Data & Knowledge Engineering 146 (2023) 102183

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

A framework for investigating the dynamics of user and community sentiments in a social platform

Gianluca Bonifazi [a,1], Francesco Cauteruccio [a,1], Enrico Corradini [a,1], Michele Marchetti [a,1], Giorgio Terracina [b,1], Domenico Ursino [a,*,1], Luca Virgili [a,1]

[a] DII, Polytechnic University of Marche, Italy
[b] DEMACS, University of Calabria, Italy

ARTICLE INFO

Keywords:
Social platforms
User and community sentiments
Model and framework for sentiment evolution
Cross-contamination of users

ABSTRACT

Social platforms are the preferred medium for many people to express their opinions on many topics. This has led many professionals from various fields (marketing, politics, research and development, etc.) to demand increasingly advanced approaches capable of analyzing the evolution of user or community sentiments on particular topics. In this paper, we want to make a contribution to addressing this issue. Specifically, we propose a model and a framework to analyze the dynamics of user and community sentiments in a social platform. In particular, our framework currently focuses on three activities, namely: *(i)* finding users capable of creating and maintaining a community that reflects their sentiment on a topic; *(ii)* studying how a user or community sentiment on a topic evolves over time; and *(iii)* investigating the cross-contamination between a user community and its neighborhood. We tested our framework by means of an extensive experimental campaign that we describe in the paper. Our framework is extremely scalable, and further activities can be easily implemented in it in the near future.

1. Introduction

Nowadays, social platforms are the most adopted tools for expressing one's opinions. In fact, millions and millions of users all over the world adopt such tools to discuss on a wide variety of topics (concerning politics, economics, environment, job, and so on and so forth) [1–4]. As a result, more and more marketers, as well as politicians, journalists, decision makers and other professionals, would like to use social platforms to understand people thoughts and sentiments on particular topics, and how these evolve over time.

The specific nature of social platforms, based on value homophily rather than status homophily [5], implies that communities, which share certain topics and, more importantly, certain sentiments on some topics, can arise in them. These communities often originate and evolve around one or more leaders [3,6–10].

Knowing how the sentiment of a person or a community on a topic evolves over time, understanding how and to what extent the evolution of those sentiments can be influenced by one or more opinion leaders, evaluating how and to what extent the sentiments of a community can influence the one of its virtual "neighborhood" (i.e., users who, although not belonging to it, have opinion exchanges with some of its members), and vice versa, can represent a valuable knowledge in all the application fields mentioned above.

* Corresponding author.
E-mail addresses: g.bonifazi@univpm.it (G. Bonifazi), f.cauteruccio@univpm.it (F. Cauteruccio), e.corradini@pm.univpm.it (E. Corradini), m.marchetti@pm.univpm.it (M. Marchetti), terracina@mat.unical.it (G. Terracina), d.ursino@univpm.it (D. Ursino), luca.virgili@univpm.it (L. Virgili).

[1] Their contribution is equal and this is testified by the alphabetical order used in the Author List.

https://doi.org/10.1016/j.datak.2023.102183
Received 1 September 2022; Received in revised form 21 February 2023; Accepted 5 May 2023
Available online 13 May 2023

This paper aims to provide a contribution in this setting. First of all, it proposes a network-based model to represent the users of a social platform, the topics of their interest, the sentiments they have on these topics, the communities they create around certain topics, the sentiments characterizing these communities, the community neighborhoods, etc. Our model also takes the time factor into account to support the analysis of the temporal evolution of all these features. Then, it leverages that model to define a framework that can support the study of the evolving dynamics of sentiments on specific topics by users and communities of a social platform.

Our framework, thanks to the underlying model, is extremely general and could support many sentiment investigation activities. In this paper, we will focus on three of them. The first activity concerns the detection of determinant users. By this term we mean those users who can determine the sentiment of a community on a certain topic, or even build a new community around themselves, reflecting their sentiment on a certain topic. The second activity aims to assess how the sentiment of a user or a community on a topic evolves over time. The third activity concerns the study of the cross-contamination between a community and its neighborhood. By the term "cross-contamination", we mean the ability of a user community to influence the sentiment of its neighborhood on a topic, or vice versa. By the term "community" we mean a group of users interested in the same topic, or set of topics, on which they make posts and comments. In addition, community users interact strongly with each other by commenting on each other's posts.[2] By the term "neighborhood", we mean the set of users who are not members of the community but have made at least one post/comment on the same topic with at least one user of the community. In each of the three activities, we define one or more parameters that can serve as quantitative support in the various evaluations in which they are involved.

These activities are certainly three contributions by themselves. However, we also present them to show the potential of our framework and the underlying model. This potential is such that, in the future, it will easily be possible to enrich our framework with new activities concerning the investigation of the interactions between users, communities and neighborhoods with regard to topics and sentiments.

Compared with the current state-of-the-art in this area, our paper introduces several contributions. In particular:

- In defining determinant users, we employ new properties not considered in the past. Specifically, we consider both time and user sentiments to identify influential users. Moreover, due to the network-based representation we adopt, the computational effort is reduced, if compared to existing approaches. Finally, the notion of determinant user introduced in this paper can be profitably employed as a measure of trust to be provided as input to a wide variety of existing frameworks handling opinion dynamics. In this way, we provide a new approach to address this latter issue.
- As for the analysis of the evolution of sentiment on topics, our approach differs significantly from existing ones tracking the evolution of communities in social platforms. In fact, instead of focusing on structural changes in communities within a social platform, we analyze the evolution of user and community sentiments on one or more topics. This also allows us to track the evolution of sentiments of a whole community, which is an issue generally not considered in related literature.
- We introduce the concept of cross-contamination between communities, which is a topic never considered in the past literature and, as such, an important advance in the state-of-the-art. Specifically, our framework is able to assess the ability of a community to influence the sentiment of its neighborhood on a topic. We believe that this feature is a challenging issue to investigate.
- Our network-based model represents an advancement of the state-of-the-art in the representation of the temporal change of a sentiment on a topic for all those scenarios in which a huge number of topics are discussed every day. In fact, our model explicitly avoids expressing all potential combinations of topics and sentiments over time. Because of such a characteristic, we are able to handle those scenarios in which topics can suddenly burst and disappear in a few time slots.
- Our model for representing a community, based on the content posted by users and the sentiments they felt, rather than on structural information alone, allows the temporal evolution of communities to be tracked more holistically. Moreover, it allows supporting both the representation of successive snapshots of communities over time and an aggregate view of the various communities with respect to the social platform to which they belong. This makes it possible to reconstruct the temporal changes characterizing communities in that social network.

This paper is structured as follows: In Section 2, we review related literature. In Section 3, we first provide a formalization of the reference scenario and then use it to present our model and framework. In Section 4, we describe the experimental campaign we conducted to test the validity of the proposed model and framework. Finally, in Section 5, we draw our conclusions and outline some possible future developments of this research.

2. Related literature

As highlighted above, our framework handles three main activities, namely: *(i)* identifying determinant users, *(ii)* analyzing the evolution of a user or community sentiment on a topic, and *(iii)* analyzing the cross-contamination between a community and its neighborhood. To the best of our knowledge, there is no approach that handles these three activities simultaneously in the past literature. Our paper also proposes a network-based model underlying our framework, which we believe is an additional distinguishing point. In what follows, we review the related literature focusing in detail on each of the three contributions mentioned above and providing a comparison between the corresponding approaches and ours.

[2] A formal definition of the concepts of community and neighborhood is provided in the next sections, after introducing the model employed in our framework.

Identifying determinant users. This task aims to identify those users who can determine a community sentiment on a topic or, even, build a community that reflects their sentiment on a topic. To achieve this goal, a variety of models studying opinion dynamics have been proposed in the literature. They include the DeGroot model [11,12], the voter model [13,14], the Sznajd model [15,16], the majority rule model [17,18], the Friedkin and Johnsen model [19], the bounded confidence model [20], and the continuous opinions and discrete actions model [21]. All of them study whether and how a community of users can reach consensus on a given topic. In investigating these dynamics, they rely on the notion of trust (or one of its variants) of a user in other users in a community. These approaches are generally orthogonal to the notion of trust adopted in this paper. Therefore, the metric we propose here, which measures how determinant a user is in leading the sentiment of other users on a certain topic, can be exploited as a measure of trust to be given as input to the above approaches.

In this setting, our framework is characterized by: *(i)* the reduced computational effort, and *(ii)* the type of properties it analyzes to identify influential users. Specifically, the proposed determinance metric is based on a combination of factors (such as the degree of a node) that can be easily computed. Therefore, it does not represent a computational bottleneck. This result is also achieved thanks to the network-based representation we decided to adopt as underlying model. Regarding the second feature, in order to identify influential users, we take into account both time and user sentiments on topics. These two peculiarities are generally not addressed together by related approaches.

As for the latter, in a recent paper, Chen et al. show that the influence of opinion leaders in social networks, especially for advertising purposes, is crucial in the evolution of follower opinions [6]. In doing so, they demonstrate the importance of identifying determinant users by constructing an integrated bounded confidence model. Actually, the approach described by them does not address how such users can be identified. This last issue is addressed by Li and Wei, who introduce the notion of popularity degree of an expert and build a consensus model based on opinion evolution [22]. Specifically, they start from the assumption that experts are more likely to accept the opinions of other experts having a strong influence relationship with them. Then, they define the popularity degree based on the proximity degree of a node in the underlying graph model. In this case, popularity is computed based on the shortest paths, whose calculation, unlike our approach, is a computationally heavy task. Recently, Chang and Wang showed that influential users identified by conventional measures often do not have a substantial impact on changes in the direction of discussion [23]. They assume that the change in the direction of discussion of a community follows the animal flocking model, i.e., the dynamics characterizing the migration of an animal flock. Differently from our approach, the one proposed by Chang and Wang is based on the structure of the underlying social network and does not take into account user sentiments on topics.

Analyzing the evolution of sentiments on a topic. The second main activity considered by our approach is the evolution of a user or community sentiment on a topic. In fact, tracking the evolution of a community on social networks is a widely studied problem; the interested reader can refer to the survey of Dakichi et al. [7] for a comprehensive overview on this issue.

In this setting, our approach differs significantly from existing ones by providing several contributions. First, existing approaches focus on the study of structural changes within social networks, while our framework focuses on the analysis of the evolution of user and community sentiments on one or more topics, which is a relatively understudied problem (see the surveys of Chang and Wang [24] and Basal et al. [25]). In addition, our framework allows the homogeneous tracking of the evolution of the sentiments of a whole community, which is an issue that is not generally investigated by existing studies. This allows for views on the sentiments of both individuals (and, thus, local views) and communities (and, thus, global views) on a set of topics. This aspect is important because, as we will discuss below, users and communities are often not taken into account in the existing analyses.

Two interesting works related to our approach in this setting are the studies presented by Yin et al. [26] and Xu et al. [27]. In the former, the authors propose a framework to analyze the dynamics of COVID-19 from tweets related to this topic collected over two weeks. They extract topics and related sentiments from online posts and study their evolution over time, but do not consider users and communities in their analysis. Additionally, they focus on computing the distribution of the sentiment on a topic on specific days, but not its change over time for users and communities. The study of Xu et al. [27] concerns the evolution of sentiments on topics in the context of online news, by exploiting a manifold learning-based model. The authors first introduce a method to identify topic-sentiment pairs, then they define measures to assess the evolution of sentiments on topics over time. However, this approach does not consider the evolution of sentiments over time periods and does not investigate the role of users and communities in the analysis they perform.

Cross-contamination between communities. The third activity that characterizes our framework is the cross-contamination between a community and its neighborhood. Regarding it, we point out that, in this connotation, this issue has never been addressed in the past literature. In fact, no approach designed specifically for this topic has been proposed in the past.

A partially related topic is the one concerning sentiment diffusion in social networks, which differs from the more widely studied area of information diffusion (see the survey of Bhattacharya and Sarkar [28]). Xu et al. [8] analyze textual information and sentiment diffusion patterns to predict the polarities of sentiments expressed in Twitter messages. The core of their approach is the notion of sentiment reversal; it indicates that a tweet and its retweet have different sentiment polarities. The authors exploit such a notion to identify sentiment diffusion patterns. Actually, in this approach, the concept of time is implicitly represented. Furthermore, the main goal of Xu et al. is the prediction of sentiment polarity and not the measure of the ability of a community to influence the sentiment of its neighborhood on a topic.

Final remarks. As previously pointed out, existing related approaches can be considered orthogonal to the one proposed in this paper. As a further remark, we consider it important to briefly highlight two additional contributions of this paper. In fact, *(i)* the network-based model we have adopted here allows us to take into account the content posted by users over time, as well as the sentiment they expressed on it; *(ii)* our model for managing communities in social networks supports both the representation of snapshots and the computation of an aggregated view in a single network encompassing the temporal changes occurred over a certain time interval.

As for the first contribution, several models capable of representing how the sentiment on a topic changes over time have been proposed in the literature. As an example, Dermouche et al. [29] presented the Time-aware Topic-Sentiment (TTS) model. In it, each topic has both a sentiment polarity and a time facet information, whose presence allows different analyses. However, time management in TTS is not capable of handling a scenario in which some topics can appear and disappear in a few time slots, which is exactly the reference scenario for our framework.

As for the second contribution, we point out that, in past approaches, the structure of a community generally derives from the ones of the underlying social network [7]. Instead, we consider a community as a group of users who show common interests and sentiments on specific topics. This implies that our notion of community is based on the content posted by users on the social platform and the sentiment they show on that content. This is very different from tracking the relationship structure of the underlying social network. Since our framework deals with the temporal evolution of communities (see the overview proposed by Giatsoglou and Vakali in [30]), it can be classified among the ones handling a dynamic representation of social networks [31]. Furthermore, it differs from most of the traditional methods addressing this issue, which model dynamic changes through a time series of static networks [31,32].

3. The proposed model and framework

In this section, we present the proposed model and framework. First, we provide a formal description of the scenario of interest (Section 3.1). Then, we describe the proposed model (Section 3.2). Finally, we present the proposed framework (Section 3.3).

3.1. A formal description of the scenario of interest

Before introducing our model and framework, we need to describe and formalize the scenario in which they operate. Such a scenario is that of a typical social platform where users can publish news-related posts that can be commented by other users. Both posts and comments consist of textual content. A user can publish a comment in response to a post or another comment.

Let $\mathcal{U} = \{u_1, \ldots, u_l\}$ be the set of users of interest. Let $\mathcal{P} = \{p_1, \ldots, p_m\}$ and $\mathcal{C} = \{c_1, \ldots, c_n\}$ be the sets of posts and comments published by the users of $\mathcal{U}$ on the reference platform in a given time interval T. Given a user $u_i \in \mathcal{U}$, we denote by $\mathcal{P}_i \subseteq \mathcal{P}$ (resp., $\mathcal{C}_i \subseteq \mathcal{C}$) the subset of posts (resp., comments) published by them.

In our scenario, the time factor is important because we want to investigate how various phenomena evolve over time. To model time, we assume that the overall time interval T of our interest can be divided into an ordered sequence of z time slices, $T = T_1, \ldots, T_z$. For instance, T could be a given month, say June 2022, and it could be divided in 30 time slices, one for each day.

It is useful to be able to index this sequence, that is, to select only a particular interval of contiguous time slices of T (think, for instance, of the second decade of June 2022). For this reason, we use the notation $T[x..y]$, $1 \leq x \leq y \leq z$, to indicate the interval of contiguous time slices in T beginning at T_x and ending at T_y. If $x = y$, then $T[x..x]$ refers to a single time slice and we will use the abbreviated notations $T[x]$ or T_x to represent it. If $x = 1$ and $y = z$, then the whole interval T is considered and, in this case, we will use the abbreviated notation T, instead of $T[1..z]$, to indicate it.

In the following, we will extend the previous notation concerning time intervals and slices to the other sets of our model. In particular, we will use the notation $\mathcal{P}[x..y] \subseteq \mathcal{P}$ (resp., $\mathcal{C}[x..y] \subseteq \mathcal{C}$) to denote the subset of posts (resp., comments) published in the time interval $T[x..y]$. Clearly, we will use the abbreviated notation $\mathcal{P}$ (resp., $\mathcal{C}$) to denote the total set of posts $\mathcal{P}[1..z]$ (resp., comments $\mathcal{C}[1..z]$) published in the whole time interval T.

In the scenario of our interest, two other important concepts are the ones of topic and sentiment tag. A topic is an abstract concept discussed in one or more posts and comments. Topic modeling and topic extraction are important tasks in the field of Natural Language Processing (NLP) [33,34]. A sentiment tag is a keyword that indicates the sentiment expressed on a particular topic; for instance, a sentiment tag could be pos, indicating a positive sentiment. In our model, we first extract topics from the textual content of posts and comments and then associate sentiment tags with them. Let $\mathcal{T} = \{t_1, \ldots, t_q\}$ be the set of topics extracted from the posts of $\mathcal{P}$ and the comments of $\mathcal{C}$. Let $\mathcal{S} = \{s_1, \ldots, s_r\}$ be the set of available sentiment (tags).[3] Given a topic $t_j \in \mathcal{T}$ and a sentiment $s_k \in \mathcal{S}$, the pair (t_j, s_k) indicates that t_j is tagged with s_k. A topic that has associated at least one sentiment is called sentiment-tagged topic.

