	C
1	Community ID	ProfileID	Score
2	0	5b1ce9f1d2d8c00001d02707	82
3	0	5af2c2f4ef68690001c9c701	72
4	0	5c2196128bdc2e0001e080e6	59
5	0	5af0bc07ef68690001a6b84c	39
6	0	5ca4123e26c7ae0001874170	12
7	0	5e8e0893b848ff00018f6f77	5
8	0	5b7309adec724f00017b2b64	3
9	0	5d6dd5edda1a3d00012a9d45	1
10	0	5b2da59090b32300012caab9	1

FIGURE 11 Results showing a list of users and their scores in a particular topic/community

window size is based on the number of observations, that is, 1000 edges. However, a time window such as edges from recent 1 day/1 month can also be considered. The resolution parameter in community detection for all the methods is set to 1.0.

7.1 | Topics analysis

The detected clusters are shown in Figures 6, 7, and 8. The figures represent sample snapshots in the end of stream. Each cluster with a different colour in the figures represents a topic. Sliding windows and biased sampling considers repetitive edges as the frequency or weight of an edge, which is depicted as thick arrows or lines in the figures. The thicker edge represents stronger connection between two hashtags. The hashtags with thicker edges are considered top hashtags in their cluster as they are most frequent.

The choice of sampling algorithm has different trade-offs. For finding the most frequent or trending topics from the stream over time, space saving is a relevant choice; however, it is computationally expensive compared to the other two though it is space efficient and the fastest one of its genre. The one with least time complexity among the three is biased sampling but lacks in terms of structure in this case, with a very sparse graph.

We see that the clusters in the figures clearly make sense in terms of synonymy and polysemy (e.g., in Figure 7, synonyms such as Covid, Covid19, and so on, are grouped in one blue cluster on the right and polysemy words are sharing two clusters green and orange in the centre).

The clusters formed by sliding windows are more denser than the other two. Quantitative metrics of these graphs are displayed in Table 1. The bias to low degree nodes increases the number of components and decreases density. Nevertheless, a large cluster of the popular topic 'covid19' can only be seen in space saving because sliding window and biased sampling collect data from the end of stream that is from the month of May, where it has low occurrence in our data. Moreover, the top trending hashtags from Figure 4 in analysis are found inside the communities of space saving (Figure 7). However, they are also found in sliding window and biased sampling as we increase the sample size.

The results also show correlated clusters that share common hashtags as seen in Figure 7, the green and orange cluster. Another important feature is that tags in different languages (Portugues, Spanish, or more) are still clustered semantically, such as in Figure 7 'confinamiento' belongs to covid cluster. Moreover, acronyms such as 'ai' belongs to artificial intelligence cluster. Each post is associated with multiple hashtags, therefore each post can be assigned to a number of topics.

7.2 | Correlation between SNA measures

It is vital to understand the correlation between the measures applied above to ensure the results are not redundant and one method would suffice for the akin. For this, we calculated Kendall's τ correlation coefficient to check the similarity between the ranks of all the users obtained by the measures in Section 6. Kendall's τ is a statistic which is used to measure the ordinal association between two variables (Kendall, 1938). The equation for calculating the Kendall's coefficient is given below. The coefficient lies between -1 and 1. τ is said to be more meaningful when the variables have a large number of tied ranks (Nie et al., 1975), which is the case with our data. The Kendall's τ coefficients comparing pairs of results from different measures are shown in Table 2. Besides, we also found that the distribution of resultant scores followed a power law distribution with more number of users having a score of zero. Therefore, to examine if the above correlations are not affected by the long tail of power law, we considered analysing the lists of top 30 users with Jaccard Similarity; which is a commonly used similarity measure between binary variable sets. As observed in Table 3, there are some measures, which are closely correlated while others are not. For brevity, we only mentioned the results of top trending topic (in the data) Coronavirus. However, the results of other topics are also similar (as shown in Tables C1 and C2).

$$\tau = \frac{\text{Number of concordant pairs-Number of discordant pairs}}{n(n-1)/2} \tag{6}$$

A maximum correlation between Degree centrality and Laplacian centrality is seen, followed by PageRank. While Closeness centrality, Clustering coefficient and Acceptance score order the nodes quite differently to other methods as well as each other. We also see that Acceptance score, which is based on the content quality of user post, has quite different popular users to the Dissemination ranks (centralities) based on friends. For instance, Degree centrality has a minimal correlation (Table 2) or no correlation considering top 30 key nodes (Table 3) to the Acceptance score based ranks. Therefore, it shows that the users who have more friends are significantly different from the users who are actively discussing and whose posts are highly 'liked' and 'shared' by other users. In other words, the users who have fewer friends are having quality/popular posts or maybe the friendship data is missing those relations; this needs further investigation.

