GENE: Graph generation conditioned on named entities for polarity and controversy detection in social media

Marcelo Mendoza[a,c,*], Denis Parra[b,c], Álvaro Soto[b,c]

[a] *Universidad Técnica Federico Santa María, Santiago, Chile*
[b] *Pontificia Universidad Católica de Chile, Santiago, Chile*
[c] *Instituto Milenio Fundamentos de los Datos, Chile*

Abstract

DRAFT, to be published in IPM `https://doi.org/10.1016/j.ipm.2020.102366`.

Many of the interactions between users on social networks are controversial, specially in polarized environments. In effect, rather than producing a space for deliberation, these environments foster the emergence of users that disqualify the position of others. On news sites, comments on the news are characterized by such interactions. This is detrimental to the construction of a deliberative and democratic climate, stressing the need for automatic tools that can provide an early detection of polarization and controversy. We introduce **GENE** (**g**raph g**e**neration conditioned on **n**amed **e**ntities), a representation of user networks conditioned on the named entities (personalities, brands, organizations) which users comment upon. **GENE** models the leaning that each user has concerning entities mentioned in the news. **GENE** graphs is able to segment the user network according to their polarity. Using the segmented network, we study the performance of two controversy indices, the existing Random Walks Controversy (RWC) and another one we introduce, Relative Closeness Controversy (RCC). These indices measure the interaction between the network's poles providing a metric to quantify the emergence of controversy. To evaluate the performance of **GENE**, we model the network of users of a popular news site in Chile, collecting data in an observation window of more than three years. A large-scale evaluation using **GENE**, on thousands of news, allows us to conclude that over 60% of user comments have a predictable polarity. This predictability of the user interaction scenario allows both controversy indices to detect a controversy successfully. In particular, our introduced **RCC** index shows satisfactory performance in the early detection of controversies using partial information collected during the first hours of the news event, with a sensitivity to the target class exceeding 90%.

Keywords: Graph-based representations, controversy detection, polarity dynamics

1. Introduction

Social networks are the new media through which people share and acquire information. It is a more complex medium than traditional media such as radio, television, and the written press. Along with direct interaction with the information, social network users interact with each other, commenting on the information they receive, and even posting new information. Social networks have allowed us to consolidate a new paradigm in the information age, guaranteeing not only free access to the media but also the freedom to publish and share new content.

All these benefits of social networks also bring about risks, being polarization one of the most important. In his work on information flow, Zachary (Zachary, 1977) showed that social networks tend towards polarization, reducing the chance of deliberation among people. The bias of social networks towards polarization

*Corresponding author
Email addresses: `mmendoza@inf.utfsm.cl` (Marcelo Mendoza), `dparras@uc.cl` (Denis Parra), `asoto@ing.puc.cl` (Álvaro Soto)

is due, among other factors, to the fact that people tend to follow users with affinities in their system of beliefs (Bessi et al., 2014). A consequence of this phenomenon is the consolidation of homophylic relations, that is, the sharing of information mostly between users with a high level of ideological coincidence (Grevet et al., 2014). By exposing these groups to new information, there is a high level of coherence in their opinions, which are reinforced and strengthened among them, leading the group towards the adoption of less deliberative positions (Flaxman et al., 2016). This phenomena, known as echo chambers, helps to explain the rise of spontaneous polarization (Garimella et al., 2017). Recent studies reinforce the idea that the bias towards polarization has a structural root and suggests that an effective way to reduce this effect is to expose communities to contrary ideas, fostering diversity, and plurality of opinions (Garimella et al., 2018).

Another risk in social networks that emerges with polarization is controversy. Controversy is the interaction between polarized users who seek to impose their views on those of others. When users are polarized, the possibilities of deliberation are minimal: the social network becomes the space for personal disqualification and controversy arises (Nelimarkka et al., 2018). Recent studies have highlighted polarization dynamics and controversy processes as one of the negative aspects of social networks (Quraishi et al., 2018). Under polarized environments, it is easier to influence people's opinions, leading them to antagonistic positions. Under these circumstances, people adopt extreme positions that foster manipulation. As an example, studies suggest that through social networks, specific groups could be manipulated, achieving a hinge effect in political elections (Guimaraes et al., 2017) (Morales et al., 2015).

The study of polarization has been addressed using natural language processing methods, such as sentiment analysis (Mejova et al., 2014). These methods have proven to be accurate and useful in detecting polarization in comments. There are some recent efforts to quantify controversy (Popescu & Pennacchiotti, 2010; Garimella et al., 2016a). These methods require the construction of a network of interactions among users, measuring the level of coupling between groups of users with different positions. These methods have shown that it is possible to quantify the level of controversy produced in a network, finding correlations between the topology of the network and the type of comments made by users. Both, polarization analysis and quantification of controversy, are ex-post analyses, that is, they operate on events that occurred in the past. These analyses require a complete characterization of the event, extracting information from conversational threads, determining the users involved, and defining relationships between users either at the level of connections in the social network or direct interactions during an event.

There are clues that indicate that the ideological biases of social network users are so explicit that it would be possible to anticipate the emergence of controversy by observing people's interaction in the discussion of news (Akoglu, 2014; Coletto et al., 2016; Guimaraes et al., 2017). If this measurement is made during the discussion, it would be possible to anticipate the result of potential polarization and controversy.

In this article, we propose a new model for early prediction of polarization and controversy in social networks. Specifically, we introduce a network representation that is capable of simultaneously modeling the bias of each user and their interactions with specific information. Key elements of our proposal are the named entities (people, organizations, brands, countries). We use the comments that users make to entities named in the news to infer their tendency towards these entities when mentioned in the future.

From our point of view, the dynamics of social networks are mainly determined by two factors: i) The first is related to which users interact in a discussion, ii) The second one is related to which entities are involved in a given story. Consequently, we exploit the identification of these two factors as the main elements behind our model to detect polarization and controversy in a social network. Specifically, our method is based on 3 main steps. A first step uses a corpus to build a user-entity model that allows us to quantify the bias of each user with respect to the most common entities appearing in the corpus. A second step takes advantage of the user-entity model to build a multi-relational graph that allows us to identify the interaction among users in a discussion that involves a specific entity. Then, a graph generation method is used to generate a representation of the network of users involved in the comments of a news item. This graph generation method produces a polarized view of the network of intervening users. Finally, a third step analyses the interactions in the graph, leading to the identification of polarization and controversy. We refer to our model as **GENE**, **g**raph g**e**neration conditioned on **n**amed **e**ntities, a new representation of a user network that is conditioned on a set of entities. **GENE** provides a user network conditioned on relevant entities, allowing us to represent the engagements of the users with these entities along with their bias. Having a representation

of the network conditioned to the entities of interest allows us to analyze interactions between users that have opposing views concerning these entities. We show that **GENE** has predictive capabilities and that, through its use, a polarization dynamic can be anticipated in a user network, predicting the emergence of controversy before it occurs.

Main contributions of this work are the following:

- We introduce **GENE**, a method for generating user networks conditioned on the joint representation of users, entities, and the inclination of users towards those entities (positive, neutral or negative).
- Extensive experiments indicate that for the task of controversy detection, **GENE** creates a rich representation of user networks, conditioned on polarized entities, that outperforms both lexicon-based methods as well as graph representations which do not exploit the entities and polarization contexts.
- In an early controversy detection setting, **GENE** also shows superior performance compared to other methods, favoring the early forecast of polarization in a user network and the characterization of a scenario for the emergence of controversy.
- We release a new dataset that includes news and conversational threads from which this study was conducted.

The article is organized as follows. In Section 2, we review related work. **GENE** is introduced in Section 3. The materials and methods used in this study are presented in Section 4. The validation of **GENE** is presented in Section 5. Section 6 is devoted to the discussion of the results, their scope, and limitations. The article concludes in Section 7, highlighting findings and main results of this work, as well as outlining future work.

2. Related Work

The dynamics and consequences of polarization in human groups have been studied for decades. In a pioneering work, Zachary (1977) observed that the homogenization of views in human groups leads to the stabilization of these groups over time. He also observed that the flow of information between different groups leads to conflict.

The consolidation of the web as the universal information sharing platform crystallized new ways of human relationships. With the emergence of online social networks, it was expected that free access to information and freedom to publish would lead to dialogue and deliberation. However, online social networks reproduce the weaknesses and limitations of human interactions, now amplified on a larger scale.

During the U.S. presidential election in 2004, Adamic & Glance (2005) observed that there were differences in linking patterns between users of conservative and liberal blogs. The study showed that users who shared the same political orientation interacted more and cited more the contents of those users. On the other hand, the interaction between groups with different points of view was low. The need to quantify the impact of the conflict in social networks led to a sequel of works whose objective was to improve the understanding of this phenomenon. For example, Choi et al. (2010) used topic models to study which specific issues produced the most significant conflict in networks. Using sentiment analysis, the authors labeled documents with a strong emotional charge and then applied topic models to identify queries that would retrieve those documents. The work showed that the queries were proper descriptors of the conflicting topics. Popescu & Pennacchiotti (2010) proposed classifiers of topics on Twitter that distinguished between controversial and non-controversial events. For this, they used linguistic features based on lexicons, such as a sentimental lexicon from OpinionFinder, a lexicon of controversial Wikipedia words, and a lexicon of bad words. Together with the linguistic characteristics, they studied the effect of the volume of tweets and the spatio-temporal location of the tweets, showing that all these features were useful for the task. Mejova et al. (2014) revisited these results, showing that controversial issues correlated with the use of negative affect and biased languages. Hamad et al. (2018) corroborated that the presence of negative words or slang is useful for detecting social media controversy. However, Kaplun et al. (2018) noted that while it is clear

how to distinguish between controversial and non-controversial words, the use of lexicon-based classifiers is not useful in controversy detection.

Network topology was at the center of the Conover et al. (2011) study, who observed that the network of retweets in the 2010 U.S. congressional midterm elections showed high partisan segregation, with minimal connectivity between liberals and conservatives. On the other hand, the network of mentions showed a single giant component, with much higher interaction between users with different points of view. Calais Guerra et al. (2013) showed that the use of modularity measures was not useful to detect polarized communities since non-polarized networks could also be partitioned using modularity. They observed that polarized communities had a more significant presence of high degree users in their boundaries, and therefore they introduced a metric that quantified the presence of highly connected nodes at the boundaries of the candidate communities. Ruan et al. (2013) introduced a measure of signal strength between two nodes in a network by fusing their link strength with content similarity. Then, they showed that the improved representation of the network helps detect polarized communities.

Several works have put their efforts in diminishing the possibility of a conflict. However, these efforts have been unsuccessful. Graells-Garrido et al. (2013) showed that the recommendation of content with opposing views had a negative emotional effect on users. Munson et al. (2013) showed that the elicitation of the political leaning of an editorial column led to a modest improvement in the balance of information exposure. The reason why these attempts were unsuccessful is that social media users tend to share information that confirms their belief system (Bessi et al., 2014). This behavior has risks. Partisan sharing (that is, to share like-minded information) produces a distorted perception of reality. The intensity of partisan sharing is directly dependent on the perceived importance of the commented topic (An et al., 2014). Grevet et al. (2014) showed that users who perceived more differences with their friends engaged less on Facebook than those who perceived more homogeneity. To bridge the ideological difference, it is good to make common ground visible while friends converse. This finding has been re-studied by Chen et al. (2018), who observe that opinions are affected by our immediate neighborhood. In social networks, some changes in the structures that define neighborhoods can reduce the risks of conflicts.

The political leaning explains behavior patterns in the adoption of positions. Akoglu (2014) modeled the political leanings of U.S. Congress members from their opinions on specific topics using signed bipartite networks. He conducted a node labeling task on U.S. congressional records, showing that the voting intent of the congressmen was predictable from this model. The effect of political cyberactivism has also been studied. A case study of Venezuela (Morales et al., 2015) confirms that just a small group of influential individuals propagating their opinions through a social network are needed to produce polarization. In a polarized scenario, opinions that show greater conviction dominate the dynamics of a network (Guimaraes et al., 2017). Political leaning can also be useful to describe an event of interest. The classification of users, according to political leaning, is used by Coletto et al. (2016) to track keywords that describe political views in conversational threads. Then, many emergent keywords can be detected during a political campaign. Using this method, it is also possible to identify polarized words based on comments on controversial topics from previously classified users (Coletto et al., 2017).

The quantification of the controversy in Twitter has been in the focus of the work of Garimella et al. (2016b). Using hashtags to retrieve the network of retweets of a particular topic, they have proposed the use of graph partition algorithms such as METIS (Lasalle & Karypis, 2013) to detect communities with opposing views. This methodology is supported by the work of Conover et al. (2011), who showed that the networks of retweets show high partisan segregation. Garimella et al. (2016a) proposed a metric based on random walks that quantified the level of coupling between two partitions induced by METIS (Karypis & Kumar, 1998). The method claims that the controversy is related to the level of interaction between influencers with opposite views. The authors have shown that the recommendation of links to bridge both poles of the network decreases the controversy score defined by them (Garimella et al., 2018). Using their methodology, the authors have analyzed controversial events revealing important findings. For example, they have observed that collective attention accelerates the dynamics of polarization increasing the volume of interactions within each pole of the network (Garimella et al., 2017). In a longitudinal study on controversial events on Twitter, they have observed that polarization has progressively increased in recent years (Garimella & Weber, 2017). Retweet networks have been used in some case studies. Retweet networks were used to track the effect of

echo chambers during the 2017 French election (Gaumont et al., 2018), showing that these networks are recomposed during external events as televised debates or primaries.

Although the networks of retweets have shown utility in the quantification of controversy, some studies indicate that not always a retweet should be interpreted as an endorsement. Calais Guerra et al. (2017) noted the presence of retweets with quotes that involved criticisms. Besides, specific linguistic features are relevant to describe a controversial event. For example, Han et al. (2017) noted that at the beginning of a conflict, words tend to be more sensational. In more advanced stages of a conflict, the proportion of words with political connotation and calls to action tend to increase. Jang et al. (2017) noted that there is a strong dependence between the specific topic and the users involved in a discussion. His study argues that controversy is a trait rooted in a population deeply related to ideological bias. Therefore, it can be observed in a specific population rather than held as a fixed universal quantity. Accordingly, the quantification of controversy requires models that jointly consider text and user interaction networks. In this line of research, Lahoti et al. (2018) use joint non-negative matrix factorization for learning ideological leaning in Twitter using the social graph and the news consumption information. The method shows good performance in user classification. Various case studies show the need to combine text and link analysis. The intensity of the Ukranian conflict was tracked using random walk graph polarization and observing the shift of word meanings through time (Rumshisky et al., 2017). The results show that the combination of text and links drive to a better tracking of the intensity of the crisis.

The role of neutrality is highlighted by Matakos et al. (2017), who observe that polarization can be limited in its effects thanks to the intervention of neutral users that promote dialogue. However, determining which users will maximize information diversity instead of polarization is a complex problem. The authors show that the problem of information diversity maximization in social networks is inapproximable and that, therefore, it can only be addressed using heuristics (Matakos & Gionis, 2018). Nelimarkka et al. (2019) observed that the recommendation of contents based on information diversification could decrease polarization but creating additional complexities through which users question information. The role of neutrality was also studied in Tumblr (Warmsley et al., 2019). The polarization on this platform was analyzed using a tripartite graph approach that includes in the analysis the presence of neutral users. Community detection algorithms improved their performance in this scenario, showing that the need to model neutrality is a key factor for the success of this kind of algorithms. Napoles et al. (2017) address the automatic identification of deliberative conversations. The authors show that linguistic characteristics, such as the use of pronouns, concordance, and certainty of sentences, as well as the use of discursive connectives, are useful for the detection of deliberative conversations. The authors note that using their method, the presence of deliberative conversations in forums is more significant than in online news sites.

Controversial issues have crossed the barriers of conventional OSN platforms as Twitter and Facebook. A recent study (Kane & Luo, 2019) shows that Reddit groups show a strong correlation between the group's topic and politically skewed and segregationist language. Political skews have also been observed in search engine results (Kulshrestha et al., 2019), indicating the presence of ideological bias in the data collected and also the presence of algorithmic bias in the ranking.

From the previous description, one can observe that strategies such as Mejova et al. (2014) and Garimella et al. (2018) have been the focus of attention to detect polarization and controversy in social networks. In contrast, in this work we claim that the identification of the position of a user with respect to an entity and the interactions among users with respect to the discussion about an entity are key sources of information to detect polarization and controversy in social networks. Users engage in a discussion depending on their willingness to comment on a specific entity. Our definition moves the focus of literature, in which the importance of topics to explain user engagements is highlighted. We believe that there are more fundamental units of information than the topics that would allow us to understand this phenomenon in a better way. These units are the entities (people, organizations, brands, among others). We argue that the predisposition of users to comment on social networks has to do with a positive or negative perception of the entities named in a piece of information. Consequently, we believe that the unit of information that allows capturing the mutual dependence between polarity and user engagements are the entities.