3.1.1. Identifying the topics characterizing user opinions

In our scenario, we assume that posts, and especially comments, consist mainly of text. Our model and framework are orthogonal to the approach chosen to build the set $\mathcal{T}$ of topics. As a consequence, any approach for identifying topics from texts proposed in the past literature can be used to build $\mathcal{T}$ (see the surveys of Jelodar et al. [34], Vayanski and Kumar [33], and Qiang et al. [35]). In the following, we simply assume that, given a text associated with a post $p \in \mathcal{P}$ (resp., a comment $c \in \mathcal{C}$), our model and framework are equipped with an approach capable of extracting the topics of p (resp., c), adding them into the overall set $\mathcal{T}$ of topics, and associating them with the post p (resp., comment c) from which they were derived.

[3] In the following, to simplify the discussion, we will use the term "sentiment" instead of the term "sentiment tag".

3.1.2. Identifying the sentiments characterizing user opinions

Similarly to what was done for topic identification, also for sentiment detection our model and framework are orthogonal with respect to the approach used to address this issue. The identification of the sentiments associated with a text is investigated primarily in the context of sentiment analysis, where several approaches have been proposed to define, characterize and extract the sentiment expressed in a text (see [36–38] for some surveys about this topic). In this research field, the term "sentiment tag" is synonymous with terms like "sentiment value" or, simply, "sentiment" [39].

In our reference scenario, the sentiment tagging activity aims to examine each post $p \in \mathcal{P}$ (resp., comment $c \in \mathcal{C}$) for identifying the sentiments emerging in it and to associate them with the corresponding topics of $\mathcal{T}$ referring to p (resp., c). More specifically, this activity proceeds as follows: let p (resp., c) be a post (resp., a comment) of $\mathcal{P}$ (resp., $\mathcal{C}$). It could be a simple text, expressing a single sentiment, or an articulate text, expressing several sentiments, in extreme cases even conflicting with each other. Since the first hypothesis can be seen as a special case of the second, in the following we will directly consider the latter. Therefore, we assume that p (resp., c) consists of a succession $p_1, p_2, \dots, p_v$ (resp., $c_1, c_2, \dots, c_v$) of textual contents such that, for each of them, a single sentiment can be identified. We use the term "fragment" to denote each textual content p_k (resp., c_k), $1 \leq k \leq v$, and we use the symbol f_k to generalize p_k and c_k.

Given the fragment f_k, it is possible to apply the approach described in Section 3.1.1 to build the set $\mathcal{T}_{f_k}$ of topics treated in f_k. Then, any of the approaches for identifying the sentiment of a textual content proposed in the past literature (such as one of the approaches described in the surveys of Jelodar et al. [34], Vayanski and Kumar [33], and Qiang et al. [35]) can be applied on f_k. Let s_k be the sentiment that this approach associates with f_k. As a consequence of this, for each topic $t_j \in \mathcal{T}_{f_k}$, our activity will return a pair (t_j, s_k) indicating that s_k is the sentiment on t_j expressed in f_k. The set of all the sentiments extracted from all the fragments of all the posts of $\mathcal{P}$ and all the comments of $\mathcal{C}$ will form the set $\mathcal{S}$ of sentiment tags available in our scenario (see Section 3.1).

3.2. The proposed model

Having formalized the scenario of interest, we are now able to propose a model for supporting its analysis. We begin by defining a support bipartite network $\mathcal{B}$, which allows us to store all the key information of the scenario of interest. In the next step, by operating appropriately on that network, we can define our proposed model. Furthermore, other models useful for investigations beyond the scope of this paper could be derived in the future starting from $\mathcal{B}$. $\mathcal{B}$ is defined as:

$$\mathcal{B} = \langle N' \cup N'', E' \rangle \tag{3.1}$$

In this case, N' and N'' represent the two subsets that collectively form the set of nodes of $\mathcal{B}$. N' is the set of nodes associated with users. There is a node $n_i \in N'$ for each user $u_i \in \mathcal{U}$. Since there is a biunivocal correspondence between a node of N' and a user of $\mathcal{U}$, we will employ these two terms interchangeably in the following. There is a node $n_{jk} \in N''$ if there exists a pair (t_j, s_k) such that $t_j \in \mathcal{T}$ and $s_k \in \mathcal{S}$ and t_j has been tagged with s_k in at least one post of $\mathcal{P}$ or comment of $\mathcal{C}$.

E' is the set of edges of $\mathcal{B}$; there is an edge $(n_i, n_{jk}) \in E'$ between a node $n_i \in N'$ and a node $n_{jk} \in N''$ if the user u_i published a post or a comment where they expressed the sentiment s_k on the topic t_j at least once. Since, in the time interval T of interest, u_i may have carried out this task more times, we associate a label l_{ijk} with (n_i, n_{jk}). It represents the list of the timestamps of posts and/or comments published by u_i in which they expressed the sentiment s_k on the topic t_j.

The network $\mathcal{B}$ defined above contains all the potentially useful information to allow the investigation of the scenario of our interest. However, its two-mode nature does not make it easy its investigation and manipulation. Therefore, as it often happens in Social Network Analysis [5], it seems reasonable to construct single-mode networks from it, which focus on the single aspects we want to study. Now, in this paper, we want to focus on the dynamics of opinion dissemination from the perspective of the users involved, their posting activity, their ability to build communities, and so on; therefore, it is reasonable to build a user-centered single-mode network starting from $\mathcal{B}$. On the other hand, if we had wanted to study the dynamics of opinion dissemination from the perspective of topics involved (a research activity that we plan to do in the future), it would have been reasonable to build a topic-centered single-mode network starting from $\mathcal{B}$.

The user-centered single-mode network that we propose here is defined as:

$$\mathcal{A} = \langle N, E \rangle \tag{3.2}$$

N is the set of nodes of $\mathcal{A}$. There is a node $n_i \in N$ for each user $u_i \in \mathcal{U}$. Also in this case, since there is a biunivocal correspondence between a user $u_i \in \mathcal{U}$ and a node $n_i \in N$, we will employ these two terms interchangeably in the following. Clearly, N is equivalent to the set N' of $\mathcal{B}$.

E is the set of edges of $\mathcal{A}$. There exists an edge $e_{ih} \in E$ between two nodes n_i and n_h if u_i and u_h published at least one post/comment on the same topic and, at least once, n_i published a comment on a post or a comment of n_h, or vice versa. It is possible to associate a label l_{ih} with each edge $e_{ih} \in E$. The value of this label should represent the strength of the relationship between u_i and u_h. This could be computed by an appropriate function $\omega()$ that receives the network $\mathcal{B}$ and two nodes n_i and n_h of N and returns a number representing the strength of the relationship between u_i and u_h. For example, $\omega()$ might indicate the number of topics in common between u_i and u_h. One could think of much more sophisticated functions, which could add some refinement degree to the computation of the strength of the relationship between u_i and u_h. However, they would require very high computational costs against a low level of further refinement they might provide, which, in our opinion, does not justify the cost required (recall that $\mathcal{U}$ could consist of thousands or millions of users).

Finally, we point out that, in some of the following investigations, we may need to retrieve data present in $\mathcal{B}$, but we do not consider it necessary to report it in $\mathcal{A}$ in order not to burden the latter (e.g., because such data is needed only rarely). In this case, we will define ad-hoc functions that perform such a task and complement our model. For example, if it were necessary to know all the users who published posts and comments on a given topic t_j in a time interval $T[x..y]$, one could define a function that receives t_j and returns the desired set of users. The functions that support the activities of our framework and complement our model are presented in Section 3.2.1.

Similarly to what we have seen for other previous cases, given the network $\mathcal{A}$ and the time interval $T[x..y]$, we denote by $\mathcal{A}[x..y]$ the "projection of $\mathcal{A}$" in $T[x..y]$:

$$\mathcal{A}[x..y] = \langle N[x..y], E[x..y] \rangle \tag{3.3}$$

A node n_i belongs to $N[x..y]$ if the corresponding user u_i published at least one post/comment in $T[x..y]$. An edge $e_{ih} \in E[x..y]$ indicates that u_i and u_h published at least one post/comment on the same topic in the time interval $T[x..y]$ and, in the same interval, at least once n_i published a comment on a post or a comment of n_h, or vice versa. The label l_{ih} is the result of the computation of the function $\omega()$ in the time interval $T[x..y]$. Clearly, as in other cases above, $\mathcal{A}[x] = A[x..x]$ is the "projection of $\mathcal{A}$" in the time slice T_x and $\mathcal{A}[1..z]$ is equivalent to $\mathcal{A}$.

As expressed in the Introduction, the analyses conducted in this paper cover both user and community behaviors. Therefore, it is necessary to identify a way to model user communities. Regarding this aspect, we relied on Social Network Analysis theory and chose the concept of clique [5] to model a community. This concept is very strict and has the advantage of obtaining strongly connected communities, although their number may be very small. Clearly, concepts less strict than clique could be adopted for this purpose, such as the concepts of n-clique, k-truss, n-clan, k-plex [5]. In this case, we will certainly have a higher number of communities, but these might not be too connected, and might be potentially less cohesive. Since the main objective of our study is sentiment, we thought it appropriate to consider strongly connected communities and thus use the concept of clique.

A clique in $\mathcal{A}$ is a maximal subset of totally connected nodes, and thus a maximal subset of users interested in the same topic t_j. In the following, we will use the symbol Q to denote a generic clique of $\mathcal{A}$, and the symbol Q_{t_j} to represent a clique of $\mathcal{A}$ in which all its nodes correspond to users who submitted at least one post or comment on t_j. Furthermore, similarly to what happened for $\mathcal{A}$, we use the notation $Q[x..y]$ (resp., $Q_{t_j}[x..y]$) to represent the projection of Q (resp., Q_{t_j}) in the interval $T[x..y]$, and the notation $Q[x]$ (resp., $Q_{t_j}[x]$) to denote $Q[x..x]$ (resp., $Q_{t_j}[x..x]$). Clearly, Q (resp., Q_{t_j}) is equivalent to $Q[1..z]$ (resp., $Q_{t_j}[1..z]$).

As we will see in the following, given a clique Q, it could be interesting to study its neighborhood Γ_Q. It is defined as a network such that: *(i)* each node is connected to at least one node of Q but is not part of the set of nodes of Q; *(ii)* each edge connects a node of Γ_Q with a node of Q or two nodes of Γ_Q.

3.2.1. Functions complementing our model

In this section, we present some support functions that complement our model. They will be used to formalize the activities performed by our framework. Before describing them, we feel it is appropriate to introduce some concepts concerning the prevalence or ambivalence of the sentiment of a user or a community on a topic.

Let u_i be a user of $\mathcal{U}$ and let t_j be a topic of $\mathcal{T}$; we say that there exists a predominantly positive (resp., negative) sentiment of u_i on t_j if at least 75% of the posts and comments published by u_i and having t_j as topic show a positive (resp., negative) sentiment. Finally, we say that there exists an ambivalent sentiment of u_i on t_j if there exists neither a predominantly positive nor a predominantly negative sentiment of u_i on t_j.

Let Q_{t_j} be a clique of $\mathcal{A}$ and let t_j be a topic of $\mathcal{T}$; we say that there exists a predominantly positive (resp., negative) sentiment on t_j in Q_{t_j} if the number of users of Q_{t_j} with a predominantly positive (resp., negative) sentiment on t_j is higher than both the number of users of Q_{t_j} with a predominantly negative (resp., positive) sentiment on t_j and the number of users of Q_{t_j} with an ambivalent sentiment on t_j. Finally, we say that there exists an ambivalent sentiment on t_j in Q_{t_j} if there exists neither a predominantly positive nor a predominantly negative sentiment on t_j in Q_{t_j}.

Having made these premises, we can now introduce our complement functions. They are:

- $degree(u_i)$: It receives a user u_i and returns their degree in $\mathcal{A}$.
- $f^+(u_i, t_j)$: It receives a user u_i and a topic t_j and returns the number of posts and comments on t_j in which u_i expressed a positive sentiment. $f^+(u_i, t_j)[x..y]$ represents the projection of $f^+(u_i, t_j)$ in the time interval $T[x..y]$; it performs the same computation as $f^+(u_i, t_j)$ but considers only the posts and comments published in the time interval $T[x..y]$.
- $f^-(u_i, t_j)$: It receives a user u_i and a topic t_j and returns the number of posts and comments on t_j in which u_i expressed a negative sentiment. $f^-(u_i, t_j)[x..y]$ represents the projection of $f^-(u_i, t_j)$ on the time interval $T[x..y]$.
- $f^=(u_i, t_j)$: It receives a user u_i and a topic t_j and returns the number of posts and comments on t_j in which u_i expressed an ambivalent sentiment. $f^=(u_i, t_j)[x..y]$ represents the projection of $f^=(u_i, t_j)$ on the time interval $T[x..y]$.
- $size(net)$: It receives a network net (which might be the whole network $\mathcal{A}$, a clique Q, a neighborhood Γ_Q, etc.) and return its size, i.e., the number of its nodes.
- $g^+(net, t_j)$: It receives a network net and a topic t_j and returns the number of its users being predominantly positive on t_j. $g^+(net, t_j)[x..y]$ represents the projection of $g^+(net, t_j)$ on the time interval $T[x..y]$.
- $g^-(net, t_j)$: It receives a network net and a topic t_j and returns the number of its users being predominantly negative on t_j. $g^-(net, t_j)[x..y]$ represents the projection of $g^+(net, t_j)$ on the time interval $T[x..y]$.

Table 1
Summarizing table of the notations introduced in Section 3.2.

Notation	Description
$\mathcal{B}$	The support bipartite network
$\mathcal{A}$	The user-centered single-mode network
Q	A clique of $\mathcal{A}$
Γ_Q	The neighborhood of Q
$degree(\cdot)$	A function that returns the degree of a user in $\mathcal{A}$
$f^+(\cdot), f^-(\cdot), f^=(\cdot)$	Functions returning the number of positive, negative or ambivalent posts/comments on a topic published by a user
$g^+(\cdot), g^-(\cdot), g^=(\cdot)$	Functions returning the number of predominantly positive, predominantly negative or ambivalent users on a topic in a network
$pp(\cdot), pn(\cdot), pa(\cdot)$	Functions returning `true` if the majority of a community is predominantly positive, predominantly negative or ambivalent
$e\text{-}size(\cdot)$	A function that returns the size of the ego network of a user in $\mathcal{A}$
$cc(\cdot)$	A function that returns the clustering coefficient of the ego network of a user in $\mathcal{A}$
$nts(\cdot)$	A function that returns the number of time slices in which a user published posts/comments on a topic

- $g^=(net, t_j)$: It receives a network *net* and a topic t_j and returns the number of its users being ambivalent on t_j. $g^=(net, t_j)[x..y]$ represents the projection of $g^+(net, t_j)$ on the time interval $T[x..y]$.
- $pp(net, t_j)$: It receives a network *net* and a topic t_j and returns `true` if the majority of users are predominantly positive, i.e., if $g^+(net, t_j) > g^-(net, t_j)$ and $g^+(net, t_j) > g^=(net, t_j)$. It returns `false` otherwise.
- $pn(net, t_j)$: It receives a network *net* and a topic t_j and returns `true` if the majority of users are predominantly negative, i.e., if $g^-(net, t_j) > g^+(net, t_j)$ and $g^-(net, t_j) > g^=(net, t_j)$. It returns `false` otherwise.
- $pa(net, t_j)$: It receives a network *net* and a topic t_j and returns `true` if $pp(net, t_j) = \texttt{false}$ and $pn(net, t_j) = \texttt{false}$. It returns `false` otherwise.
- $e\text{-}size(u_i)$: It receives a user u_i and returns the size of the ego network of u_i in $\mathcal{A}$, intended as the number of its edges.
- $cc(u_i)$: It receives a user u_i and returns the clustering coefficient of the ego network of u_i in $\mathcal{A}$.
- $nts(u_i, t_j)$: It receives a user u_i and a topic t_j and returns the number of the time slices of T in which u_i published at least one post or one comment on t_j.

3.3. The proposed framework

After providing a formal description of the scenario of interest and proposing a model to represent it, in this section we present a framework for analyzing the behavior over time of users and communities interacting with each other on a social platform. There are many activities that could be conducted regarding this issue. In this paper, we will focus on three of them, namely: *(i)* identifying users who can determine the sentiment of a community on a certain topic or, even, build communities reflecting their sentiment on it; *(ii)* analyzing the evolution of the sentiment of users or communities on a particular topic; *(iii)* analyzing the cross-contamination that can arise when a community in which a certain sentiment on a topic prevails is in contact with a neighborhood in which a different sentiment on the same topic is dominant, and vice versa. However, the formalization of the scenario proposed in Section 3.1 and the model presented in Section 3.2 is general enough to allow the extension of our framework so that approaches for solving additional problems of the same nature can be defined and implemented in it in the future.