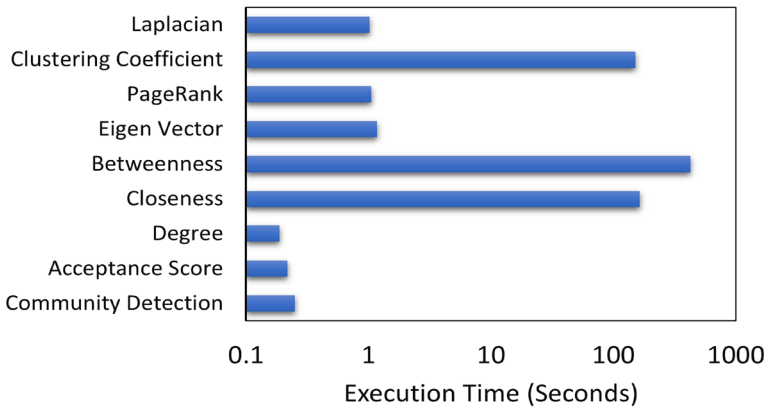


FIGURE 12 Execution time for centralities of all users in the given friendship network and acceptance score on like, share and Comment's network per day and community detection on hashtag network per day

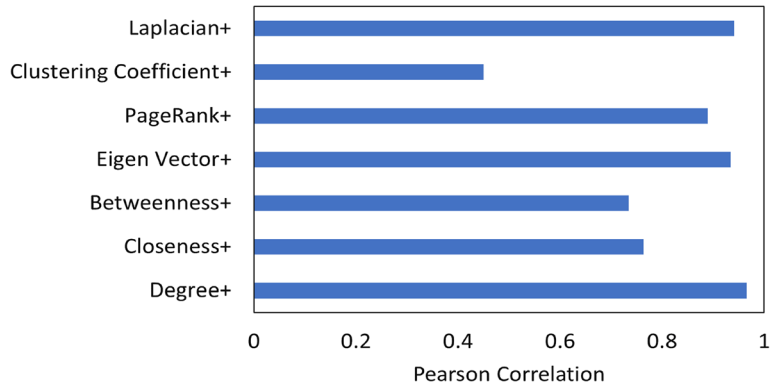


FIGURE 13 Comparison of improved centralities by calculating Pearson correlation of their top K ranks with the reference rank

7.3 | Efficiency in ranking influencers

How good are the methods in ranking influencers? To investigate this we compared the ranks generated by the improved methods Degree+, Betweenness+, Closeness+, PageRank+, Laplacian+, EigenVector+ and Clustering Coefficient+, against a Reference rank. The Reference rank is an average ranking created from the ranks produced by improved measures given in Section 6.3. A user having a high rank in most of the measures will achieve a high average rank. The correlation of resultant ranks from different methods (noted above) and the reference rank is illustrated in Figure 13. Degree+ earns maximum ranks that are supported by other measures, followed by Laplacian+, as Laplacian centrality is correlated to Degree centrality, which was also observed in Tables 2 and 3. Clustering coefficient has least ranks that correlate with the Reference rank.

7.4 | Execution time

The experiments had been carried out on a computer with Intel Core i7-6500U CPU @ 2.50 GHz and a RAM of 8 GB on the data set discussed. Figure 12 displays the execution time of SNA metrics computed over the Friendship Network from the friendship file, the Acceptance score metric computed over a sliding window of 1 day over the posts file on average and lastly shows the average time taken for detecting communities/clusters over the Hashtag Network from posts' file per day. It is to be noted that the execution times in the above figure does not include the time for building Hashtag Network, identifying user measures related to communities or topics, and writing results to files, and so on. Degree centrality is fast to compute and increment; besides is also effective as seen in above section. Acceptance score on the other hand is a contrast measure with similar complexity.