3. GENE: Graph generation conditioned on named entities

3.1. Polarity and controversy

The scientific literature shows that there are diverse approaches to address the study of controversy. Some of them aim to identify characteristics of the content, and by extracting linguistic characteristics, they aspire to make a characterization of the most controversial topics and events. Other approaches highlight the role that social network users play. For these approaches, the polarization of the network plays a fundamental role during the process.

We will follow an approach that defines controversy as a complex phenomenon. Timmermans et al. (2017) introduces the concept of computational controversy, which we will follow in this work to address the phenomenon under study. Computational controversy comprehends many factors as the presence of many actors, polarized viewpoints, an open space where it happens (e.g., a forum), persistence through time, and the presence of strong sentiments and emotions during the discussion. This definition points in a sense that brings together all the approaches discussed in the related work. Computational controversy states that the study of this matter requires the consideration of all the above factors.

Computational controversy assumes that the emergence of a controversy requires a polarized scenario. We believe that the existence of a polarized scenario is defined by the concurrence of two critical factors: the users involved in the discussion and the entities mentioned during the discussion. In our analysis, these two factors have a strong dependence. We claim that users engage in a discussion depending on their willingness to comment on a specific entity. Our definition moves the focus of literature, in which the importance of topics to explain user engagements is highlighted. We believe that there are more fundamental units of information than the topics that would allow us to understand this phenomenon in a better way. These units are the entities (people, organizations, brands, among others). We argue that the predisposition of users to comment on social networks has to do with a positive or negative perception of the entities named in a piece of information. Consequently, we believe that the unit of information that allows capturing the mutual dependence between polarity and user engagements are the entities.

We claim that a polarization dynamic is conditioned to the leaning of users towards the entities named in a story. Haters and partisans will master polarization dynamics since they have a predisposition to be actively involved in discussions. However, users alternate these roles. An activist can be partisan of specific entities and hater of others. In our proposal, we will model this leaning towards specific entities at the user level to characterize a polarized scenario.

We distinguish between polarization and controversy. A polarized scenario is characterized by the cohesive presence of user groups with opposite views. However, if there is no interaction between these groups, there is no controversy. We understand polarization as a precondition for the emergence of a controversy. Therefore, controversy is characterized by the level of interaction between groups with opposite views. A high level of cross interactions between opinion poles indicates the presence of conflict in the context of an event that involves specific entities. Under this approach, controversy can be detected using representations that account for scenarios subject to polarization over time.

We introduce **GENE**, **g**raph g**e**neration conditioned on **n**amed **e**ntities, which is a representation that brings together the factors mentioned above. **GENE** produces views of user networks conditioned on named entities according to the leaning that users have to comment positively or negatively on these entities. **GENE** jointly models the tendency of each user to comment on the entity (user engagement) and its leaning towards this entity (polarity). During the development of a discussion, **GENE** allows detecting the factors that make up a polarized scenario. **GENE** produces views of the user network according to the users who are involved in a discussion. The interaction between haters and partisans is captured by **GENE**, allowing to detect a polarized scenario. We will show that by computing coupling measures between opinion poles, controversy can be anticipated.

*3.2. The overall look of **GENE***

GENE is based on 3 main steps. A first step builds a user-entity model that allows to quantify the bias of each user with respect to the most common entities appearing in the corpus. A second step takes advantage of the user-entity model to build a multi-relational graph. Using embeddings computed from

the graph, **GENE** generates a polarized network of intervening users in a discussion that involves specific entities. Finally, a third step analyses the interactions in the generated graph, leading to the identification of controversy. To implement these steps, **GENE** considers several data processing modules. Its overall architecture is described in Figures 1 and 2. In the following sections we will describe each of these modules. The description of them involves several variables. For the sake of simplicity, in Table 1 we summarize them.

Table 1: Notation

$\vec{u}_i$: vector embedding for user i.
$\vec{e}_i$: vector embedding of polarized name entity i.
$\mathbf{W}_{in}$: user-entity model user matrix.
$\mathbf{W}_{out}$: user-entity model entity matrix.
S: number of training instances.
k: number of latent factors in user-entity model.
$\vec{v}^{(u)}$: representation of user u in user-entity model.
$\vec{v}^{(e)}$: representation of entity e in user-entity model.
$\pi^{(i)}$: i-th probability vector of user-entity model projected on user network.
$G^{(i)}$: weighted directed graph generated according to $\pi^{(i)}$.
$< \Theta_s, \Theta_d, \Theta_r^{(i)} >$: graph model parameters for nodes and relations.
$(s, r(i), d)$: graph triplet corresponding to nodes s and r connected by relation $r^{(i)}$.

3.3. User-entity model

In Figure 1, a first **GENE** component is described, indicated with the letter a, which shows the processing pipeline that defines the user-entity model. The user-entity model is the **GENE** component that models the mutual dependence between users and entities. Figure 1 shows that the user-entity model takes a corpus of news along with its comments from an online news site or forum. **GENE** applies named entity recognition (NER) to the titles and subheads of the news to identify the named entities. To each comment, **GENE** applies a sentiment analysis method to identify its polarity. Then, a sliding window is applied on each conversational thread to determine the multi-user contexts of each news, imputing the polarity of the comment to the entities named in the news that triggers the thread. A sequencer builds tuples formed by named users and entities, which are used to compute the user-entity model. At the bottom of Figure 1, the user entity-model is described as a bi-factorial model defined from the mutual dependencies between users and named entities. The parameters of the model are stored in the matrices indicated with the letter b.

The input of the user-entity model is a set of one-hot encoded vectors of users. This fact means that once one user comments about a named entity, the component of the vector that represents the user is one, and all other components are zero. Then, the user set is represented in the user-entity model by a collection of one-hot encoded vectors $\{\vec{u}_1, \ldots, \vec{u}_n\}$. The output of the user-entity model is a collection of one-hot encoded vectors $\{\vec{e}_1, \ldots, \vec{e}_m\}$ that represents polarized named entities. A particularity of the user-entity model is that it establishes a relationship between users and named entities according to polarity classes of sentiment analysis. The polarity is defined in three classes, being these positive, negative, or neutral. The type of polarization is defined by a comment that a user made about a named entity. Then, strictly speaking, the multi-user context of an entity corresponds to the group of users recovered from a conversational thread that agrees in polarity concerning the named entity. Consequently, for each named entity, the user entity model defines three vectors that model it, according to the three polarity classes explained above.

Both collections of vectors are connected by the parameters of the user-entity model. The user-entity model considers two matrices of parameters, $\mathbf{W}_{in}$ and $\mathbf{W}_{out}$. Each row in $\mathbf{W}_{in}$ is a d-dimensional vector representation $\vec{v}^{(u)}$ of a user u and can be computed as $\mathbf{W}_{in}^t \vec{u}$. Analogously, each column in $\mathbf{W}_{out}$ is a d-dimensional vector representation $\vec{v}^{(e)}$ of a named entity and can be computed as $\mathbf{W}_{out}\vec{e}$.

In the user-entity model, the context of a named entity e in the network corresponds to users who comment about e and agrees in polarity. The user-entity model captures polarized multi-user contexts of named entities using the average of the user vectors:

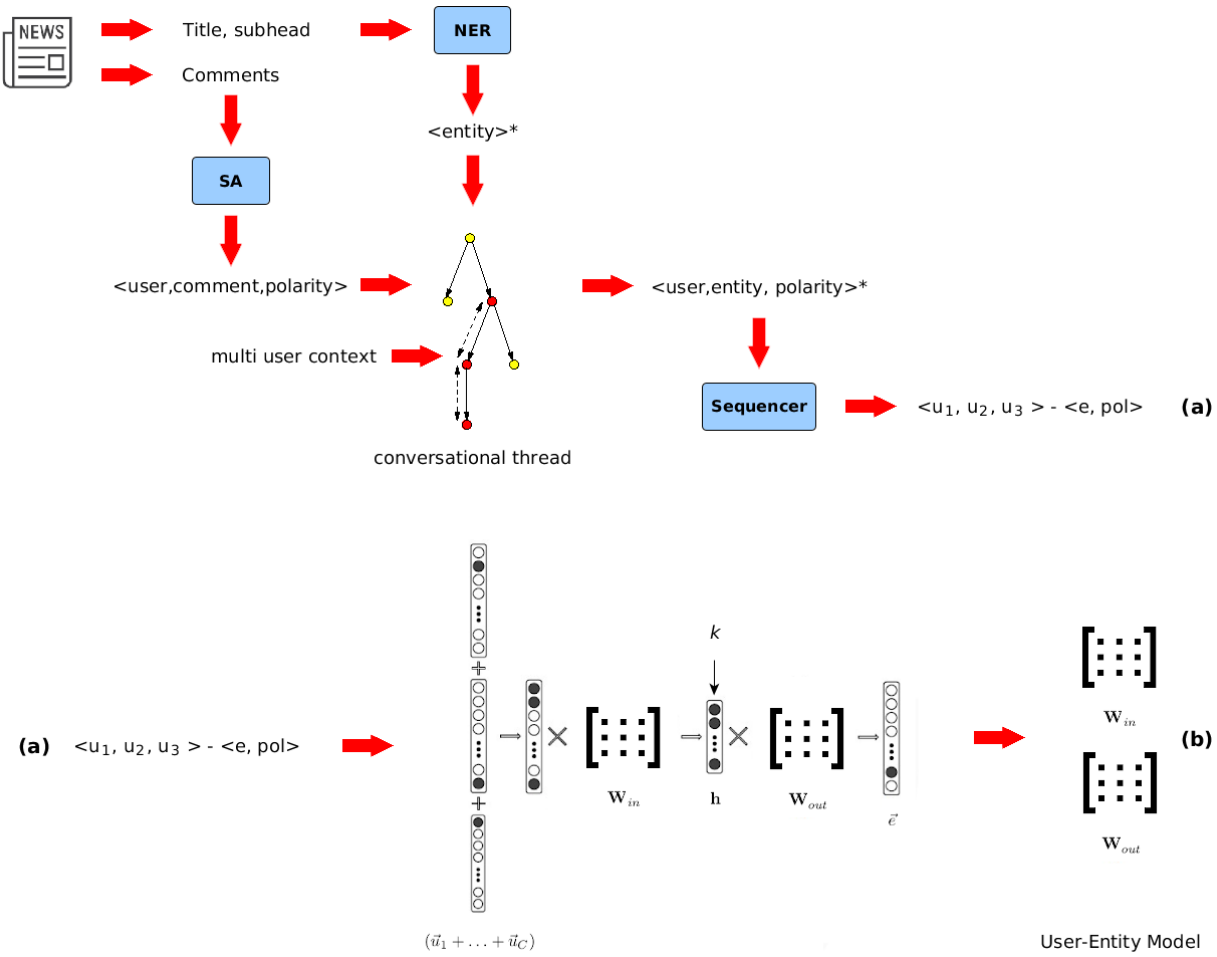


Figure 1: **GENE** models the mutual dependence between users and entities in the user-entity model. **GENE** applies sentiment analysis and NER to adjust the model parameters using news comments retrieved from online news sites. The model output is used by **GENE** to build views of the network of users conditioned on named entities.

$$\mathbf{h} = \frac{1}{C}\mathbf{W}_{in}^{T} \cdot (\vec{u}_1 + \ldots + \vec{u}_C),$$

where C is the number of users that belongs to the context of e and $\mathbf{h}$ is a latent vector of the model that represents the multi-user context of e. The context of e is defined from the conversational threads that emerge from the network in which e is named. In the user-entity model, the polarized multi-user context is retrieved using a fixed-sized sliding window that runs through each conversational thread, retrieving multi-user contexts corresponding to named entities. The sliding window is carried out in-order according to the timestamps of the comments, capturing the temporal dependencies among them.

Note that the latent vector $\mathbf{h}$ can be written in terms of the vector representations $\vec{v}^{(u)}$:

$$\mathbf{h} = \frac{1}{C}(\vec{v}_1^{(u)} + \ldots + \vec{v}_C^{(u)})$$

The user-entity model can be computed using a feed-forward neural network with a softmax layer at the output. The training task that allows fitting the parameters of the model corresponds to the forecast of e conditioned to a multi-user context C. This training task resembles the task defined to train the CBOW

model for word embeddings (Mikolov et al., 2013), in which the context of a target word is given by the set of words that belong to its surrounding text. The loss function of the user-entity model is:

$$\mathcal{L} = -\frac{1}{|S|} \cdot \sum_{S} \log(p(e|u_1, \ldots, u_C))$$

where S is the set of training instances. We can rewrite the loss function in terms of the latent vector:

$$\mathcal{L} = \frac{1}{|S|} \cdot \sum_{S} [\log(\sum_{C} \exp(v^{(u)t} \cdot \mathbf{h} - v^{(e)t} \cdot \mathbf{h}))]$$

User-entity model outputs. The user-entity model can be used to recover a scoring function over the network. Given a polarized named entity e, we define the projection of e on the user network:

$$\mathbf{h}_e = \mathbf{W}_{in} \vec{v}^{(e)}$$

The projection of e on the network indicates the willingness of each user to comment on a given polarity orientation about e. Analogously, we can define the projection of a user u over the set of polarized entities:

$$\mathbf{h}_u = \mathbf{W}^t_{out} \vec{v}^{(u)},$$

which indicates the willingness of u to comment about each entity in a given polarized orientation. Accordingly, we define a scoring function s that measures the relation between a user u and an entity e according to a given polarity:

$$s(e, u) = \vec{v}^{(u)^t} \cdot \vec{v}^{(e)}.$$

The function s corresponds to the product between the vectors that represent u and e in the user-entity model. The product corresponds to the inner product of both vectors and therefore s measures the angle between u and e in the user-entity model.

3.4. Learning a multi-relational graph

Figure 2 shows that **GENE** takes the parameters of the user-entity model to learn a multi-relational graph. A min-max scaler applies to each column in $\mathbf{W}_{in}$. **GENE** uses each column vector as a distribution of stationary probabilities over the user network. **GENE** feeds each distribution $\pi^{(i)}$ in a graph generator according to a preferential attachment generative model computed using random walks. Two parameters define the graph generator, the number of steps of the random walks, and the dumping factor that defines the probability of process restart. The graph generator produces as many graphs as latent variables the user-entity model has. These graphs are used to generate a multi-relational graph, where each relationship corresponds to a graph generated from the user-entity model. Then, the multi-relational graph is vectorized using PyTorchBigGraph (PBG), a graph-based representation learning framework released by Facebook AI (Lerer et al., 2019). On the parameters trained by PBG, a second fitting process is conducted, which recomputes the parameters of the graph considering the strength of the relationships between users detected in the user-entity model. The output of this process releases vector representations of users, conditioned on the latent factors of the user-entity model, as it is shown in letter c. These vectors are indexed to efficiently compute proximity relationships between users according to each latent factor of **GENE**.

Let $\mathbf{W}^{(i)}_{in}$ be the i-th column of $\mathbf{W}_{in}$. We define $\pi^{(i)} = \frac{\mathbf{W}^{(i)}_{in}}{\sum_j \mathbf{W}^{(i)}_{in}[j]}$ as the i-th probability vector of the user-entity model projected on the user network. Note that i ranges from 1 to k, and k indicates the number of latent factors of the user-entity model. Note that if the user-entity model is computed using a feed-forward neural network with one hidden layer, k corresponds to the number of neurons of the hidden layer.