A list of the main notations introduced in Section 3.2 is reported in Table 1.

3.3.1. Overall behavior

A flowchart describing the behavior of our framework is shown in Fig. 1. The goal of our framework is to understand the dynamics of user or community sentiments on a specific topic. As shown in the figure, the first step is the identification of the sentiment of the users of a community on the topic under consideration. This is an important step as it helps to establish a baseline for understanding the overall community sentiment on the topic.

The next step is the identification of the determinant users within the community. This identification is done through two parallel tasks, which aim to find the users who best represent the sentiment on the topic and those who have a significant influence on the community sentiment on the topic. By identifying these users we can gain a better understanding of the factors driving the community sentiment on the topic.

The third step is the analysis of the time evolution of the sentiment on the topic. This step also involves two parallel tasks. The first analyzes the evolution of the sentiment of determinant users on the topic. The second studies the evolution of the overall community sentiment on the topic. This step is important as it can help identify trends or patterns in community sentiment on the topic and understand how such a sentiment evolves over time.

The last step is the study of the cross-contamination of the sentiment on the same topic among communities and their neighborhoods. This step is important because it can help identify how the sentiment on the topic is spreading between different communities and neighborhoods and how this may impact the overall sentiment on the topic itself. By studying cross-contamination we can gain a better understanding of how a community sentiment on the topic is influenced by external users, and vice versa.

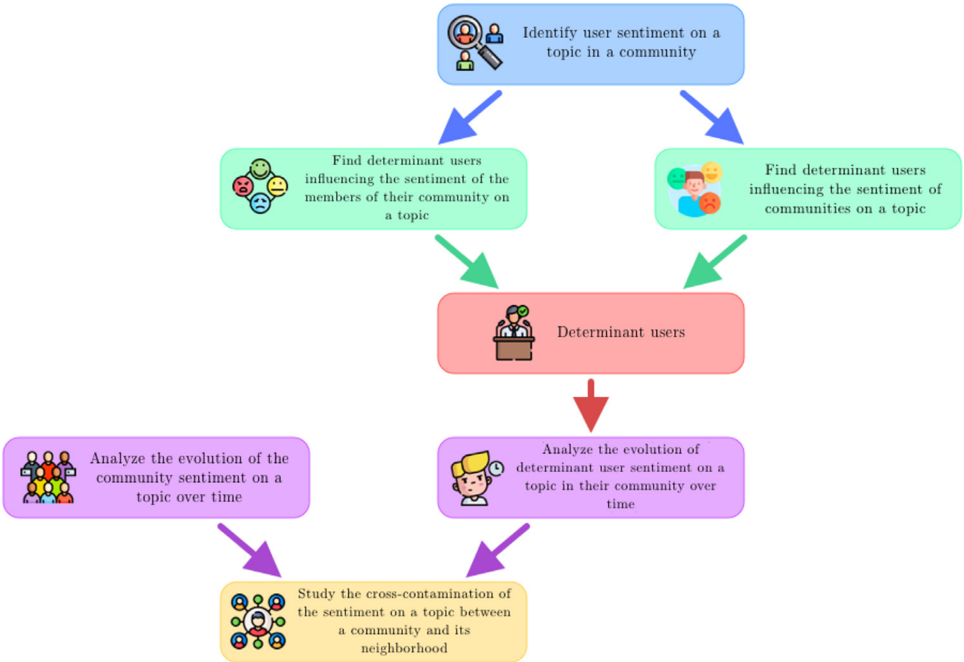


Fig. 1. Representation of the behavior of our framework.

3.3.2. Identifying determinant users

The first activity our framework deals with is the identification of determinant users, that is, those users who can determine the sentiment of a community on a certain topic or, even, build a new community reflecting their sentiment on a certain topic.

In order to establish which users are determinant, we planned to define a metric. In performing this task, we started from the consideration that a user u_i, in order to be determinant in establishing the sentiment of a community on a topic t_j, must:

- Have an unambivalent sentiment on t_j. In fact, if u_i does not have a firm sentiment on t_j, they certainly cannot be convincing for other users.
- Share with many other users their opinions about t_j. In fact, if u_i does not widely share their opinions with other users, they cannot hope to influence them.
- Be potentially a "seed" for creating new communities.
- Be constantly active in expressing their opinions and sentiments.

Using the formalization of the scenario of interest described in Section 3.1, the model illustrated in Section 3.2 and its complement functions presented in Section 3.2.1, the previous four conditions can be formulated as follows. Let u_i be a user of $\mathcal{U}$ and let t_j be a topic of $\mathcal{T}$ and assume that u_i has a predominantly positive sentiment on t_j (a reasoning dual to the one below could be made in case of a user with a predominantly negative sentiment on t_j). In order for u_i to be determinant in spreading and maintaining their sentiment on t_j in a community Q_{t_j}, it will have to happen that:

- The fraction of posts and comments containing t_j in which u_i expressed a positive sentiment must tend to 1.
- The degree of u_i in $\mathcal{A}$ must be as high as possible.
- The size of the ego network of u_i in $\mathcal{A}$ should be as large as possible. In addition, the clustering coefficient of the same ego network must tend to 1. The latter condition denotes that the alters of the ego network are already strongly connected to each other and, thus, are inclined to create cliques.
- The number of time slices of T in which u_i published a post or comment on t_j should be as large as possible.

Based on these considerations, the parameter $det(u_i, t_j)$, which indicates how determinant u_i is in conditioning a community with their sentiment on t_j, can be computed as:

$$det(u_i, t_j) = \alpha \cdot f^+(u_i, t_j) + \beta \cdot \frac{degree(u_i)}{degree_{max}} + \gamma \cdot \frac{e{-}size(u_i)}{e{-}size_{max}} + \delta \cdot cc(u_i) + \varrho \cdot \frac{nts(u_i, t_j)}{z} \tag{3.4}$$

In this formula: *(i)* $degree_{max}$ is the maximum degree of a user in $\mathcal{A}$; *(ii)* $e\text{-}size_{max}$ is the size of the maximum ego network in $\mathcal{A}$; *(iii)* z is the number of time slices; *(iv)* $\alpha, \beta, \gamma, \delta$, and ϱ are parameters in the real interval $[0, 1]$ such that their sum is equal to 1; they allow us to specify to which component of $det(u_i, t_j)$ we want to assign more importance. In this definition, we assume that

z, $degree_{max}$ and $e\text{-}size_{max}$ are greater than 0. Such assumptions are reasonable for any scenario that is not degenerate. Under these assumptions, $det(u_i, t_j)$ ranges in the real interval $[0,1]$; the higher its value, the more determinant u_i will be in conditioning the sentiment on t_j.

3.3.3. Analyzing the evolution of the sentiment of a user or a community on a topic

The second activity in our framework aims to understand how the sentiment of a user or a community of users on a topic evolves over time. Specifically, let t_j be a topic of $\mathcal{T}$ and let Q_{t_j} be a community of users who submitted posts and comments regarding t_j. We have said before that the sentiment of a user u_i on t_j can be predominantly positive, predominantly negative, or ambivalent. Some good metrics for measuring it could be the functions $f^+(u_i,t_j)$, $f^-(u_i,t_j)$ and $f^=(u_i,t_j)$ introduced in Section 3.2.1. Switching from a single user to a community of users, the functions $g^+(Q_{t_j},t_j)$ $g^-(Q_{t_j},t_j)$ and $g^=(Q_{t_j},t_j)$, defined in the same section, could be adopted to measure the sentiment of a community Q_{t_j} on t_j.

Starting from these considerations, and taking into account that the time factor plays a key role in an evolution analysis, we define some metrics that allow us to study the evolution of the sentiment of u_i or Q_{t_j} on t_j over time.

The first metric is an indicator of the interest that t_j stimulates in Q_{t_j}. It starts from the idea that, regardless of the sentiment expressed, the more interesting a topic is, the more it attracts people and, therefore, the more the community interested in it grows. Therefore, we can define the change in interest of Q_{t_j} with respect to t_j over the time interval $T[x..y]$ as:

$$\imath(Q_{t_j},t_j)[x..y] = \frac{size(Q_{t_j}[x..y])}{size(Q_{t_j}[1..x])} \tag{3.5}$$

Observe that $size(Q[1..x])$ is always greater than 0 for any non-degenerate network. Consequently, the value of $\imath(Q,t_j)[x..y]$ ranges in the real interval $[0,+\infty)$. The higher its value, the higher the increase in the interest of Q_{t_j} on t_j.

The next three metrics represent indicators of the evolution of the sentiment of u_i on t_j over the time interval $T[x..y]$. Specifically, $\sigma^+(u_i,t_j)[x..y]$ (resp., $\sigma^-(u_i,t_j)[x..y]$, $\sigma^=(u_i,t_j)[x..y]$) is an indicator of the evolution of the positive (resp., negative, ambivalent) sentiment of u_i on t_j in the time interval $T[x..y]$. They are defined as:

$$\sigma^+(u_i,t_j)[x..y] = \frac{f^+(u_i,t_j)[1..y] - f^+(u_i,t_j)[1..x]}{f^+(u_i,t_j)[1..y]} \tag{3.6}$$

$$\sigma^-(u_i,t_j)[x..y] = \frac{f^-(u_i,t_j)[1..y] - f^-(u_i,t_j)[1..x]}{f^-(u_i,t_j)[1..y]} \tag{3.7}$$

$$\sigma^=(u_i,t_j)[x..y] = \frac{f^=(u_i,t_j)[1..y] - f^=(u_i,t_j)[1..x]}{f^=(u_i,t_j)[1..y]} \tag{3.8}$$

In particular, $\sigma^+(u_i,t_j)[x..y]$ (resp., $\sigma^-(u_i,t_j)[x..y]$, $\sigma^=(u_i,t_j)[x..y]$) is an indicator of the fraction of the number of posts and comments containing t_j published in the time interval $T[x..y]$ that elicited a positive (resp., negative, ambivalent) sentiment in u_i. Observe that $\sigma^+(u_i,t_j)[x..y]$ (resp., $\sigma^-(u_i,t_j)[x..y]$, $\sigma^=(u_i,t_j)[x..y]$) ranges in the real interval $[0,1]$. The higher its value, the greater the evolution toward positivity (resp., negativity, ambivalence) of the sentiment of u_i on t_j.

Finally, the last three metrics, i.e., $\sigma^+(Q_{t_j},t_j)[x..y]$, $\sigma^-(Q_{t_j},t_j)[x..y]$ and $\sigma^=(Q_{t_j},t_j)[x..y]$, are indicators of the evolution of the positive, negative and ambivalent sentiment of Q_{t_j} on t_j in the time interval $T[x..y]$. They are defined as:

$$\sigma^+(Q_{t_j},t_j)[x..y] = \begin{cases} \frac{g^+(Q_{t_j},t_j)[1..y]-g^+(Q_{t_j},t_j)[1..x]}{g^+(Q_{t_j},t_j)[1..y]} & if\, g^+(Q_{t_j},t_j)[1..y] \geq g^+(Q_{t_j},t_j)[1..x] \\ 0 & \text{otherwise} \end{cases} \tag{3.9}$$

$$\sigma^-(Q_{t_j},t_j)[x..y] = \begin{cases} \frac{g^-(Q_{t_j},t_j)[1..y]-g^-(Q_{t_j},t_j)[1..x]}{g^-(Q_{t_j},t_j)[1..y]} & if\, g^-(Q_{t_j},t_j)[1..y] \geq g^-(Q_{t_j},t_j)[1..x] \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

$$\sigma^=(Q_{t_j},t_j)[x..y] = \begin{cases} \frac{g^=(Q_{t_j},t_j)[1..y]-g^=(Q_{t_j},t_j)[1..x]}{g^=(Q_{t_j},t_j)[1..y]} & if\, g^=(Q_{t_j},t_j)[1..y] \geq g^=(Q_{t_j},t_j)[1..x] \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

In particular, $\sigma^+(Q_{t_j},t_j)[x..y]$ (resp., $\sigma^-(Q_{t_j},t_j)[x..y]$, $\sigma^=(Q_{t_j},t_j)[x..y]$) is an indicator of the variation of the number of users of Q_{t_j} who are predominantly positive (resp., predominantly negative, ambivalent) on t_j in the time interval $T[x..y]$. Observe that $\sigma^+(Q_{t_j},t_j)[x..y]$ (resp., $\sigma^-(Q_{t_j},t_j)[x..y]$, $\sigma^=(Q_{t_j},t_j)[x..y]$) ranges in the real interval $[0,1]$. The higher its value, the greater the increase of the number of predominantly positive (resp., predominantly negative, ambivalent) users of Q_{t_j} on t_j.

3.3.4. Analyzing the cross-contamination between a community and its neighborhood

In this section, we aim to study the cross-contamination between a community and its neighborhood. By this term we mean the ability of a community to influence the sentiment of its neighborhood on a topic, and vice versa.

Let t_j be a topic of $\mathcal{T}$, let Q_{t_j} be a clique of users interested in Q_{t_j}, and let $\Gamma_{Q_{t_j}}$ be the neighborhood of Q_{t_j}. In the following of this section, to avoid burdening the formalism, we will use the symbols Q and Γ_Q, instead of Q_{t_j} and $\Gamma_{Q_{t_j}}$, because there is no risk of misunderstanding.

The first metric we consider measures the attractiveness of Q for Γ_Q in a time interval $T[x..y]$. It evaluates the fraction of nodes of Γ_Q that have been incorporated into Q in the time interval $T[x..y]$. It can be defined as:

$$\rho(Q, \Gamma_Q)[x..y] = \frac{passing(\Gamma_Q, Q)[x..y]}{size(\Gamma_Q)[1..x]} \tag{3.12}$$

In this formula, $passing(\Gamma_Q, Q)[x..y]$ denotes the number of nodes belonging to Γ_Q in the time slice T_x that became nodes of Q in the time slice T_y. Here, we assume that Γ_Q is non-degenerate, and thus consists of at least one node. Note that $\rho(Q, \Gamma_Q)[x..y]$ ranges in the real interval $[0, 1]$; the higher its value, the greater the attractiveness of Q for Γ_Q.

The second metric we consider is dual to the previous one and measures the attractiveness of Γ_Q for Q in the time interval $T[x..y]$. It evaluates the fraction of nodes of Q that have been incorporated into Γ_Q. It can be defined as follows:

$$\rho'(\Gamma_Q, Q)[x..y] = \frac{passing(Q, \Gamma_Q)[x..y]}{size(Q)[1..x]} \tag{3.13}$$

Also in this case, we assume that Q is non-degenerate; $\rho'(\Gamma_Q, Q)[x..y]$ ranges in the real interval $[0, 1]$; the higher its value, the greater the attractiveness of Γ_Q for Q.

Moving a node from Q to Γ_Q, or vice versa, implies a strong change in the scenario regarding Q and Γ_Q and, thus, a strong form of cross-contamination. Actually, it is possible to think of weaker, and possibly slower, forms of cross-contamination. They occur when elements of Γ_Q, after interacting with elements of Q, change their sentiment on a topic, and vice versa. This happens when Q (resp., Γ_Q), which has the majority of its members with a certain sentiment (i.e., predominantly positive, predominantly negative, ambivalent) on t_j, pushes toward the same sentiment some elements of Γ_Q (resp., Q) that previously had a different sentiment on t_j.

Based on this reasoning, we can introduce a third metric of cross-contamination. It measures the influence exerted by Q on Γ_Q and is defined as follows:

$$\varphi(Q, t_j)[x..y] = \begin{cases} \frac{g^+(\Gamma_Q, t_j)[1..y] - g^+(\Gamma_Q, t_j)[1..x]}{size(\Gamma_Q[1..x])} & if\ \ pp(Q[x], t_j) = \texttt{true} \text{ and } pp(\Gamma_Q[x], t_j) = \texttt{false} \text{ and} \\ & (g^+(\Gamma_Q, t_j)[1..y] > g^+(\Gamma_Q, t_j)[1..x]) \\ \frac{g^-(\Gamma_Q, t_j)[1..y] - g^-(\Gamma_Q, t_j)[1..x]}{size(\Gamma_Q[1..x])} & if\ \ pn(Q[x], t_j) = \texttt{true} \text{ and } pn(\Gamma_Q[x], t_j) = \texttt{false} \text{ and} \\ & (g^-(\Gamma_Q, t_j)[1..y] > g^-(\Gamma_Q, t_j)[1..x]) \\ \frac{g^=(\Gamma_Q, t_j)[1..y] - g^=(\Gamma_Q, t_j)[1..x]}{size(\Gamma_Q[1..x])} & if\ \ pa(Q[x], t_j) = \texttt{true} \text{ and } pa(\Gamma_Q[x], t_j) = \texttt{false} \text{ and} \\ & (g^=(\Gamma_Q, t_j)[1..y] > g^=(\Gamma_Q, t_j)[1..x]) \\ 0 & \text{otherwise} \end{cases} \tag{3.14}$$

The rationale behind this formula is the following: if, at the time slice T_x, Q is predominantly positive (resp., predominantly negative, ambivalent) and Γ_Q is not, and, at the time slice T_y, $y > x$, the number of predominantly positive (resp., predominantly negative, ambivalent) users in Γ_Q increases, then it is possible to conclude that Q has really exerted an influence on Γ_Q. An indicator of this influence can be the fraction of the users of Γ_Q who became predominantly positive (resp., predominantly negative, ambivalent) during the interval $T[x..y]$. In all the other cases, the influence exerted by Q on Γ_Q is null. Here, we assume that Γ_Q is non-degenerate. Observe that $\varphi(Q, t_j)$ ranges in the real interval $[0, 1]$; the higher its value, the greater the influence that Q exerted on Γ_Q.

