Joint theme and event based rating model for identifying relevant influencers on Twitter: COVID-19 case study

Ali Srour [a], Hakima Ould-Slimane [b], Azzam Mourad [a,d],∗, Haidar Harmanani [a], Cathia Jenainati [c]

a *Department of Computer Science and Mathematics, Lebanese American University, Beirut, Lebanon*
b *Department of Software Engineering and IT, École de Technologie Supérieure, Montreal, QC, Canada*
c *Department of English, Lebanese American University, Beirut, Lebanon*
d *Division of Science, New York University, Abu Dabi, United Arab Emirates*

∗ Corresponding author at: Department of Computer Science and Mathematics, Lebanese American University, Beirut, Lebanon.
E-mail address: azzam.mourad@lau.edu.lb (A. Mourad).

ABSTRACT

The continuous proliferation of social media platforms and the exponential increase in users' engagement are impacting social behavior and leading to various challenges, including the detection and identification of key influencers. In fact the opinions of these influencers are at the core of decision-making strategies, and are leading trends on the virtual social media landscape. Moreover, influencers might play a crucial role when it comes to misinformation and conspiracy during sensitive, controversial and trending events. However, due to the dynamic and unrestricted nature of social media, and diversity of targeted topics and audiences, identifying and ranking key influencers that are impactful, credible, and knowledgeable about their specialist topic or event remains an evolving and open research paradigm. In this paper, we address the aforementioned problem by proposing a novel influence rating and ranking scheme to identify key and highly influential users for a certain event over Twitter using a mixed theme/event based approach while considering historical data and profile reputation. We further apply our approach to a global pandemic case study, the novel Coronavirus, and conduct performance analysis. The presented experimental results and theoretical analysis explore the relevance of our proposed scheme for identifying and ranking reputable and theme/event related influencers.

Keywords:
Social Network Analysis

Twitter

Influence rating

Influencers

Reputation

Social Listening

Computational Social Science

User impact

Big data

Data science

COVID-19

Infodemic

Natural Language Processing

1. Introduction

Social media platforms have been one of the most prominent ways for connecting people, offering a medium for sharing publicly individual experiences and stories with an estimated 3.4 billion people in 2019 using social media worldwide said *STATISTA*.[1] Furthermore, these platforms have the potential to enhance people's quality of life by offering easy accessibility to information and providing social support. Users from different backgrounds can discuss trending topics, share thoughts and initiatives thus creating communities of shared experiences. Therefore, they are no longer a medium for basic social interaction; rather they have evolved into hubs for sharing news, advertising campaigns, promoting entertainment and media events, and exchanging health and lifestyle tips. They also offer different features that can be tailored for different purposes. Accordingly, research firms and organizations develop different strategies for the various platforms to understand user interaction and to assess the reliability of information being disseminated. Such a process is complicated by the fact that data generated by users over social media platforms evolves in real time and is noisy, unstructured and redundant.

Moreover, news creation and consumption have been adapting and evolving since the advent of social media. Most platforms are typically used to transmit relevant news, guidelines and precautions to people. However, according to the World Health Organisation (WHO), in the context of epidemics, uncontrolled conspiracy theories and propaganda are spread faster than the pandemic events themselves, thus creating an infodemic and causing panic and disseminating misleading medical advice and economic disruption said the Director-General of WHO.[2] Consequently, determining which news sources can be trusted is essential to validate information. While medical and research centers can provide major insights into the flow of infectious diseases, the opinions and discussions of social media users and especially influencers during pandemics are worth observing as they provide valuable insight into the psychological and sociological impact on followers. Influencers over social platforms are known to explore users' opinions from predicting election results to offering advice on medical practices. The impact of influencers in times of crisis is a hotly debated topic. Identifying and predicting influencers on social networks have many applications including viral marketing [1], searching [2], and expert recommendation [3]. The ability of influencers to spread information has been well studied on Twitter [4,5] where influence was used for different goals such as human mobility [6], rumor spreading [7] and epidemiology [8] among others. However, capturing public unfiltered opinions in the digital ocean of social media data presents a key challenge through which

[1] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.
[2] https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

fine-grained filtering and proper extraction should be made to detect relevant, reputable and theme/event related influencers [9–11].

Identifying and ranking influencers on Twitter is becoming an appealing and challenging problem [12], which we address in this paper. The quantitative assessment of influencer behavior is considered a key challenge whereby different solutions and approaches are being proposed. For instance, predicting influencers in micro-blogging critical events requires measurements that accurately capture this influence and does not tolerate errors. In fact, the definition of influence varies from one context to another. The influence of a user can be defined as the user's capacity to influence others with their opinion, while it can also be defined as the ability of a user to spread a certain message to others with whom they typically interact [13]. While detecting influencers in a social network can be approached at a global or a general level [14], we focus in this paper on providing a solution that identifies influencers in particular events [15,16]. We define an event as an incident that occurs in a certain time frame and specified location that lead users to tweet and initiate various hashtags and trends while the event is ongoing. Several approaches have been proposed in the literature like [17–23], and [24] targeting finding influence level while using different approaches of analyzing user tweets and connections. However, to the best of our knowledge, none of them have addressed finding influence level within specific events regardless of their influence on their networks, which was clear when the authors in [25,26], and [27–30] used singular user meta data analysis without connecting event based knowledge or factors.

In this context, we elaborate in this paper an approach that entails ranking of influencers by assessing their level of influence based on theme, event and reputation contexts. Our proposed model correlates multiple influence ratio calculations to achieve this objective. We aim at capturing all event influencers and at the same time all theme influencers with similar time, location, and sample size criteria, and then correlating and analyzing historical influence and other attributes to obtain maximum accuracy of influence rating and ranking. To elaborate more, we first aim at calculating influencers ratings and then selecting those influencers in the context of an event while jointly maximizing influence rating accuracy and minimizing spam user irrelevance by joining similar calculations and findings about the selected theme to which the event belongs. Second, we aim at selecting, out of the first resulting selected list of influencers, the set of influencers with the best reputations by analyzing their content and profiles credibility and impact. We measure the influence of each user from multiple angles and using a combination of influence ratios to achieve the aforementioned objectives. The main contributions of the paper is summarized as follows:

- Identifying twenty two different data points and twelve calculated ratios to calculate the Event's Influence and Users' Historical Influence rates for maximizing the influence rate calculations of a certain event on Twitter.
- Elaborating a Lexicon-based approach to measure the content and profile credibility of a selected list of influencers based on their profiles and tweeting activities.
- Optimizing the accuracy of influence rating through a Joint Theme and Event based content analysis model considering the impact and reputation influence ratio.

The paper is organized as follows. Section 2 provides the literature review. Section 3 describes the approach overview including a road map of the entire analysis. Section 4 presents the system model with a detailed problem and mathematical formulation. Section 5 illustrates the approach and model implementation followed by comprehensive experiments and discussion. Finally, Section 6 discusses the conclusion and future research directions.

2. Literature review

In this section, we provide a literature review related to reputation and credibility analysis, and influence ranking in social networks.

2.1. User-centered and content-based reputation and credibility analysis

Many authors worked on sorting and ranking users based on their influence ratios. Mainly, the underlying techniques were based on two major categories: (1) focusing on tweets content to calculate the reputation level using ML techniques [19,31,32], and (2) focusing on the network analysis of each user's graph network to assess his/her level of influence [17–19]. In order to achieve a higher accuracy, some authors have combined the two mentioned categories. Jain et al. [20] used a class of graph theory algorithm in order to score users based on their centrality for classifying and identifying universal leaders opinions in later stages. But in terms of social distancing, authors of [33] provided an important analysis on how this can affect trust and trustworthiness among users. Mohammadinejad et al. [21] in his article presented a framework that derives the users' influence levels from their analyzed personalities. The authors of [22] identified user behaviors based on their frequent activities in order to score their credibility. Wang et al. [23] correlated tweets credibility to user profiles. In addition to that, Tsikerdekis et al. [34] highlighted more important user habits or actions like creating multiple accounts. Also, Ahmad et al. [23] article presented a survey on different approaches used for detecting social networks rumors. Buzz et al. [35] in his article, used sentiment analysis for Arabic language tweets in order to classify content and users based on their sentiment score. Finally, Alrubaian et al. [24] deduced user credibility in order to measure the level of fake and malicious news

spreading. Some of the aforementioned approaches have correlated tweets credibility to user profiles and have considered trust and centrality during specific events. These approaches helped us in defining deeper insights about users. However, to the best of our knowledge, none of them offered a model correlating event topic credibility and user reputation findings to identify relevant influencers.

2.2. Influence ranking in social networks

Influence detection in events and public conversations has become one of the most challenging research direction in the field of Social Network Analysis especially in Twitter event. But finding influencers and ranking users based on their influence ratio can be done using different methods and techniques. Authors in [25–30] found that user meta data like number of followers, number of tweets, and number of followees in addition to tweets meta data like retweet and favorite counts are enough to find users influence ratios. On the other hand, authors in [36] rank user influence based on their actual relationships. While [37] links users influence with their social activity during the selected event. The authors of [38] calculated a potential social networking ratio (SNP) for each user in among a number of most popular Twitter accounts in Austria using their accounts meta data like followers and followees counts. While Bakshy, Eytan et al. [39] used diffusion trees techniques in order to calculate the influence ratios of users whom their tweets has URLs by calculating the reach of those URLs in other platforms. In [40], M. Anjaria and R. M. R Guddeti used Incremental Learning algorithms with NLTK sentiment analysis to predict the presidential elections in the US. While other authors such as in [41] classified and categorized users into four influence levels categories based on their interactions and activities. In [42], Y. Mei, Y. Zhong and J. Yand in order to calculate the popularity for each user, they used eight data points in addition to NewFollowers and NewMentions features to find the top hundred users in Australia. Riquelme et al. [31] targeted the propagation ratios of users' tweets using two different linear centrality approaches. Similarly, Li et al. [43] proposed an eigenvector centrality based approach to measure users' influence rate as well. Lahuerta-Otero et al. [44] applied behavioral analysis techniques among special kind of twitter users in which those techniques can increase user's influence ratios. However, Sharma et al. [45] proposed a novel approach that combines users tweets and their trend scores in order to elect potential influence ratios. Moreover, Huynh et al. [46] focused on analyzing tweets tags and their correlation with the speed of their propagation. In [47], the authors propose Weighted Correlated Influence (WCI) approach which combines the relative impact of timeline-based and trend-specific features of social media users. The proposed approach merges both the profile activity and underlying network topology to calculate the influence score for each user. The authors differentiate trends related to COVID-19 over Twitter to generate their results. In [48], the authors identify information influencers during the COVID-19 pandemic and present analysis over a dataset of collected Arabic tweets during the COVID-19 pandemic. The study uses the network topology to identify influencers during the month of March, 2020 and then implements both HITS and

PageRank algorithms to analyze and compare the ranking among users. The results show that both HITS and PageRank algorithms have 40% similar influencers. Although the aforementioned approaches assessing user and tweets meta data and activity records are very helpful to discover user influence, they might not be enough to conclude influence in a specific topic. To the best of our knowledge, none of them have correlated those assessments to specific topic to increase the probability of identifying the intended relevant influencers.