Let $G^{(i)}$ be a weighted directed graph generated using a preferential attachment model consistent with $\pi^{(i)}$. To generate a graph whose distribution of stationary probability is given by $\pi^{(i)}$, we run a random walk process with restarts. The process starts from a node chosen at random and then it continues along

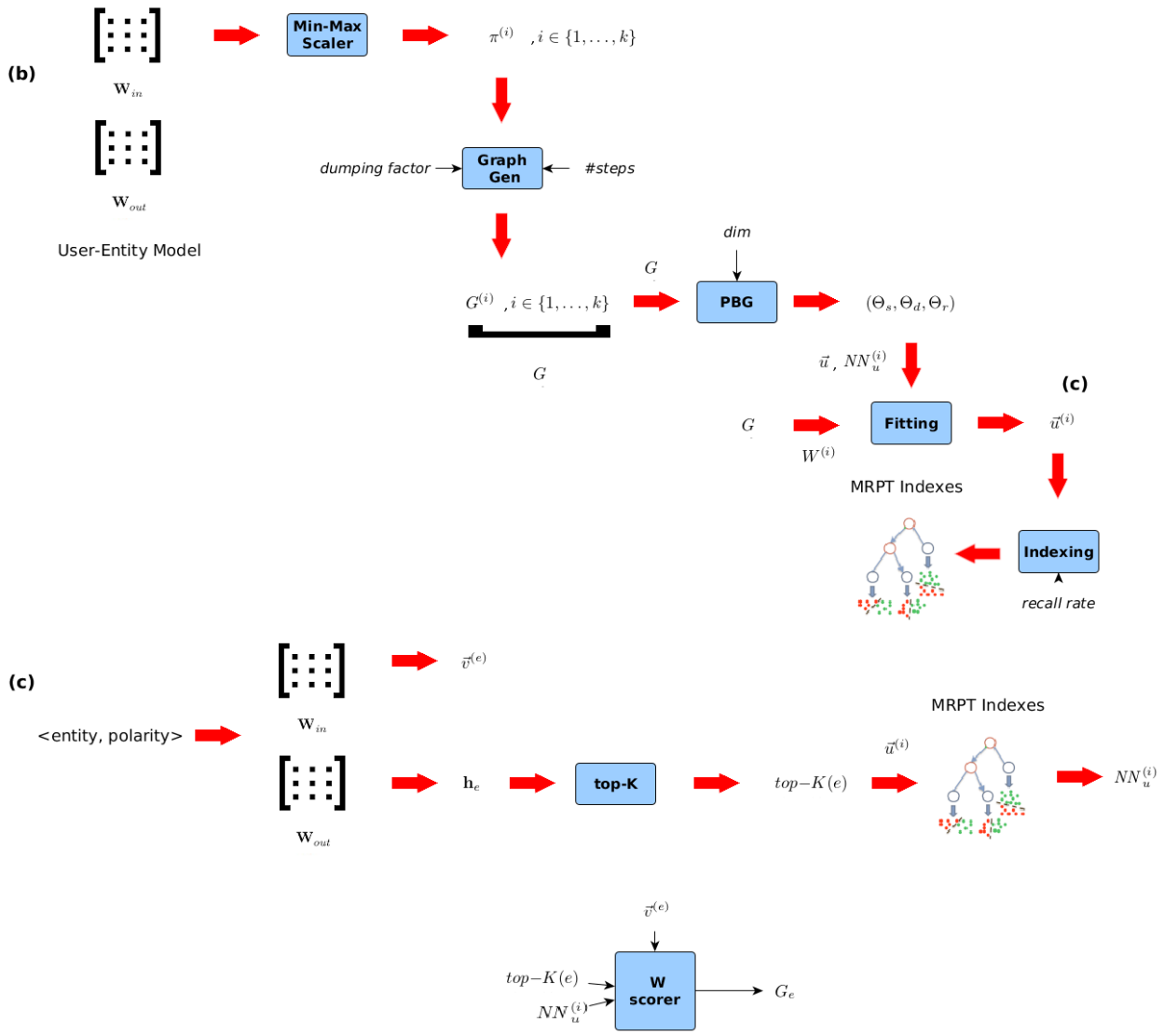


Figure 2: **GENE** produces proximity graphs between users conditioned to the latent factors of the user-entity model. Proximity graphs are used to generate polarized views of user networks involved in online discussions.

the graph, choosing the next node according to $\pi^{(i)}$, considering a probability of restart given by a damping factor. The process is defined by a finite number of transitions that are determined according to the number of nodes of $G^{(i)}$. During the process, each edge of $G^{(i)}$ records the number of times the random walk passed through it. Once the process concludes, a process is driven so that the weights of $G^{(i)}$ are transformed into probabilities of transition between nodes.

Note that the networks that are generated from each latent factor of the user-entity model do not necessarily represent interactions between users. By constructing a graph using the latent variable as the stationary probability of the network, greater visibility in the network is given to users who have a significant presence in multi-user contexts related to prevalent entities for that latent factor. The effect that occurs in the network is to cluster users who share the same orientation around entities in common. Consequently, each network represents communities with coherent orientations towards the same group of entities.

We built an embedding that encodes the set of graphs $G^{(i)}$ in a single and coherent representation using

PBG (Lerer et al., 2019). PBG offers the chance to learn representations of nodes (node embeddings), of relationships (edge embeddings), or both (hybrid embedding). Since the networks we work with are multi-relational, the type of representation we learn provides hybrid embeddings. To learn a hybrid embedding, we define a multi-relational graph G, which encompasses all the graphs $G^{(i)}$, $i \in \{1, \ldots, k\}$ learned from the user-entity model. We define G as follows. Let $G(V, R, E)$ be a multi-relational graph, where V is the set of nodes of G and it represents the users of the user-entity model. R is a collection of relationships between users, so that each relation $r^{(i)} \in R$ corresponds to the relations represented in the graphs $G^{(i)}$, $i \in \{1, \ldots, k\}$ learned from the user-entity model. E is a set of edges that connects pairs of users through relations $r^{(i)} \in R$. An edge $e \in E$ is defined by a triplet $(s, r^{(i)}, d)$, where s and d are nodes $\in V$ and $e = (s, d)$ belongs to $G^{(i)}$. For each triplet in E, PBG defines model parameters denoted by $(\Theta_s, \Theta_d, \Theta_r)$ and a fitness function f whose purpose is to reach a maximum if the triplet e belongs to E, and a minimum otherwise. To do this, f encodes the representation learned for s and d through $r^{(i)}$: i.e., $f(\Theta_s, \Theta_d, \Theta_r^{(i)}) = sim(g_{(s)}(\Theta_s, \Theta_r^{(i)}), g_{(d)}(\Theta_d, \Theta_r^{(i)}))$. The function g encodes the relation between the node and the relation. PBG combines both functions to learn the embeddings. Then, the embeddings are trained defining a task that minimizes the triplet loss function:

$$\mathcal{L} = \sum_{e \in E} \sum_{e' \in S'_e} \text{MAX}(f(e) - f(e') + \lambda, 0),$$

where S'_e is a set of edges obtained using negative sampling. PBG samples half of negatives according to their prevalence in the training data (i.e., corrupting a node of an existing edge), generating the other half of them uniformly at random. PBG optimizes the margin-based ranking objective between each edge in the training data and the set of negative sampling edges. Therefore, PBG maximizes the inner product (minimizes the distance) between the embedding parameters that correspond to edges in E and minimizes the inner product (maximizes the distance) between the embedding parameters that correspond to edges in S'_e. The inclusion of S'_e allows to use a type of training that is similar to the one used in word2vec with negative sampling (Mikolov et al., 2013).

PBG allows defining f and g in different ways depending on the type of relationships that are modeled. As the graph is multi-relational and the expected number of relationships is high, we choose to implement g for s and d using $\Theta_s \odot \Theta_r$ and $\Theta_d \odot \Theta_r$, respectively. Then, the function f is defined using the relation-dependent scoring function:

$$f(e) = \langle \Theta_s \odot \Theta_r, \Theta_d \odot \Theta_r \rangle,$$

which corresponds to the DistMult learning representation strategy (Yang et al., 2015). DistMult is very useful in capturing rich structures hidden in the multi-relational data, the reason why we chose this option instead of others implemented in PBG. Once the model parameters are learned, the user embeddings are retrieved, combining them with the parameters of each relationship. The effect produced on the parameters of s and d is to have specific user embeddings for each relation of the graph.

It can be noted that each $G^{(i)}$ defines a type of relationship in G. Since $G^{(i)}$ is a weighted graph, we can use these values to fit the embeddings to $G^{(i)}$ according to its distribution of weights. Let $W^{(i)}$ be the distribution of weights defined in $G^{(i)}$ over E. Note that if $e \in E$ does not belong to $G^{(i)}$, $W^{(i)}(e) = 0$. In another case, the weight corresponds to the value learned during the generation of $G^{(i)}$.

Formally, let $\vec{u}^{(i)}$ be the vector representation learned using PBG for a given user u and a given relation $r^{(i)}$. We define the neighborhood $NN_u^{(i)} = \{u_1, \ldots, u_{n_u}\}$ as the users in V that are connected with u in $G^{(i)}$. To make efficient use of computing resources, we will constraint $NN_u^{(i)}$ to the nearest neighbors of u. Each neighborhood is computed using the encoding of the relationship. The scoring function used to retrieve the nearest neighbors is $\langle \Theta_u^{(i)}, \Theta_r^{(i)}, \Theta_v^{(i)} \rangle$, which corresponds to the inner product of the triplet. The triplet scoring function used is defined by $\sum_{d=1}^{D} \Theta_{u,d} \cdot v_{r,d} \cdot \Theta_{v,d}$, where d ranges the dimensionality of the vector of parameters that represents the relation. The relations are limited to diagonal matrices represented by the vector $v_{r,d}$. As the operator is commutative (we can exchange u and v obtaining the same result), the model represents symmetric relationships. This restriction leads to less over-fitting.

Given the distribution of weights $W^{(i)}$ in $G^{(i)}$ computed in the user-entity model, the vector representation $\vec{u}^{(i)}$ of u can be fitted to its neighborhood by adding to $\vec{u}^{(i)}$ the vectors $\{\vec{u}_1, \ldots, \vec{u}_{n_u}\}$ according to $W^{(i)}$. To keep the scale learned from the original embedding, the new vector $\vec{u}^{(i)'}$ is scaled by the weights of its neighbors:

$$\vec{u}^{(i)'} = \frac{1}{n_u + \sum_{u_j \in NN_u^{(i)}} W^{(i)}[u_j, u]} \cdot \left(n_u \cdot \vec{u}^{(i)} + \sum_{j=1}^{n_u} \vec{u}_j^{(i)'} \cdot W^{(i)}[u_j, u] \right). \quad (1)$$

Equation (1) shows how the vector representation of u is fitted to its neighborhood. To consider the cross effect of the fitting of each vector to its neighborhood, we use an iterative algorithm that conduct a scan through each user vector. After several iterations, all vectors converge in the sense of the least-squares. The vector recalculation algorithm minimizes the function of loss of least-squares between $\vec{u}^{(i)}$ and the new embedding $\vec{u}^{(i)'}$:

$$\mathcal{L}^{(i)} = \sum_{u \in V} \left[n_u \cdot \|\vec{u}^{(i)'} - \vec{u}^{(i)}\|^2 + \sum_{j=1}^{n_u} W^{(i)}[u_j, u] \cdot \|\vec{u}_j^{(i)'} - \vec{u}^{(i)}\|^2 \right], i \in \{1, k\} \quad (2)$$

To favor the efficient use of user embeddings, we create metric indexes on the representations, through which we can run nearest neighbors queries. As the user embeddings are relation-dependent, we created one index per relation. We use multiple random projection trees (MRPT) (Hyvönen et al., 2016) to index the user embeddings, an index that uses an approximate algorithm providing guarantees on recall rates. MRPT is adjustable according to a user-defined recall rate, which establishes the sizing of the index necessary to accomplish the guarantee. In our experience, a 95% recall rate establishes a fair compromise between the size of the index and the running time of each query.

3.5. Computing graphs conditioned on named entities

The third component of **GENE** is responsible for computing the views of the user network according to the participants in a discussion. For this stage, there are two approaches. The offline approach shows the historical leaning of users towards entities. The online approach retrieves a view of the network to characterize the scenario in which a discussion takes place.

The offline approach retrieves the top-K users with a greater leaning towards the entities of interest and a given polarity, using the polarized entity vector of the user-entity model $\mathbf{h}_e$. Using **GENE** indexes, the relation-dependent neighbors of these top-K users are retrieved. Then, we compute the embeddings of these users, which corresponds to the linear combination of their relation-dependent embeddings, according to the prevalence of the entity $\mathbf{v}_e$ in each relation $r^{(i)}$. The online approach, instead of retrieving top-K users using $\mathbf{h}_e$, retrieves users from a conversational thread during the development of a discussion. For these users, their relation-dependent neighborhoods are retrieved using the indices of **GENE**, and their embeddings are computed by combining their relation-dependent embeddings according to the prevalence vectors of the entities $\mathbf{v}_e$ named in the news that triggered the conversational thread. As $\mathbf{v}_e$ corresponds to a polarized representation of an entity, both approaches provide a view of the user network conditioned to the named entities and a given polarity, as indicated in letter c of Figure 2. As we work with three polarities, we achieve several user networks, one for each entity and polarity. Accordingly, we conduct a consolidation step the merge these networks computing a proximity graph G_e.

For the offline analysis approach of **GENE**, the graph G_e is defined by a set of nodes $V_e = \{u \in \text{top-}K(e)\} \cup \{v \in NN_u^{(i)} | u \in \text{top-}K(e), i \in 1, \ldots, k\}$, a set of edges $E_e = \{(u, v) | (u, v) \in V_e \times V_e\}$, and a function of weights:

$$W[u, v] = \sum_{i=1}^{k} \text{Cos}(\vec{u}^{(i)}, \vec{v}^{(i)}) \cdot \vec{v}^{(e)}[i], \quad (3)$$

where k is the number of relations in G (i.e., number of factors defined in the user-entity model). Equation (3) calculates the strength of each edge in G_e considering the cosine similarity between u and v in each relation of G, weighted with the prevalence of that relation on e. Note that the vector of prevalence correspond to the vector representation $\vec{v}^{(e)}$ of the entity. In this way, the edges with the largest weights in G_e correspond to the pairs $\langle u, v \rangle$ that are more strongly connected in G and that are more relevant to describe the mentions of e in the user-entity model.

For the online mode of **GENE**, the computation of G_e corresponds to a slight variation of the offline mode. Instead of retrieving the top-K users of the user-entity model related to an entity of interest, in the online mode, **GENE** retrieves the users involved in a discussion in progress. The neighborhoods and embeddings of these users are retrieved using the indices of **GENE**, and then they are combined using the weight scorer defined in Equation (3)

4. Materials and methods

4.1. Data

The study was conducted using data retrieved from an online news site in Chile named Emol [1]. The data is openly published by Emol, using JSON format. The data was downloaded directly from the site. To comply with data privacy protocols, user identities were anonymized. The dataset is fully available [2].

Emol users can create an account on the site or associate a Facebook or Twitter account within site. All users with Emol publishing rights do so with authentication. There are no anonymous comments.

Emol provides to its users with mechanisms to comment each news. Users can comment directly on a story. The site keeps a record of the comments related to each news item published by Emol. Emol provides a voting mechanism for comments, which considers likes and dislikes. Site users widely use this mechanism. There are other interaction mechanisms provided by Emol, such as blocking users and reporting comments, which have less use. In addition to commenting on news, Emol users can keep a wall. In this place, opinions of general interest can be published. Emol users can follow other users to access their walls. However, the social graph associated with this interaction mechanism has low density, so it is of little interest to our study.

The dataset comprises a total of 143340 news retrieved from April 1, 2016, to April 20, 2019. Of this set, 122778 news have comments.

4.2. Exploratory analysis

Only some news captures the attention of Emol users. In Figure 3, we show that a significant fraction of news produces very few comments (see Figure 3 top-left). On the other hand, some news produces many comments (see Figure 3 bottom-right). The most commented news on Emol produces in the order of thousands of comments. While some news has a much more significant impact than others, we observe that the Emol user community is active, producing news with a significant amount of comments.

The total number of users who have posted comments is 192551. In Figure 4, we show the distribution of comments by users and the support of comments on the population of users of the site. Although some users are much more active than others, the comments have significant support in the population of users of the site. 80% of the comments have support in 10,000 users or more. At the bottom of figure 4a, we observe that several users have a long record of comments in the site, ranging from 500 to 10,000. On the other hand, a significant fraction of users shows sporadic interactions. This fact indicates that the collective attention of the community is sporadic and focuses on a few high impact events. A considerable proportion of the activity is carried regularly by the users at the bottom of figure 4b, who show traits of cyberactivism due to the volume of comments produced.

The site provides users with mechanisms to express their consent or disagreement with the comments related to the published news. Emol users widely use the mechanism of likes and dislikes as it is shown in Figure 5. Each user can emit only one like or dislike per comment.

[1] https://www.emol.com/

[2] https://doi.org/10.6084/m9.figshare.c.4834062.v1

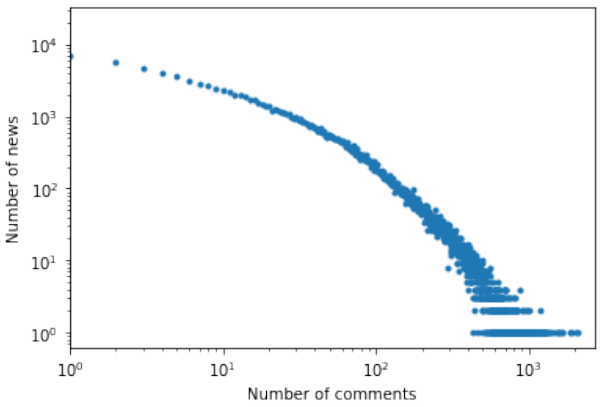
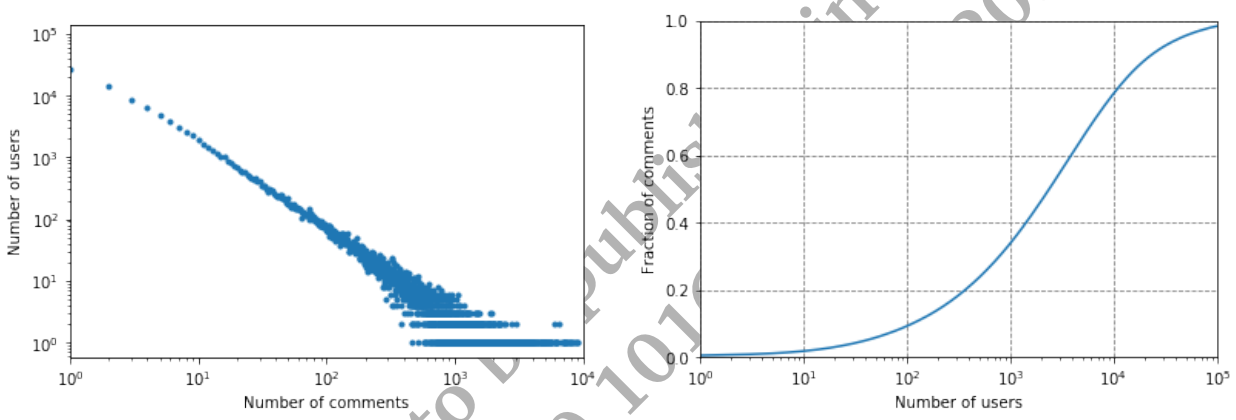


Figure 3: Some news produces many more comments than others. It is observed that the Emol user community is active, producing news with a significant amount of comments.