The fourth and last cross-contamination metric is the dual of the third. It measures the influence exerted by Γ_Q on Q and is defined as follows:

$$\varphi'(\Gamma_Q, t_j)[x..y] = \begin{cases} \frac{g^+(Q, t_j)[1..y] - g^+(Q, t_j)[1..x]}{size(Q[1..x])} & if\ \ pp(\Gamma_Q[x], t_j) = \texttt{true} \text{ and } pp(Q[x], t_j) = \texttt{false} \text{ and} \\ & (g^+(Q, t_j)[1..y] > g^+(Q, t_j)[1..x]) \\ \frac{g^-(Q, t_j)[1..y] - g^-(Q, t_j)[1..x]}{size(Q[1..x])} & if\ \ pn(\Gamma_Q[x], t_j) = \texttt{true} \text{ and } pn(Q[x], t_j) = \texttt{false} \text{ and} \\ & (g^-(Q, t_j)[1..y] > g^-(Q, t_j)[1..x]) \\ \frac{g^=(Q, t_j)[1..y] - g^=(Q, t_j)[1..x]}{size(Q[1..x])} & if\ \ pa(\Gamma_Q[x], t_j) = \texttt{true} \text{ and } pa(Q[x], t_j) = \texttt{false} \text{ and} \\ & (g^=(Q, t_j)[1..y] > g^=(Q, t_j)[1..x]) \\ 0 & \text{otherwise} \end{cases} \tag{3.15}$$

Also in this case, we assume that Q is non-degenerate. $\varphi'(\Gamma_Q, t_j)$ ranges in the real interval $[0, 1]$; the higher its value, the greater the influence that Γ_Q exerted on Q.

A list of the main notations introduced in Section 3.3 is reported in Table 2.

3.4. Analysis of the computational complexity of the proposed framework

In this section, we present a brief discussion of the computational complexity characterizing the proposed framework. In carrying out this task, we consider both spatial and time complexity. We define the computational complexity of a single application of our framework, which is the scenario we presented in Section 3.3. For this purpose, we first investigate the space and time complexity of the model computation. Then, we study the space and time complexity of the tasks performed by our framework. Without loss of generality, we assume that all the data required for the scenario of interest are available and preprocessed.

Model construction consists of three tasks. The first one is the building of the networks $\mathcal{B}$ and $\mathcal{A}$. These two networks model users, topics and sentiment tags. The time complexity of their construction is $\mathcal{O}(q \cdot l \cdot r + l^2)$, where l is the number of users, q is

Table 2
Summarizing table of the notations introduced in Section 3.3.

Notation	Description
$det(\cdot)$	A function that returns how determinant a user is in conditioning a community with their sentiment
$\alpha, \beta, \gamma, \delta, \varrho$	Weights of the five components of $det()$
$degree_{max}$	Maximum degree of a user in $\mathcal{A}$
$e\text{-}size_{max}$	Size of the maximum ego network in $\mathcal{A}$
$\iota(\cdot)$	A function that returns the change in interest of a community on a topic
$\sigma^{+}(\cdot)$, $\sigma^{-}(\cdot)$, $\sigma^{=}(\cdot)$	Functions returning the evolution of the positive, negative and ambivalent sentiment of a community on a topic
$passing(\cdot,\cdot)$	A function that returns the number of users passing from a community to another
$\rho(\cdot)$	A function that returns the attractiveness that a community exerts on its neighborhood
$\rho'(\cdot)$	A function that returns the attractiveness that a community experiences from its neighborhood
$\varphi(\cdot)$	A function that returns the influence that a community exerts on its neighborhood
$\varphi'(\cdot)$	A function that returns the influence that a community undergoes from its neighborhood

the number of topics and r is the number of sentiment tags. The second task carries out the computation of the ego network for a given user; its time complexity is $\mathcal{O}(l^2)$. The third task performs the computation of the clustering coefficient of the ego network of a given user; its time complexity is $\mathcal{O}(l^2)$. Aggregating the three tasks we obtain that the overall time complexity is $\mathcal{O}(q \cdot l \cdot r + l^2)$. Instead, the overall space complexity for model construction is $\mathcal{O}(q \cdot l^2 \cdot r)$.

Our framework performs three main activities. The first one concerns the identification of determinant users. The time complexity of this activity can be traced back to the definition of $det(\cdot,\cdot)$, reported in Eq. (3.4). Assuming that the results of the functions complementing our model are pre-computed, the time complexity is $\mathcal{O}(n)$, where n is the size of available comments. This holds for each pair of user and topic. The second task involves analyzing the evolution of a user or community sentiment on a topic. Its time complexity depends linearly on the size of the time interval of interest. Accordingly, it is $\mathcal{O}(z)$, where z is the number of time slices in the scenario of interest. The third activity regards the analysis of the cross-contamination between a community and its neighborhood. This depends on both the size of the community and that of its neighborhood. Therefore, its time complexity is equal to $\mathcal{O}(l^2)$. Finally, by aggregating these three activities, we can express the overall time complexity as $\mathcal{O}(n + z + l^2)$. As for the overall space complexity, we point out that these activities are carried out on the model. Therefore, they do not require any additional resources, except for the third activity that requires the computation of the attractiveness. The latter was defined in Section 3.3.4 and its space complexity is $\mathcal{O}(l^2)$.

4. Experimental campaign

In this section, we propose a series of experiments based on the model and the framework introduced above. In particular, we start by describing the dataset on which our experiments have been performed (Section 4.1). Then, we explain the process behind the identification of the topics and the related sentiments (Section 4.2). After that, we present an exploratory analysis of the users and communities characterizing our dataset (Section 4.3). Finally, we illustrate the experiments associated with each of the three activities implemented in our framework (Sections 4.4–4.6).

4.1. Dataset description

In order to show how our model and framework work and highlight their potential in analyzing the behavior over time of users posting opinions on a social platform, we carried out an experimental campaign. To this end, we chose Reddit as our reference social platform. The reasons for this choice lie in the fact that: *(i)* it is one of the most popular social platforms in the world (it currently ranks 20th among the most visited sites according to Visual Capitalist[4];) *(ii)* it allows users to publish posts and comments on any topic; *(iii)* its data can be easily accessed through `pushshift.io` [40], which provides an API to download posts and comments from it.

In our experimental campaign, we focused on posts and comments published on the subreddit /r/worldnews. We decided to focus on it both because it has already been used as a reference in other studies [41–43] and because it is considered one of the most comprehensive and neutral news-related subreddits. Thanks to `pushshift.io`, we retrieved all the posts and corresponding comments, along with their related metadata, published on this subreddit from February 25, 2022 to March 25, 2022. The number of posts composing our dataset is 9,884 while the number of comments is 633,371.

To refine the initially collected data, we performed a preliminary ETL (Extraction, Transformation and Loading) activity on it. During this task, we decided to remove all the posts and comments published by users who had left Reddit. In addition, we removed all the posts and comments that did not have text content or were written in a language other than English. As for the discussion theme, we selected a specific one, but which had many facets (and thus many topics). In particular, we chose the armed conflict in Ukraine that began on February 24, 2022.

[4] www.visualcapitalist.com

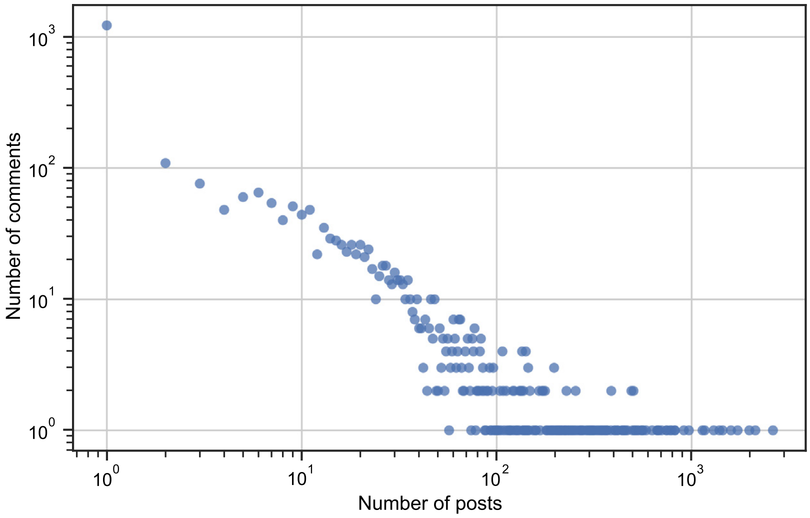


Fig. 2. Distribution of comments against posts (log–log scale).

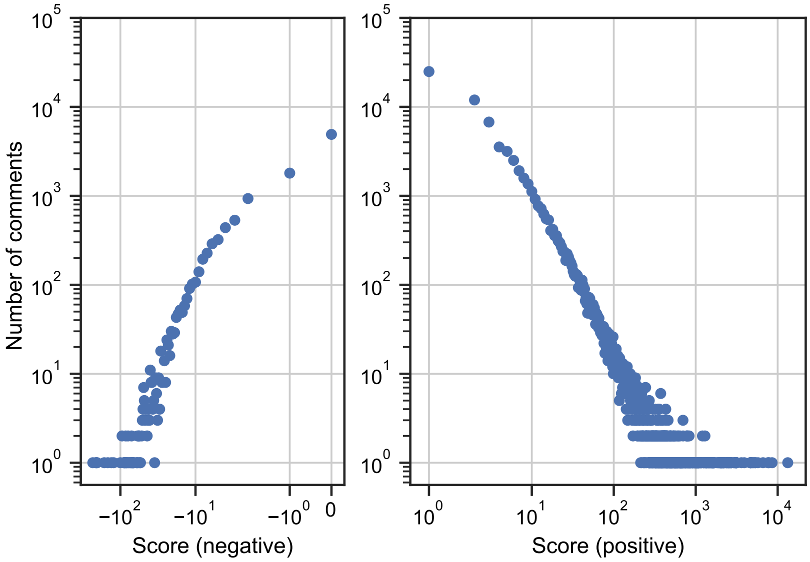


Fig. 3. Distribution of comments against score (log–log scale).

As a result of all the activities described above, the final number of available posts is 2,703, representing 27.12% of the initial ones. In addition, the final number of comments was 82,617, corresponding to 13.21% of the initial ones. In Table 3, we report some of the main characteristics of the final dataset. This table already highlights some interesting details. In fact, it indicates that the number of interacting authors in our dataset is 41,191. A very interesting feature is the fact that the number of authors publishing both posts and comments is very low. In fact, it is equal to 814, which is 18.35% of the authors publishing posts and 2.17% of the authors publishing comments.

Fig. 2 shows the distribution of comments against posts, while Fig. 3 reports the distribution of comments against score. These figures are in log–log scale. As can be seen from them, all the reported distributions follow power laws. In Table 4 we show the values of the coefficients α and δ related to this distribution.

4.2. Identification of topics and sentiments

As specified in Section 3.1.1, our model and framework are orthogonal to the approach that can be selected for constructing the set $\mathcal{T}$ of topics. In our experimental campaign, we adopted BERTopic [44] for this purpose. BERTopic is based on BERT (Bidirectional Encoder Representation from Transformers), a powerful deep learning based framework to perform NLP tasks on texts. It is a topic

Table 3
Some main parameters of the considered dataset.

Parameter	Value
No. of posts	2,703
No. of comments	82,617
No. of (distinct) authors	41,191
No. of (distinct) authors publishing posts	4,435
No. of (distinct) authors publishing comments	37,570
No. of (distinct) authors publishing both posts and comments	814

Table 4
Values of α and δ of the power law distributions for the considered dataset.

Distribution	α	δ
Fig. 2	1.8408	0.0419
Fig. 3 (left)[a]	2.9262	0.0418
Fig. 3 (right)	2.0383	0.0136

[a]These values were computed considering the absolute values of scores

Table 5
Some examples of the topics and their descriptions extracted by BERTopic.

Topic	Description
t_1	$\{\mathsf{invasion}, \mathsf{invade}, \mathsf{mission}\}$
t_2	$\{\mathsf{nato}, \mathsf{defence}, \mathsf{member}, \mathsf{treaty}\}$
t_3	$\{\mathsf{bunker}, \mathsf{underground}\}$

Table 6
Some examples of fragments and their extrapolated sentiments obtained by roBERTa-base (swear words are partially masked).

Fragment	Sentiment
"It makes me hopeful too. We need to find a way to get NATO forces engaged."	pos
"But its a f***ing kid that got killed by that c**t"	neg
"Anyone know when this interview took place? NBC has no time stamp on the video"	neu

modeling technique that relies on transformers [45] and c-TF-IDF [46] to create dense clusters that allow for easily interpretable topics. Starting from a set of documents, BERTopic returns a list of topics and, for each of them, its description and count. The description of a topic is a set containing the most important words characterizing it. Therefore, it can be intended as a human-readable interpretation of the topic itself. The count of a topic indicates the number of documents mentioning it. Given a document, it is always possible to retrieve the topics mentioned in it.

Applying BERTopic to the 2,703 posts and 82,617 comments in the dataset, we obtained a set $\mathcal{T}$ of 101 topics. Some examples of the extracted topics are reported in Table 5.

After constructing the set $\mathcal{T}$ of topics, we proceeded to consider the sentiments that characterized user opinions. In more detail, to extract the sentiment expressed in a given textual content, we used roBERTa-base [47]. This system was trained on approximately 124 million tweets published from January 2018 to December 2021. Next, it was expressively fine-tuned for sentiment analysis using the TweetEval benchmark [48]. The rationale behind the adoption of roBERTa-base is to be found in the shape similarity between the textual content of comments (or portion of comments) in Reddit and tweets in Twitter. In fact, both of them represent fast-paced messages employed to express opinions and thoughts in general. Furthermore, the subreddit we used for our dataset, namely /r/worldnews, deals with news-related content. In this context, users tend to discuss and post comments in a fast-paced way.

The set of sentiments that can be extrapolated is $\mathcal{S} = \{\mathsf{neg}, \mathsf{neu}, \mathsf{pos}\}$. In Table 6, we show some examples of fragments with the correlated sentiment extracted by roBERTa-base. Given a comment, or a fragment of it, say f_k, characterized by a unique sentiment, and given the set $\mathcal{T}_{f_k}$ of topics discussed in it as identified by BERTopic, we can associate each topic of $\mathcal{T}_{f_k}$ with the sentiment s_k that roBERTa-base extracted for f_k. This association leads to the construction of sentiment-tagged topics expressed as pairs (t_j, s_k). They indicate that the sentiment s_k was associated with the topic t_j in f_k. As mentioned above, 101 topics were identified for our dataset. From them, 302 sentiment-tagged topics were obtained.

4.3. Some exploratory data analyses on users and communities

In this section, we present some exploratory data analyses that we conducted on the users and communities of our dataset. Our ultimate goal was to understand the variation of some of their distributions over time. Indeed, such knowledge is preliminary to the tests that we will present in the next sections, which cover the three main activities of our framework. Recall that the time interval

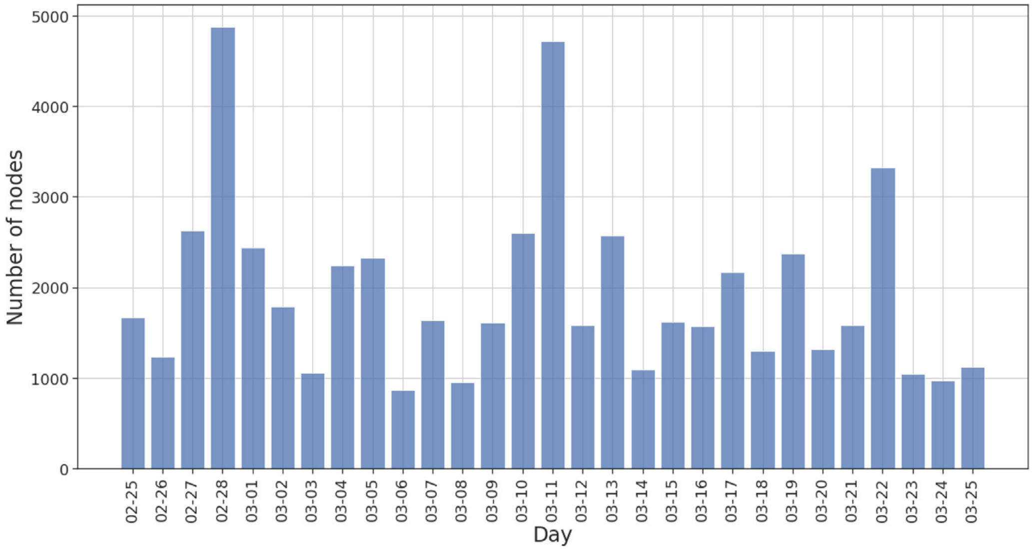


Fig. 4. Number of nodes per day in $\mathcal{A}$.