3. Approach overview

In this section, we describe the approach architecture and main metrics in addition to the analysis roadmap. We are using the term ‘‘accuracy’’ in our approach to distinguish from other user influence rating approaches. The more the methodology can classify event real influencers, the more accurate it is. The proposed approach is used to measure and calculate users influence rates as accurate or non-accurate, providing an accuracy threshold or percentage after validating the results. Moreover, in order to maximize the accuracy of finding influencers in social media events, we infer the actual event influencers based on a model that can reduce the gap between the user’s calculated influence rates and their actual influence. Finally, influence rating accuracy can be clearly defined as the level of match between a user’s given influence rate and his actual influence value. Each user is considered a key driver in social media platforms. Hence, studying users’ behaviors entails a set of metrics that specifies their profiles and their synergies. To start with, we describe in the following the main user social metrics used in our approach:

- **Reputation:** Reputation can be defined as an attitude composed of an emotional and a rational aspect of a certain content or user [49]. Reputation can be used for different purposes on Twitter, such as political activities, human mobility and epidemiology [50].
- **Influence:** In the context of social media platforms, various definitions for influence have been defined and measured [51]. Influence can be generally defined as the effect induced by a certain person on the ideas, thoughts or behaviors of other people [52]. Indeed, Katz at al. [53] claim that influencers are able to produce, using word-of-mouth, a chain-reaction of influence resulting in high reach.
- **Credibility:** Credibility can be defined as trustworthiness and inherent persuasiveness. Utilizing credible sources and information over social media is essential for deriving accurate conclusions. However, evaluating credibility is a challenging problem. Usually, some users are not reliable and thus there is no guarantee regarding the validity of their content. Twitter provides the username as the only data about those users and thus their profiles can be fake and may generate false information. Other accounts are verified accounts and refer to

legitimate source that is authoring the account's tweets.[3] However, this is only a small group of users whose accounts are verified. Hence, measuring the credibility of a certain social media user is essential to depict the credibility of the given piece of information. Different levels of credibility are defined and various research interests are being explored to measure the credibility for each level:

- Post Credibility: defines whether a certain post represent relevant and accurate information about a specific topic [54].
- User Credibility: quantifies the reliability of a certain user and is typically associated with a certain score [55].
- Topic Credibility: corresponds to the acceptance of a certain topic or event [56].

The scheme architecture, depicted in Fig. 1, represents the main modules and flowchart about the analysis mechanisms used to find top influencers for a selected event. To maximize the accuracy of finding such a class of users, a correlation is performed between the influence metrics calculation according to the selected event with the encompassing theme having the same location and time-frame. Moreover, adding a layer of reputation maximization is a major enhancement improving the final rating results. For a better measurement, we chose to select 1000 users as initial user datasets selected for both event and theme. Then, we investigated the results while sorting and displaying the final insights of the top 10 percent (100 users) of the initial user dataset in the optimization layer-2, which is intended to be supporting our approach by looking at the each users history of activities to find their general content credibility as displayed in Fig. 1 and explained in the following modules (11 to 17):

The flowchart in Fig. 1 provides a step-by-step description of our approach. After collecting and cleaning the initial datasets, the following processes take place by order in two stages. **Stage-1** embeds the main approach of accuracy maximization (for the selected 1000 users dataset) including Engagement Calculation, Event Influencers Selection, Theme Influencers Selection, Event Influence Rate Calculation, Theme General Influence Rate Calculation, Event/Theme Influencers Classification, Event Influence Rating Maximization, Event Influence Irrelevance Minimization, Event/Theme Activity Calculation, and Event Influencers Credibility Aggregation. **Stage-2** embeds the second layer of reputation and credibility (for the selected 100 users dataset) including Joint Influencers Selection for Reputation, Influencers Impact and Profile Credibility Measurement, Influencers General Impact and Content Credibility Measurement, Influencers Reputation Calculation, Influencers Reputation Maximization, Influencers Sorting, and Influencers Exported List. In the sequel, we describe each module presented in the architecture assuming that the data is already collected, cleaned and prepared

[3] https://twitter.com/verified.

for analysis. In the sequel, we describe each module in details presented in the architecture assuming that the data is already collected, cleaned and prepared for analysis.

(1) *Engagement Calculation:* Calculates the Engagement Rate of all users involved in the selected event.

(2) *Event Influencers Selection:* Selects the top 1000 event influencers based on the calculated engagement rates.

(3) *Theme Influencers Selection:* Selects and sorts the top 1000 influencers from among the theme users based on their number of followers.

(4) *Event Influence Rate Calculation:* Calculates and sorts the Event Influence Rate for the selected 1000 event users.

(5) *Theme General Influence Rate Calculation:* Calculates and sort the Theme General Influence Rate for the selected 1000 theme users.

(6) *Event/Theme Influencers Classification:* Classifies both the event and theme influencers lists into Master, General, and Micro influence categories.

(7) *Event Influence Rating Maximization:* Maximizes the Influence Rate calculation by correlating both the Event and Theme results into general influence.

(8) *Event Influence Irrelevance Minimization:* Minimizes the irrelevance by re-arranging the users in both event and theme influencers list who might appear as spams, malicious and/or blacklisted users.

(9) *Event/Theme Activity Calculation:* Calculates and differentiates activity ratios for both event and theme influencers

(10) *Event Influencers Credibility Aggregation:* Finds credibility aggregations for the event influencers based on their profiles, tweets, and types.

(11) *Joint Influencers Selection for Reputation:* Selects top 100 influencers from the joint result of theme and event influencers/users list.

(12) *Influencers Impact and Profile Credibility Measurement:* Calculates the impact and profile credibility ratios for the selected 100 influencers based on their profile biographies.

(13) Influencers General Impact and Content Credibility Measurement: Calculates the general impact and content credibility ratios for the selected 100 influencers based on their historical activities.

(14) *Influencers Reputation Calculation:* Calculates the influencers reputation rates and the general reputation rates for the selected 100 influencers.

(15) *Influencers Reputation Maximization:* Correlates the previously calculated rates to maximize the reputation rates for the selected 100 influencers.

(16) *Influencers Sorting:* Sorts and displays the result of the maximized reputation rates.

(17) *Influencers Exported List:* The result shows the top 100 influencers in the selected event using the proposed scheme.

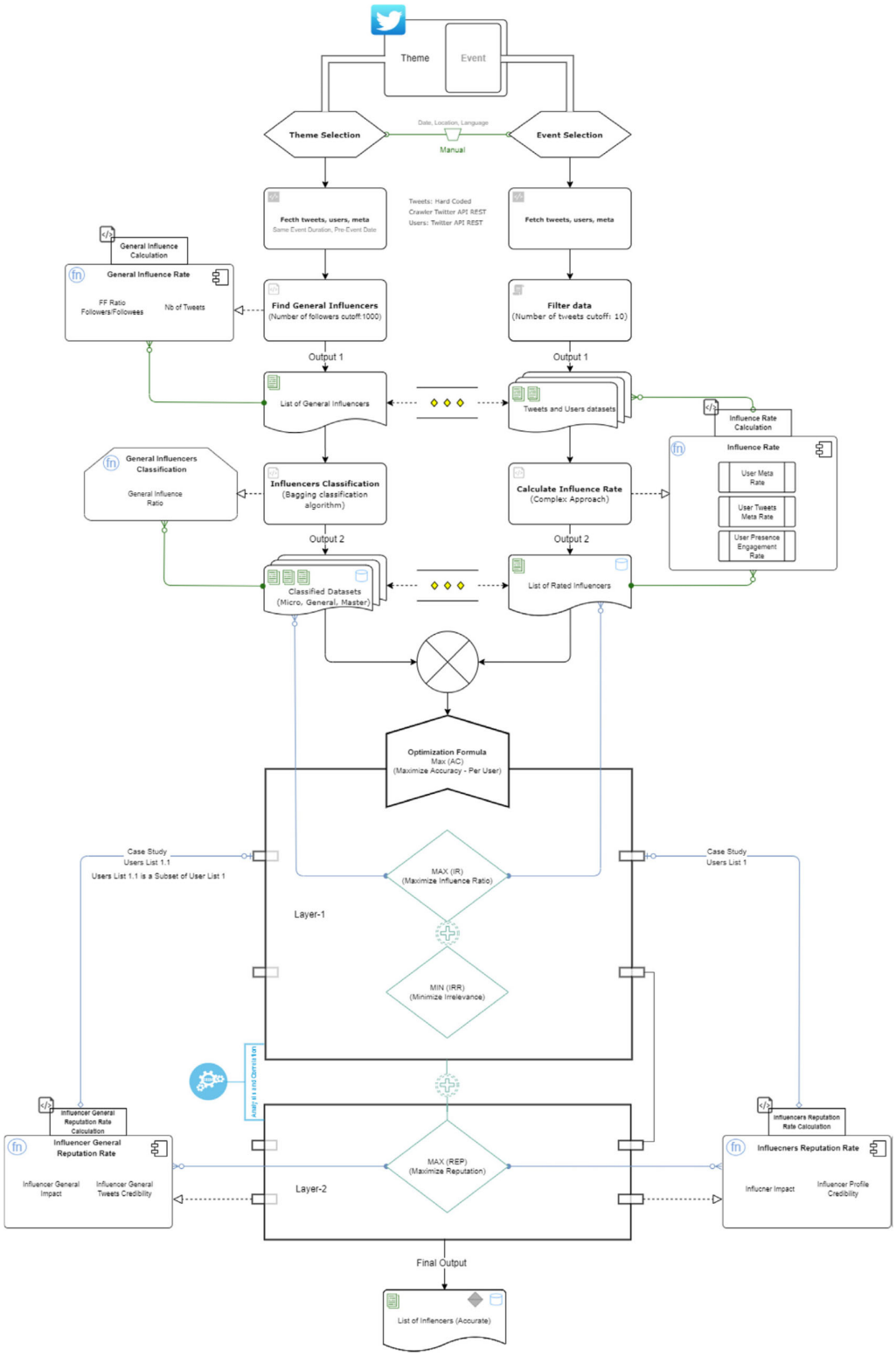

Figure 1 Scheme architecture.