(a) Distribution of users and number of comments.

(b) Fraction of comments and their support in the site.

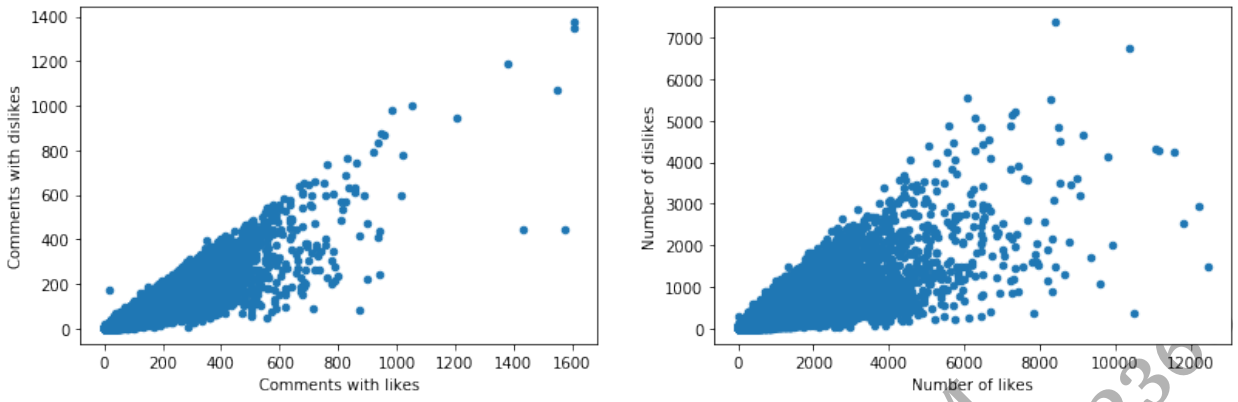
Figure 4: Although some users are much more active than others, comments have significant support in the population of users of the site. 80% of the comments have support on 10,000 users or more.

As Figure 5 shows, there is a kind of balance between likes and dislikes. However, some news produces comments with many more likes and dislikes than others. This finding indicates that some news has a much more significant polarization effect than others, generating more polarizing comments.

4.3. Labeling controversial news

We will use the likes and dislikes votes as controversy proxies to label the dataset. We claim that news that produces controversy generates many comments with likes and dislikes. This way of imputing controversy models the coupling of two factors, the collective attention that news produced, and the polarization that the comments of the news produced in the community. Controversial news, first, will capture the attention of many users with opposite views. These points of view will materialize in the likes and dislikes that the comments associated with the news will receive.

Likes and dislikes measure the level of implicit interaction between users in the conversational thread of a news. Unlike most of the studies performed on platforms such as Twitter or Facebook, the social graph loses relevance in an online news site. In this context, users interact directly in the conversational thread triggered by a news. What loses relevance at this point is the direct interaction between users. Instead, the interaction with the news and the conversational thread triggered by it becomes more relevant. We believe



(a) Comments with likes and dislikes.

(b) Total amount of likes and dislikes per news.

Figure 5: The comment voting mechanism is widely used by Emol users. The voting mechanism exhibits a balance between likes and dislikes. It is observed that some news produces comments with many likes and dislikes, uncovering the level of polarization that news and its comments can trigger in the community.

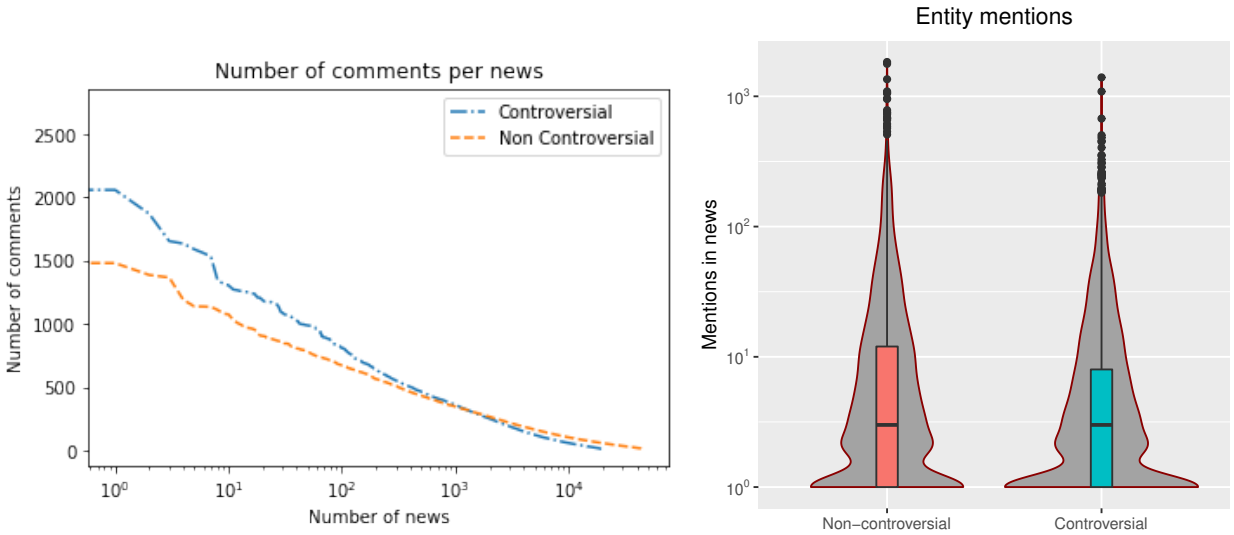
that this is a more transparent way to relieve information about the content itself, reducing the effect that direct interaction between users introduces in the controversy imputation process.

To impute the controversy of each news, we retrieved the conversational thread of each of them, recording the number of likes and dislikes of each comment. In this way, each comment was processed as a sample of controversy. The number of likes and dislikes of each sample was counted to compute a vote density function for each news. Then, each comment corresponds to an observation of the voting density function of the news. The number of likes and dislikes was represented in the positive (likes) and negative (dislikes) axes to approximate the density function, respectively. To limit the effect of the noise that this data usually has, we established a minimum number of votes to process a valid observation. The minimum number of votes per comment was set at 5. Therefore, a controversy label was imputed to that news that generated at least ten valid comments, to have at least ten observations of the news density function. Accordingly, the imputable set of news was reduced to 79728 instances.

To each density function, we applied a unimodality test, retrieving its p-value. For this purpose, we used the unidip algorithm (Maurus & Plant, 2016), which is a robust noise-clustering algorithm for one-dimensional numerical data. The algorithm recursively extracts density peaks in the data using the Hartigan's Dip-test of unimodality (Hartigan & Hartigan, 1985). The Hartigan's Dip-test was performed at a sensitivity level for the p-value set at 0.05, using 1000 trials of the test for each density function. The test indicates unimodality if the p-value > 0.1. If the vote density function corresponds to a unimodal function, it is understood that there is a homogeneous population that consistently generated the votes. If the p-value is ≤ 0.1, it is assumed that the data was generated from two or more modes produced by two or more groups with divergent voting patterns.

There are cases in which the algorithm does not find enough verifiable information. This is because there were no clear voting patterns in the data. Of the 79728 news, 66036 were verifiable by the algorithm, determining that of these 46451 were unimodal. These news were labeled as non-controversial, and the remaining 19686 as controversial. We compare the number of messages per news according to the imputed classes. These plots are shown in Figure 6a.

As Figure 6a shows, controversial news produces more comments, which confirms that collective attention is a determining factor for the occurrence of a conflict. This fact happens for the news that produces the most significant impact, which is the first thousand most commented news of each class, each one with at least 500 comments on the site. From the one thousand news onwards, the tails of both distributions are quite similar, and, at the end of the distribution, the non-controversial news of lower interest produces more comments than the controversial news.



(a) (b)

Figure 6: (a) Among the news that attracts more attention, controversial news produces more comments than non-controversial ones. However, the tail of both distributions is quite similar, and, among the news with the lowest impact, the non-controversial ones produce more comments than the controversial ones. (b) As the dataset has more non-controversial news, we record more mentions to entities in this type of news. However, both distributions are heavy-tailed and share the same median.

The dataset shows differences in terms of the most named entities in both classes. The NER algorithm detected 16,912 different named entities in the news of EMOL. The total number of entity mentions in non-controversial and controversial news are 103,032 and 47,428, respectively. Accordingly, the number of entities mentioned in each news, on average, is 2.28 and 2.45, respectively. The presence of more entities in controversial news indicates that the emergence of the controversy bears some relation to a higher number of entities mentioned in the description of a news story. As Figure 6b shows, some entities have many mentions, but most entities have few occurrences in the dataset. In both types of news, the distribution of the number of mentions per entity is skew, with a long-tail of few-mentioned entities. Entities have more mentions in non-controversial news because the dataset has more news of this type. In the heavy-tail of both distributions, both types of news share the median. As a significant fraction of the mentions in the news corresponds to the same entities, it is interesting to see what happens in the head of each distribution.

Table 2 shows the top-20 most mentioned entities along with the number of news items in which they have been mentioned. The first columns show the entities related to non-controversial news, while the last columns show those that correspond to controversial news.

Table 2 shows that some entities are mentioned in both classes, such as Colo-Colo, a famous soccer team in Chile, or Sebastián Piñera, the current president of Chile. The presence of soccer players in non-controversial news, such as Alexis Sánchez, Arturo Vidal, and Claudio Bravo, suggests that a significant fraction of the non-controversial news is about sports. On the other hand, among the most named entities in controversial news are government, President Bachelet, president and minister, which correspond to the most important political named entities of the Chilean government. The presence of politicians such as Sebastián Piñera, Alejandro Guillier, Cristina Fernández, Nicolás Maduro, and Evo Morales among the entities with most mentions in controversial news is striking, indicating that a significant fraction of these news is about politics. It is important to note that in this type of news, the presence of institutions is salient. Several of these institutions correspond to government institutions such as the court, the public prosecutor, or the deputies.

Emol news are classified into news categories. A crew of journalists makes this classification for the site. Table 3 shows the fractions of controversial and non-controversial news in each informational category.

Table 2: This table of named entities shows clear differences between non-controversial and controversial news. While in the non-controversial news, the entities with the most mentions are soccer players, in the controversial the presence political leaders is predominant. Some entities were translated from Spanish to English, which are shown with italic fonts, and others were kept in Spanish so as not to lose their meaning, as in the case of Roja, DC, and Nueva Mayoría, which correspond to the names given to the Chilean national soccer team, to a political party, and a political alliance, respectively.

Ranking	Entity	Non-controversial news	Entity	Controversial news
1	Sebastián Piñera	1835	*government*	1397
2	*government*	1776	*President* Bachelet	1089
3	Donald Trump	1352	Sebastián Piñera	675
4	Colo-Colo	1091	*president*	503
5	*president*	1066	Nicolás Maduro	481
6	Roja	1053	Alejandro Guillier	454
7	Alexis Sánchez	953	*minister*	448
8	Arturo Vidal	783	DC	404
9	Michelle Bachelet	744	Evo Morales	351
10	*minister*	676	*deputies*	264
11	Santiago	658	*public prosecutor*	253
12	Claudio Bravo	589	*opposition*	250
13	Jorge Sampaoli	509	Nueva Mayoría	234
14	Alejandro Guillier	459	*court*	197
15	*soccer team*	420	Roja	196
16	Chile Vamos	415	Cristina Fernández	191
17	Juan Antonio Pizzi	399	Colo-Colo	190
18	Champions	393	Araucanía	188
19	Lionel Messi	389	*justice*	185
20	*public prosecutor*	354	*reform*	180

Table 3: Sports has an important presence on non-controversial news, more than double what is observed in the controversial news. On the other hand, politics is pervasive in the controversial news.

Category	Non controversial news	Fraction	Controversial news	Fraction
National (including politics)	16079	0.346	10439	0.530
Sports	13337	0.287	2949	0.149
International	6348	0.136	2871	0.145
Economy	4039	0.086	1500	0.076
Entertainment	3777	0.081	1181	0.059
Trends	1974	0.042	542	0.027
Technology	897	0.019	204	0.010

Table 3 shows that most of the news belongs to the category of national news. This category includes news about politics. The presence of sports in non-controversial news is significant (0.28), and it is of lesser relative importance in controversial news (0.14). The relative importance of national is crucial in the controversial news, with more than half of the class devoted to this issue (0.53). While national news also occupies the first place in the ranking of non-controversial news categories, its relative importance decreases (0.34).

4.4. Validation methodology

The validation methodology considers two complementary approaches. The first one studies the problem of detecting controversy a posteriori, that is, with complete data of the event once it has already occurred. This approach allows us to compare **GENE** with related work methods that take full data of the event. The purpose of these methods is to produce a ex-post characterization of the event for data analytics purposes.

They do not have an anticipatory predictive purpose, which is why they work with full data a posteriori. The second methodological approach considers controversy forecasting during an event and allows us to evaluate the predictive capacity of **GENE**. This approach considers data acquisition during the development of an event. The study of the early detection of controversy is crucial for this methodological approach.

The news dataset was partitioned into two folds: one for training and one for testing. The dataset partition was done by sorting the news according to timestamp and determining the date around which the training partition (news prior to the date) and the testing partition (news after the date) produced an 80/20 partition, respectively. Of the 66036 labeled news, the first 53145 were used in the training partition, reserving the remaining 12891 for testing. The date around which the partition was made was October 30, 2018, which corresponds to a training partition with news covering a horizon of two years and a half, reserving six months of news for testing. Of these testing examples, 3901 correspond to controversial news, and the remaining 8990 to non-controversial.

Controversy Detection (ex-post)

In this methodological approach, we evaluate the ability to detect controversy using different methods on complete data, that is, once the event occurred. The related work shows several approaches that seek to detect controversy by analyzing the comments produced during the event of interest (Popescu & Pennacchiotti, 2010; Mejova et al., 2014; Hamad et al., 2018). These works use lexicographic features of the text to build controversy classifiers based on the messages. To study the performance of these methods, we build classifiers based on linguistic characteristics using conversational threads. We used two variants for this purpose. The first was a text classifier that, for each story, build a vector of words associated with each comment. To create one vector associated with the news, the comment vectors were averaged. The word vector was built using a bag-of-words (**BOW**) Tf-Idf vector representation, similar to the works conducted by Popescu & Pennacchiotti (2010) and Mejova et al. (2014). A second variant considers a sentiment analysis preprocessing of each comment to infer its polarity, which we call **BOW-SA**. For this, we used a sentiment analysis method in Spanish [3]. The polarity scores were concatenated into the feature vector of each news item using the mean and variance of the set of values. This method allows evaluating the capacity that sentiment analysis has in the detection task, similar to the study conducted by Hamad et al. (2018). **SVM** classifiers were used for this task, with linear kernel and balanced class weight to avoid overfitting the majority class. Random forest (**RF**) was also used, with balanced class weight. Additionally, an architecture based on recurrent neural networks was considered, for which the vector of each comment was ingested according to a sequence ordered by time. The variant evaluated was a bidirectional LSTM (**bi-LSTM**), with hidden units of 32 dimensions and a softmax at the output. To work with variable length sequences, we used padding. The input sequence in the **bi-LSTM** network was fixed using the average length of the sequences in the training partition. To study a text representation based on word embeddings, word vectors computed using **BERT** (Devlin et al., 2019) were also considered, which is regarded as state of the art in word embeddings.