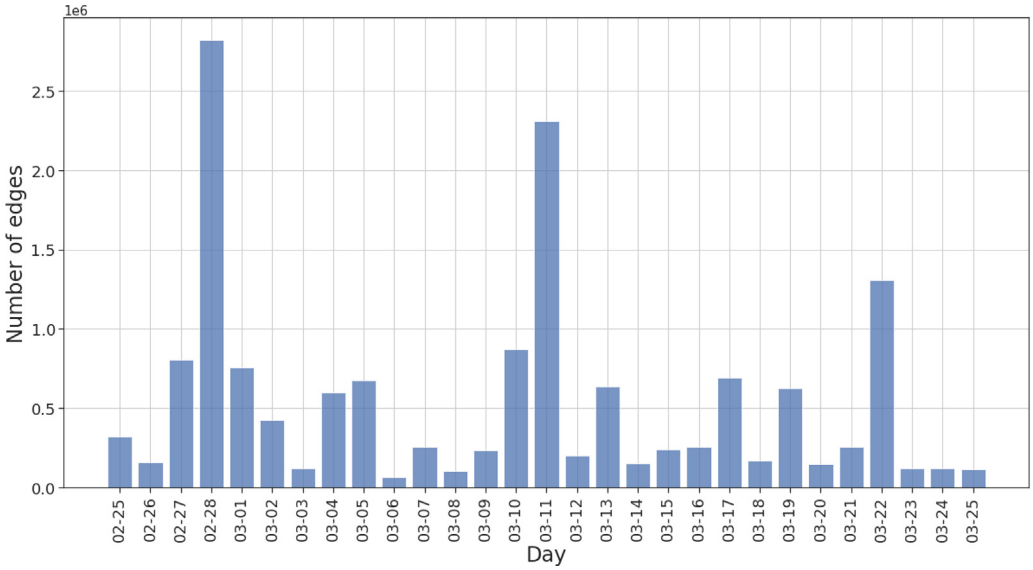


Fig. 5. Number of edges per day in $\mathcal{A}$.

T to which our data refer is from February 25, 2022 to March 25, 2022. In our case, each time slice corresponds to one day; hence, for instance, $T[1]$ represents the first day, $T[1..7]$ the first week, $T[11..20]$ the second decade, and so on.

We begin our analysis by considering the distribution of users in the time interval T. It is shown in Fig. 4. In that figure, the x-axis reports the single time slices, and thus the single days, while the y-axis reports the number of nodes of the network $\mathcal{A}$, and thus the number of active users. Some interesting considerations can be drawn from the analysis of this figure. First, it can be seen that the distribution is irregular, in that it does not exhibit any periodicity over time. The number of users on one day seems to be totally unrelated to the number of users on other days. Note that, in the time slices corresponding to February 28, March 11 and March 22, there is a higher number of users than on the other days.

To complement these observations, in Fig. 5, we show the distribution of the number of edges of the network $\mathcal{A}$ against time. Recall that the edges in this network represent user interactions. As it might be expected, this figure shows a pattern similar to the one of Fig. 4. In fact, the distribution is irregular, and on February 28, March 11 and March 22 there is a higher number of edges than on the other days.

After looking at users and their relationships over time, we also want to take a look at communities. In particular, we want to analyze how they change over time and whether they have users in common. In Fig. 6, we show the variation in the number of

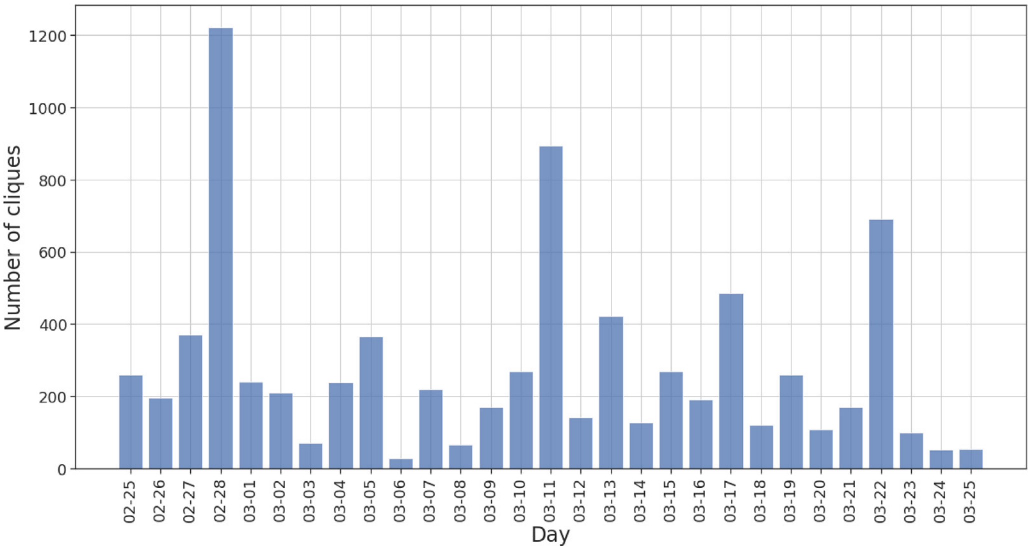


Fig. 6. Number of cliques per day in $\mathcal{A}$.

communities over time. As in the previous cases, on the x-axis we report time, specifically the single days of the period T under consideration, while on the y-axis we report the number of communities. Again, we can observe that the distribution of communities is irregular and that on February 28, March 11 and March 22 the number of communities is higher than on the other days. Actually, we can see that the distribution of communities fairly closely reflects that of nodes and edges. This tells us that, in our scenario, the presence of more users tends to generate more interconnections (which was not obvious since there could have been a lot of users not communicating with each other) and, therefore, more communities.

In order to try to give an explanation for spikes in Figs. 4–6, we investigated to see whether there were any particular events in the war in Ukraine during the corresponding three days. Indeed, this research gave us some interesting confirmations. In fact:

- On February 28, 2022, there was the first round of peace talks between Russia and Ukraine in Gomel, Belarus. These were the first peace talks since the beginning of the war. On the same day, the city of Kharkhiv, which is the second most popular city in Ukraine, was bombed. The bombing left dozens dead and hundreds wounded.
- On March 10, 2022, there was the first meeting between the foreign ministers of Russia and Ukraine in Turkey.
- On March 22, 2022, Russia launched the first Kinzhal hypersonic missile on Kiev. On the same day, the Ukrainians liberated the town of Makarov. Finally, in Mariupol, the Russians almost completed the occupation of the city, and house-to-house fighting took place.

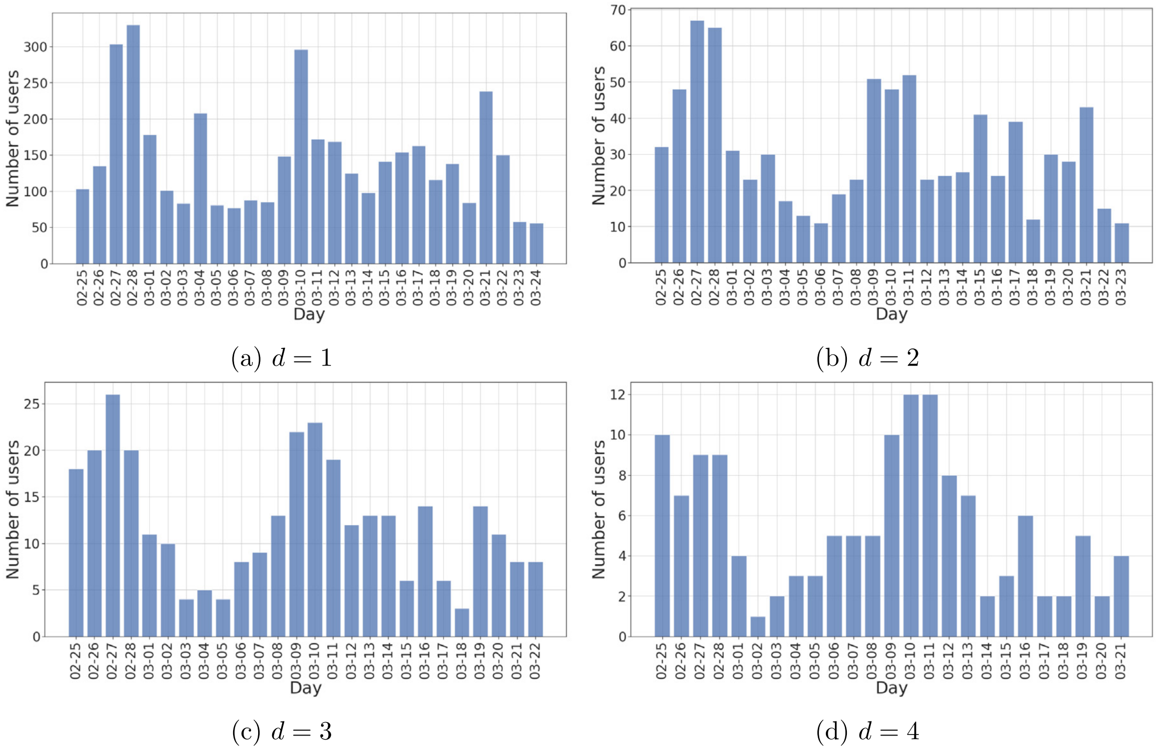
The above events are all extremely important facts that stimulate posts and comments. Therefore, they represent a possible explanation for the increase in the number of nodes, arcs and cliques in $\mathcal{A}$, and more generally, for the debates of the /r/worldnews users regarding the war in Ukraine.

In our opinion, it may be interesting to see whether the communities on different days are made up of the same users; indeed, this would indicate a certain stability of these communities. In order to investigate this phenomenon, we analyzed whether the cliques on different consecutive days have intersections or not. To this end, for each time slice, we determined the number of users in common between a community of that day and a community of the next d days, with $d = 1..4$. The results of these computations are shown in Fig. 7. From the analysis of this figure we can observe that the peaks on the three days February 28, March 11 and March 22 are still visible for $d = 1$ while, for the next values of d, they tend to disappear. The presence of the peaks around the three days highlighted above when $d = 1$ makes us say that the cliques characterizing the communities on the three days involve users operating both before and after the peak day. Thus, these are regular users or, in other words, users who did not specifically enter only on the peak day. It is also interesting to note the magnitude of common users against d. In fact, for $d = 1$, the maximum number of common users is close to 350; this number quickly drops to 12 for $d = 4$. For $d > 4$ it is very low; therefore, we do not report it because it carries no relevant information.

Therefore, we can conclude that the presence of common users among different communities is significant only in the very short period, that is, if we consider communities enclosed in a time interval of 2–3 days.

4.4. Analysis of determinant users

The first activity of our framework concerns the identification of the most determinant users. In Section 3.3.2, we introduced the function $det(\cdot,\cdot)$, which receives a user u_i and a topic t_j as input and returns a value in the real interval $[0, 1]$, which indicates



(a) $d = 1$ (b) $d = 2$

(c) $d = 3$ (d) $d = 4$

Fig. 7. Number of users in common between the communities of a day and the next d day(s).

how determinant u_i is in influencing a community regarding its overall sentiment on t_j. The higher the value returned by $det(\cdot,\cdot)$, the more decisive u_i will be in conditioning the community's sentiment on t_j.

As we have seen in Section 3.3.2, the function $det(\cdot,\cdot)$ consists of a weighted mean of five factors representing five aspects that come into play in establishing how determinant u_i is. In the following, we call $det_\alpha(\cdot,\cdot)$ (resp., $det_\beta(\cdot,\cdot)$, $det_\gamma(\cdot,\cdot)$, $det_\delta(\cdot,\cdot)$, $det_\varrho(\cdot,\cdot)$) the component of $det(\cdot,\cdot)$ associated with the weight α (resp., β, γ, δ, ϱ). The value of $det_\alpha(u_i,t_j)$ (resp., $det_\beta(u_i,t_j)$, $det_\gamma(u_i,t_j)$, $det_\delta(u_i,t_j)$, $det_\varrho(u_i,t_j)$) can be obtained by applying the formula for computing $det(u_i,t_j)$ - see Eq. (3.4) - and setting α (resp., β, γ, δ, ϱ) to 1 and all the other weights to 0.

In the analysis of determinant users, it might be extremely interesting to compare the five components that contribute to $det(\cdot,\cdot)$. In fact, this could provide a better understanding of the role each of them plays in identifying the most determinant users. Furthermore, it might be extremely interesting to see how much the users being determinant on the basis of one of these components are also determinant based on other components, as well as based on the overall $det(\cdot,\cdot)$.

In order to proceed with this analysis, we defined the set ψ_α (resp., ψ_β, ψ_γ, ψ_δ, ψ_ϱ) of the top determinant users identified by considering only $det_\alpha(\cdot,\cdot)$ (resp., $det_\beta(\cdot,\cdot)$, $det_\gamma(\cdot,\cdot)$, $det_\delta(\cdot,\cdot)$, $det_\varrho(\cdot,\cdot)$). Furthermore, we defined the set ψ_{det} of the most determinant users identified by taking the overall function $det(\cdot,\cdot)$ into consideration. In order to determine the value of the weights in the formula of $det(\cdot,\cdot)$ when computing ψ_{det}, we considered a series of preliminary tests with the aim of finding their optimal values. Due to space limitations, we cannot report these experiments in detail in this paper. However, we report that the optimal values finally obtained are as follows: $\alpha = 0.30$, $\beta = 0.20$, $\gamma = 0.15$, $\delta = 0.15$, $\varrho = 0.20$.

Once all these details regarding the experiment were settled, the last decision to be made concerned the size of ψ_α (resp., ψ_β, ψ_γ, ψ_δ, ψ_ϱ, ψ_{det}).[5] In order to consider several scenarios potentially different from each other, we chose different values for this size. Specifically, the values we chose were: 100, 500, 1,000 and 10,000 (recall that the total number of users in our dataset is 41,191).

As mentioned above, our goal is to compare the contribution of the five components of $det(\cdot,\cdot)$. One way to do this is the computation of the intersection between the various sets of the most determinant users defined above. In Table 7, we report the values obtained from the intersections focusing on the ones that we consider most significant.

From the analysis of this table, we can draw several interesting knowledge patterns. First, we note the small intersection existing between all pairs related to ψ_α, ψ_β, ψ_γ, ψ_δ and ψ_ϱ, and between each of them and ψ_{det}. This is the case for the top 100, the top 500 and the top 1,000 users. This result represents a confirmation that the various components of det capture different, but all interesting, facets related to a user's ability to be determinant for a sentiment on a topic. A separate discussion deserves the top 10,000 users, where the intersections between the various sets are much greater. This can be explained by taking into account that,

[5] It is worth pointing out that, in order to make a comparison, the size of ψ_α, ψ_β, ψ_γ, ψ_δ, ψ_ϱ, and ψ_{det} must be equal to each other.

Table 7
Size of the intersections between several sets of top determinant users.

	Top 100	Top 500	Top 1,000	Top 10,000
$\psi_\alpha \cap \psi_\beta$	12	72	180	3,956
$\psi_\alpha \cap \psi_\gamma$	10	65	158	2,953
$\psi_\alpha \cap \psi_\delta$	11	78	161	3,541
$\psi_\alpha \cap \psi_\varrho$	9	52	148	2,845
$\psi_\beta \cap \psi_\gamma$	18	106	265	5,824
$\psi_\beta \cap \psi_\delta$	14	83	208	4,578
$\psi_\beta \cap \psi_\varrho$	17	105	260	5,718
$\psi_\gamma \cap \psi_\delta$	34	206	512	8,424
$\psi_\gamma \cap \psi_\varrho$	25	149	371	7,148
$\psi_\delta \cap \psi_\varrho$	24	143	356	6,984
$\psi_{det} \cap \psi_\alpha$	33	201	502	8,221
$\psi_{det} \cap \psi_\beta$	28	170	424	7,918
$\psi_{det} \cap \psi_\gamma$	26	158	395	7,724
$\psi_{det} \cap \psi_\delta$	24	146	365	7,221
$\psi_{det} \cap \psi_\varrho$	25	149	372	7,418
$\psi_{det} \cap \psi_\alpha \cap \psi_\beta \cap \psi_\gamma \cap \psi_\delta \cap \psi_\varrho$	8	41	125	2,642

in this case, we are considering as top users a large set of the total ones (i.e., about 25% of them). In such a scenario, it is expected that the overlaps between the various sets are large.