4. Problem formulation & system model

In this section, we describe the problem formulation and system model of the proposed solution after showing the motivation of our intended formulation methodology. While most of the reviewed approaches tackling influence and influencer rating on Twitter events use user-centric and tweet-centric data points like followers, followees, retweets, interactions, and other data points to assess the level of influence a user might have within a specific network or a specific event (discussed in details in Section 2), we believe that there are multiple event-centric and theme-centric data points that could contribute to the influence rating formulation or calculation (the main motivation of our approach) should be addressed as well.

‘‘Are-influencers-really-influencing-anyone’’,[4] an article has been published in early 2021 talking about country influencers failing to influence people on COVID-19-related topics, and raised a question about if those influencers are not influencing people, who is? This was a trigger for us to support our motivation and try to find what we are calling ‘‘real-event influencers’’ by unlocking more data points and trying to apply our weighting methodology (which is a hypothesis we are trying to prove in this article) to contribute to the influence rating research. For that reason, we split the importance factor (weight) of all data points into a theme and event-based ones, and to formulate them into a mathematical model, we decided to evaluate the event at first and later combine theme-based findings to optimize the results.

Before proposing our methodology, we started by tracking a local topic related to ‘‘fashion’’ and manually tried to find the event influencers (using some of the approaches described in Section 2) after collecting a list of tweets that use a specific hashtag, and we were able to select a set of influencers which are celebrates and very well-known social media activists that contributed to the event using the same hashtag. However, their contribution was very short and came on the last days of the event without being contributed to the event evolution on Twitter before they start tweeting, in addition to that they do not belong to the same community of the event ‘‘fashion’’. That is the reason why we believed that those influencers are still influencers due to their profiles, however, they might not be influencers in this specific event due to the mismatch between their profiles and their historical interests or influencing topics. From this point of view, we decided to perform our approach by assigning all users who engaged with a specific event, different weights for their data points by reducing the importance of being influencers in their networks, and increasing the importance of being engaged in the specific topic and having created more impact on their spread content, and

[4] https://beirut-today.com/2021/01/26/are-influencers-really-influencinganyone/

finally accelerating the importance for those having previous activities on similar topics to be selected as credible users by combining their historical data points. The following metrics rules and mathematical formulas are intended to differentiate our work from other user and content influence rating approaches while looking at the problem from different angles. Enumerating formulas with weights was very helpful to explain the features importance while formulating the main optimization solution. All mathematical formulas are being weighted and added up to 1 (scale of 0–1 equivalent to 0%–100%). Given a set of users V disseminating a set of tweets T, regarding a set of events ε, belonging to a set of themes H, the function $f^{\rightarrow}, f^{\leftarrow}: U \rightarrow 2^{V}$ defines respectively the set of followers and followees of a given user. Our joint theme and event based rating approach aims to identify the most influential users list $L_V \subset V$ while maximizing the content accuracy $ACC(u)$ and reputation $REP(u)$ of the selected influencers as described in the following:

$$\max ACC(u), \forall u \in L_V \quad (1)$$

$$\max REP(u), \forall u \in L_V \quad (2)$$

We consider a sample of users $U_{e,h} \subset V$ engaged in a certain event $e \in \varepsilon$, belonging to a certain theme $h \in H$. Our objective consists of identifying the key influencers having the most significant impact on other users during this event. To achieve this goal, we use a joint theme and event based approaches for ranking the given sample of users. While the event based approach aims at rating the influence of each user, the theme based approach gets general list of influencers and apply classification algorithm to get general influencers. We specify the irrelevance of a user contribution in terms of spam accounts while eliminating the irrelevance as much as possible. We further optimize the list of influencers, crossed between both approaches, to get the list of influencers $L_{e,h}$ related to an event e belonging to the theme h. We then optimize the reputation of those influencers through selecting top influencers in $L_{e,h}$. In the following subsections, we divide the metrics into 2 (two) main layers. The layer of maximizing the accuracy which covers the initial influence calculation for theme datasets and the three user-centric and tweets-centric measures for the event datasets. Moreover, we add another layer of "Reputation Maximization" which covers the measurement of users and content credibility and popularity.

4.1. Accuracy maximization

We aim at maximizing the accuracy for each selected influencer in a certain event by maximizing the influence rate IR while minimizing the irrelevance IRR. To achieve this objective, we follow a joint approach composed of Theme and Event based models. While theme-based selection strategy focuses on obtaining influencers with general and historical perspective in this theme, the event-based

selection strategy aims at obtaining influencers particularly related to this special event using multiple metrics.

(I) Theme based Influence Rate Calculation: After fetching the sample of users U using Twitter API, we aim at identifying general influencers with at least 1000 followers within the given theme. Our general influence rate GIR model entails the historical influence achieved for each influencer by adding two ratios each with a predefined weight. The first one is the Followers to Followees percentage (the followers to followees ratio per user from the total ratio of all the selected users). The second one is the percentage of tweets count for each user (the count of tweets per each user from the total count of all tweets provided by all the selected users). Both percentages are calculated during the selected period of the selected theme.

$$G\ I\ R(U) = w_1.F\ F\ R\ (U) + w_2.T\ C\ R\ (U), \qquad \sum_{i=1}^{2} w_i = 1 \tag{3}$$

where FFR corresponds to the followers followee rate and TCR corresponds to the Tweet count rate of user sample. Due to the generality of this approach, multiple influential users with different backgrounds and roles are obtained within the given theme. As a result, we apply bagging algorithm in order to classify this heterogeneous list of influencers into three consistent groups *(Micro, General, Master)*. The general influence classification criteria is being investigated between four main interval thresholds (25, 50, 75, and 100) representing the percentages of the general influence rate.

$$C\ (GIR) = \begin{cases} C_{mic} = 26 < GIR < 50 \\ C_{Gen} = 51 < \ GIR < 75 \\ C_{Mas} = 76 < GIR < 100 \end{cases}$$

The list of influencers L_h related to a certain theme h is then obtained for this theme.

(II) Event based Influence Rate Calculation: The goal of event based model is to select the list of influencers L_e with the highest influence during a certain event e. Fig. 2 depicts the different components of this process (User Meta Rate, User Presence Rate, and Tweets Meta Rate), and the displayed circle sizes reflect the importance or weight of each component. However, those weights may differ from one event to another depending on different factors including topic, importance, category, location, and many other factors. User-centric measures are of key importance as they essentially help to quantify user's engagement. Therefore, we define two types of user-centric measures: (1) User Meta Rate (*UMR*) that represents the meta data points like account age, number of

followers and followees, and (2) User Presence Engagement Rate ($UPER$) that captures the engagement of a user in the particular event e in terms of tweeting frequency and duration. Hence, for the same sample of users U, we aim for detecting the set of users U_e that are mostly engaging in this particular event without having previously been influencers (i.e. before the event e). We fetch all the necessary data corresponding to tweets, users and meta, and then filter this data according to a certain tweet cutoff. In addition to the user-centric measures, the tweets meta rate for a selected user $UTMR$ is another part of this model since the meta rates (retweets, favorites, and replies) that are related to all the tweets initiated by a certain user can define the popularity of this user during a certain event.

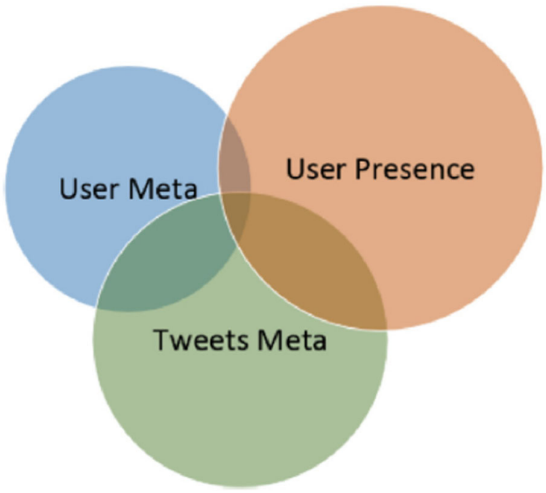

Figure 2 Venn diagram of the multiple measures of the UIR: User meta, user presence, and tweets meta.

(1) User Meta Rate UMR is calculated as follows:

$$UMR(u) = \alpha.SYR(u) + \beta.UFOR(u) + \gamma.UFER(u) + \kappa.ULR(u) \quad (4)$$

where α, β, γ and κ correspond to the weights given to each term respectively. $SYR(u)$ corresponds to the User Since Years Rate, $UFOR(u)$ corresponds to Users Followers Rate, $UFER(u)$ corresponds to the User Following Rate and $ULR(u)$ corresponds to the User Listed Rate. Theses rates are defined as follows:

- User Since Year Rate $SYR(u)$: depicts how long this user has been using this account. A user with a long period of time may have a higher level of engagement or a larger number of

connections than a user with a new registered account. Hence, this user might likely be more influential than a newly registered user. To calculate the User Since Year Rate of a certain user $SYR(u)$, we consider the number of years since the registration of this user over the total number of years since Twitter platform has been created (age of twitter) as follows:

$$SYR(u) = \frac{Number\ of\ Years\ Since\ User\ Registration}{Number\ of\ Years\ Since\ 2006} * 100 \tag{5}$$

- User Followers Rate $UFOR(u)$: Generally, the number of followers can be considered as a key indicator for measuring the influence of a certain user. Since we aim at minimizing the irrelevance IRR of users content, we use this metric for eliminating spams and fake accounts. We calculate this metric by considering the ratio of the number of followers for this particular user over the total number of followers engaged in this particular event:

$$UFOR(u) = \frac{User\ Number\ of\ Followers}{Total\ Number\ of\ Followers\ for\ All\ Users\ engaged\ in\ e} * 100 \tag{6}$$

- User Following Rate $UFER(u)$: is the number of the user followings divided by the total number of all users followings:

$$UFER(u) = \frac{User\ Number\ of\ Followings}{Total\ Number\ of\ Followings\ for\ All\ Users\ engaged\ in\ e} * 100 \tag{7}$$

- The User Listed Rate $ULR(u)$ To obtain this metric, the two lists L_e and L_h are crossed to derive the list of influencers:

$$ULR(u) = \frac{Number\ of\ Listed\ Times}{Total\ Number\ of\ Listed\ Times\ of\ all\ Users} * 100 \tag{8}$$