To evaluate features of the user network, we used detection measures applied to **GENE**. It should be noted that studies on retweet networks or networks of mentions are not directly comparable with this study (Conover et al., 2011; Calais Guerra et al., 2013; Ruan et al., 2013). Instead, the most relevant interaction mechanism on Emol is done by directly commenting on the news. To conduct a fair comparison with the methods of related work that have considered user network features, therefore, we study the predictive capacity of controversy detection metrics on the graphs generated by **GENE**. We used the measure of detection of coupling between communities based on random walks (**RWC**) proposed by Garimella et al. (2016b) to conduct this comparison, which is considered state-of-the-art in detecting ex-post controversy on user networks. If the results are satisfactory, it will be shown that the controversy analysis conducted on Twitter is transferable to online news sites with weak social graphs using **GENE**. Also, we used a measure proposed by us to compute a score of controversy, called controversy based on relative closeness (**RCC**). Both measures were also used in the ex-ante analysis and will be explained in the next sections.

[3] `https://github.com/aylliote/senti-py`

Graph-based baselines

We studied two additional ways to generate graphs, making it possible to have graph-based baselines with which to compare the performance of **GENE**. These graphs are not conditioned to named entities, which allows quantifying the contribution of **GENE**. To generate these graphs, we will follow two ways of using news engagement data, establishing relationships between users who comment the same news.

User engagement graph: The user engagement graph connects users who comment on the same news. We have two versions of this undirected graph, the subsequent engagement graph, which connects two users if they subsequently comment on the same news item, and the full engagement graph, which connects all users who commented on the same news item. Both versions of the graph allow us to evaluate the usefulness of this data to generate graph representations with different densities, the subsequent engagement graph being less dense than the full engagement graph. One difficulty with using this data is that it is not directly partitionable, and then they are not useful for controversy analysis. While the graph of subsequent comments produces on average a connectivity of degree 2 (each user connects with the previous and subsequent user in the thread), the dense graph generates a k-clique, where k is the number of users who commented on the news item. Neither of these graphs is partitionable. To extract the most utility from this data, we use these graphs to build user embeddings. The graphs were combined, generating a single weighted graph, where the weight of each edge represents the number of news engagements in common. To produce reliable relations between users, we created an edge between two users only if there were at least two news engagements per edge. The rest of the arcs were removed to avoid spurious relationships and a high level of noise. The subsequent and full graphs comprise 159,788 and 161,391 users, with 4,761,841 and 125,463,737 edges each one. We computed user embeddings using PBG. We did not use an operator to model relationships. The node embedding method implemented in PBG learns from pairs of nodes, doing similar the parameters of nodes that are connected in the graph, and doing dissimilar if they are not. Unconnected node pairs are produced using negative sampling. We set the dimensions of the embeddings at 200. The embeddings were learned in 10 epochs, with a learning rate of 0.001. We denote by **Eng-sub** and **Eng-full** each one of these graph-based representations of users.

User-user graph: The user-user graph is obtained from a variant of the user-entity model. For each news item, the comment threads are retrieved, and a fixed-length sliding window is moved over the thread, retrieving sequences of users who have subsequently commented on the same news item. For each user window $X = (u_{i-w}, \ldots, u_i, \ldots, u_{i+w})$, the target $Y = u_i$ corresponds to the user who occupies the central position in the window. On all the pairs (X, Y) of the dataset, a user-user model is trained, in a similar way to that used to create the user-entity model. The user-user model solves the same task addressed by CBOW (Mikolov et al., 2013), in which the distances between embeddings of words that occur in the same contexts are minimized. The user-user model was trained with windows of length 5, generating 6,005,385 training instances. One hundred neurons were used in the hidden layer, using categorical cross-entropy as a loss function and rmsprop as an optimizer. The model under this configuration considered a total of 32,439,591 parameters, with 161,391 users. The training was done with 32-size batches. The network was trained in 10 epochs. We denote by **UU-graph** this graph-based representation of users.

The user embeddings calculated using the methods introduced above allow us to generate a proximity graph between users. For each news, we connected every pair of users who commented on it with an arc whose weight corresponds to the cosine similarity of their embeddings. We use a graph partitioning algorithm to produce two partitions in each proximity graph. After partitioning a graph, it is possible to use both indexes of controversy, either **RWC** or **RCC**. To partition each graph, we use METIS (Karypis & Kumar, 1998) , a state-of-the-art k-direct graph partitioning algorithm that can efficiently work on weighted graphs. METIS is faster than other algorithms, such as spectral clustering or recursive bisection partitioning. METIS has been successfully used to partition graphs in controversy analysis (Lasalle & Karypis, 2013; Garimella et al., 2016b).

The graph-based baselines introduced can be used in both evaluation scenarios, that is, in *ex-ante* or *ex-post* scenarios.

Controversy detection (ex-ante)

To study the early detection of controversy using **GENE**, we stripped the data acquired from the site on an hourly basis. Thus, in each hour, the predictive capacity of **GENE** was carried out. Two predictive tasks were analyzed, the polarization of the users and the controversy imputed to the news.

The task of polarity forecasting is to determine the fraction of intervening users that adopt a specific polarity in their comments concerning a given news. To conduct this task, a node labeling procedure was implemented using **GENE**, which works as follows:

Labeling the nodes of GENE. For each analyzed news, the entities mentioned in the title and subhead of the news are retrieved. Then, for a group of intervening users, the node embeddings of these users computed in **GENE** are retrieved, and three graphs are built, one for each polarity (positive, negative, neutral), according to the method described in Section 3.5. For each user involved, all their neighbors are retrieved in the three graphs, and its polarity is imputed by comparing the weights of their neighbors' edges. The comparison is made by ordering the edges retrieved from the three graphs according to their weight. Each triplet is solved by comparing the weights and imputing to the triplet the label with the highest value. Then, the majority function is applied to the resulting labels, imputing the label to the user. It can be observed that the graphs that correspond to different entities can be overlapped, which allows producing graphs conditioned to two or more entities named in the same news.

The controversy detection task operates on the graph labeled by the node labeling process. The level of controversy is measured by computing the level of coupling between the three poles of the network (positive, negative, and neutral). We used two approaches to measure the coupling between the poles, random walks (**RWC**), like that proposed by Garimella et al. (2016a), and relative closeness (**RCC**), which corresponds to a measure based on cluster cohesion.

Controversy based on random walks. Garimella et al. (2016a) defines the random walk controversy score (**RWC**) by computing several random walks and counting how many of them end at the same pole in which they started and how much of them end at the opposite pole. These accounts allow them to estimate transition probabilities between both poles and transition probabilities to the same pole. The controversy score based on random walks (**RWC**) corresponds to the difference between both quantities:

$$\mathrm{RWC}_{x,y} = P_{x,x}P_{y,y} - P_{x,y}P_{y,x}$$

The **RWC** score is close to one when the coupling between both poles is low and is close to zero when the coupling is high. This measure is pairwise, so to define an **RWC** score for **GENE**, we need to extend it to three poles. To fulfill this purpose, we measure the level of coupling of the positive and negative poles with the neutral pole, computing their relative difference with the coupling between the positive and negative poles:

$$\mathrm{RWC} = \frac{2 \cdot \mathrm{RWC}_{+,-}}{\mathrm{RWC}_{+,N} + \mathrm{RWC}_{-,N}}$$

If the coupling between the positive and negative poles is low, and the coupling with the neutrality is high, **RWC** will have values greater than one, indicating the absence of controversy. On the other hand, a high coupling between the positive and negative poles and a low coupling with neutrality will show an **RWC** with values less than one, indicating the presence of controversy.

Controversy based on relative closeness. This controversy score is based on the measure of cluster cohesion called relative closeness, which measures the ratio between the similarity between two clusters and the internal similarity of each cluster. The similarity between two nodes in **GENE** can be measured using the weight of the edge that joins two nodes. The internal similarity of a cluster is the sum of the weights of all its edges. The similarity between two clusters is the sum of the weights of all the edges that connect nodes in both clusters. Therefore, relative closeness (RC) is defined according to:

$$\mathrm{RC}_{x,y} = \frac{2 \cdot \textsc{Sim}_{x,y}}{\textsc{Sim}_{x,x} + \textsc{Sim}_{y,y}}$$

The relative closeness measure gets a ratio higher than one if the clusters are coupled, and less than one if they are decoupled.

The relative closeness measure is pairwise. Since the node labeling process provides three graphs, one for each polarity, we need to extend the definition of relative closeness. We do this by measuring the relative closeness of each pole to the neutrality graph, and computing its ratio with the closeness between the positive and negative poles. Accordingly, we define the Relative Closeness Controversy (**RCC**) score, as follows:

$$\mathrm{RCC} = \frac{2 \cdot \mathrm{RC}_{+,-}}{\mathrm{RC}_{+,N} + \mathrm{RC}_{-,N}}$$

The **RCC** score indicates the presence of controversy for values greater than one, and the absence of controversy for values less than one.

Confidence of the scores. It can be observed that controversy scores can introduce artifacts in the measurement when the sizes of the three poles are unbalanced. For example, the neutral pole could be very small but have a high coupling with the positive and negative poles. This fact would cause a false negative. On the other hand, the neutral pole could be much larger than the positive and negative poles, but if there is a strong coupling between these last two poles, a controversy score will produce a false positive. To reduce the rate of false positives and false negatives due to size imbalance, we introduce a confidence measure.

Given three poles, we measure the support of each of them. Let N be the number of nodes comprised in the sample, and N_1, N_2, and N_3, the number of nodes at poles 1, 2, and 3, respectively. The support of each pole corresponds to the fraction of N that each pole has given by $s_1 = \frac{N_1}{N}$, $s_2 = \frac{N_2}{N}$, and $s_3 = \frac{N_3}{N}$, respectively. The confidence function will obtain its maximum if the support of the three poles is equal, and therefore $s_1 = s_2 = s_3 = \frac{1}{3}$, and will decrease as the imbalance between the supports increases. To achieve this purpose, we define the confidence of a controversy score using a $\textsc{MultivariateNormal}$ density function with means $\mu_1 = \mu_2 = \mu_3 = \frac{1}{3}$ and $\sigma_1 = \sigma_2 = \sigma_3 = \frac{1}{9}$, so that the $6 \cdot \sigma$ support of the function is in $[0, \frac{2}{3}] \times [0, \frac{2}{3}] \times [0, \frac{2}{3}]$. Covariance 0 is assumed between the variables, producing a symmetric density function around its maximum. Therefore, we define the confidence of the controversy score as $C = \frac{\textsc{MultivariateNormal}(s_1, s_2, s_3)}{\textsc{MultivariateNormal}(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})}$. This function will reach the maximum only if the support of the three poles is identical.

Inference based on dipoles. A scenario that our analysis also contemplates is the absence of a neutral pole. Although it is expected that in many cases, there are healthy conversations that consolidate the existence of a neutral pole, in other cases, the level of polarization could be stronger. Therefore, the relative size of the neutral pole concerning the positive and negative poles could be marginal. In these cases, the measurement of controversy should be redirected to the study of the level of coupling between the positive and negative poles.

If the confidence of a controversy score based on three poles is low, the presence of a strong imbalance in the supports of the three poles is evident. If this is the case, we recalculate the scores of controversy using only the positive and negative poles, that is $\mathrm{RWC} = \mathrm{RWC}_{+,-}$ and $\mathrm{RCC} = \mathrm{RCC}_{+,-}$. To measure the confidence of this new scenario, we use a $\textsc{BivariateNormal}$ density function, that will obtain its maximum if the support of both poles is identical. Therefore, we define the density function by $\mu_1 = \mu_2 = \frac{1}{2}$, and $\sigma_1 = \sigma_2 = \frac{1}{3}$, so that the $6 \cdot \sigma$ support of the function is in $[0, 1] \times [0, 1]$. Covariance 0 is assumed between both supports. Therefore, we redefine the confidence of the controversy score as $C = \frac{\textsc{BivariateNormal}(s_1, s_2)}{\textsc{BivariateNormal}(\frac{1}{2}, \frac{1}{2})}$, where $s_1 = \frac{N_1}{N}$, $s_2 = \frac{N_2}{N}$, and $N = N_1 + N_2$.

5. Experimental results

*5.1. **GENE** training*

GENE training considered several stages. The first was to train the user-entity model. We used the training partition for this purpose. The user-entity model was trained in Keras, on a server with GPUs (GeForce RTX 1070 card). As the news and comments are in Spanish, specific NLP libraries were used to work in this language. For NER, the Spanish library of Spacy was used [4]. For SA, a Spanish sentiment analyzer was used [5]. To train the user-entity model, the sequencer of multiuser contexts was used using windows of size 5 (1 target plus two users on the left and right contexts, according to timestamps) for users with comments consistent with the entity named according to polarity. The sequencer produced a total of 7,764,461 training tuples. One hundred neurons were used in the hidden layer, and the network was fitted using categorical cross-entropy as a loss function and *rmsprop* as an optimizer. The model under this configuration considered a total of 17,847,212 parameters, with 161,391 users in the input layer and 16,912 entities in the output layer. The training was done with 32-size batches, which corresponds to 242,639 steps per epoch. The network was trained in 10 epochs, in a total of 5 hours and 43 minutes.

The second stage of **GENE** training consisted of the generation of graphs from the user-entity model using GraphGen. A dumping factor of 0.85 was used. GraphGen was tested with different number of steps The convergence between the PageRank of the graph generated by GraphGen and the vectors retrieved from the user-entity model was verified. For different vectors of the user-entity model and different numbers of steps, it was found that with 1,000,000 steps, convergence was achieved. Therefore, each graph was generated in 37 minutes, so the 100 graphs took almost 60 hours to compute. Since there is no computation dependence between the graphs, they can be computed in parallel, significantly reducing the computation time involved in this process. For each graph, a relationship was defined, consolidating a multi-relational graph.

PBG [6] was trained using DistMult as a relational operator. Internal units of 200 dimensions were used in PBG. The training considered ten epochs and took less than two hours in total. The parameters of nodes and relationships delivered by PBG were used to recompute the node embeddings using the weights released by GraphGen. Two iterations were conducted to recompute the vectors.

The embeddings were indexed using MRPT [7] using a recall rate of 0.95 for 100 near neighbors. Autotuning was used to optimize the index. The entire process of creating the **GENE** indexes took approximately 40 minutes.

5.2. Controversy detection (ex-post)

The ex-post performance study was conducted on data of each news item, including conversational threads. Both text-based approaches (**BOW** and **BOW-SA**) described in section 4.4 were evaluated using two classic classifiers (**SVM** and **RF**). An architecture based on sequential learning was also evaluated, using a **bi-LSTM** architecture. In this last variant, the text vectors that represent each post were ingested in the machine in the same order in which they were produced in the conversational threads. To study a text representation based on word embeddings, word vectors computed using **BERT** (Devlin et al., 2019) were also considered, which is considered state of the art in word embeddings. Using this representation, all previous architectures were evaluated. In the case of **BERT** with **SVM** and **RF**, we use averaged word embeddings (AWE) to build a single vector for each news item, averaging the word vectors of each conversational thread. In the case of **bi-LSTM**, averaged word embeddings were used at the post level, obtaining one vector per post. Each of these vectors was fed according to the order indicated by the timestamp recorded in the conversation thread. To work with variable length sequences, training sequences were generated using padding, adjusted to the longest sequence of the training set.

[4] https://spacy.io/models/es
[5] https://github.com/aylliote/senti-py
[6] https://github.com/facebookresearch/PyTorch-BigGraph
[7] https://github.com/vioshyvo/mrpt

The graphs conditioned to the named entities in each news of the testing set were constructed using the **GENE** model built on the training set. In the ex-post evaluation, the graph obtained by **GENE** was built using all the users involved in the conversational threads related to each evaluated news. Then, the nodes of each graph were labeled according to polarity following the method described in section 4.4. Then, in each partitioned graph, the **RWC** and **RCC** measures were evaluated to infer the presence of controversy.

To evaluate the performance of the graph-based baselines **Eng-sub**, **Eng-full**, and **UU-graph**, the proximity graph of the intervening users of each news item in the testing set was partitioned using METIS (Karypis & Kumar, 1998). The analysis in these representations was conducted on dipoles, segmenting each graph into two disjoint partitions. For each baseline, the detection of controversy was analyzed using both **RCC** and **RWC** scores.

The results on the testing set of news were evaluated using precision, recall, and F_1-score metrics, disaggregated per class. The measurements obtained globally for the testing partition are also reported, using accuracy, and macro and micro F_1 scores. These results are shown in Table 4.

Table 4: Results obtained using different ex-post methods of controversy detection. The methods were evaluated using different text representations. In the case of graph-based methods using the baselines or **GENE**, the **RWC** and **RCC** measurements were evaluated. Performance measures are disaggregated per class. Global measures are reported, including F_1-score metrics at the micro and macro levels. Controversy and non controversy partitions are reported in columns true and false, respectively. Bold fonts indicate the best results per measure.