Regarding the single rows of Table 7, we can see that the intersection between ψ_γ and ψ_δ is always larger than the others. This is justified by the fact that the corresponding components refer to two aspects that quantify the same phenomenon, namely the ability of users to be seeds in creating new communities. In ψ_γ this ability is measured through the size of ego networks, while in ψ_δ the same is measured through the clustering coefficient (see Section 3.3.2).

Finally, an interesting result is given by the last row of Table 7. In it, we consider the intersection between all the sets of determinant users. In this way, we aim to find out the "power determinant" users, that is, those users who, regardless of the parameters and viewpoints chosen to evaluate a user's ability to be determinant, are selected as top users. As we can see, there are users with such characteristics. Their number is very low if we consider the top 100 and top 500 users; it starts to grow if we consider the top 1,000 users. Finally, it becomes significant if we take the top 10,000 users into account.

4.5. Analysis of the evolution of the sentiment

The second activity of our framework is the analysis of the evolution of a user or community sentiment on a topic. In this section, we illustrate some experiments regarding this activity. The first function we wanted to analyze was the function $\iota(Q_{t_j}, t_j)$, defined in Section 3.3.3, which indicates the change in interest of the community Q_{t_j} on the topic t_j in a given time interval. In particular, for each day, we considered the change in interest of Q_{t_j} on t_j with respect to the previous day. To perform this experiment, given a time slice of interest, we sorted all cliques in that time slice in descending order of size. Then, we constructed the set Q_{top} (resp., Q_{mid}, Q_{bot}) comprising the cliques that were in the first (resp., fifth and sixth, tenth) decile. For each clique Q_{t_j} of Q_{top} (resp., Q_{mid}, Q_{top}) we calculated the value of $\iota(Q_{t_j}, t_j)$. We repeated this calculation by randomly selecting 30 of the 101 topics in our dataset. Finally, we computed the average of the values of $\iota(Q_{t_j}, t_j)$ with respect to the elements of Q_{top} (resp., Q_{mid}, Q_{bot}) and the selected topics. In Fig. 8, we show the results obtained by examining the usual time interval of our dataset, i.e., the one ranging from February 25 to March 25. Note that in this figure, as well as in all the next ones in which the parameter value shown is obtained as the difference between what happens on two consecutive days, the first value reported is the one corresponding to February 26.

From the analysis of this figure we can deduce the following observations:

- The function $\iota(Q_{t_j}, t_j)$ has a fairly irregular trend for Q_{bot}. When we move to Q_{mid}, it begins to be more regular. Finally, when Q_{top} is considered, it is fairly regular. This can be explained by taking into account that $\iota(Q_{t_j}, t_j)$ represents a relative change in the size of a community from its size at the beginning of the current time slice. Therefore, the same absolute value of variation affects much more (causing irregularities) in small communities than in large ones.
- The average values of the function $\iota(Q_{t_j}, t_j)$, and thus the average changes in the interest of a community, are generally greater in Q_{bot} than in Q_{mid}, which, in turn, are greater than in Q_{top}. This trend can be explained by taking into account, once again, the fact that $\iota(Q_{t_j}, t_j)$ calculates a relative change in the size of a community from its size at the beginning of the current time slice. Clearly, the same absolute value affects small communities much more than medium and large ones.
- The average value of the function $\iota(Q_{t_j}, t_j)$ tends to decrease over time. This can be explained by considering that communities consolidate as time passes, and, thus, it is difficult for them to continue to grow further. Moreover, as time passes, the interest possibly generated by a certain topic fades, and thus the push for community growth reduces.

The second group of functions that we analyzed in these tests are $\sigma^+(u_i, t_j)$, $\sigma^-(u_i, t_j)$ and $\sigma^=(u_i, t_j)$. Recall that these functions indicate the evolution of the positive, negative or ambivalent sentiment of the user u_i on the topic t_j. Similarly to the previous test, in the computation of the values of $\sigma^+(u_i, t_j)$ (resp., $\sigma^-(u_i, t_j)$ and $\sigma^=(u_i, t_j)$) for one day, we considered the changes of the values of

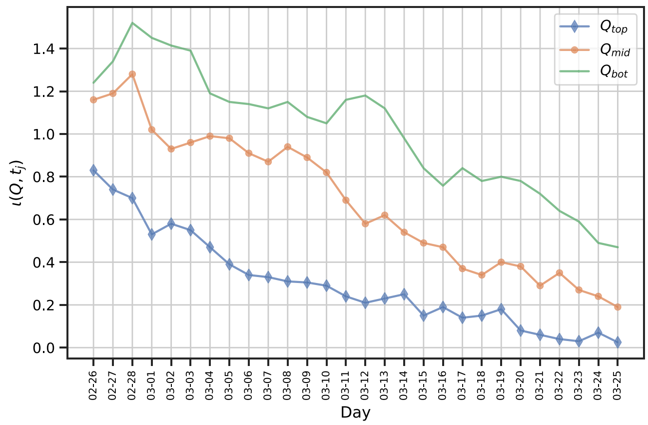


Fig. 8. Variation of the average value of $\iota(Q_{t_j}, t_j)$ over time for Q_{top}, Q_{mid} and Q_{bot}.

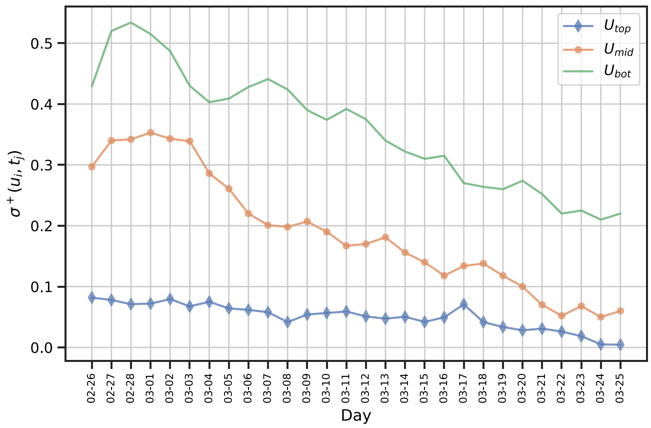


Fig. 9. Variation of the average value of $\sigma^+(u_i, t_j)$ over time for U_{top}, U_{mid} and U_{bot}.

the functions $f^+(u_i, t_j)$ (resp., $f^-(u_i, t_j)$, $f^=(u_i, t_j)$) in that day and in the previous one. In the following, due to space limitations, we consider only $\sigma^+(u_i, t_j)$, although the same reasoning can be repeated for $\sigma^-(u_i, t_j)$ and $\sigma^=(u_i, t_j)$. To carry out this experiment, given a time slice of interest, we sorted all users in that time slice in descending order of the corresponding parameter det (and, thus, sorted them from the most determinant to the least determinant ones). Then, we constructed the set U_{top} (resp., U_{mid}, U_{bot}) by selecting users who were in the first (resp., fifth and sixth, tenth) decile. For each user u_i of U_{top} (resp., U_{mid}, U_{bot}), we calculated the value of $\sigma^+(u_i, t_j)$. We repeated this calculation by randomly selecting 30 of the 101 topics of our dataset. Finally, we averaged the values of $\sigma^+(u_i, t_j)$ over the elements of U_{top} (resp., U_{mid}, U_{bot}) and the selected topics. In Fig. 9 we show the results obtained for the time interval from February 25 to March 25.

From the analysis of this figure, we can deduce the following conclusions:

- The function $\sigma^+(u_i, t_j)$ has an irregular trend for U_{bot}. This trend begins to become more regular for U_{mid} and becomes substantially regular for U_{top}.
- The average value of the function $\sigma^+(u_i, t_j)$ is greater in U_{bot} than in U_{mid}, which, in turn, is greater than in U_{top}.
- The average value of the function $\sigma^+(u_i, t_j)$ tends to decrease over time.

These trends can be explained by considering that $\sigma^+(u_i, t_j)$ represents a relative increase in the posts and comments related to t_j that received a positive sentiment. The users of U_{bot} are the least determinant. Based on the formula of the parameter det defined in Section 3.3.2, this implies that these users tend to be ambivalent, to post little and to be inconsistent. Consequently, the sentiment variation tends to be frequent, and even small variations have large repercussions on $\sigma^+(u_i, t_j)$ because the denominator of the formula defining it tends to be small. When we move to U_{mid}, things tend to stabilize and so the variations are less noticeable and smaller in values. This trend becomes even stronger when moving from U_{mid} to U_{top}. Finally, as time passes, user sentiment tends to stabilize and this explains the decrease in the value of the functions $\sigma^+(u_i, t_j)$ over time.

The third group of functions we analyzed are $\sigma^+(Q_{t_j}, t_j)$, $\sigma^-(Q_{t_j}, t_j)$ and $\sigma^=(Q_{t_j}, t_j)$. They indicate the evolution of the positive, negative or ambivalent sentiment of the community Q_{t_j} on the topic t_j. In what follows, due to space limitations, and in a

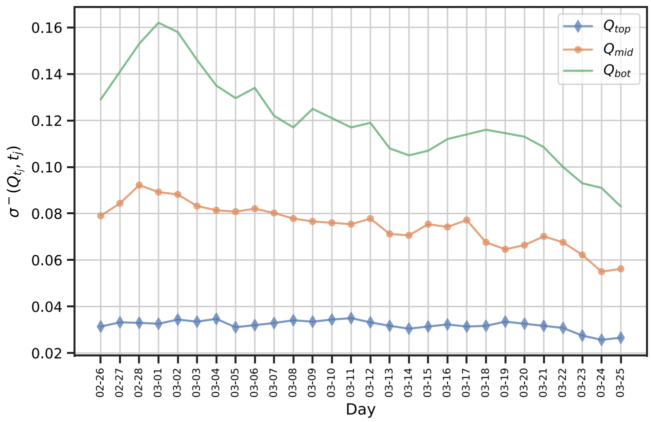


Fig. 10. Variation of the average value of $\sigma^-(Q_{t_j}, t_j)$ over time for Q_{top}, Q_{mid} and Q_{bot}.

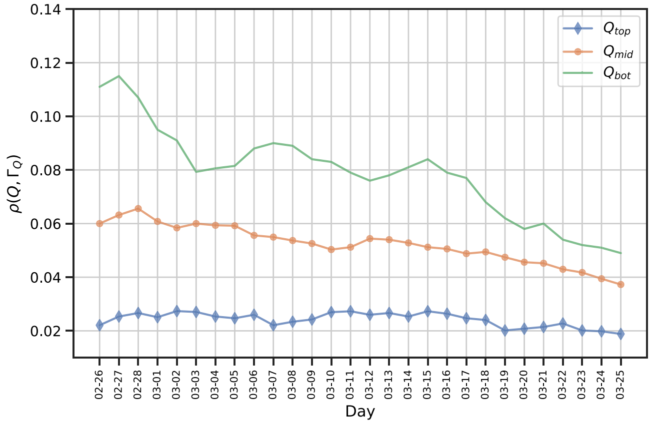


Fig. 11. Variation of the average value of $\rho(Q, \Gamma_Q)$ over time for Q_{top}, Q_{mid} and Q_{bot}.

complementary way with the previous case, we will analyze the trend of $\sigma^-(Q_{t_j}, t_j)$, although the same reasoning applies to $\sigma^+(Q_{t_j}, t_j)$ and $\sigma^=(Q_{t_j}, t_j)$. Also in this experiment, we considered the three sets Q_{top}, Q_{mid} and Q_{bot} defined above and proceeded in a similar way as shown previously. In Fig. 10 we present the results obtained for the time interval from February 25 to March 25.

From the analysis of this figure, we can deduce the same considerations that we had drawn for Fig. 9. The explanations for these results take into account the size of the cliques and the fact that $\sigma^-(Q_{t_j}, t_j)$ represents a relative variation of the size of Q_{t_j} from its actual size. They are similar to those presented in Fig. 8. Therefore, we do not report them for space reasons and not to burden the discussion.

4.6. Analysis of the cross-contamination

The third activity in our framework is the analysis of the cross-contamination between a community and its neighborhood. In this section, we describe the experiments regarding this activity. The first two functions we want to analyze are $\rho(Q, \Gamma_Q)$ and $\rho'(\Gamma_Q, Q)$, which measure the attractiveness of a community for its neighborhood, and vice versa. To carry out this experiment, we used a similar approach as in Section 4.5. Specifically, we considered the time interval from February 25 to March 25 in which each time slice represents one day. In addition, we constructed Q_{top}, Q_{mid} and Q_{bot} using the approach described in Section 4.5. Furthermore, we considered the differences between the current day and the previous one. Finally, we built Γ_{top}, Γ_{mid} and Γ_{bot} by applying on $\Gamma_{Q_{top}}$, $\Gamma_{Q_{mid}}$ and $\Gamma_{Q_{bot}}$ the same procedure used to construct Q_{top}, Q_{mid} and Q_{bot}. In Fig. 11 we show the variation of the average value of $\rho(Q, \Gamma_Q)$ over time for Q_{top}, Q_{mid} and Q_{bot}.

From the analysis of this figure, we can deduce the following observations:

- All communities show some attractiveness to neighborhoods. In the case of Q_{top} this is apparently low; it is higher in the case of Q_{mid} and even higher in the case of Q_{bot}. This seemingly counter-intuitive result is related to the fact that the size of the communities of Q_{top} is large, so whatever variation occurs will still be small compared to the original size of the communities of Q_{top}. The weight of each single variation becomes greater in the case of Q_{mid} and even greater in the case of Q_{bot}.

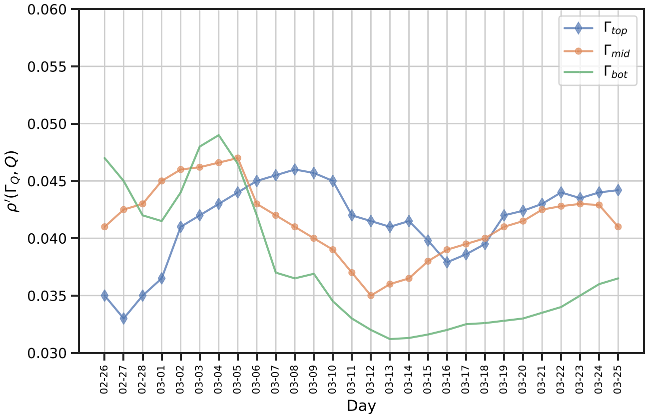


Fig. 12. Variation of the average value of $\rho'(\Gamma_Q, Q)$ over time for Γ_{top}, Γ_{mid} and Γ_{bot}.

- In the case of Q_{top}, the attractiveness remains roughly constant over time; in the case of Q_{mid}, it seems to decrease gradually; finally, in the case of Q_{bot}, it seems to decrease substantially. Also this result should actually be read taking into account the definition of $\rho(Q, \Gamma_Q)$ and the dynamics of communities. In fact, as a community attracts new people it becomes larger, so the weight of each attraction will gradually decrease. This weight was already small in the case of Q_{top}, because the communities were already large, so this difference is not noticeable. This explains the constant trend of $\rho(Q, \Gamma_Q)$ over time for Q_{top}. Instead, in the other two cases, the decrease in the relative weight of each attraction becomes more noticeable and this leads to the decrease in the values of $\rho(Q, \Gamma_Q)$ for both Q_{mid} (in which case, values tend to approach the initial values of $\rho(Q, \Gamma_Q)$ for Q_{top}) and Q_{bot} (in which case, values tend to be close to the initial values of $\rho(Q, \Gamma_Q)$ for Q_{mid}).

In Fig. 12 we show the variation of the average value of $\rho'(\Gamma_Q, Q)$ over time for Γ_{top}, Γ_{mid} and Γ_{bot}.
From the analysis of this figure, we can observe that:

- The trends of the average value of $\rho'(\Gamma_Q, Q)$ over time for Γ_{top}, Γ_{mid} and Γ_{bot} tend to overlap, so there is no substantial differences between the three cases.
- These trends are very irregular over time, with continuous increases and decreases.
- Generally, the values of $\rho'(\Gamma_Q, Q)$ are rather lower than the ones of $\rho(Q, \Gamma_Q)$.

All this leads us to conclude that the attractiveness of neighborhoods to communities is generally lower than the one of communities to neighborhoods. Moreover, the attractiveness of neighborhoods to communities, when it exists, is limited to specific instances and is unlikely to continue over time. In other words, we can say that a well-organized community can exert an attractiveness to its neighborhood. Instead, any reverse process, though possible, is to be considered occasional and related to specific episodes.