8 | CONCLUSION AND FUTURE WORK

The proposed work advances in the method and approach for finding experts/influencers in specific topics derived from the micro-blogs' hashtags while taking into account the evolving nature of topics and their relationships. So, it has potential benefits in real-world applications; for instance identifying domain experts/authority nodes for authenticity of information dissemination on specific events or social awareness activities or finding influencers in a particular niche for marketing, campaigns and more. For that reason, our experiments are carried out on a real-world large data set of micro-blogs from the most popular social media activity applications. Moreover, our approach uses techniques for handling large data which is beneficial for such data-intensive applications. The results show that the obtained clusters of topics clearly make sense and learn the latest connections between words/tags from a number of different languages. We also observe that not all users who have many friends are active influencers. It shows that friendship popularity does not reflect acceptability by other users. Ergo, our proposed Influence ranking takes into account the dissemination power of the user and the acceptance of it by other users that reflects the quality of his/her posts and expertise in that topic/product. As a result, we discover users who are active influencers with high popularity and acceptability.

In summary, we have presented a fast and memory efficient approach for incrementally identifying top K influencers and categorizing posts into topics using hashtags. Our topic modelling technique uses information rich hashtags network rather than complex and time-consuming models that rely on unstructured post data and a fixed set of word relations. Whereas, our dynamic model captures evolving word relations by applying dynamic sampling methods. Consequently, we discussed on how the given sampling algorithms can effect the outcome and compared them in terms of semantics and structure of clusters. Further, considered their biases and trade-offs. We analysed the seasonality and trending

TABLE 1 Network properties

	Average degree	Avg. weighted degree	Density	Modularity	#Clusters
Sliding window	6.6	6.667	0.087	0.723	25
Space saving	3.413	3.413	0.022	0.806	33
Biased sampling	2.687	2.747	0.015	0.872	61

TABLE 2 Kendall's τ coefficient between ranks from different measures of all users posting about coronavirus

	Degree	Closeness	Betweenness	PageRank	Clustering	Laplacian	Eigen vector	Acceptance
Degree	1	0.556	0.609	0.748	0.352	0.785	0.75	0.014
Closeness	0.556	1	0.424	0.527	0.197	0.69	0.659	−0.06
Betweenness	0.609	0.424	1	0.669	0.08	0.522	0.495	0.033
PageRank	0.748	0.527	0.669	1	0.217	0.568	0.527	0.048
Clustering	0.352	0.197	0.08	0.217	1	0.337	0.352	−0.027
Laplacian	0.785	0.69	0.522	0.568	0.337	1	0.918	−0.045
Eigen Vector	0.75	0.659	0.495	0.527	0.352	0.918	1	−0.045
Acceptance	0.014	−0.06	0.033	0.048	−0.027	−0.045	−0.045	1

TABLE 3 Jaccard similarity between top 30 influential users posting about coronavirus by applying different measures

	Degree	Closeness	Betweeness	PageRank	Clustering	Laplacian	Eigenvector	Acceptance
Degree	1.0	0.0	0.176	0.224	0.0	0.875	1.0	0.0
Closeness	0.0	1.0	0.0	0.0	0.071	0.0	0.0	0.017
Betweenness	0.176	0.0	1.0	0.579	0.0	0.176	0.053	0.071
PageRank	0.224	0.0	0.579	1.0	0.0	0.2	0.071	0.053
Clustering	0.0	0.071	0.0	0.0	1.0	0.0	0.0	0.034
Laplacian	0.875	0.0	0.176	0.2	0.0	1.0	0.622	0.0
Eigenvector	1.0	0.0	0.053	0.071	0.0	0.622	1.0	0.0
Acceptance	0.0	0.017	0.071	0.053	0.034	0.0	0.0	1.0

hashtags in the data. Also compared quite a number of SNA measures against each other for ranking influencers and found new insights. Additionally, we proposed an Acceptance measure, which determines the popularity or quality of post and combined with the SNA measures for complementing Acceptance with Dissemination scores. Finally, we compared the improved measures to a reference rank to evaluate their efficiency of spread. To facilitate comprehensibility we preferred network visualization layouts over the conventional presentation using tables for word clusters.

As a future plan, there are many possible advancements of this work. One potential line of work would be about studying the evolution of topics and the time varying influencers. On the availability of posts text, we can implement other topic models and improve them using our approach. Further, we intend to analyse the trend of topics overtime and the evolution of communities. Additionally, predicting hashtags for the missing ones using our topic model.