	True			**False**			**Global**		
Method	P	R	F_1	P	R	F_1	Acc	$F_1{}^{(mi)}$	$F_1{}^{(ma)}$
BOW + SVM	0,47	0,57	0,51	**0,79**	0,72	0,76	0,67	0,68	0,63
BOW + RF	0,56	0,28	0,37	0,74	**0,90**	**0,82**	0,72	0,68	0,59
BOW-SA + SVM	0,49	0,56	0,52	0,77	0,74	0,75	0,68	0,68	0,63
BOW-SA + RF	0,53	0,34	0,41	0,76	0,87	0,81	0,71	0,69	0,61
BOW + bi-LSTM	0,51	0,43	0,46	0,77	0,82	0,79	0,70	0,69	0,63
BERT + SVM	0,45	0,59	0,51	0,74	0,67	0,70	0,65	0,64	0,61
BERT + RF	0,50	0,33	0,39	0,76	0,68	0,71	0,57	0,62	0,56
BERT + bi-LSTM	0,56	0,62	0,59	0,75	0,69	0,72	0,67	0,68	0,66
Eng-sub + RWC	0,52	0,59	0,55	0,64	0,71	0,67	0,61	0,63	0,61
Eng-sub + RCC	0,52	0,58	0,54	0,62	0,69	0,65	0,59	0,62	0,60
Eng-full + RWC	0,45	0,52	0,48	0,61	0,72	0,66	0,56	0,61	0,57
Eng-full + RCC	0,46	0,51	0,48	0,59	0,69	0,63	0,55	0,58	0,56
UU-graph + RWC	0,63	0,60	0,61	0,60	0,72	0,65	0,62	0,63	0,63
UU-graph + RCC	0,64	0,59	0,61	0,58	0,73	0,64	0,60	0,63	0,62
GENE + RWC	**0,85**	**0,68**	**0,75**	0,65	0,83	0,73	0,74	**0,74**	**0,75**
GENE + RCC	**0,85**	**0,68**	**0,75**	0,65	0,83	0,73	**0,75**	**0,74**	**0,75**

Table 4 corroborates that the problem of detecting controversy is difficult even in ex-post scenarios. All the text-based methods evaluated in the controversial news partition get low precision and recall rates. On the other hand, it can be seen that most of these methods performs well in the non-controversial news class. It is for this reason that while F_1 measures appear to be competitive at the micro-level, they show a different performance at the macro-level. Of the text-based methods, the one that shows the best performance is **bi-LSTM**. Specifically, the representation of text based on **BERT** vectors ingested in a **bi-LSTM** architecture obtains a competitive F_1 of 0,66 at the macro-level, outperforming all other text-based methods. The methods that use **BERT** and construct a vector averaging all the words in each conversation thread show a significant deterioration in performance, evidencing a significant loss of information produced by averaging the word vectors along threads. Therefore, given the nature of the problem addressed, it is clear that the use of sequential learning strategies offers advantages over other text-based classification techniques. This fact confirms that the temporal dimension of the data is relevant to the task of controversy detection and that, therefore, a learning architecture that considers time in the training process can better reflect the nature of the studied phenomena.

The results obtained using the baselines based on user engagement graphs (**Eng-sub** and **Eng-full**) are unsatisfactory. The **Eng-sub** graph performs better than the **Eng-full** graph, illustrating that the use of k-cliques introduces spurious relationships between users, which affects the calculation of the user embeddings obtained from these data. The **UU-graph** gets interesting results, performing well in the true controversy class, at the cost of lower performance in the non-controversy class. Results in the true controversy class rank second in these experiments, both in precision and recall. This fact indicates that the **UU-graph** encodes important information that CBOW uses effectively when generating user embeddings.

The results obtained using **GENE** are satisfactory. Table 4 shows that **GENE** detects controversies more accurately than any of the other methods evaluated, showing a clear advantage in the target class, reaching a precision of 0.85 and recall 0.68 in both scores (**RCC** and **RWC**), accounting the best result in F_1 score (0.75). The cost that **GENE** pays is to show slightly lower performance in the non-controversial news class, with a precision of 0.65. Despite having lower precision, **GENE**'s recall in the non-controversial news class is very high, with a recall rate of 0.83. Consequently, the overall results shown by **GENE** are the strongest of the evaluation, with F_1 at the macro and micro levels of 0.75 and 0.74, respectively. There are no differences between **RCC** and **RWC** in an ex-post scenario, both scores obtaining practically identical results, with only a slight difference in favor of **RCC** in terms of accuracy. That indicates that in news characterization with complete information (that is, after the first 24 hours), both scores measure the same.

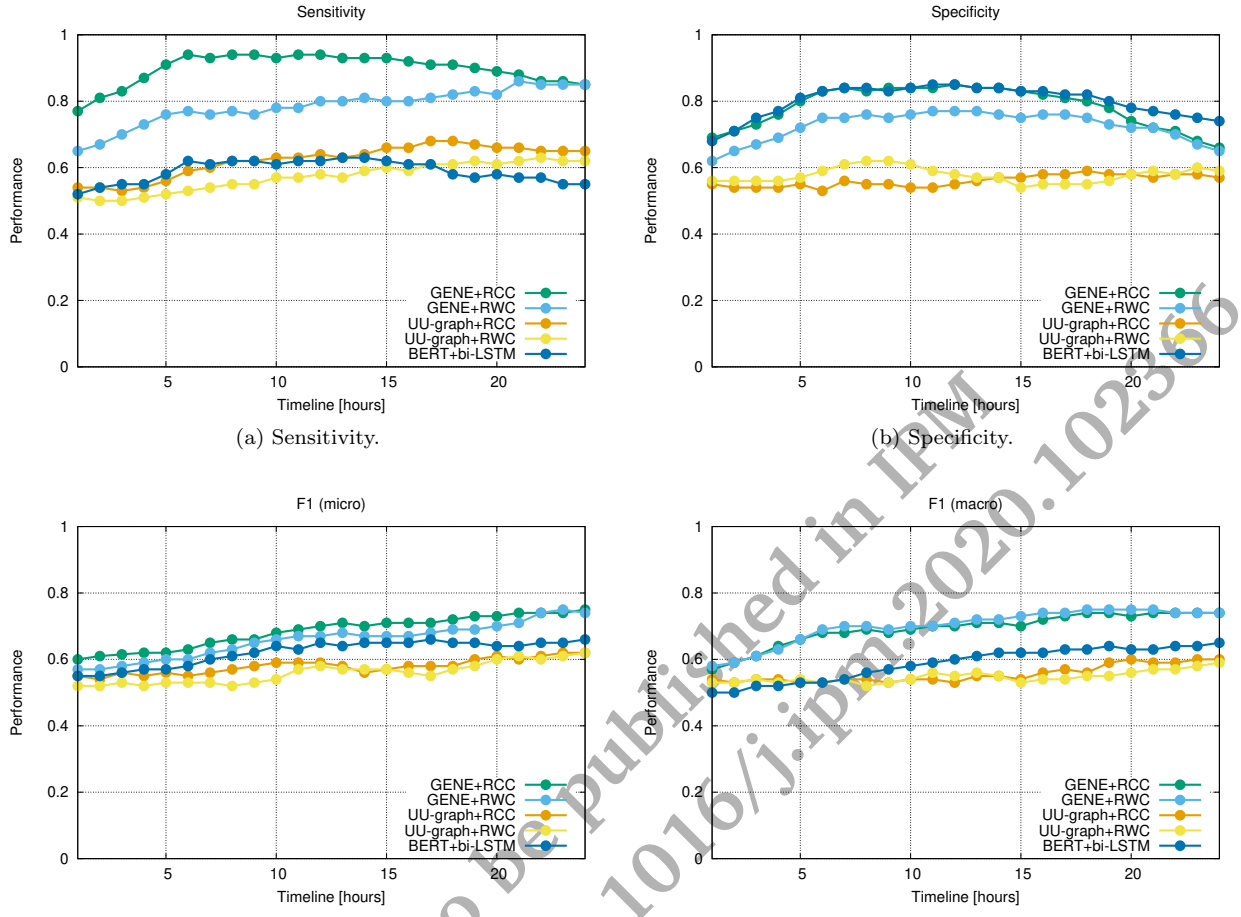
The differences that **GENE** makes compared to the other graph-based methods are important. Its closest competitor, the **UU-graph**, is surpassed by almost 15 points in F1 score in the target class, and by nearly 10 points in F1 in the non-controversy class. These important differences in favor of **GENE** indicate that the use of a representation conditioned to named entities is beneficial in this problem. Since **GENE** uses a graph partitioning strategy that is conditioned on entities and polarization, the partition strategy works at a finer granularity than the **UU-graph**, which is agnostic to entities and polarization. This fact produces a significant difference in favor of **GENE**.

GENE has several components, among them the most crucial are the user-entity model and the Graph-Gen generative method of graphs. We explore the impact of each of these components in a model ablation study to elucidate what effect each of them has on the performance of **GENE**.

5.3. Controversy detection (ex-ante)

For the evaluation in the ex-ante scenario, that is, with partial information of the event, techniques that can make time-sensitive forecasts were evaluated. In this sense, models based on sequential learning, using **BERT** vectors at the post level, on **bi-LSTM** architectures were used. To evaluate the performance of these machines in an ex-ante scenario, the posts of each thread of conversation were ingested for each news of the testing partition in the order indicated by the timestamp, up to a timeout defined as an evaluation parameter. The lags used in this scenario were parameterized on an hourly basis, during the first 24 hours since the publication of the news that triggered each conversation thread. In order to evaluate **GENE**, the same time setting was used. This fact means that for each time window, the representation of the network given only by the intervening users was built. We also included in the ex-ante scenario the evaluation of the UU-graph with graph partition based on METIS, which obtained good results in the target class in the ex-post setting. These results are shown in Figure 7.

Figure 7 shows the results for **GENE+RCC**, **GENE+RWC**, **UU-graph+RCC**, **UU-graph+RWC**, and **BERT+bi-LSTM** in four performance measures: Sensitivity, specificity, and F_1 scores at micro and macro levels. In the case of sensitivity and specificity, these measures correspond to the precision in the controversial and non-controversial news classes, respectively. We can see that the sensitivity of the methods evaluated has disparate behavior over time. While **GENE+RCC** shows the best predictive ability in the target class, **BERT+bi-LSTM** and both variants of **UU-graph** show poor performance. This fact indicates that **GENE** has better predictive capabilities in an ex-ante scenario, and therefore, is better conditioned for early detection of controversy. Another interesting observation that emerges from the sensitivity assessment is the difference in favor of **RCC** over **RWC**. Both scores maintain a gap higher than ten percentage points during the first 18 hours of the event. This gap disappears when full information is available (first 24 hours). This fact is because **RWC** improves its results in direct proportion to the amount



(a) Sensitivity. (b) Specificity.

(c) F_1 (micro). (d) F_1 (macro).

Figure 7: Results obtained in controversy detection in ex-ante scenario. The evaluation considers the first 24 hours after the news were published on the website. Precision results are shown in controversial news (sensitivity) and in non-controversial news (specificity). Results are also shown using F1 scores at micro and macro level.

of information it has. With few users involved, **RWC** shows a deterioration in performance but shows progressive improvements as there are more users involved. On the other hand, **RCC** does not require so many users and shows its best performance during the first hours of the event. That suggests that **RCC** draws on the intervention of users with sturdy engagement to make the forecast, which typically occurs during the first hours of the event. As more users intervene in the conversation, **RCC** begins to lose precision in the target class.

Regarding specificity, **GENE+RCC** and **BERT+bi-LSTM** show the best results during the first 20 hours of the event. While the performance of **GENE+RWC** improves and the results narrow between the three methods, **RCC** maintains the best performance. Around hour 18, the number of users involved in the threads confuses **RCC**, which causes a deterioration in its performance. **RWC** shows progressive improvements in this class, and as it has more data, it improves its performance. In fact, during the last hours of the first day of the event, **RWC** reaches **RCC**. **BERT+bi-LSTM** also benefits from having more data, achieving progressive improvements in specificity as the event passes, and outperforming **RCC** during the last hours of the event. Both variants of **UU-graph** show poor performance.

The results in F_1 scores show that **GENE** is the best method evaluated for early detection of controversy. During the first 20 hours of the event, **GENE** maintains a gap of more than 10 points in F_1 at macro level

over its most direct competitor, **BERT-bi-LSTM**. Both variants of the **UU-graph** have a performance comparable to that of **BERT-bi-LSTM** during the first hours in F1 macro, but they fail to maintain performance when more users are added to the representation.

Our results show that in a polarized scenario, the controversy can be detected early. Both **RWC** and **RCC** show good results in identifying controversy, although **RCC** shows better results in the early stages of an event of interest. This fact is because **RWC** shows a deterioration in its performance when the graphs have few users, something that **RCC** manages to handle well due to the confidence score introduced by us. Both **RWC** and **RCC** are applied to graphs with three or two poles, but our confidence score allows us to focus the analysis on two poles when the relative size of the neutral pole is smaller than that of the positive and negative poles. This fact allows **RCC** to display a false positive rate lower than RWC.

Since **GENE** requires the labeled network, the predictive capacity in this specific task was also evaluated. Then, the polarity imputed using **GENE** was compared with that given by the analysis of sentiment analysis at the post level, which allowed measuring the accuracy of **GENE** in this specific task. Using the same temporal parameterization for the experiment, the temporal accuracy of **GENE** was measured. In this evaluation, the accuracy indicates the fraction of users of the total number of intervening users whose polarity was correctly predicted by **GENE**. Therefore, each accuracy point represents the level of hits per hour, across the news. To indicate the dispersion between news in this predictive task, we will also show the variance around the average accuracy. These results are shown in Figure 8.

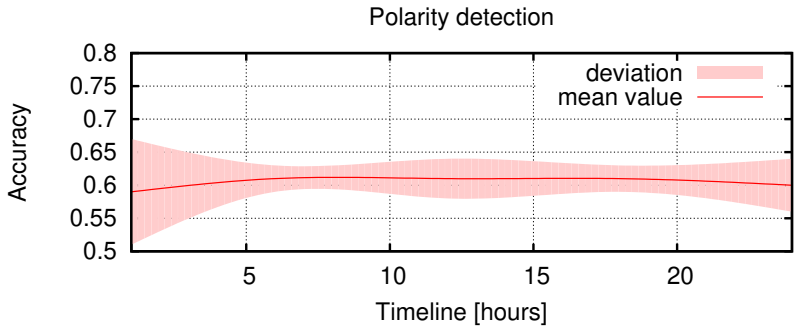
Figure 8a shows the results in terms of accuracy for the polarity detection task in the news testing set, disaggregated during the first 24 hours of the event. The plot shows an accuracy of around 60%, which indicates that **GENE** tagged about 60% of the intervening users with a polarity that matches that of the post sent. It is expected that in this task, which is more complicated and, therefore, less predictable because it operates at a very tight level of granularity (polarity at the user level over three possible classes), the results are lower than those obtained in the detection of controversy. However, surprisingly, the results indicate high predictability of polarity in more than half of the users involved. This fact has two implications. First, given that the scenario for discussing a story is partially predictable, **GENE** can make use of this information to infer the presence of controversy. Another implication is related to how predictable the polarities of user opinions are. Figure 8a shows that many of these comments are predictable in polarity. The best adjustment that **GENE** achieves in this task occurs around the 6^{th} hour of the event, which reinforces the idea that **GENE** exploits the information provided by users with a high level of engagement on the site. The most challenging scenario for **GENE** is shown at the beginning of the event, where the variance gets its maximum value. Figures 8b and 8c show the results disaggregated by class. In the case of controversial news, **GENE** obtains a superior performance than non-controversial news. The dispersion around the average at hour 6 for controversial news is narrow, and this increases as the event passes. The smallest dispersion around the average for non-controversial news occurs around the 18^{th} hour of the event. This observation explains that Figure 8a has two minimums in variance, one around hour 6 and one for hour 18, the first due to controversial news and the second to non-controversial news. The variance in accuracy for polarity is higher in the case of non-controversial news, and its average performance is lower than for controversial news.

5.4. Ablation study

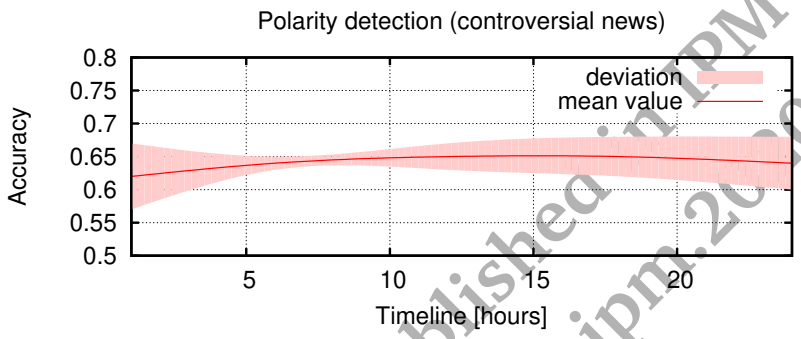
GENE has components whose effect on controversy detection performance must be evaluated. Conducting an ablation study on a component of a pipeline poses challenges since the removal of a module requires its replacement by another that performs a similar function. Since the goal of the study is to determine the impact of a component on the performance of the system, the rest of the system should not be modified, where possible.