As we mentioned in Section 3.3.4, the functions $\rho(Q, \Gamma_Q)$ and $\rho'(\Gamma_Q, Q)$ are indicators of strong forms of cross-contamination. In the same section, we also introduced two indicators of more attenuated forms of cross-contamination. These are represented by the functions $\varphi(Q, t_j)$ and $\varphi'(\Gamma_Q, t_j)$. Therefore, we find it interesting to consider the trends of the average values of the function $\varphi(Q, t_j)$ (resp., $\varphi'(\Gamma_Q, t_j)$) over time for Q_{top}, Q_{mid} and Q_{bot} (resp., Γ_{top}, Γ_{mid} and Γ_{bot}) in the time interval from February 25 to March 25 and compare these trends with the corresponding ones of the functions $\rho(Q, \Gamma_Q)$ (resp., $\rho'(\Gamma_Q, Q)$). The trend of the function $\varphi(Q, t_j)$ (resp., $\varphi'(\Gamma_Q, t_j)$) is shown in Fig. 13 (resp., Fig. 14).

Comparing Fig. 13 with Fig. 11 we can observe that everything we saw for the function $\rho(Q, \Gamma_Q)$ is confirmed, and even amplified, for the function $\varphi(Q, t_j)$. Instead, comparing Fig. 14 with Fig. 12 we can observe that:

- The average values of $\varphi'(\Gamma_Q, t_j)$ are generally greater than the ones of $\rho'(\Gamma_Q, Q)$.
- The trends of $\varphi'(\Gamma_Q, t_j)$ are much less irregular than those of $\rho'(\Gamma_Q, Q)$, in the sense that there are fewer alternations of increases and decreases, which, in turn, are less large.
- Also for the trends of $\varphi'(\Gamma_Q, t_j)$, analogously to what happened for the ones of $\rho'(\Gamma_Q, Q)$, there are no substantial differences between Γ_{top}, Γ_{mid} and Γ_{bot}.

These observations lead us to conclude that, when moving to a weaker form of attractiveness, it becomes even more evident that communities are able to attract neighborhoods. Instead, as far as the opposite case is concerned, we can glimpse some possibility for neighborhoods to exert a weak attraction on communities, but it is more an influence than an attraction. In other words, it can lead to variations in the strength of the sentiment but can hardly result in a change in its polarity.

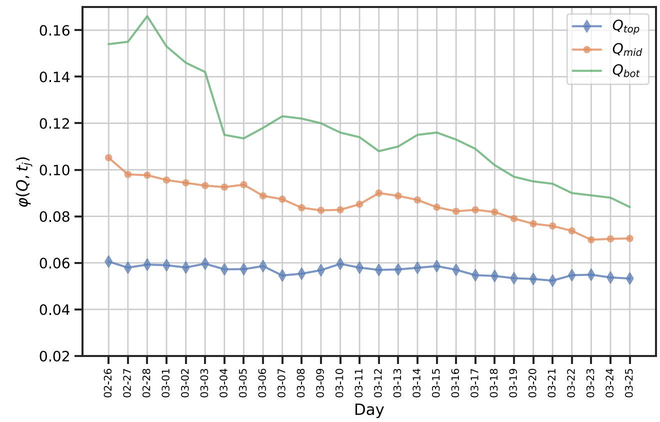


Fig. 13. Variation of the average value of $\varphi(Q, t_j)$ over time for Q_{top}, Q_{mid} and Q_{bot}.

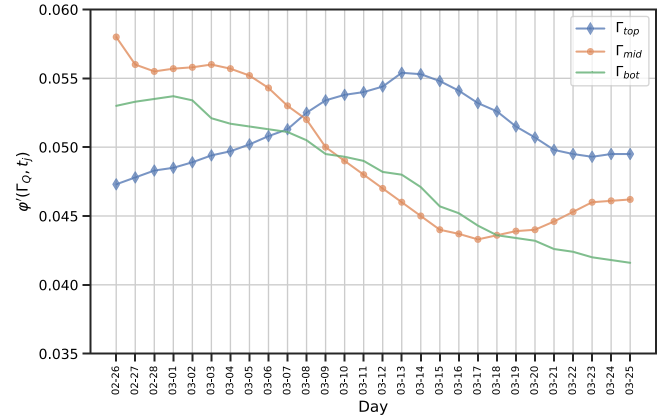


Fig. 14. Variation of the average value of $\varphi'(\Gamma_Q, t_j)$ over time for Γ_{top}, Γ_{mid} and Γ_{bot}.

4.7. Extension from reddit to another social network

All the experiments described above were performed on a dataset derived from Reddit. In fact, the working mechanism of this social network (involving the presence of posts and comments) fits perfectly to be represented by means of our model, and then to be investigated by means of our framework. In this section, we want to test whether the experiences made on Reddit, and the results thus obtained, also apply to other social networks, possibly with a working mechanism different from the one based on posts and comments. To this end, we thought of applying our model and framework to Twitter. Therefore, we derived a dataset of tweets following the same steps seen in Section 4.1 for Reddit.

Using the API of Twitter, we downloaded all the tweets with the hashtag `#UkraineRussianWar` published from February 25, 2022 to March 25, 2022 (which is the same time interval chosen for the previous experiments). After that, we divided the resulting tweets into *original tweets*, i.e., standalone tweets that did not respond to other tweets, and *response tweets*, i.e., tweets published to respond to other tweets. Twitter has much more published content than Reddit. To reduce the size of this content, we filtered the dataset; to this end, we removed original tweets without responses, as they did not allow us to build a community of users. In addition, we removed all tweets with less than three words, as well as those containing only emoji or only hashtags. Table 8 shows the main parameters of the resulting dataset.

We followed the same steps described in Sections 4.2 and 4.3 to determine topics and sentiments. Due to space limitations, we do not describe these steps in detail.

As in the case of Reddit, the first thing to analyze is how determinant a user u_i is in influencing a community on Twitter in terms of its sentiment on a topic t_j. For this purpose, we used the same procedure as described in Section 3.3.2 and, again, considered the top 100, 500, 1,000 and 10,000 users of each of the sets ψ_α, ψ_β, ψ_γ, ψ_δ, ψ_o and ψ_{det} defined in Section 3.3.2. However, we observe that the number of users in the Twitter dataset (i.e., 197,385 - see Table 8) is much higher than the number of users in the Reddit dataset (i.e., 41,919 - see Table 3).

In Table 9, we report the values of the intersections between the sets of top determinant users already considered for Reddit in Table 7. From examining this table we can see that it confirms the various trends seen in Table 7 for the Reddit dataset. Apparently,

Table 8
Some main parameters of the Twitter dataset.

Parameter	Value
No. of original tweets	51,468
No. of response tweets	205,874
No. of (distinct) authors	197,385
No. of authors publishing original tweets	37,503
No. of authors responding to tweets	159,882
No. of authors publishing tweets and responding	8,965

Table 9
Size of the intersections between several sets of top determinant users in the Twitter dataset.

	Top 100	Top 500	Top 1,000	Top 10,000
$\Psi_\alpha \cap \Psi_\beta$	72	358	716	7,160
$\Psi_\alpha \cap \Psi_\gamma$	48	274	548	5,481
$\Psi_\alpha \cap \Psi_\delta$	66	370	740	7,410
$\Psi_\alpha \cap \Psi_\varrho$	42	376	752	7,528
$\Psi_\beta \cap \Psi_\gamma$	86	392	784	7,841
$\Psi_\beta \cap \Psi_\delta$	83	388	776	7,761
$\Psi_\beta \cap \Psi_\varrho$	85	371	742	7,428
$\Psi_\gamma \cap \Psi_\delta$	91	391	782	7,819
$\Psi_\gamma \cap \Psi_\varrho$	93	375	750	7,511
$\Psi_\delta \cap \Psi_\varrho$	90	422	844	8,444
$\Psi_{det} \cap \Psi_\alpha$	97	391	782	7,821
$\Psi_{det} \cap \Psi_\beta$	93	413	826	8,265
$\Psi_{det} \cap \Psi_\gamma$	91	453	906	9,062
$\Psi_{det} \cap \Psi_\delta$	89	481	962	9,622
$\Psi_{det} \cap \Psi_\varrho$	90	362	724	7,243
$\Psi_{det} \cap \Psi_\alpha \cap \Psi_\beta \cap \Psi_\gamma \cap \Psi_\delta \cap \Psi_\varrho$	39	221	488	4,121

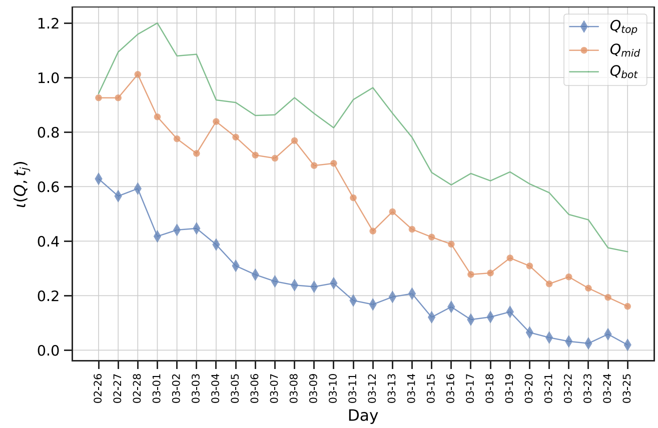


Fig. 15. Variation of the average value of $\iota(Q_{t_j}, t_j)$ over time for Q_{top}, Q_{mid} and Q_{bot} in Twitter.

it seems that the number of users involved in each intersection is higher. Actually, if we look at the percentages of users, instead of the absolute values, we will see that they are comparable or, even, lower. In the top 10,000 columns, the percentages of users involved in the Twitter dataset are much smaller than those observed in the Reddit dataset. However, this can be easily explained. In fact, while in the Reddit dataset the top 10,000 users represent 24.28% of the sample, in the Twitter dataset the top 10,000 users represent 5.07% of the sample. Therefore, if we want to consider percentages instead of absolute values, the percentages of Twitter users referring to the fourth column of Table 9 must be compared with the percentages of Reddit users referring to the first two columns of Table 7. If we proceed with this kind of comparison, we can see that the percentages of users are comparable.

The next step in the framework is the analysis of the sentiment evolution in Twitter, similar to what was done for Reddit in Section 4.5. Fig. 15 (resp., 16, 17) shows the evolution of the function $\iota(Q_{t_j}, t_j)$ (resp., $\sigma^+(u_i, t_j)$, $\sigma^-(Q_{t_j}, t_j)$) for Q_{top} (resp., U_{top}, Q_{top}), Q_{mid} (resp., U_{mid}, Q_{mid}) and Q_{bot} (resp., U_{bot}, Q_{bot}) in Twitter. If we compare Fig. 15 (resp., 16, 17) with the corresponding Fig. 8 (resp., 9, 10) valid for Reddit, we can see that the trends are very similar and the function values are comparable.

The last activity in our framework concerns the analysis of cross-contamination. To perform this task, we followed the same steps described in Section 4.6 for Reddit. In Fig. 18 (resp., 19, 20, 21), we show the trend of the function $\rho(Q, \Gamma_Q)$ (resp., $\rho'(\Gamma_Q, Q)$, $\varphi(Q, t_j)$, $\varphi'(\Gamma_Q, t_j)$) over time for the reference interval of the dataset. Comparing Fig. 18 (resp., 19, 20, 21) with the corresponding

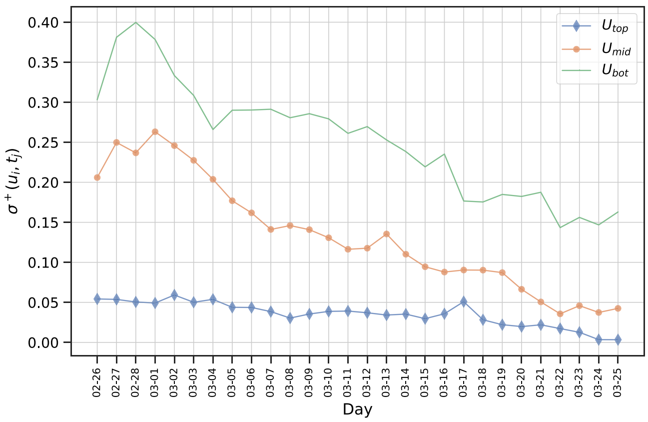


Fig. 16. Variation of the average value of $\sigma^+(u_i, t_j)$ over time for U_{top}, U_{mid} and U_{bot} in Twitter.

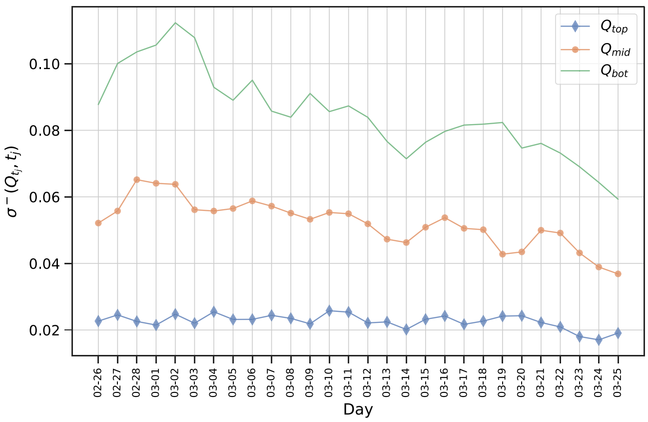


Fig. 17. Variation of the average value of $\sigma^-(Q_{t_j}, t_j)$ over time for Q_{top}, Q_{mid} and Q_{bot} in Twitter.

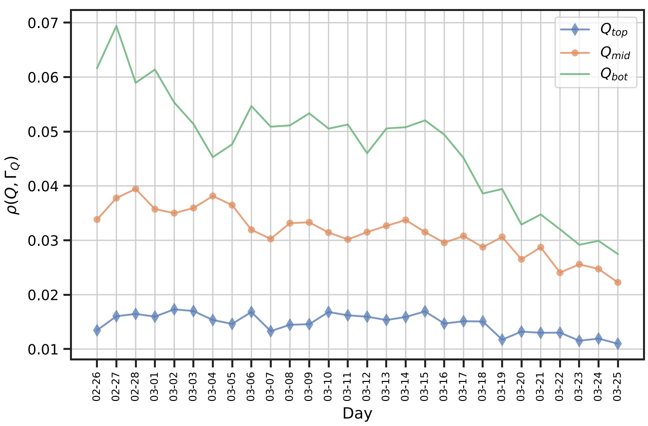


Fig. 18. Variation of the average value of $\rho(Q, \Gamma_Q)$ over time for Q_{top}, Q_{mid} and Q_{bot} in Twitter.

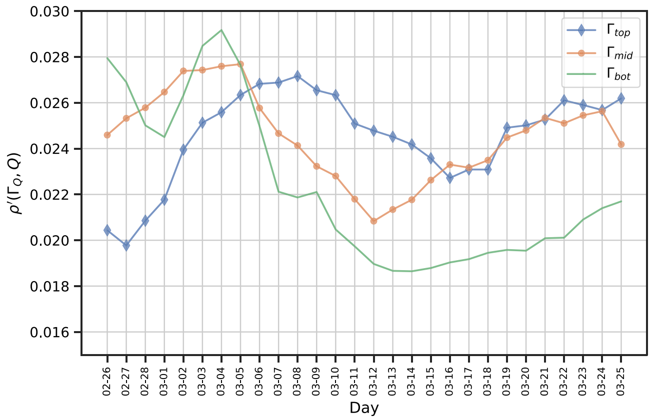


Fig. 19. Variation of the average value of $\rho'(\Gamma_Q, Q)$ over time for Γ_{top}, Γ_{mid} and Γ_{bot} in Twitter.

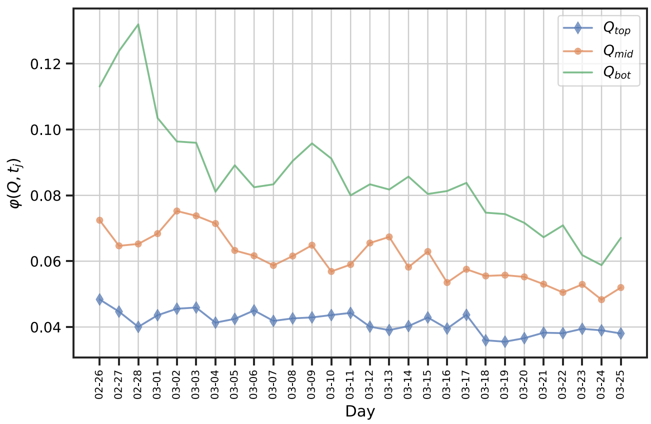


Fig. 20. Variation of the average value of $\varphi(Q, t_j)$ over time for Q_{top}, Q_{mid} and Q_{bot} in Twitter.