(2) User Presence Engagement Rate $UPER$: In general, influencers are usually active but not all active users are influencers. The engagement can be considered as a very important factor that affects the total influence. We develop this metric in order to quantify the productivity of a certain user. Posting a large number of tweets is one of the key indicators about the intense level of engagement of the user in an event. We calculate the User Presence Engagement Rate $UPER$ as follows:

$$UPER(u) = \frac{User\ Number\ of\ Tweets * UER(u)}{Total\ Number\ of\ Tweets\ for\ All\ Users} \tag{9}$$

where *UER* corresponds to User Engagement Rate which is defined as follows:

$$UER(u) = \frac{User\ Duration\ of\ Tweeting}{Total\ Duration\ of\ Tweeting\ for\ All\ Users} * 100 \tag{10}$$

where the User Duration of Tweeting per the selected event e is calculated as follows:

$$User\ Duration\ of\ Tweeting = User\ Last\ Tweet\ Time - User\ First\ Tweet\ Time \tag{11}$$

and the total duration of tweeting per the selected event e is calculated as follows:

$$Total\ Duration\ of\ Tweeting = Last\ Tweet\ Time - First\ Tweet\ Time \tag{12}$$

(3) User Tweet Meta Rate UTMR: Quantifying the importance of tweet generated by a certain user is very critical in determining the influence of this user and hence, we define User Tweet Meta Rate *UTMR* in terms of retweets and favorites rate achieved by a certain initiated tweet as follows:

$$UTMR(u) = 60\%URR(u) + 40\%UFR(u) \tag{13}$$

where the User Retweets Rate *URR* and User Favorites Rate *UFR* are calculated as follow:

$$UFFR(u) = \frac{User\ Number\ of\ Retweets}{Total\ Number\ of\ Retweets\ for\ All\ Users} * 100 \tag{14}$$

$$UFR(u) = \frac{User\ Number\ of\ Favorites}{Total\ Number\ of\ Favorites\ for\ All\ Users} * 100 \tag{15}$$

To calculate the user influence rate $IR(u)$, we combine *UMR*, *UPER s*and *UTMR* as follows:

$$IR(u) = \omega_1.UMR(u) + \omega_2.UPER(u) + \omega_3.UTMR(u) \qquad \sum_{i=1}^{3} w_i = 1 \tag{16}$$

where ω_1, ω_2 and ω_3 corresponds to the weights of each user meta rate. Crossing each of the event-based model and theme-based model leads to a list of influencers $\mathcal{L}_{e,h}$.

4.2. Reputation maximization

We further consider sampling out of the selected influencers $\mathcal{L}^t_e$ to apply a second step with the objective of maximizing the selected influencers. We select a list of influencers (please refer to Fig. 1 *Case Study User List 1.1*) out of the user list $\mathcal{L}_{e,t}$. The reputation rate of user u is defined as follows:

$$RepR(u) = \omega_1.IGRepR(u) + \omega_2.IRepR(u) \quad \sum_{i=1}^{2} \omega_i = 1 \tag{17}$$

where *IGRepR* corresponds to the Influencer General Reputation Rate and *IRepR* corresponds to the Influencer Reputation Rate. The Influencer General Reputation Rate *IGRepR* specifies the reputation rate of the selected influencer before the occurrence of the event and thus gets a historical reputation about this user. We calculate the *IGRepR* in terms of Influencer General Impact rate *IGIR* and Influencer General Tweets Credibility rate *IGCR* based on user theme tweets as follows:

$$IGRepR(u) = \left(\omega_1.IGIR(u) + \omega_2.IGCR(u)\right) \quad \sum_{i=1}^{2} \omega_i = 1 \tag{18}$$

IGIR is calculated as follows:

$$IGIR(u) = \frac{\left|T_u^h\right| * |f^{\rightarrow}(u)| + rt(u,h) + fv(u,h)}{Total\ Impact\ of\ All\ Users} * 100 \tag{19}$$

where $|T_u^h|$ corresponds to the User number of tweets per theme h, $|f^{\rightarrow}(u)|$ corresponds to the number of user followers, $rt(u, h)$ corresponds to the number of user retweets per theme and $fv(u, h)$ corresponds to the number of user favorites per theme.

IGCR is calculated as follows:

$$IGCR(u) = \begin{cases} 1 * \left(\frac{Number\ of\ user's\ tweets\ per\ theme}{Total\ number\ of\ tweets}\right) * 100 & if\ u\ is\ credible \\ 0 & otherwise \end{cases} \tag{20}$$

We also aim at assessing the reputation of this selected influencer in the context of this event and hence, we further calculate the Influencer Reputation Rate *IRepR* as follows:

$$IRepR(u) = \omega_1 . IIR(u) + \omega_2 . IPCR(u) \qquad \sum_{i=1}^{2} \omega_i = 1 \tag{21}$$

where IIR corresponds to the Influencer Impact rate and IPCR corresponds to the Influencer Profile Credibility rate.

IIR is defined as follows:

$$IIR(u) = \frac{|T_u^e| * |f^{\rightarrow}(u)| + rt(u,e) + f\ v(u,e)}{Total\ Impact\ of\ All\ users} * 100 \tag{22}$$

where $|T_u^e|$ corresponds to the User Number of tweets per event e,$|f^{\rightarrow}(u)|$ corresponds to the number of user's followers, $rt(u, e)$ corresponds to the number of user's retweets per event e and $fv(u, e)$ corresponds to the number of user's favorites per event.

IPCR is defined as follows:

$$IPCR(u) = \begin{cases} 1 * \left(\frac{Number\ of\ user's\ tweets\ per\ event}{Total\ number\ of\ tweets} \right) * 100\ if\ u\ is\ credible \\ 0 \qquad\qquad otherwise \end{cases} \tag{23}$$

5. Implementation & experiments

In this section, we provide the implementation, data processing and different performance analysis realizing the proposed methodology and scheme. COVID-19 is chosen as a main event. We show the performance analysis of the experiments performed individually for each model followed by the overall findings and results.

5.1. Topic selection

The adopted selection methodology started by selecting the event topic and choosing its relevant hashtags and keywords that were used to build the basic twitter search query in order to get the targeted results. Afterwards, the system collected one million unique tweets with all their relevant

unique user profiles. Once the tweets and user profiles were fetched, additional aggregations and labels were added to each tweet and/or profile, which indicates a few additional attributes and classifications. In the sequel, we elaborate on the aforementioned action steps based on the proposed methodology.

- One million public tweets were collected using a hard-coded python script that uses Tweepy [57]. All tweets contain at least one of the following keywords (terms not hashtags) ‘‘corona’’, ‘‘covid’’, or ‘‘sarscov2’’, which are the most used keywords during the COVID-19 global event. After fetching all tweets, the unique users were listed to be collected using Twitter REST API V1 [58] access tokens in order to be analyzed and classified at later stages.
- In order to classify tweets whether they are corona related or non-corona related based on their contexts, an ontology/lexicon of common related keywords/terms was produced and used within the NLP entity extractor modules. Below are the different lexicons used during our study to label tweets and profiles based on the NLP results: Corona Top Used Hashtags Lexicon, Corona Social Media Context Lexicon for Tweets, Occupation Lexicon for Grouping Users Based on their Biographic Information, Medical Occupation Lexicon for Users and Virus Specialty Occupation Lexicon for Users.
- After building all the aforementioned lexicons, each single tweet was labeled as ‘‘isCorona’’ tweet (i.e. the tweet content contains Corona related keywords) or ‘‘isNotCorona’’ tweet (i.e. the tweet content does not contain Corona related keywords), and each user profile was labeled as ‘‘isMedical’’ (i.e. user profile contains medical keywords) or ‘‘isSpecialty’’ (i.e. user profile contains virus and vaccine specialization keywords) based on the users claimed biography information. In addition to that, a specific occupation field was added to each user based on their claimed biography details as well.
- An enriched dataset was next designed to help aggregating and creating more relationships between all the data points and the available classifications. The obtained dataset is structured according the following features: TweetID, TweetHashtagCount, TweetFavorates, TweetRetweets, TweetMentions, TweetTotalInteractions (i.e. favorites plus retweets), TweetTotalReach (i.e. number of followers for the tweet’s user), UserID, User-ClaimedLocation, UserOccupation (extracted from the user’s profile), isCoronaTweet, isMedicalProfileUsers, and isSpecialtyProfileUser.

5.2. Data processing

We present in the following additional description of the proposed scheme presented in Fig. 1 in the following:

- *Event Selection:* The event parameters such as the event, location, language, and duration were selected. We selected ‘‘COVID-19’’ as our event with global location and restricted the language to English. The data was collected from Jan 25 2020 to March 20 2020 with a size equal to 1 million unique tweet and 288.4 thousand unique user.
- *Event Data Collection:* This is done through querying all tweets containing at least one of the following terms: ‘‘covid, corona, sarscov’’ over Twitter public blog.
- *Event Data Cleaning and Filtering:* GCP DataFlow and BigQuery were used to manage and execute sequence of steps related to cleaning and filtering the tweets.
- *Event Data Analysis:* GCP DataFlow, BigQuery Analysis, and custom python scripts were used to analyze the tweets.
- *Launching Core Event-Based Algorithms:* The core implemented algorithms to calculate the Influencer Rating, Classifying Influencers, and Influencers Credibility and Impact Measurements were launched in the form of scripts.
- *Event Exported Datasets:* The inferred results included a dataset of 1 million unique tweets, dataset of top 1000 Influencers (ordered by number of event-tweets), and a sorted list of 1000 Event Influencers.
- *Theme Selection:* The theme parameters such as the event, location, language, duration were selected. We chose ‘‘Medical, Virus’’ as our event with global location and restricted the language to English. Our data was collected from Oct 01 2020 to Dec 05 2020 with a size equal to 850 thousand unique tweets and 401.7 thousand unique user.
- *Theme Data Collection:* We all tweets containing at least one of the keywords: ‘‘medic, virus, vaccine’’ were extracted from Twitter Public Blog.
- *Theme Data Cleaning and Filtering:* GCP DataFlow and BigQuery were used to manage and execute sequence of steps related to cleaning and filtering the tweets.
- *Theme Data Analysis:* GCP DataFlow, BigQuery Analysis, and custom python scripts were used to analyze the tweets.
- *Launching Core Theme-Based Algorithms:* The core implemented algorithms to calculate the General Influence Rates, classify General Influencers, and calculate activity ratios for all theme influencers are launched in the form of scripts.
- *Theme Exported Datasets:* The inferred results included a dataset of 1 million unique tweets, dataset of top 1000 General Influencers ordered by number of followers and a sorted list of 1000 General Influencers.
- *Joint Theme-Event Exported Datasets:* After calculating the influence rates and maximizing the reputation for the selected influencers, the following joint theme-event datasets were obtained: a dataset of all event influencers with maximized influence ratios after merging theme-based general influence rates, a dataset of the selected 100 influencers with maximized

reputation ratios, and finally, a sorted dataset of 100 influencers based on the final Reputation Rate from the selected event.