In this ablation study, we focus on two components of **GENE**: the user-entity model and the graph generative method. By replacing the user-entity model with another component, we assess the impact of using a conditional representation on named entities. The replacement of the graph-generative model allows us to evaluate the effect that the generation of a multi-relational graph has on the performance of the system.

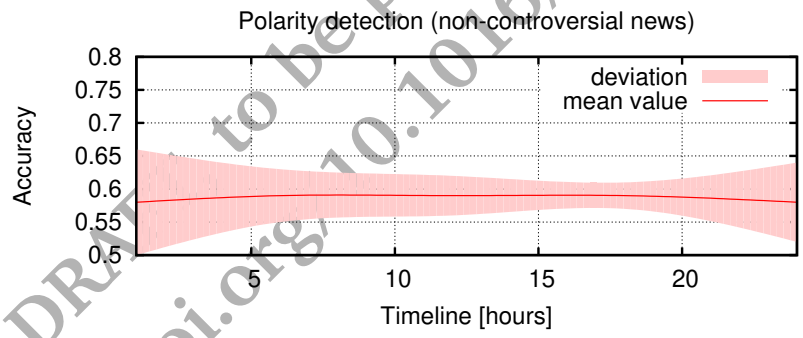
User-entity model replacement: The user-entity model was replaced by the user-user model, introduced in the graph-based baselines section (see Section 4.4). The user-user model was trained using the CBOW



(a) Polarity detection in both classes.



(b) Polarity detection in controversial news.



(c) Polarity detection in non-controversial news.

Figure 8: Results obtained in polarity detection in ex-ante scenario. The evaluation considers the first 24 hours after the news were published on the website. Accuracy results are shown in controversial news and in non-controversial news. The chart at the top shows the results obtained in both classes.

architecture on the news engagement data and allows obtaining user embeddings not conditioned to entities and polarization. The $\mathbf{W}_{in}$ weights of the user-user model were ingested in the **GENE** multi-relational graph generative method. The user embeddings obtained using this method were indexed in MRPT. As there are no entity embeddings, **GENE** can only construct a proximity graph between the intervening users of each news item. This graph was partitioned into two disjoint partitions using METIS. Finally, the presence of controversy was evaluated using the **RCC** and **RWC** scores.

Replacement of the generative method of graphs: The user embeddings of the user-entity model were directly indexed in MRPT, skipping the generative process of multi-relational graphs. For each news item, the entity

embedding of each named entity was retrieved from the user-entity model. Then, each entity was projected on the user network, retrieving its top-K most engaged users. It should be noted that the user-entity model encodes an entity embedding for each polarity. Accordingly, the method explained above allows us to generate a graph for each polarity. Using the set of engaged users and the intervening users in each news item, **GENE** created one graph per polarity. On the union of these three graphs, the analysis was conducted using the **RCC** and **RWC** scores.

The evaluation of these two variants of **GENE** was studied in an ex-post setting. The results are shown in Table 5.

Table 5: Ablation study results in an ex-post experimental setting. The variants of **GENE** with the replacement of the user-entity model and the generative method of graphs are indicated as $\mathbf{GENE}^{\setminus \mathbf{UE}}$ and $\mathbf{GENE}^{\setminus \mathbf{MG}}$, respectively.

	True			**False**			**Global**		
Method	P	R	F_1	P	R	F_1	Acc	$F_1^{(mi)}$	$F_1^{(ma)}$
$\mathbf{GENE}^{\setminus \mathbf{UE}}$**+ RWC**	0,65	0,62	0,63	0,63	0,75	0,68	0,65	0,67	0,66
$\mathbf{GENE}^{\setminus \mathbf{UE}}$**+ RCC**	0,66	0,61	0,63	0,61	0,76	0,67	0,67	0,67	0,66
$\mathbf{GENE}^{\setminus \mathbf{MG}}$**+ RWC**	0,72	0,65	0,68	0,64	0,68	0,66	0,70	0,66	0,67
$\mathbf{GENE}^{\setminus \mathbf{MG}}$**+ RCC**	0,74	0,66	0,69	**0,65**	0,69	0,67	0,71	0,68	0,68
GENE + RWC	**0,85**	**0,68**	**0,75**	**0,65**	**0,83**	**0,73**	0,74	**0,74**	**0,75**
GENE + RCC	**0,85**	**0,68**	**0,75**	**0,65**	**0,83**	**0,73**	**0,75**	**0,74**	**0,75**

The results in Table 5 show that **GENE** outperforms both variants of the method. The underlined results indicate which of these variants ranked second, behind **GENE**. In the target class, the variant $\mathbf{GENE}^{\setminus \mathbf{MG}}$ obtained the best results. This result shows that the substitution of the generative method of graphs does not affect the performance in the target class. In this variant, the user embeddings are generated using the user-entity model. In the controversy analysis phase, since entity embeddings are available, it is possible to create three graphs, one per polarity, which is used effectively by the **RCC** and **RWC** scores. This finding indicates that the performance in the target class is mainly due to the inclusion of entities and polarities. In the non-controversy class, the variant $\mathbf{GENE}^{\setminus \mathbf{UE}}$ performed better. The results of this variant in recall are especially good. This result indicates that replacing the user-entity model with the user-user model does not affect the performance in the non-controversy class. In this variant, the multi-relational graph is generated in the same way as in **GENE**, using the user embeddings of the user-user model. This **GENE** component is the one that explains the performance of the system in the non-controversy class, especially in recall. Since $\mathbf{GENE}^{\setminus \mathbf{UE}}$ does not model entities, the detection of controversy is done by partitioning the graph of intervening users in each news item using METIS. This constraint affects the performance of the model, especially in the target class.

6. Discussion of results

To understand the results obtained by **GENE**, that is, to interpret the factors that condition a case of success and failure, we will analyze some conversational news threads considered in the testing set. Some examples of success and failure are shown in Figure 9.

The news described in Figure 9 shows some of the most clarifying comments for the detection of controversy. The agreement between users is depicted with colors. The entities detected in **GENE** and from which the graph representation is built are indicated in blue color. The news in Figure 9a shows two entities named in the headline: FA, a left-wing political alliance in Chile, and Sebastian Piñera, president of Chile during the period covering the news of the testing set. The news mentions a comment from the leaders of the FA, in which they are disenchanted with the result of their meeting with Piñera. While user 1 sympathizes with the FA, users 2 and 3 argue with user 1. User 4 supports the opinion of user 1. Likes and dislikes show that EMOL users' preferences move from side to side, with a tendency towards rejection of the FA and support for Piñera. As there are comments that support both positions, which evidences

Title: *FA "disappointed" after an appointment with Piñera: "It is very difficult to have a dialogue when there is nothing to propose from the counterpart"*

- User 1: *Easy. Piñera only sees them as the modern young rebels. Nothing to do.* {likes: 8, dislikes: 51}
- User 2: *@User 1 Nothing to do with those who have never done anything and believe they know everything. Young people from high school to the congress.* {likes: 117, dislikes: 4}
- User 3: *@User 2 and earning millions at the expense of those who work.* {likes: 55, dislikes: 1}
- User 4: *@User 2 then you say you have to have a university degree to steal? Because your president is more corrupt than all middle school students together.* {likes: 2, dislikes: 38}

(a)

Title: *Central Station Municipality prohibits putting tents in squares and the Alameda avenue*

- User 1: *It is expected that the measure will be imitated by other communes and that it will be applied to all those who take public spaces as their own.* {likes: 104, dislikes: 7}
- User 2: *What do you suggest for homeless people? , for example, those who live on the train line, on Route 68, in front of the central post, under the Mapocho bridges, etc ... because they cannot evaporate.* {likes: 57, dislikes: 58}
- User 3: *@User 2 release your balcony. Or that they sleep in one of the thousands of properties of the socialist or communist party.* {likes: 79, dislikes: 49}
- User 4: *@User 2 as always our authorities thinking of hiding what bothers and shows the inability to realize solutions to a problem.* {likes: 12, dislikes: 12}

(b)

Title: *Positive perception of the economy falls strongly and only 15% consider it "good"*

- User 1: *They showed the other day, as people earned 800 thousand pesos a day at lunches in a business city ..... and it was not just one, there were several, this by far has been a good year.* {likes: 14, dislikes: 71}
- User 2: *@User 1 They must have been businessmen in that place ... because if they go to Patronato ... they cook behind the bowling alley and don't earn 300 thousand pesos* {likes: 37, dislikes: 14}
- User 3: *@User 2. After having worked for more than 40 years in a good company and with a salary above average, to maintain my standard of living, I must be 9 hours a day behind a counter* {likes: 40, dislikes: 3}
- User 4: *@User 1 Good name for a fake profile! Congratulations.* {likes: 9, dislikes: 1}

(c)

Title: *Frei criticizes detractors of TPP11: "Today everyone speaks what they want and lies shamelessly"*

- User 1: *His government was the worst moment ... he himself is to blame for one of the worst reforms in favor of the AFPs* {likes: 31, dislikes: 8}
- User 2: *@User 1 and Bachelet's ... that the average life of Chileans is 105 years ..... ???* {likes: 15, dislikes: 4}
- User 3: *@User 2 sorry, 110. How the waddle of laughter will squeeze these scoundrels.* {likes: 10, dislikes: 0}
- User 4: *@User 1 This man gave EMOS to the Spaniards and here we are, paying up to 4 times what we paid before!* {likes: 6, dislikes: 0}
- User 5: *TPP11 does not give us any additional commercial advantage over the treaties we already have with these countries.* {likes: 14, dislikes: 13}

(d)

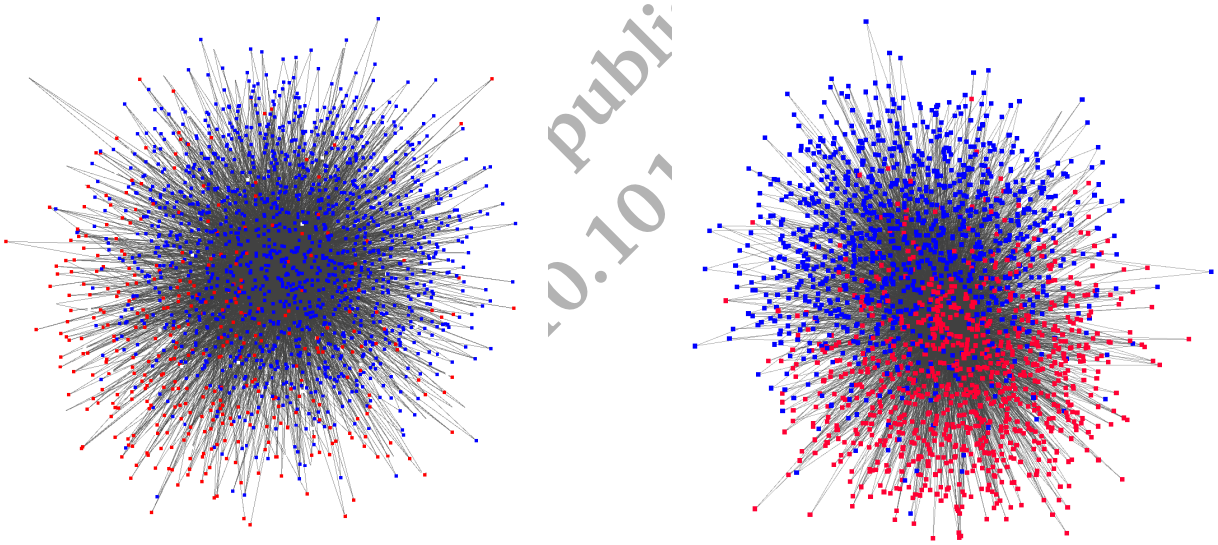
Figure 9: Examples of controversial and non-controversial news from EMOL. Examples a) and b) are controversies while c) and d) are not. Both a) and c) were correctly detected by **GENE**, while b) and d) were erroneously classified. Colors indicate agreement between users. Named entities detected by **GENE** are depicted in blue.

the controversy, both likes and dislikes receive a significant amount of votes. This news is labeled as controversial in our dataset and correctly predicted by **GENE** because the polarities of the users concerning these entities are evident and, therefore, predictable. In the case of the news in Figure 9b, which is also labeled as controversial, **GENE** fails to detect it. This fact is because these named entities produce less polarization in the users, and therefore, the controversy is more difficult to predict. It is understood that in this case, the controversy occurs, preferably by the nature of the message indicated by user 1, which makes users 2 and 4 polemize. In the news indicated in Figure 9c, labeled as non-controversial, **GENE** produces a representation for the entity "the economy", which produces low polarization in people. The comments show user 1 issuing an opinion that generates rejection, and users 2, 3, and 4 polemicizing. Users set a clear trend towards rejection of user comment 1, and the support of user comments 2, 3, and 4. A vision is finally imposed, which indicates the absence of controversy. **GENE** correctly tags the news as non-controversial,

due to the low interaction between poles in a graph that is also poorly polarized. Finally, in the case of the news of Figure 9d, messages coinciding with a posture (rejection of former President Frei) are accounted, so it is labeled as non-controversial. **GENE** tags it as controversial because former President Frei produces an intense polarization in the network. However, in this case, some of the users (for example, user 1) who appear to show a rejection of the TPP-11 and the AFPs and could be classified as a left-wing user, also rejects Frei. Consequently, although users polarize, there is such low support for this political figure that controversy does not occur. **GENE** is confused in this case and tags the news as controversial.

As illustrated in the news in Figure 9, the success of **GENE** depends on the engagement that the news produces in the intervening users. Some users have an evident bias concerning individual political personalities. Other entities produce a less evident bias, and therefore the polarity produced by them is less predictable. A success factor of **GENE** lies in its dependence on users with evident bias since these users are generally the ones who also have the highest engagement when commenting on the news. Indeed, polarization dynamics are dominated by the interventions of users with higher bias. The bias of these users is the reason why they show a tendency to emit more polarized opinions, which in turn produce greater polarization in the network expressed through likes and dislikes. The named entities condition this pattern.

GENE shows that some entities have a favorable scenario for the rise of conflict, while others have a more open scenario in EMOL. To illustrate this phenomenon, we will show the **GENE** graphs for all users involved in EMOL for two very relevant entities in Chile: Sebastián Piñera, current president of Chile, and Michelle Bachelet, the former president. Colored graphs according to positive and negative polarity are shown in Figure 10.



(a) GENE graph for Sebastián Piñera.

(b) GENE graph for Michelle Bachelet.

Figure 10: GENE graphs for Sebastián Piñera and Michelle Bachelet. The users in favor of the entity are colored in blue, while the users against it are colored in red. Users with neutral positions are indicated with white nodes and are the minority in both graphs. The presence of two poles in the case of Michelle Bachelet is evident, conditioning the emergence of controversies in the news that mention it. In the case of Sebastián Piñera, the EMOL community shows a tendency in favor of this person.

The period considered for the construction of **GENE** considered the electoral campaign of Sebastián Piñera of 2017, who finally won the presidential elections. EMOL is a newspaper with a community and right-wing editorial line, which explains the results found. While the scenario for Piñera shows a graph with more users of pro-Piñera trend (marked with blue color), in the case of Bachelet, the graph shows a higher polarization, with many users for and against her. For Michelle Bachelet, **GENE** ranks 42% of users in favor, 55% against her, and only 3% with a neutral position. In the case of Sebastian Piñera, 78% are

in favor, 11% against him, and the remaining 11% with a neutral position. Given these positions, **GENE** is much more likely to detect controversy for these popular entities, since they produce a high level of polarization in the network, with a very low presence of users holding a neutral position. Other less popular entities produce less polarization. In these cases, for **GENE** it is more difficult to detect controversy.

To illustrate the effect of the number of mentions of an entity in **GENE**, we partitioned the testing news set according to the number of mentions in EMOL of its most popular named entity. Then, we built ten folds, at intervals of length 100. In each of these partitions, we evaluated the performance of **GENE** using **RCC** in terms of sensitivity, specificity, and F1 measures at a micro and macro level. For each news in each fold, we build two graphs conditioned on named entities: a graph generated using all the entities named in the news, and a graph generated using the most popular entity mentioned in the news. This allows us to have two performance values per news item. The Figure 11 shows a band around the average performance on each fold.