ACKNOWLEDGEMENTS

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020. The authors also acknowledge SKORR for providing data.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Paulo J. Azevedo https://orcid.org/0000-0002-0877-3070

Mario Cordeiro https://orcid.org/0000-0001-8846-9484

REFERENCES

Alash, H. M., & Al-Sultany, G. A. (2020). Improve topic modeling algorithms based on twitter hashtags. *Journal of Physics: Conference Series*, *1660*, 012100.

Alp, Z. Z., & Öğüdücü, Ş. G. (2018). Identifying topical influencers on twitter based on user behavior and network topology. *Knowledge-Based Systems*, *141*, 211–221.

Al-Shargabi, A. A., & Selmi, A. (2021). Social network analysis and visualization of arabic tweets during the covid-19 pandemic. *IEEE Access*, *9*, 90616–90630.

Argyrou, A., Giannoulakis, S., & Tsapatsoulis, N. (2018). Topic modelling on instagram hashtags: An alternative way to automatic image annotation? In *2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization* (*SMAP*) (pp. 61–67).

Bhakdisuparit, N., & Fujino, I. (2018). Understanding and clustering hashtags according to their word distributions. In *2018 5th International Conference on Business and Industrial Research* (*ICBIR*) (pp. 204–209).

Bhattacharya, S., & Sarkar, D. (2021). Study on information diffusion in online social network. In *Proceedings of International Conference on Frontiers in Computing and Systems* (pp. 279–288).

Bi, B., Tian, Y., Sismanis, Y., Balmin, A., & Cho, J. (2014). Scalable topic-specific influence analysis on microblogs. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining* (pp. 513–522).

Blei, D., & Lafferty, J. (2006). Correlated topic models. *Advances in Neural Information Processing Systems*, *18*, 147.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.

Bond, E. U., III, Walker, B. A., Hutt, M. D., & Reingen, P. H. (2004). Reputational effectiveness in cross-functional working relationships. *Journal of Product Innovation Management*, *21*(1), 44–60.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, *25*(2), 163–177.

Chouchani, N., & Abed, M. (2020). Enhance sentiment analysis on social networks with social influence analytics. *Journal of Ambient Intelligence and Humanized Computing*, *11*(1), 139–149.

Cordeiro, M., Sarmento, R. P., Brazdil, P., & Gama, J. (2018). Evolving networks and social network analysis methods and techniques. *Social Media and Journalism-Trends, Connections, Implications*, 101–134.

Corradini, E., Nocera, A., Ursino, D., & Virgili, L. (2021). Investigating negative reviews and detecting negative influencers in yelp through a multi-dimensional social network based model. *International Journal of Information Management*, *60*, 102377.

Cruickshank, I. J., & Carley, K. M. (2020). Characterizing communities of hashtag usage on twitter during the 2020 covid-19 pandemic by multi-view clustering. *Applied Network Science*, *5*(1), 1–40.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, *101*-(Suppl. 1), 5220–5227.

Gama, J. (2010). *Knowledge discovery from data streams*. CRC Press.

Hajian, B., & White, T. (2011). Modelling influence in a social network: Metrics and evaluation. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* (pp. 497–500).

Hinz, O., Skiera, B., Barrot, C., & Becker, J. U. (2011). Seeding strategies for viral marketing: An empirical comparison. *Journal of Marketing*, *75*(6), 55–71.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM Sigir Conference on Research and Development in Information Retrieval* (pp. 50–57).

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics* (pp. 80–88).

Ishfaq, U., Khan, H. U., Iqbal, S., & Alghobiri, M. (2022). Finding influential users in microblogs: State-of-the-art methods and open research challenges. *Behaviour & Information Technology*, *41*(10), 2215–2258.

Jain, S., & Sinha, A. (2020). Identification of influential users on twitter: A novel weighted correlated influence measure for covid-19. *Chaos, Solitons & Fractals*, *139*, 110037.

Jianqiang, Z., Xiaolin, G., & Feng, T. (2017). A new method of identifying influential users in the micro-blog networks. *IEEE Access*, *5*, 3008–3015.

Kaple, M., Kulkarni, K., & Potika, K. (2017). Viral marketing for smart cities: Influencers in social network communities. In *2017 IEEE Third International Conference on Big Data Computing Service and Applications* (*Bigdataservice*) (pp. 106–111).

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81–93.

Mao, G.-J., & Zhang, J. (2016). A pagerank-based mining algorithm for user influences on micro-blogs. In *PACIS 2016 Proceedings* (p. 226).