5.3. Event based analysis

We first extracted around 1 million tweets for a total number of users equal to 288,439, a total number of followers equal to 21,393,493 and a total number of interactions equal to 36,964,431,381. Fig. 3 depicts influencers in the context of COVID event. As the figure shows, top influencers include *EclipseMist*, *evankirstel*, and *myamigouk* with INR equal to 41.92, 37.21, and 35.27 respectively.

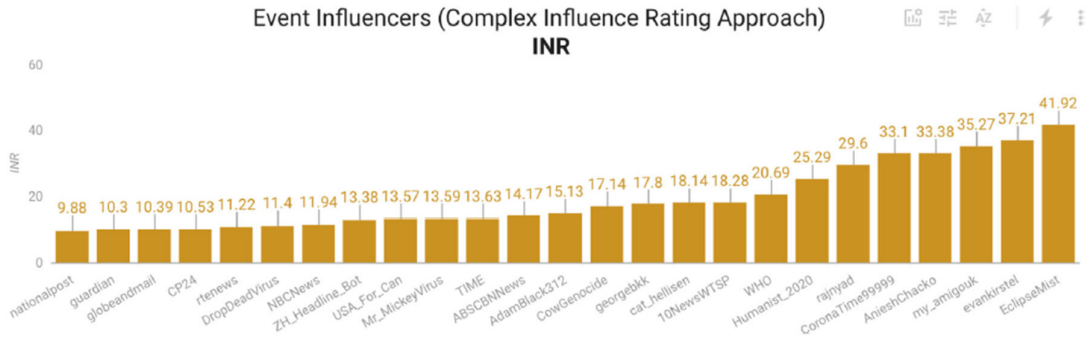

Figure 3 Event influencers using INR proposed calculations.

We further show in Fig. 4 the top 1000 influencers with their corresponding UMR, UPER, and UTMR ratios (refer to Section 4 for more details). The figure illustrates the non-linear and dense relationship among those three components.

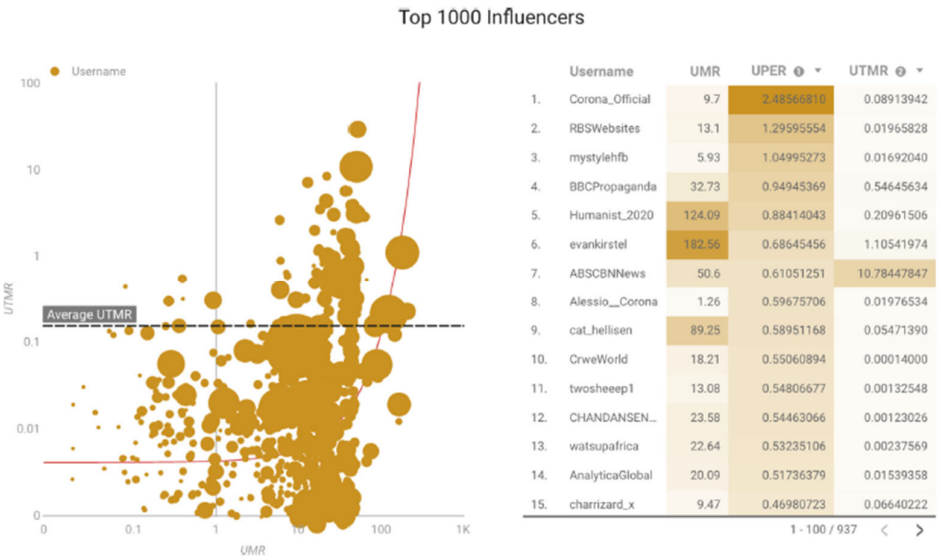

Figure 4 Top 1000 influencers.

As Fig. 4 illustrates, the distribution of the UMR, UPER and UTMR vary from one user to another, recording different username for the three different maximums. For example, UPER is the highest for user numbered 1, the UMR is the highest for user numbered 6 and the UTMR is the highest for user number 7. This due to the fact that each user has different tweeting behavior and characteristics.

5.4. Theme based analysis

We extract around 1 million tweets with a total number of users equal to 391,795, a total number of followers equal to 16,973,449,114 and a total number of interactions equal to 6,950,323. Fig. 3 depicts influencers in the context of COVID event. As shown in the figure, the top influencers include *UberFacts*, *VoceNaoSabiaQ*, and *ANI* with INR equal to 335.6, 90.39, and 83.35 respectively (see Fig. 5).

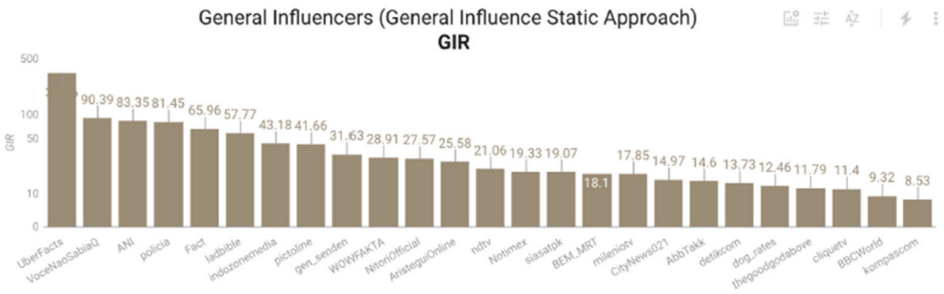

Figure 5 Selecting influencers using the general approach

We further show in Fig. 6 the top 1000 influencers with their corresponding Followers, Tcr and FFr. A non-linear relationship is also depicted when using this approach but with less density.

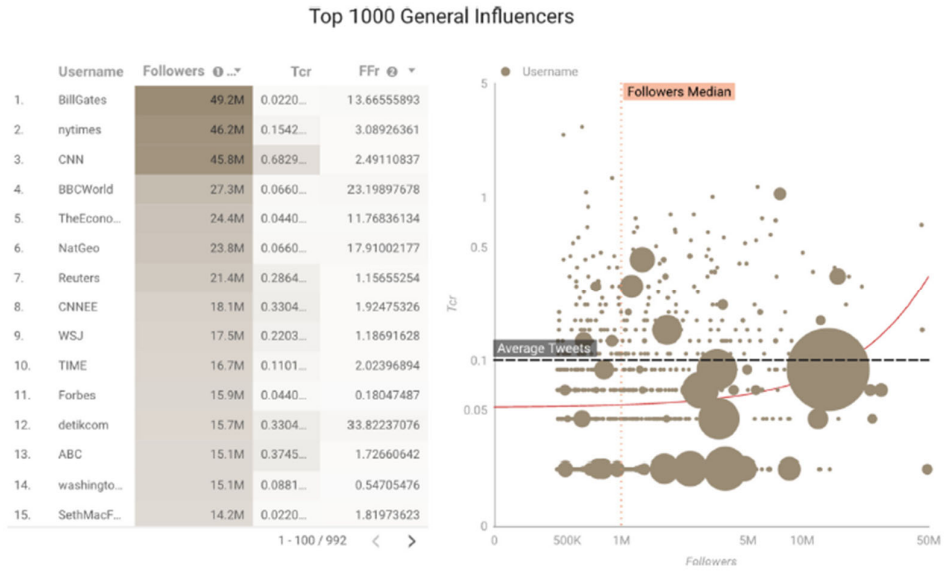

Figure 6 Top 1000 general influencers.

5.5. Joint event and theme based analysis

The resulting list of influencers using the joint theme and event based scheme is depicted in Fig. 7. As depicted, the GIR alone fails to capture influencers in the context of COVID while the INR achieves a higher level of accuracy. The figure shows further analysis on the scale of tweets achieved by influencers captured by each of the event and theme based approaches individually. Selected influencers in the theme based approach reach an average of 4.54 tweets per influencer (0.45%), which is less than that of the event based approach with 143.27 tweets per influencer (13.56%). This signifies the impact of influencers selected when tweeting in terms of the total number of users.

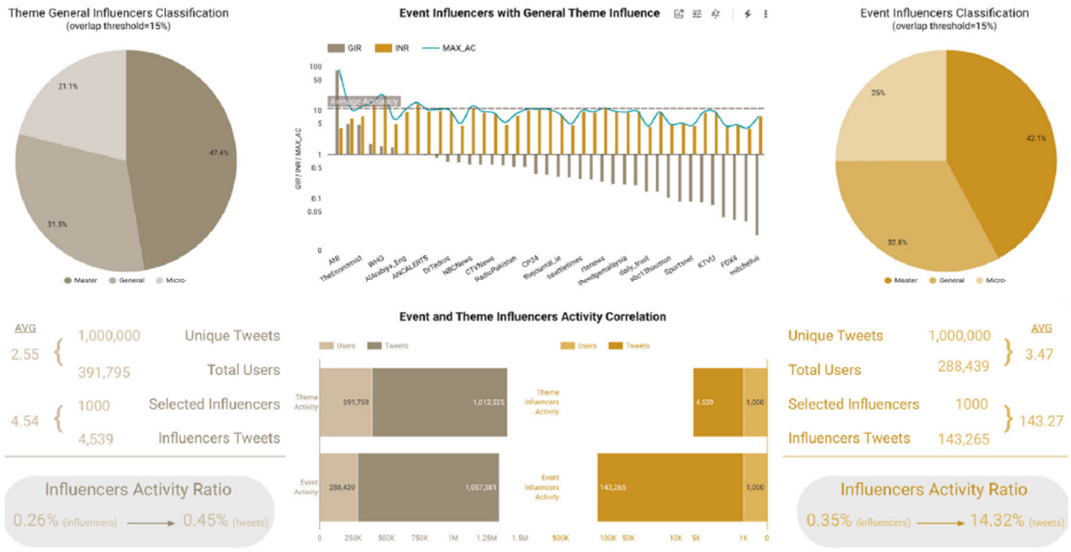

Figure 7 List of influencers of the proposed joint model.

Furthermore, the credibility of those top influencers is measured in terms of content credibility, profiles, virus clinical profile and company in Fig. 8. The percentage of content credibility is low in comparison with the total content released about COVID. This shows that most influencers did not have credible content. Similarly, medical profiles record a low percentage in the set of influencers. This might be due to the fact that most influencers pertain to the group of journalists and news agencies that release information.