Fig. 11 (resp., 12, 13, 14) valid for Reddit, we can see that, again, the trends of the functions are similar and their values are comparable.

5. Conclusion

In this paper, we have presented a model and a framework to support analyses of the dynamics of user and community sentiments in a social platform. In particular, we have seen that our model is network-based and employs two networks. The first is a bipartite network that allows the storage of all available information. The second is a user-centric network that allows the study of the evolution of the user and community sentiment on a certain topic over time. Our framework employs our model as a support for its analyses. In this paper, we focused on the implementation of three activities concerning the framework, namely: *(i)* finding users capable of creating and maintaining a community that reflects their sentiment on a topic; *(ii)* studying how a user or community sentiment on a topic evolves over time; and *(iii)* investigating the cross-contamination between a user community and its neighborhood. We have also clarified that our framework is scalable and further activities can be easily implemented in it thanks to its scalability and the one of the underlying model.

Along with the strengths seen above, our framework also has some limitations. First, it is designed for social platforms whose content is mainly textual. Therefore, it cannot currently be easily used in all social platforms based mainly on images and videos. In fact, in such platforms, topic and sentiment extraction would have to be conducted differently from the way proposed in this paper. Handling non-textual data, like images and videos, can be a challenging task, since it requires additional preprocessing steps, such as image and video captioning, object recognition and sentiment analysis on captions. These tasks generally require the adoption of computer vision and Natural Language Processing approaches. However, the former are not currently included in our framework. As for a second limitation, the communities on which our framework focuses are cliques. In fact, it does not consider other forms

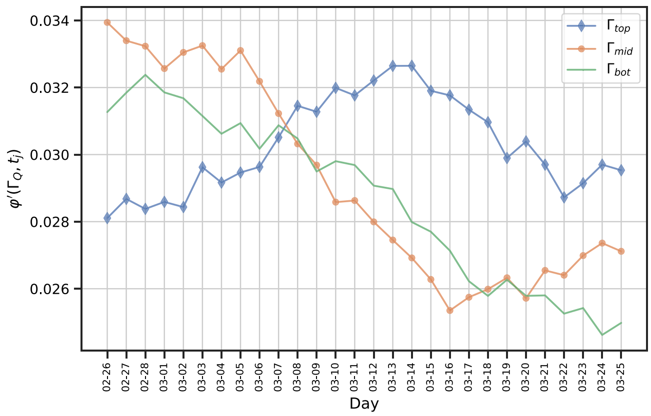


Fig. 21. Variation of the average value of $\varphi'(\Gamma_Q, t_j)$ over time for Γ_{top}, Γ_{mid} and Γ_{bot} in Twitter.

of subnets, or even other forms of communities, such as those based on modularity. In addition, our approach could be sensitive to content automatically generated via bots. In fact, since it does not identify the latter ones, their action could have an impact on its activities. For example, in analyzing determinant users, bots could produce content specifically tailored to influence the outcome of the analysis. It would also be interesting to enhance our approach with filtering techniques so as to remove noise, that is, content that should not be considered for analysis. Another limitation concerns the content language; in fact, only English is currently supported. Finally, our approach does not currently perform semantic analysis on topics and sentiment tags, which could be useful to identify controversial topics and to check whether they are able to influence the results returned.

In the future, we plan to continue this research along several directions and to enrich our framework with new activities. For example, we would like to study which mechanisms can lead to the disaggregation of a community that was previously cohesive with respect to a sentiment on a topic. Furthermore, we would like to investigate (if they exist) the various possible "life cycles" of a community, i.e., given a sentiment on a topic, we want to understand the ways in which a community of users sharing that sentiment can be generated, evolve and die. Understanding this could allow for predictions and, in the most advanced case, even prescriptions. The latter could indicate a set of posts or comments that could be published in the reference social platform to favor or counter certain dynamics of one or more users or communities regarding sentiments on topics.

CRediT authorship contribution statement

Gianluca Bonifazi: Interacted with eachother in all the tasks connected with the presented research. **Francesco Cauteruccio:** Interacted with eachother in all the tasks connected with the presented research. **Enrico Corradini:** Interacted with eachother in all the tasks connected with the presented research. **Michele Marchetti:** Interacted with eachother in all the tasks connected with the presented research. **Giorgio Terracina:** Interacted with eachother in all the tasks connected with the presented research. **Domenico Ursino:** Interacted with eachother in all the tasks connected with the presented research. **Luca Virgili:** Interacted with eachother in all the tasks connected with the presented research.

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] K. Cortis, B. Davis, Over a decade of social opinion mining: a systematic review, Artif. Intell. Rev. 54 (2021) 4873–4965, Springer.
[2] V. Basile, F. Cauteruccio, G. Terracina, How dramatic events can affect emotionality in social posting: the impact of covid-19 on reddit, Future Internet 13 (2) (2021) 29, MDPI.
[3] M. Lai, M. Tambuscio, V. Patti, G. Ruffo, P. Rosso, Stance polarity in political debates: A diachronic perspective of network homophily and conversations on Twitter, Data Knowl. Eng. 124 (2019) 101738, Elsevier.
[4] D. Ramachandran, R. Parvathi, A novel domain and event adaptive tweet augmentation approach for enhancing the classification of crisis related tweets, Data Knowl. Eng. 135 (2021) 101913, Elsevier.
[5] M. Tsvetovat, A. Kouznetsov, Social Network Analysis for Startups: Finding Connections on the Social Web, O'Reilly Media, Inc., Sebastopol, CA, USA, 2011.
[6] J. Chen, G. Kou, H. Wang, Y. Zhao, Influence identification of opinion leaders in social networks: an agent-based simulation on competing advertisements, Inf. Fusion 76 (2021) 227–242, Elsevier.

[7] N. Dakiche, F.B. Tayeb, Y. Slimani, K. Benatchba, Tracking community evolution in social networks: A survey, Inf. Process. Manage. 56 (3) (2019) 1084–1102, Elsevier.
[8] Z. Xu, X. Rui, J. He, Z. Wang, T. Hadzibeganovic, Superspreaders and superblockers based community evolution tracking in dynamic social networks, Knowl.-Based Syst. 192 (2020) 105377, Elsevier.
[9] B. Zhang, L. Zhang, C. Mu, Q. Zhao, Q. Song, X. Hong, A most influential node group discovery method for influence maximization in social networks: a trust-based perspective, Data Knowl. Eng. 121 (2019) 71–87, Elsevier.
[10] A. Almars, X. Li, X. Zhao, Modelling user attitudes using hierarchical sentiment-topic model, Data Knowl. Eng. 119 (2019) 139–149, Elsevier.
[11] R.L. Berger, A necessary and sufficient condition for reaching a consensus using DeGroot's method, J. Amer. Statist. Assoc. 76 (374) (1981) 415–418, Taylor & Francis.
[12] M.H. DeGroot, Reaching a consensus, J. Amer. Statist. Assoc. 69 (345) (1974) 118–121, Taylor & Francis.
[13] E. Ben-Naim, L. Frachebourg, P.L. Krapivsky, Coarsening and persistence in the voter model, Phys. Rev. E 53 (4) (1996) 3078–3087, APS.
[14] J.T. Cox, Coalescing random walks and voter model consensus times on the torus in Zd, Ann. Probab. 17 (4) (1989) 1333–1366, IMS.
[15] K. Sznajd-Weron, J. Sznajd, Opinion evolution in closed community, Internat. J. Modern Phys. C 11 (6) (2000) 1149–1156, World Scientific Press.
[16] F. Slanina, H. Lavicka, Analytical results for the sznajd model of opinion formation, Eur. Phys. J. B 35 (2003) 279–288, Springer.
[17] P.L. Krapivsky, S. Redner, Dynamics of majority rule in two-state interacting spin systems, Phys. Rev. Lett. 90 (2003) 238701, APS.
[18] C.J. Tessone, R. Toral, P. Amengual, H.S. Wio, M. San Miguel, Neighborhood models of minority opinion spreading, Eur. Phys. J. B 39 (2004) 535–544, Springer.
[19] N.E. Friedkin, E.C. Johnsen, Social influence networks and opinion change, Adv. Group Process. 16 (1) (1999) 1–29, JAI Press Inc..
[20] J. Lorenz, Continuous opinion dynamics under bounded confidence: A survey, Internat. J. Modern Phys. C 18 (12) (2007) 1819–1838, World Scientific Press.
[21] A.C.R. Martins, Discrete opinion models as a limit case of the CODA model, Physica A: Statist. Mech. Appl. 395 (2014) 352–357, Elsevier.
[22] S. Li, C. Wei, Modeling the social influence in consensus reaching process with interval fuzzy preference relation, Int. J. Fuzzy Syst. 21 (2019) 1755–1770, Springer.
[23] J. Chang, G. Wang, Formulating the dynamics of an online community and identifying the influential members using the animal flocking model, Electron. Commer. Res. Appl. 49 (2021) 101085, Elsevier.
[24] R. Cazabet, F. Amblard, Dynamic community detection, in: Encyclopedia of Social Network Analysis and Mining, Springer New York, 2014, pp. 404—414.
[25] S. Bansal, S. Bhowmick, P. Paymal, Fast community detection for dynamic complex networks, in: Complex Networks, Springer, 2011, pp. 196—207.
[26] H. Yin, S. Yang, J. Li, Detecting topic and sentiment dynamics due to COVID-19 pandemic using social media, in: Proc. of International Conference on Advanced Data Mining and Applications, ADMA'20, in: Lecture Notes in Computer Science, Springer, Foshan, China, 2020, pp. 610–623.
[27] Y. Xu, Y. Li, Y. Liang, L. Cai, Topic-sentiment evolution over time: a manifold learning-based model for online news, J. Intell. Inf. Syst. 55 (2020) 27–49, Springer.
[28] S. Bhattacharya, D. Sarkar, Study on information diffusion in online social network, in: Proc. of the International Conference on Frontiers in Computing and Systems, COMSYS'21, AISC, Springer, Meghalaya, India, 2021, pp. 279–288.
[29] M. Dermouche, J. Velcin, L. Khouas, S. Loudcher, A joint model for topic-sentiment evolution over time, in: Proc. of the International Conference on Data Mining, ICDM'14, IEEE, Shenzhen, China, 2014, pp. 773–778.
[30] M. Giatsoglou, A. Vakali, Capturing social data evolution using graph clustering, IEEE Internet Comput. 17 (1) (2013) 74–79, IEEE.
[31] B. Qiu, K. Ivanova, J. Yen, P. Liu, Behavior evolution and event-driven growth dynamics in social networks, in: Proc. of the IEEE International Conference on Social Computing, SocialCom'10, IEEE, Minneapolis, MN, USA, 2010, pp. 217–224.
[32] R.S. Caceres, T. Berger-Wolf, Temporal scale of dynamic networks, in: Temporal Networks, Springer, 2013, pp. 65–94.
[33] I. Vayansky, S.A.P. Kumar, A review of topic modeling methods, Inf. Syst. 94 (2020) 101582, Elsevier.
[34] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jian, Y. Li, L. Zhao, Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey, Multimedia Tools Appl. 78 (2019) 15169–15211, Springer.
[35] J. Qiang, Z. Qian, Y. Li, Y. Yuan, X. Wu, Short text topic modeling techniques, applications, and performance: A survey, IEEE Trans. Knowl. Data Eng. 34 (3) (2022) 1427–1445, IEEE.
[36] L. Yue, W. Chen, X. Li, W. Zuo, M. Yin, A survey of sentiment analysis in social media, Knowl. Inform. Syst. 60 (2) (2019) 617–663, Springer.
[37] K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: Tasks, approaches and applications, Knowl.-Based Syst. 89 (2015) 14–46, Elsevier.
[38] A. Yadav, D.K. Vishwakarma, Sentiment analysis using deep learning architectures: a review, Artif. Intell. Rev. 53 (6) (2020) 4335–4385, Springer.
[39] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, A. Gelbukh, Sentiment and sarcasm classification with multitask learning, IEEE Intell. Syst. 34 (3) (2019) 38–43, IEEE.
[40] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, J. Blackburn, The pushshift Reddit dataset, in: Proc. of the International AAAI Conference on Web and Social Media, ICWSM'20, vol. 14, AAAI Press, Atlanta, GA, USA, 2020, pp. 830–839.
[41] R.A. Mills, Reddit.com: A census of subreddits, in: Proc. of the International Web Science Conference, WebSci'15, ACM, Oxford, UK, 2015, pp. 49:1–49:2.
[42] A. Guimaraes, O. Balalau, E. Terolli, G. Weikum, Analyzing the traits and anomalies of political discussions on reddit, in: Proc. of the International Conference on Web and Social Media, ICWSM 2019, AAAI, Munich, Germany, 2019, pp. 205–213.
[43] B.D. Horne, S. Adali, The impact of crowds on news engagement: A reddit case study, in: Proc. of the International AAAI Conference on Web and Social Media, ICWSM'17, Montréal, Québec, Canada, 2017, p. 11(1).
[44] M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure, 2022, arXiv preprint arXiv:2203.05794.
[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) Curran Associates, Inc..
[46] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, G. Varoquaux, API design for machine learning software experiences from the scikit-learn project, in: Proc. of the International Workshop on Languages for Data Mining and Machine Learning, Springer, Prague, Czech Republic, 2013, pp. 108–122.
[47] D. Loureiro, F. Barbieri, L. Neves, L.E. Anke, J. Camacho-collados, TimeLMs: Diachronic language models from Twitter, in: Proc. of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL'22, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 251–260.
[48] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, L. Neves, TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification, in: Findings of the Association for Computational Linguistics (EMNLP'20), Association for Computational Linguistics, 2020, pp. 1644–1650, Virtual Conference.

Gianluca Bonifazi received the M.Sc. Degree in IT Engineering from the Polytechnic University of Marche in February 2020. He is currently a Ph.D. Student in Information Engineering at the same University. His research interests include Machine Learning, Social Network Analysis, Complex Networks and Internet of Things. He is an author of 12 papers.

Francesco Cauteruccio received the Ph.D. in Mathematics and Computer Science from the University of Calabria in January 2018. From February 2018 to January 2022, he was a Research Fellow at University of Calabria. Currently, he is a Junior Assistant Professor at the Department of Information Engineering at the Polytechnic University of Marche. He is actively working in research projects, such as the "Smarter Solutions in the Big Data World", funded within the call "HORIZON2020" PON I&C 2014–2020 and he is a member of the Advanced Database research lab, hosted by University of Calabria. His research interests include Machine Learning, Time Series Analysis, Social and Complex Network Analysis, and Biomedical Applications. He is an author of more than 30 papers in international journals, conferences and book chapters. He also served with different roles (organizer, program committee member, reviewer) in international journals and conferences.

Enrico Corradini received the M.Sc. Degree in IT Engineering from the Polytechnic University of Marche in October 2019. He is currently a Ph.D. Student in Information Engineering at the same University. His research interests include Social Network Analysis, Social Internetworking, Source and Data Integration, Multiple Internet of Things scenarios, Data Lakes. He is an author of more than 15 papers.

Michele Marchetti received the M.Sc. Degree in IT Engineering from the Polytechnic University of Marche in October 2021. He is currently a Ph.D. Student in Information Engineering at the same University. His research interests include Social Network Analysis, Natural Language Processing, Blockchains and Source and Data Integration. He is an author of 5 papers.

Giorgio Terracina received the M.Sc. Degree in Computer Engineering from the University of Calabria in April 1999. He received the Ph.D. in Electronic Engineering from the University of Reggio Calabria in 2002. From December 2002 to October 2010, he was a Tenured Assistant Professor at the University of Calabria, and from November 2010 he is an Associate Professor at the same University. He has been responsible of several research units in national and international research projects and held several management positions, such as member of the board of directors in University of Calabria, and member of the board of directors in IDUM, a spin-off from University of Calabria. His research interests include Source and Data Integration, Social Network Analysis, Knowledge Representation and Reasoning, Bioinformatics. In these research areas, he published more than 140 papers.

Domenico Ursino received the M.Sc. Degree in Computer Engineering from the University of Calabria in July 1995. He received the Ph.D. in System Engineering and Computer Science from the University of Calabria in January 2000. From January 2005 to December 2017 he was an Associate Professor at the University Mediterranea of Reggio Calabria. From January 2018 he is a Full Professor at the Polytechnic University of Marche. His research interests include Social Network Analysis, Social Internetworking, Source and Data Integration, Innovation Management, Multiple Internet of Things scenarios, Knowledge Extraction and Representation, Biomedical Applications, Recommender Systems, Data Lakes. In these research fields, he published more than 220 papers.

Luca Virgili received the M.Sc. Degree in IT Engineering from the Polytechnic University of Marche in October 2018. He is currently a Ph.D. Student in Information Engineering at the same University. His research interests include Social Network Analysis, Social Internetworking, Source and Data Integration, Multiple Internet of Things scenarios, Privacy, Blockchains. He is an author of 25 papers.