Metwally, A., Agrawal, D., & El Abbadi, A. (2005). Efficient computation of frequent and top-k elements in data streams. In *International Conference on Database Theory* (pp. 398–412).

Mikolov, T., Yih, W.-T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746–751).

Mostafa, M. M. (2021). Information diffusion in halal food social media: A social network approach. *Journal of International Consumer Marketing*, *33*(4), 471–491.

Muntean, C. I., Morar, G. A., & Moldovan, D. (2012). Exploring the meaning behind twitter hashtags through clustering. In *International Conference on Business Information Systems* (pp. 231–242).

Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, *74*(3), 036104.

Nie, N. H., Hull, C. H., Jenkins, J., Steinbrenner, K., & Bent, D. H. (1975). *SPSS: Statistical Package for the Social Sciences* (2nd ed., Vol. 5., pp. 41–42). McGraw-Hill Book Co., Journal of Advertising.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web*. Stanford InfoLab.
Sanawi, J. B., Samani, M. C., & Taibi, M. (2017). # vaccination: Identifying influencers in the vaccination discussion on twitter through social network visualisation. *International Journal of Business and Society*, *18*(S4), 718–726.
Suau-Gomila, G., Pont-Sorribes, C., & Pedraza-Jiménez, R. (2020). Politicians or influencers? Twitter profiles of pablo Iglesias and albert Rivera in the spanish general elections of 20-d and 26-j. *Communications Society*, *33*, 209–225.
Tabassum, S. (2020). *Massive scale streaming graphs: Evolving network analysis and mining*. Universidade do Porto.
Tabassum, S., & Gama, J. (2016). Sampling massive streaming call graphs. In *ACM Symposium on Advanced Computing* (pp. 923–928).
Tabassum, S., Pereira, F. S., Fernandes, S., & Gama, J. (2018). Social network analysis: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(5), e1256.
Wang, Y., Liu, J., Huang, Y., & Feng, X. (2016). Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs. *IEEE Transactions on Knowledge and Data Engineering*, *28*(7), 1919–1933.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442.
Xiao, F., Noro, T., & Tokuda, T. (2014). Finding news-topic oriented influential twitter users based on topic related hashtag community detection. *Journal of Web Engineering*, *13*(5&6), 405–429.
Yang, Y., Wang, Z., Pei, J., & Chen, E. (2017). Tracking influential individuals in dynamic networks. *IEEE Transactions on Knowledge and Data Engineering*, *29*(11), 2615–2628.
Yin, X., Hu, X., Chen, Y., Yuan, X., & Li, B. (2021). Signed-pagerank: An efficient influence maximization framework for signed social networks. *IEEE Transactions on Knowledge and Data Engineering*, *33*(5), 2208–2222. https://doi.org/10.1109/TKDE.2019.2947421
Yu, Y., Mo, L., & Wang, J. (2016). Identifying topic-specific experts on microblog. *KSII Transactions on Internet and Information Systems (TIIS)*, *10*(6), 2627–2647.

AUTHOR BIOGRAPHIES

Shazia Tabassum is a researcher at the Laboratory of Artificial Intelligence and Decision Support (LIAAD), INESC TEC Porto. She obtained her PhD in Informatics Engineering from the Faculty of Engineering, University of Porto in Portugal. She also holds the degrees of Bachelor's and Master's in Computer Applications from India. During the past eight years, she contributed to research in the areas of Social Network Analysis, Evolving Data Science, Machine Learning and Explainability applied in the domains such as Telecommunications, Marketing, and Social Media by working in European research projects and industry collaborations. She authored a number of publications including book chapters, high quality journals and conference proceedings at notable venues (IEEE, Wiley, Springer, Cambridge and ACM). She served as a reviewer in several indexed Journals/conferences and also as a co-chair of the special sessions/workshops including Evolving Networks DSAA'17, Discovery Challenge EPIA'17 and MobDM IEEE MDM'16.

João Gama is a full professor at the School of Economics, University of Porto, Portugal. He received his PhD in Computer Science from the University of Porto in 2000. He is EurIA Fellow, IEEE Fellow, Fellow of the Asia-Pacific AI Association, and member of the board of directors of the LIAAD, a group belonging to INESC TEC. He is ACM Distinguish Speaker. His h-index at Google Scholar is 68. He is an Editor of several top-level Machine Learning and Data Mining journals. He served as Program Chair of ECMLPKDD 2005, DS09, ADMA09, EPIA 2017, DSAA 2017, served as Conference Chair of IDA 2011, ECMLPKDD 2015, DSAA'2021, and a series of Workshops on KDDS and Knowledge Discovery from Sensor Data with ACM SIGKDD. His main research interests are in knowledge discovery from data streams, evolving network data, probabilistic reasoning, and causality. He published more than 300 reviewed papers in journals and major conferences. He has an extensive list of publications in data stream learning.