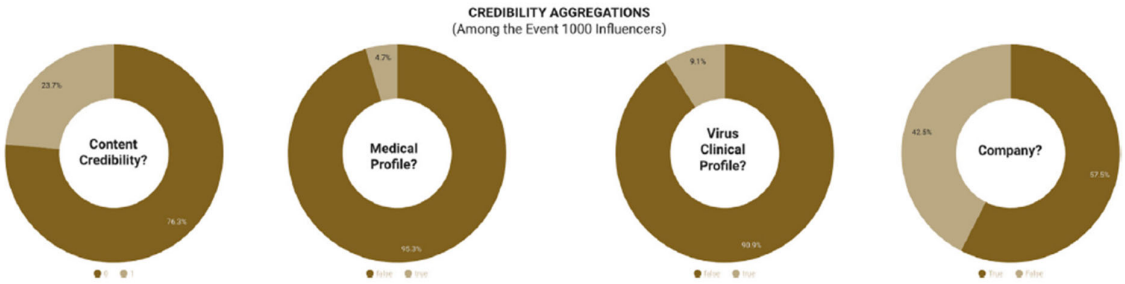

Figure 8 Credibility analysis and aggregations for selected influencers of joint approach.

5.6. Reputation based analysis

In this subsection, we compute the reputation for the Top 100 influencers. Fig. 9 shows the distribution of reputation for both Influencer General Reputation Calculation (INGEPr) and Influencer Reputation Calculation (INREPr).

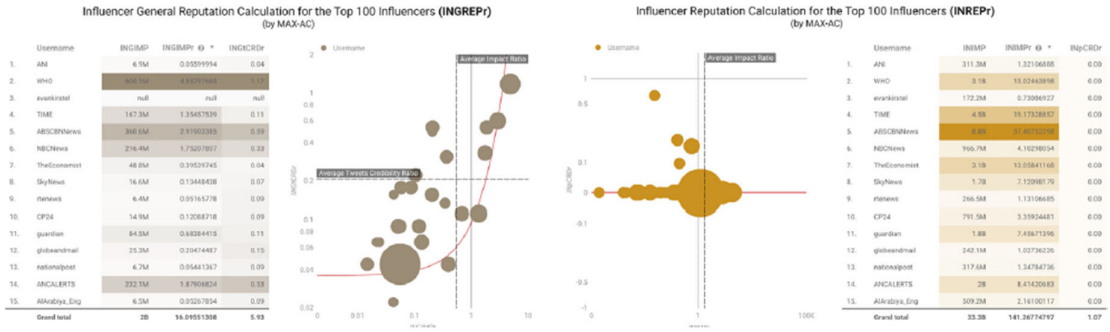

Figure 9 Comparison between list of selected influencers with influencer general reputation and influencer reputation calculation.

Figs. 10 and 11 list the main influencers while highlighting major components of the reputation based approach.

	Username	INREPr	INGREPr	REPR
1.	ABSCBNNews	14.96300919	1.52452035	14.96300919
2.	TIME	7.66931543	0.60792401	7.66931543
3.	TheEconomist	5.22336467	0.18459652	5.22336467
4.	WHO	5.20985559	2.64378550	5.20985559
5.	ANCALERTS	3.36568273	1.06887779	3.36568273
6.	guardian	2.98268558	0.33963151	2.98268558
7.	SkyNews	2.84839271	0.09345006	2.84839271
8.	htTweets	1.67028648	0.21784030	1.67028648
9.	NBCNews	1.64119222	0.89911279	1.64119222
10.	CP24	1.34369793	0.10122995	1.34369793
11.	IndiaToday	1.23668280	0.03031314	1.23668280
12.	AlArabiya_Eng	0.86440047	0.07394650	0.86440047
13.	CBCAlerts	0.61783173	0.04236259	0.61783173
14.	nationalpost	0.53913894	0.07464055	0.53913894
15.	ANI	0.52842755	0.04883752	0.52842755
		0 10	0 1 2	0 10

Figure 10 List of influencers with reputation maximization.

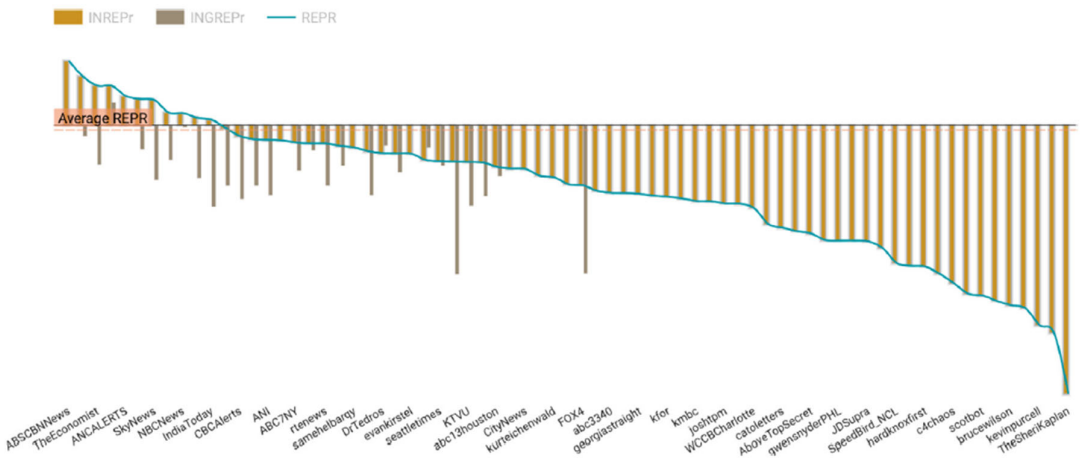

Figure 11 Multiple bar charts showing influencers with reputation maximization.

Fig. 12 compares the achieved credibility of top influencers between the event and theme approaches in terms of their provided content, medical profile, virus clinical profile and company. As the figure depicts, the credibility of event-based influencers is much higher than the credibility of theme-based influencers through four dimensions.

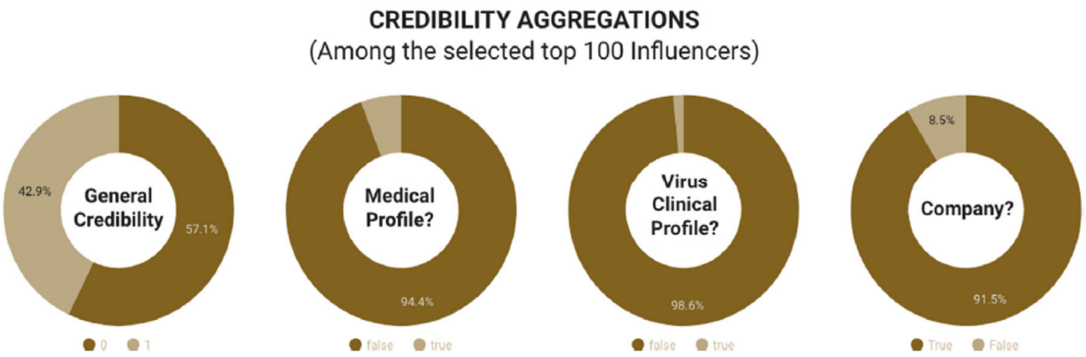

Figure 12 Credibility analysis and aggregations for list of influencers with maximized reputation.

5.7. Discussion

Finding relevant event influencers can help in understanding different user behaviors and thus recommending different strategies and further research enhancements and directions.

Referring to other approaches that use general influence data points to rate influencers on Twitter, we used some of which mentioned in the literature review to select the general influencers of the theme that our event case study belongs to. Taking into account that our approach is to find influencers of a specific event and that the event has a timeframe and a location, it is also important to mention that narrowing down into those specific event attributes means that influencers might change based on those attributes and that is hypothetically correct. However, general influencers are supposed to be always influencers based on their FF ratio or influenced network for instance, and that is hypothetically not correct from an event-based perspective. While investigating the lists of users selected and classified by the general influence rating approaches (please see the 6 steps on the left side of Fig. 1 diagram or check steps 1, 3, and 5 in Section 2)I, we were able to discover users that are influencers in general but not necessarily being influencers in our specific event (Covid-19) and yet some of which do not have any intervention within the event community discussion. In fact, that was the reason why we introduced a specific event based approach while adding more event-related meta data as constraints while finding event specific influencers, which are clearly defined in steps 2, 4, 6, 7, and 8 in Section 3. Comparing results of the mentioned classifications, we were able to distinguish the classification and the rating based on the users' interaction level with both the theme and the event topics. For instance, being a country president makes the user an influencer in that specific country. However, that does not mean that the same user is an influencer in a topic that might need specific field knowledge. Moreover, using the general influence rating methods while selecting theme influencers helped us to generate a ground truth layer to verify our model enhancements. Nevertheless, after calculating the influence ratios for the list of users that belongs to the event, we

have found that some users can be highly influencing the event community based on the interaction and the reach level of their event-based contribution (e.g. comments, posts), while having less number of followers and tweets in general. In addition, adding a layer of reputation and credibility (refer to steps 9 to 17 in Section 3) enhanced our approach by looking at the users historical engagement about the same topic (i.e., event) and its theme at the same time. In fact, this was very useful to verify the results of the maximization claim by calculating the reputation and the credibility of the top 10% of the users. The theoretical analysis and manual verification of the final exported list of event influencers explore that our approach is capable of fine tuning the results and eliminating the user inactive in the field compared to the mid-layer theme oriented list of users provided by other theme-oriented approaches. By selecting samples of influencing users based on several theme-based approaches, we were able to define through counter examples the ground-truth results used to assess our approach and determine the potential improvements.

In the sequel, we provide a through discussion and conclusions based on the presented experimental results. To start with, influencers are not just those who have large numbers of follower. Even if that has a high impact ratio in general, it does not mean that they have to be the relevant influencers and/or the leaders of a certain event. Results explore that after selecting and sorting out 1000 influencers from the selected event based on their engagement (i.e. number of tweets within the selected event time-frame), some of the names that showed at the beginning of the list continued to show at the final exported list. At the same time, most of the selected 1000 theme influencers that were at the very top of the sorted list based on their followers count, disappeared in the next maximization levels. These results illustrate that the user number of followers is not necessarily a main factor in the whole calculation and evaluation.