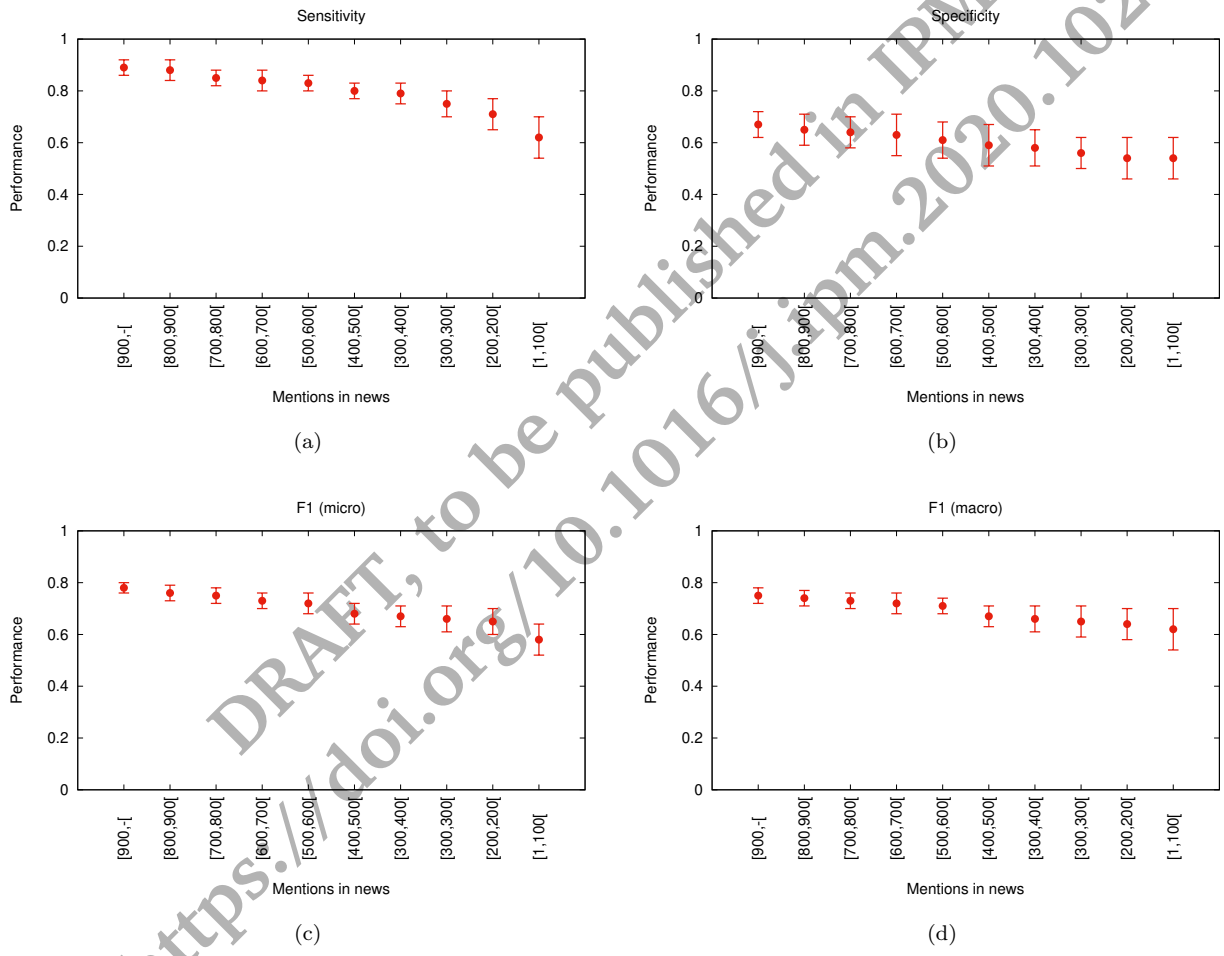


Figure 11: **GENE** performance according to the number of mentions in EMOL of each named entity. The testing dataset was partitioned into ten folds, and in each of them, the performance of **GENE** was evaluated according to (a) Sensitivity, (b) Specificity, (c) F1 (micro), and (d) F1 (macro).

Figure 11(a) shows that **GENE** performs well in the target class when the news includes entities with many mentions in EMOL. As the news includes entities with fewer mentions, the sensitivity of the method declines. **GENE** is less sensitive to this effect in the non-controversial news class, as Figure 11(b) shows. The aggregated effect of both classes, measured in the F1 scores at micro and macro level (see Figures 11(c) and

11(d)), shows that **GENE** exhibits a decrease in its performance when the analyzed news includes entities with fewer mentions in EMOL. Note that the deviation around the mean increases as the performance of **GENE** decreases.

Implications of the analysis. **GENE** has allowed us to show that the dynamics of polarization and controversy are predictable on an online news site. The fact that **GENE** shows so many success instances makes it possible to improve the understanding of the polarization phenomenon in social media. These dynamics are dominated by a few users, with a lot of activity in networks, and with very evident biases. The emergence of controversy is conditioned to the engagement that a news item is able to produce on users with a clear bias. We have shown that user engagements are conditioned to named entities. Some entities have a higher prevalence in controversial news than others. The mentions of these entities in the news capture the attention of users. Once these users intervene, the network polarizes, and the conditions for the controversy are given.

Limitations of GENE. A limitation of **GENE** is related to users outside the network. **GENE** makes use of the user-entity model, which models users considered in the training set. If users not included in the user-entity model are incorporated into the testing set, **GENE** is unable to model them. This limitation forces **GENE** to retrain periodically, keeping the list of users that it is capable of modeling updated. In the case of an online news site, such as modeling in this work, this is not so problematic, since these types of communities have slow growth. In the case of a network with a more dynamic community, **GENE** should be periodically retrained to incorporate new users into the model. Another limitation of **GENE** lies in its inability to model topic drifts. Under certain circumstances, user biases towards entities may change. **GENE** assumes that ideological orientations have little dynamism. This fact is often true, and consequently, it is expected that many users maintain their views over time, and their stance changes are gradual. In these cases, **GENE** will be able to detect these changes. However, **GENE** does not have mechanisms that can effectively model an abrupt change. These changes could be relevant in shock scenarios, and their effects could be higher in young users, who have held ideological positions for less time.

7. Conclusion

We have introduced **GENE**, a model that allows building representations of a network of users of an online news site conditioned on the entities commented by the users. Using these representations, we have studied the problem of controversy detection. Using two indices for detection, we have shown that **GENE** produces a representation of the network of intervening users that facilitates the prediction of a dispute. Notably, **GENE** can detect the emergence of controversy in an ex-ante scenario, showing its best performance during the first hours of the event.

GENE has allowed us to better understand the phenomenon of controversy. The **GENE** results show that the dynamics of controversy are dominated by the intervention of users with a high level of polarization and that these interventions occur during the first hours of the news event. Consequently, the detection of controversy is more predictable at the beginning of an event, and as more users intervene, the task becomes more complicated.

GENE also allows labeling the polarity of user comments concerning the news. The data shows that this task is complicated, but it is also predictable. This finding corroborates that the factor that dominates the dynamics of controversy is user bias. Little deliberation is evident, and instead, what is illustrated in our experiments is the confrontation between previously acquired positions. Given the high predictability of the polarities of the comments concerning the entities named in a story, **GENE** allows us to characterize a discussion scenario before the discussion takes place. **GENE** graphs show that on the EMOL site, some entities have a more favorable scenario for the emergence of controversies.

GENE can be extended in many ways. One of the possibilities is to study the problem of viralization of a discussion, that is, which network users produce guarantees for the emergence of a controversy. Using **GENE**, we could define indices conditioned to entities that measure the impact that the comments of each user have on the network. Another extension is to apply **GENE** to a more open network, such as

Twitter, evaluating the predictive capacity of **GENE** in that scenario. A difficulty for this extension is to have a dataset of labeled controversial news on Twitter since this platform does not have explicit dislikes mechanisms that allow measuring an adverse reaction of the network. Finally, the study of topic drifts is of great interest. It is well known that in a situation of shock, people tend to change their stances. How to model the dynamics of user leanings to named entities is a challenging task, which would require the definition of a dynamic modeling mechanism in **GENE**, which accounts for changes in user stance.

Acknowledgements

Mr. Mendoza, Soto, and Parra acknowledge funding support from the Millennium Institute for Foundational Research on Data. Mr. Mendoza was also funded by ANID PIA/APOYO AFB180002 and ANID FONDECYT grant 1200211. Mr. Parra was also funded by ANID FONDECYT grant 1191791.

References

Adamic, L., & Glance, N. (2005). The political blogosphere and the 2004 u.s. election: Divided they blog. In *3rd International Workshop on Link Discovery, LinkKDD 2005 - in conjunction with 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 36–43).

Akoglu, L. (2014). Quantifying political polarity based on bipartite opinion networks. In *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014* (pp. 2–11).

An, J., Quercia, D., & Crowcroft, J. (2014). Partisan sharing: Facebook evidence and societal consequences. In *COSN 2014 - Proceedings of the 2014 ACM Conference on Online Social Networks* (pp. 13–23).

Bessi, A., Caldarelli, G., Vicario, M., Scala, A., & Quattrociocchi, W. (2014). Social determinants of content selection in the age of (mis)information. In *Lecture notes in Computer Science (including subseries Lecture notes in Artificial Intelligence and Lecture notes in Bioinformatics)* (pp. 259–268). volume 8851.

Calais Guerra, P., Meira Jr., W., Cardie, C., & Kleinberg, R. (2013). A measure of polarization on social media networks based on community boundaries. In *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013* (pp. 215–224).

Calais Guerra, P., Souza, R., Assunção, R., & Meira, W. (2017). Antagonism also flows through retweets: The impact of out-of-context quotes in opinion polarization analysis. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017* (pp. 536–539).

Chen, X., Lijffijt, J., & De Bie, T. (2018). Quantifying and minimizing risk of conflict in social networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1197–1205).

Choi, Y., Jung, Y., & Myaeng, S.-H. (2010). Identifying controversial issues and their sub-topics in news articles. *Lecture notes in Computer Science (including subseries Lecture notes in Artificial Intelligence and Lecture notes in Bioinformatics)*, *6122 LNCS*, 140–153.

Coletto, M., Garimella, K., Gionis, A., & Lucchese, C. (2017). Automatic controversy detection in social media: A content-independent motif-based approach. *Online Social Networks and Media*, *3-4*, 22–31.

Coletto, M., Orlando, S., Lucchese, C., & Perego, R. (2016). Polarized user and topic tracking in twitter. In *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 945–948).

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., & Menczer, F. (2011). Political polarization on twitter. *Proc. 5th Intl. Conference on Weblogs and Social Media*, .

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (pp. 4171–4186).

Flaxman, S., Goel, S., & Rao, J. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, *80*, 298–320.

Garimella, K., De Francisci Morales, G., Gionis, A., & Mathioudakis, M. (2016a). Quantifying controversy in social media. In *WSDM 2016 - Proceedings of the 9th ACM International Conference on Web Search and Data Mining* (pp. 33–42).

Garimella, K., De Francisci Morales, G., Gionis, A., & Mathioudakis, M. (2018). Reducing controversy by connecting opposing views. In *IJCAI International Joint Conference on Artificial Intelligence* (pp. 5249–5253).

Garimella, K., Gionis, A., De Francisci Morales, G., & Mathioudakis, M. (2017). The effect of collective attention on controversial debates on social media. In *WebSci 2017 - Proceedings of the 2017 ACM Web Science Conference* (pp. 43–52).

Garimella, K., Mathioudakis, M., De Francisci Morales, G., & Gionis, A. (2016b). Exploring controversy in twitter. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* (pp. 33–36).

Garimella, V., & Weber, I. (2017). A long-term analysis of polarization on twitter. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017* (pp. 528–531).

Gaumont, N., Panahi, M., & Chavalarias, D. (2018). Reconstruction of the socio-semantic dynamics of political activist twitter networks—method and application to the 2017 french presidential election. *PLoS ONE*, *13*.

Graells-Garrido, E., Lalmas, M., & Quercia, D. (2013). Data portraits: connecting people of opposing views. *Data Portraits: Connecting People of Opposing Views*, .

Grevet, C., Terveen, L., & Gilbert, E. (2014). Managing political differences in social media. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* (pp. 1400–1408).

Guimaraes, A., Wang, L., & Weikum, G. (2017). Us and them: Adversarial politics on twitter. In *IEEE International Conference on Data Mining Workshops, ICDMW* (pp. 872–877).

Hamad, M., Skowron, M., & Schedl, M. (2018). Regressing controversy of music artists from microblogs. In *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI* (pp. 548–555).

Han, L., Han, L., Darney, B., & Rodriguez, M. (2017). Tweeting pp: an analysis of the 2015–2016 planned parenthood controversy on twitter. *Contraception*, *96*, 388–394.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics*, *13*, 70–84.

Hyvönen, V., Pitkänen, T., Tasoulis, S. K., Jaasaari, E., Tuomainen, R., Wang, L., Corander, J., & Roos, T. (2016). Fast nearest neighbor search through sparse random projections and voting. In *Proceedings of the 3rd IEEE International Conference on Big Data (BigData'16)* (pp. 881–888). IEEE.

Jang, M., Dori-Hacohen, S., & Allan, J. (2017). Modeling controversy within populations. In *ICTIR 2017 - Proceedings of the 2017 ACM SIGIR International Conference on the Theory of Information Retrieval* (pp. 141–148).

Kane, B., & Luo, J. (2019). Do the communities we choose shape our political beliefs? a study of the politicization of topics in online social groups. In *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018* (pp. 3665–3671).

Kaplun, K., Leberknight, C., & Feldman, A. (2018). Controversy and sentiment: An exploratory study. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*.

Karypis, G., & Kumar, V. (1998). Multilevel k-way partitioning scheme for irregular graphs. *Journal of Parallel Distributed Computing*, *48*, 96–129.

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M., Ghosh, S., Gummadi, K., & Karahalios, K. (2019). Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal*, *22*, 188–227.

Lahoti, P., Garimella, K., & Gionis, A. (2018). Joint non-negative matrix factorization for learning ideological leaning on twitter. In *WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining* (pp. 351–359).

Lasalle, D., & Karypis, G. (2013). Multi-threaded graph partitioning. In *27th IEEE International Symposium on Parallel and Distributed Processing, IPDPS* (pp. 225–236).

Lerer, A., Wu, L., Shen, J., Lacroix, T., Wehrstedt, L., Bose, A., & Peysakhovich, A. (2019). PyTorch-BigGraph: A Large-scale graph embedding system. In *Proceedings of the 2nd Conference on Systems and Machine Learning (SysML'19)*.

Matakos, A., & Gionis, A. (2018). Tell me something my friends do not know: Diversity maximization in social networks. In *Proceedings - IEEE International Conference on Data Mining, ICDM* (pp. 327–336).

Matakos, A., Terzi, E., & Tsaparas, P. (2017). Measuring and moderating opinion polarization in social networks. *Data Mining and Knowledge Discovery*, *31*, 1480–1505.

Maurus, S., & Plant, C. (2016). Skinny-dip: Clustering in a sea of noise. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1055–1064).

Mejova, Y., Zhang, A., Diakopoulos, N., & Castillo, C. (2014). Controversy and sentiment in online news. In *CJ'14: Computation+Journalism Symposium*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems* (pp. 3111–3119).

Morales, A., Borondo, J., Losada, J., & Benito, R. (2015). Measuring political polarization: Twitter shows the two sides of venezuela. *Chaos*, *25*.

Munson, S., Lee, S., & Resnick, P. (2013). Encouraging reading of diverse political viewpoints with a browser widget. In *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013* (pp. 419–428).

Napoles, C., Pappu, A., & Tetreault, J. (2017). Automatically identifying good conversations online (yes, they do exist!). In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017* (pp. 628–631).

Nelimarkka, M., Laaksonen, S.-M., & Semaan, B. (2018). Social media is polarized, social media is polarized: Towards a new design agenda for mitigating polarization. In *DIS 2018 - Proceedings of the 2018 Designing Interactive Systems Conference* (pp. 957–970).

Nelimarkka, M., Rancy, J., Grygiel, J., & Semaan, B. (2019). (re)design to mitigate political polarization: Reflecting habermas' ideal communication space in the united states of america and finland. *Proceedings of the ACM on Human-Computer Interaction*, *3*.

Popescu, A., & Pennacchiotti, M. (2010). Detecting controversial events from twitter. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM* (pp. 1873–1876).

Quraishi, M., Fafalios, P., & Herder, E. (2018). Viewpoint discovery and understanding in social networks. In *WebSci 2018 - Proceedings of the 10th ACM Conference on Web Science* (pp. 47–56).

Ruan, Y., Fuhry, D., & Parthasarathy, S. (2013). Efficient community detection in large networks using content and links. In *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web* (pp. 1089–1098).

Rumshisky, A., Gronas, M., Potash, P., Dubov, M., Romanov, A., Kulshreshtha, S., & Gribov, A. (2017). Combining network and language indicators for tracking conflict intensity. *Lecture notes in Computer Science (including subseries Lecture notes in Artificial Intelligence and Lecture notes in Bioinformatics)*, *10540 LNCS*, 391–404.

Timmermans, B., Kuhn, T., Beelen, K., & Aroyo, L. (2017). Computational controversy. *Lecture notes in Computer Science (including subseries Lecture notes in Artificial Intelligence and Lecture notes in Bioinformatics)*, *10540 LNCS*, 288–300.

Warmsley, D., Xu, J., & Lu, T.-C. (2019). From gamergate to fifa: Identifying polarized groups in online social media. In *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018* (pp. 3991–3995).

Yang, B., Yih, W., He, X., Gao, J., & Deng, L. (2015). Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Zachary, W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, *33*, 452–473.