Paulo Azevedo is a lecturer at the Department of Informatics at the University of Minho. He is also a researcher at HASLab/INESC TEC. His research interest focused mainly on machine learning and data mining. Occasionally, he participates in Bioinformatics research projects involving analysis of molecular dynamic simulations of protein folding/unfolding. He holds a PhD in Computing from Imperial College (University of London) where he did research in logic programming. He has been working on the development of association rules mining algorithms and novel patterns to capture distribution learning. He also has interest in social network analysis, graph mining, subgroup mining and motif discovery in time series.

Mário Cordeiro holds a master's degree in Computer Engineering and a degree in Electrical and Computer Engineering from the Faculty of Engineering of the University of Porto. Currently is a PhD candidate in the Doctoral Program in Computer Engineering researching in the areas of data mining, machine learning, data streams and dynamic networks analysis. Since 2010 he has collaborated as an invited lecturer at the University of Porto, Polytechnic of Porto - School of Engineering, and Porto Business School. Mário Cordeiro is also a Solutions Architect at Critical Manufacturing leading the design and development of MES (Manufacturing Execution Systems) solutions for the semiconductors and medical device industries.

Carlos Martins is an experienced hands-on leader with a track record of developing innovative solutions leveraging ML/AI, Big Data & Cloud technologies. He is a Head of Engineering at Mobileum and Vice President of DTx - Digital Transformation CoLAB. He was responsible for

the development of solutions in various areas including Fraud, Revenue Assurance, and Analytics applications for Carriers and large Enterprises. He manages global teams of up to 130 engineers. He has been actively participating and leading several international innovation projects (sponsored by public innovation agencies).

André Martins is the Head of Engineering at Skorr, a platform for Social and Digital Media, that enhances the experience of end-users, and brings users, even more, closer to brands. As background hold's a master degree in Formal Methods and Distributed Systems from the University of Minho. Over the years, he gained experience with different roles in sectors and industries in the market, passing through the sector of Banks, Telecommunications, Fraud, and now recently Social Media.

How to cite this article: Tabassum, S., Gama, J., Azevedo, P. J., Cordeiro, M., Martins, C., & Martins, A. (2022). Social network analytics and visualization: Dynamic topic-based influence analysis in evolving micro-blogs. *Expert Systems*, e13195. https://doi.org/10.1111/exsy.13195