Moreover, when calculating event influence rates, we took into account user meta information and assigned to it a weight representing the level of importance in which a set of data points can contribute to the main influence calculation. Presence Engagement is the most important factor which reflects the activity of the user in the event lifetime. Engagement percentages of all users add up to 100%. Users with higher percentages are the most engaging ones during the event. In addition, Tweets Meta Rate representing the level of interactions with those tweets (i.e. replies, retweets, likes) reflects their impact. In other words, the more a tweet has interactions, the more it will be impacting the social network with more spreading. This is a primary factor of the calculation, but it does not imply the nature of the impact (i.e positive, negative,..). Another highlight in the GIR results of Fig. 6 is the linear relationship between the high count of followers. In other words, the number of user followers depicts the real influence and vise versa. But what if those users have a very rare number of tweets or engagement level? There is no relationship between being influencer and being actively engaging in that scenario. Accordingly, the assumption of being influencers is still valid for having high FFr, but

must have another supporting influence indicator like the TCr. Moreover, Fig. 3 shows the non-relational connection between the three calculated rates UMR, UPER, and UTMR since they are not in a linear relationship. These results illustrates that each one of the mentioned factors has its own weight driving the user to the top or the bottom of the final sorted list of influencers. Furthermore, correlating historical influence rates with the selected event influencers can maximize the accuracy of the aforementioned assumption. For instance, some of the selected event influencers might have historical general influence in similar events or at least in the same field. That way, we could support our finding and maximize the accuracy of the calculation when both GIR and INR for a selected user U are calculated. In other words, a selected event user with a general influence rate GIR greater than zero would lead to a higher influence rate (see Fig. 7). However, finding event influencers by mixing event and theme influence findings might still lead to some inaccurate classification. To reduce that, the reputation factor was necessary to correlate both user tweets content and user profile credibility. Not to forget, the content and profile impact/popularity analysis within the event time-frame and the theme time-frame was also a credible badge to add to the model. Based on the mentioned facts, we can conclude that the influence rates for theme or event users may vary upon their credibility and impact ratios and findings. For instance, a journalist, who has a large network of followers and a high level of engagement during a specific event (e.g. Covid 19) and whom his/her tweets get a large number of interactions, is clearly considered as an event influencer. On the other hand, the same user might be giving advice about COVID-19 during the event for the first time while his/her profile does not match the content and the event topic engaging in. In this case, his/her profile is not credible enough to talk about fields such as Medical or Healthcare. Similar tweets would have higher content credibility ratios if they belong to other credible users. Hence, the final list of influencers can change based on the reputation rate sorting as depicted in Fig. 11. In addition, the reason behind choosing two time frames while investigating users' influence instead of using the same time frame for both Theme and Event data, is that when dealing with credibility, historical measurement is needed to see whether a user is credible to speak about this event from being credible when speaking with similar events (theme) in the past. This would help in assessing whether there might be a chance of bot-spam activity introduced during the event, which potentially allowed the user who is interested to join the trend and publish intentional or unintentional misleading content to join easily. This also can explain on the other hand, while assessing a ''trend'' like covid-19 back then, people who might have historical presence in similar events, were potentially having less chances of becoming spammers or irrelevant.

As a conclusion, it is clear that finding event-based influencers is still a challenging problem depending on the event itself and how wide it could be when trying to find a global influencer or else trying to find an influencer in a specific location. In addition to that, our case study meant to be using one language (English), while if we add more languages, this will make it harder to investigate the content credibility for instance. As a compound problem, correlating findings from historical user

activities was very helpful to obtain better results and reduce accuracy limitations while sorting event influencers.

6. Conclusion & future directions

This paper addressed the problem of finding relevant Twitter influencers within specific events. In this context, we proposed a novel joint theme and event based data-driven rating scheme to maximize the accuracy of identifying relevant influencers on a global event. User profile reputation and credibility in specific context were also considered in the proposed model for enhancing the inferred results. Covid-19 was adopted as a case study for a specific period on Twitter. We performed extensive experiments followed by a through analysis and discussion illustrating the relevance of our proposed solutions. Users, tweets, profiles, history, event, theme, and correlations were all dimensions considered in our approach. It is worth to mention that our proposed scheme can be adapted to diversity of events, languages, regions, and time-frames. Although Covid-19 may potentially be a wider theme of other events like ‘‘vaccination’’ campaigns and ‘‘postpandemic’’ regulations or topics being discussed on social media, it can be treated as both an Event in a wider Theme, or a Theme having sub events. For the sake of applying our methodology on a trendy event on social media, we have treated Covid-19 as an event. Finally, we present in the sequel some important analytical questions that can be answered through our proposed scheme:

- How might a user with very high engagement rate have less reputation than others in a certain event?
- Can influencers in events be classified credible without having historical activities about the event-theme?
- How credible are those influencers within their networks both in general and based on their profiles?
- How to differentiate between being a general or specific event’s influencer?
- What are the main drivers of being an event influencer?
- Can an influencer with very large network and reach level be considered as non-influencer?
- Can an influencer with very small network be considered as an event influencer?
- How can we find hidden influencers in similar events?
- How does time affect being an influencer?

As future research directions, considering NLP-Based analysis of the sentiment and content meanings in the influence identification model may form a very promising extension to improve the proposed scheme. Studying and quantifying the influence of the engaged groups (e.g. organizations, professional entities) on certain audiences may also be considered another important extension.

Moreover, embedding graph network analysis by applying interconnected relationship influence ratios may constitute a great enhancement. Finally, thinking of a continuous evaluation for such a theme-event approach, adding more events in the future while assuming that the current event will be a theme of nested events, might be very helpful to inherit influence ratios over time.

CRediT authorship contribution statement

Ali Srour: Idea, Architecture, Design, Experimentation, Elaboration, Writing of the paper, Implementation of the experiments.

Hakima Ould-Slimane: Idea, Architecture, Design, Experimentation, Elaboration, Writing of the paper.

Azzam Mourad: Idea, Architecture, Design, Experimentation, Elaboration, Writing of the paper.

Haidar Harmanani: Idea, Architecture, Design, Experimentation, Elaboration, Writing of the paper.

Cathia Jenainati: Idea, Architecture, Design, Experimentation, Elaboration, Writing of the paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. Kempe, J. Kleinberg, E. Tardos, Maximizing the spread of influence through a social network, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '03, ACM, New York, NY, USA, 2003, pp. 137–146, [Online]. Available: https://doi.org/10.1145/956750.956769.

[2] L. Adamic, E. Adar, How to search a social network, Social Networks 27 (3) (2005) 187–203, [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378873305000109.

[3] X. Song, B.L. Tseng, C.-Y. Lin, M.-T. Sun, Personalized recommendation driven by information flow, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, in: SIGIR '06, Association for Computing Machinery, New York, NY, USA, 2006, pp. 509–516, [Online]. Available: https://doi.org/10.1145/1148170.1148258.

[4] J. Clement, Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2019, 2019, Available at https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.

[5] A. Java, X. Song, T. Finin, B. Tseng, Why we Twitter: Understanding microblogging usage and communities, in: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, ACM, New York, NY, USA, 2007, pp. 56–65, [Online]. Available: https://doi.org/10.1145/1348549.1348556.

[6] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, M. Yamir, The Dynamics of Protest Recruitment through an Online Network, Scientific Reports, 2011, p. 197.
[7] J. Borge-Holthoefer, Y. Moreno, Absence of influential spreaders in rumor dynamics, Phys. Rev. E 85 (2012) 026116, [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.85.026116.
[8] G. Stilo, P. Velardi, A.E. Tozzi, F. Gesualdo, Predicting flu epidemics using Twitter and historical data, in: D. Śle¸zak, A.-H. Tan, J.F. Peters, L. Schwabe (Eds.), Brain Informatics and Health, Springer International Publishing, Cham, 2014, pp. 164–177.
[9] M.A. Abebe, J. Tekli, F. Getahun, R. Chbeir, G. Tekli, Generic metadata representation framework for social-based event detection, description, and linkage, Knowl.-Based Syst. 188 (2020) 104817.
[10] M. Arafeh, P. Ceravolo, A. Mourad, E. Damiani, E. Bellini, Ontology based recommender system using social network data, Future Gener. Comput. Syst. 115 (2021) 769–779.
[11] M. Arafeh, P. Ceravolo, A. Mourad, E. Damiani, Sampling online social networks with tailored mining strategies, in: Proceedings of Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2019, pp. 217–222.
[12] A. Mourad, A. Srour, H. Harmanani, C. Jenainati, M. Arafeh, Critical impact of social networks infodemic on defeating coronavirus COVID-19 pandemic: Twitterbased study and research directions, IEEE Trans. Netw. Serv. Manag. 17 (4) (2020) 2145–2155.
[13] V. Nebot, F. Rangel, R. Berlanga, P. Rosso, Identifying and classifying influencers in Twitter only with textual information, in: M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, F. Meziane (Eds.), Natural Language Processing and Information Systems, Springer International Publishing, Cham, 2018, pp. 28–39.
[14] F.N. Abu-Khzam, K. Lamaa, Efficient heuristic algorithms for positive-influence dominating set in social networks, in: Proceedings of IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2018, pp. 610–615.
[15] A. Nsouli, A. Mourad, D. Azar, Towards proactive social learning approach for traffic event detection based on arabic tweets, in: Proceedings of 14th International Wireless Communications Mobile Computing Conference (IWCMC), 2018, pp. 1501–1506.
[16] B. Halawi, A. Mourad, H. Otrok, E. Damiani, Few are as good as many: An ontology-based tweet spam detection approach, IEEE Access 6 (2018) 63890–63904, http://dx.doi.org/10.1109/ACCESS.2018.2877685.
[17] Y. Riyanto, J. Yeo, Directed trust and trustworthiness in a social network: An experimental investigation, J. Econ. Behav. Organ. 151 (C) (2018) 234–253, [Online]. Available: https://EconPapers.repec.org/RePEc:eee:jeborg:v:151:y:2018:i:c:p:234-253.
[18] X. Li, Y. Liu, Y. Jiang, X. Liu, Identifying social influence in complex networks: A novel conductance eigenvector centrality model, Neurocomputing 210 (2016).
[19] A. Gün, P. Karagöz, A hybrid approach for credibility detection in Twitter, in: Proceedings of Hybrid Artificial Intelligence Systems, Springer International Publishing, Cham, 2014, pp. 515–526.