APPENDIX

DEGREE DISTRIBUTION OF FRIENDSHIP NETWORK

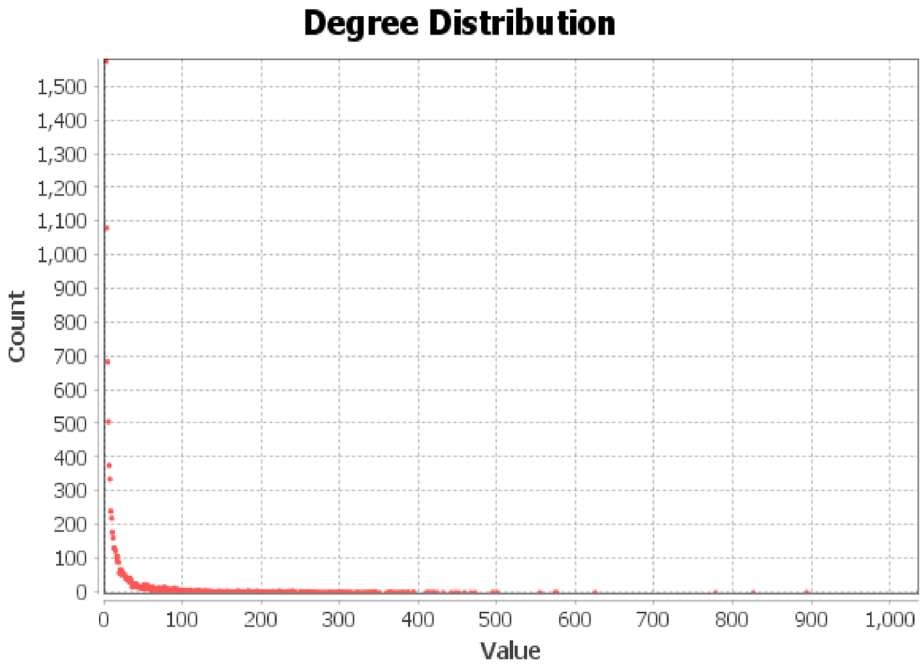
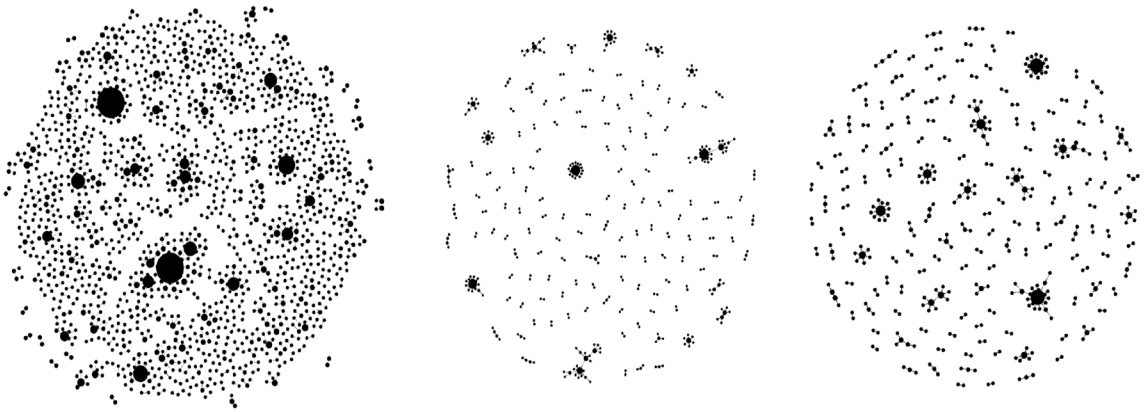


FIGURE A1 Degree distribution of the nodes in friendship network

FRIENDSHIP NETWORKS OF INFLUENTIAL NODES PER TOPIC

CORRELATION OF SNA MEASURES

The correlation of SNA measures over the next top trending topics in the data i.e. marketing and social media is given in tables below.



(a) Influential nodes of coronavirus (b) Influential nodes of marketing (c) Influential nodes of socialmedia

FIGURE B2 Friendship networks of influential nodes in a topic (size represents degree) and their dissemination/spread (one hop)

TABLE C2 Kendall's τ coefficient between ranks from different measures for users posting about socialmedia

	Degree	Closeness	Betweenness	PageRank	Clustering	Acceptance	Laplacian	Eigen vector
Degree	1	0.555	0.529	0.774	0.117	0.150	0.868	0.785
Closeness	0.555	1	0.526	0.615	−0.124	0.071	0.618	0.568
Betweenness	0.529	0.526	1	0.667	−0.189	0.176	0.481	0.414
PageRank	0.774	0.615	0.667	1	−0.28	0.165	0.681	0.593
Clustering	0.117	−0.124	−0.189	−0.28	1	−0.046	0.117	0.161
Acceptance	0.150	0.071	0.176	0.165	−0.046	1	0.125	0.107
Laplacian	0.868	0.618	0.481	0.681	0.117	0.125	1	0.9
Eigen Vector	0.785	0.568	0.414	0.593	0.161	0.107	0.9	1

TABLE C1 Kendall's τ coefficient between ranks from different measures for users posting about marketing

	Degree	Closeness	Betweenness	PageRank	Clustering	Acceptance	Laplacian	Eigen vector
Degree	1	0.549	0.555	0.773	0.124	0.141	0.854	0.774
Closeness	0.549	1	0.525	0.583	0.114	0.045	0.632	0.584
Betweenness	0.555	0.525	1	0.677	−0.161	0.096	0.512	0.449
PageRank	0.773	0.583	0.677	1	−0.003	0.119	0.666	0.579
Clustering	0.124	0.114	−0.161	−0.003	1	0.06	0.115	0.15
Acceptance	0.045	0.045	0.096	0.119	0.06	1	0.125	0.115
Laplacian	0.854	0.632	0.512	0.666	0.115	0.125	1	0.901
Eigen Vector	0.774	0.584	0.449	0.579	0.15	0.115	0.901	1