[20] L. Jain, R. Katarya, S. Sachdeva, Opinion leader detection using whale optimization algorithm in online social network, Expert Syst. Appl. 142 (2020).
[21] M. Brede, How does active participation affect consensus: Adaptive network model of opinion dynamics and influence maximizing rewiring, Complexity 2019 (2019) http://dx.doi.org/10.1155/2019/1486909.
[22] S. Zhang, Y. Cai, H. Xia, A privacy-preserving interactive messaging scheme based on users credibility over online social networks, in: IEEE/CIC International Conference on Communications in China (ICCC), 2017, pp. 1–6.
[23] J. Cetkovic, S. Lakić, M. Lazarevska, M. arković, S. Vujošević, J. Cvijović, M. Gogić, Assessment of the real estate market value in the European market by artificial neural networks application, Complexity 2018 (2018) 1–10.
[24] M. Alrubaian, M. Al-Qurishi, M.M. Hassan, A. Alamri, A credibility analysis system for assessing information on Twitter, IEEE Trans. Dependable Secure Comput. 15 (4) (2018) 661–674.
[25] S.A. Ríos, F. Aguilera, J.D. Nuñez-Gonzalez, M. Graña, Semantically enhanced network analysis for influencer identification in online social networks, Neurocomputing 326–327 (2019) 71–81.
[26] Y. Liu, J. Cao, IIRank: A Novel algorithm for identifying influencers in microblog social networks, in: International Conference on Data Mining Workshops (ICDMW), 2019, pp. 735–740.
[27] H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a social network or a news media? in: WWW '10: Proceedings of the 19th International Conference on World Wide Web, ACM, New York, NY, USA, 2010, pp. 591–600.
[28] M. Cha, H. Haddadi, F. Benevenuto, K.P. Gummadi, Measuring user influence in twitter: The million follower fallacy, in: Proceedings of ICWSM, 2010.
[29] M. Luiten, W.A. Kosters, F.W. Takes, Topical influence on Twitter : A feature construction approach, 2012.
[30] J. Weng, E.P. Lim, J. Jiang, Q. He, Twitterrank: Finding topic-sensitive influential Twitterers, in: In Proceedings of the Third ACM International Conference on Web Search & Data Mining, 2010.
[31] F. Riquelme, P. González-Cantergiani, X. Molinero, M. Serna, Centrality measure in social networks based on linear threshold model, Knowl.-Based Syst. 140 (2018) 92–102.
[32] H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a social network or a news media?, in: Proceedings of the 19th International Conference on World Wide Web, 2010, pp. 19.
[33] Y.E. Riyanto, Y.X. Jonathan, Directed trust and trustworthiness in a social network: An experimental investigation, J. Econ. Behav. Organ. 151 (2018) 234–253.
[34] M. Tsikerdekis, S. Zeadally, Multiple account identity deception detection in social media using nonverbal behavior, IEEE Trans. Inf. Forensics Secur. 9 (8) (2014) 1311–1321.
[35] S.A. Curiskis, B. Drake, T.R. Osborn, P.J. Kennedy, An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit, Inf. Process. Manage. 57 (2) (2020) 102034, [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457318307805.

[36] B.A. Huberman, D.M. Romero, F. Wu, Social networks that matter: Twitter under the microscope, 2008, CoRR, abs/0812.1045 [Online]. Available: http://arxiv.org/abs/0812.1045.
[37] R. Cappelletti, N. Sastry, IARank: RAnking users on Twitter in near real-time, based on their information amplification potential, in: International Conference on Social Informatics (SocialInformatics), IEEE Computer Society, Los Alamitos, CA, USA, 2012, pp. 70–77, [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/SocialInformatics.2012.82.
[38] I. Anger, C. Kittl, Measuring influence on Twitter, in: Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, in: i-KNOW '11, Association for Computing Machinery, New York, NY, USA, 2011, [Online]. Available: https://doi.org/10.1145/2024288.2024326.
[39] E. Bakshy, J.M. Hofman, W.A. Mason, D.J. Watts, Everyone's an influencer: Quantifying influence on Twitter, in: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, in: WSDM '11, Association for Computing Machinery, New York, NY, USA, 2011, pp. 65–74, [Online]. Available: https://doi.org/10.1145/1935826.1935845.
[40] M. Anjaria, R.R. Guddeti, Influence factor based opinion mining of Twitter data using supervised learning, in: 2014 6th International Conference on Communication Systems and Networks, COMSNETS 2014, 2014, pp. 1–8.
[41] C. Schenk, D. Sicker, Finding event-specific influencers in dynamic social networks, 2011, pp. 501–504.
[42] Y. Mei, Y. Zhong, J. Yang, Finding and analyzing principal features for measuring user influence on twitter, in: Proceedings of 2015 IEEE First International Conference on Big Data Computing Service and Applications, 2015, pp. 478–486.
[43] X. Li, Y. Liu, Y. Jiang, X. Liu, Identifying social influence in complex networks: A novel conductance eigenvector centrality model, Neurocomputing 210 (2016) 141–154, SI:Behavior Analysis In SN. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231216305860.
[44] E. Lahuerta-Otero, R. Cordero-Gutiérrez, Looking for the perfect tweet. The use of data mining techniques to find influencers on twitter, Comput. Hum. Behav. 64 (2016) 575–583, [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0747563216305258.
[45] P. Sharma, A. Agarwal, N. Sardana, Extraction of influencers across Twitter using credibility and trend analysis, in: Eleventh International Conference on Contemporary Computing (IC3), 2018, pp. 1–3.
[46] T. Huynh, I. Zelinka, X.H. Pham, H.D. Nguyen, Some measures to detect the influencer on social network based on information propagation, in: Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics, in: WIMS2019, Association for Computing Machinery, New York, NY, USA, 2019, [Online]. Available: https://doi.org/10.1145/3326467.3326475.

[47] S. Jain, A. Sinha, Identification of influential users on Twitter: A novel weighted correlated influence measure for Covid-19, Chaos Solitons Fractals 139 (2020) 110037, [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960077920304355.
[48] S. Alqurashi, A. Alashaikh, E. Alanazi, Identifying information superspreaders of COVID-19 from arabic tweets, 2020.
[49] C. Fombrun, N. Gardberg, J. Sever, The reputation quotientsm: A multistakeholder measure of corporate reputation, J. Brand Manage. 7 (2013) http://dx.doi.org/10.1057/bm.2000.10.
[50] M. El Marrakchi, H. Bensaid, M. Bellafkih, Scoring reputation in online social networks, in: Proceedings of 10th International Conference on Intelligent Systems: Theories and Applications (SITA), 2015, pp. 1–6.
[51] M. Cha, H. Haddadi, F. Benevenuto, K.P. Gummadi, Measuring user influence in twitter: The million follower fallacy, in: Proceedings of ICWSM, 2010.
[52] Y. Mei, W. Zhao, J. Yang, Influence maximization on twitter: A mechanism for effective marketing campaign, 2017, pp. 1–6.
[53] E. Katz, P. Lazarsfeld, Personal Influence, the Part Played By People in the Flow of Mass Communications, in: A Report of the Bureau of Applied Social Research Columbia University, Free Press, 1966, [Online]. Available: https://books.google.com.lb/books?id=rElW8D0D8gYC.
[54] C. Castillo, M. Mendoza, B. Poblete, Proceedings of information credibility on Twitter, in: WWW '11, Association for Computing Machinery, New York, NY, USA, 2011, pp. 675–684, [Online]. Available: https://doi.org/10.1145/1963405.1963500.
[55] M.-A. Abbasi, H. Liu, Measuring user credibility in social media, in: A.M. Greenberg, W.G. Kennedy, N.D. Bos (Eds.), Social Computing, Behavioral-Cultural Modeling and Prediction, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 441–448.
[56] S. Kwon, M. Cha, K. Jung, W. Chen, Y. Wang, Proceedings of prominent features of rumor propagation in online social media, in: 2013 IEEE 13th International Conference on Data Mining, 2013, pp. 1103–1108.
[57] tweepy, Tweepy api reference, 2020, Available at http://docs.tweepy.org/en/latest/api.html.
[58] twitter, tweepy api reference, Available at https://developer.twitter.com/en/docs/api-reference-index.

Ali Srour Received his M.S. degree in Computer Science from the Department of Computer Science and Mathematics, Lebanese American University. Ali is a Senior Data Scientist and AI research developer, an international Data Science consultant and keynote speaker, and the head of Research at SocialLab Academy of AI and Data Sciences, SocialLab, Estonia. Ali is a certified AI and Digital Transformation professional from MIT Institute, Cambridge, USA. His research interests are Social Network Analysis, Computational Social Science, Artificial Intelligence for Development, and Applied Machine Learning in Health, Media, and Climate Change.

Hakima Ould-Slimane received the Ph.D. degree in computer science from Laval University, Quebec, QC, Canada, in 2011. She is currently a Researcher and a Lecturer with the École de Technologie Supérieure, Montreal, QC, Canada. Her research interests include mainly information security, cryptography, federated learning, preserving data privacy in smart environments, reliability of collaborative computing, and formal methods.

Azzam Mourad (Senior Member, IEEE) received his M.Sc. in CS from Laval University, Canada (2003) and Ph.D. in ECE from Concordia University, Canada (2008). He is currently a Professor of Computer Science with the Lebanese American University, a Visiting Professor of Computer Science with New York University Abu Dhabi and an Affiliate Professor with the Software Engineering and IT Department, Ecole de Technologie Superieure (ETS), Montreal, Canada. He published more than 100 papers in international journal and conferences on Security, Federated Learning, Network and Service Optimization and Management targeting IoT, Cloud/Fog/Edge Computing, and Vehicular and Mobile Networks. He has served/serves as an associate editor for IEEE Transactions on Services Computing, IEEE Transactions on Network and Service Management, IEEE Network, IEEE Open Journal of the Communications Society, IET Quantum Communication, and IEEE Communications Letters, the General Chair/Vice- Chair of IWCMC2020/2022, the General Co-Chair of WiMob2016, and the Track Chair, a TPC member, and a reviewer for several prestigious journals and conferences.

Haidar Harmanani (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer engineering from the Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH, USA, in 1989, 1991, and 1994, respectively. He is currently a Professor of Computer Science with the Lebanese American University. He serves on the steering committee of the IEEE NEWCAS conference and the IEEE ICECS conference. He has also served on the program committees of various international conferences. His research interests include electronic design automation, high-level synthesis, design for testability, and parallel programming. He is a Senior Member of ACM

Cathia Jenainati is currently a Professor of English Literature and the Dean of the School of Arts and Sciences, Lebanese American University. He has previously served as the Founding Head of the School for Cross-Faculty Studies, Warwick University, U.K. Her research focuses on women's activism, oral history, the global sustainable development agenda, and the history of education missions in the Middle East. In addition, she has been recognized as a leader in innovative pedagogies especially around liberal education, and received several research grants in the field. She serves as the Associate Editor for the Journal of Coaching Practice, the Chair of the International Advisory Board of Amsterdam University College, an Educational Coach of the Growth Coaching International (Australia–U.K.–USA), and a Founding Fellow of the Warwick Higher Education Academy, U.K.