Stance polarity in political debates: A diachronic perspective of network homophily and conversations on Twitter

STANCE POLARITY IN POLITICAL DEBATES: A DIACHRONIC PERSPECTIVE OF NETWORK HOMOPHILY AND CONVERSATIONS ON TWITTER

Mirko Lai
Dipartimento di Informatica
Università degli Studi di Torino
Torino, Italy
lai@di.unito.it

Marcella Tambuscio
Dipartimento di Informatica
Università degli Studi di Torino
Torino, Italy

Viviana Patti
Dipartimento di Informatica
Università degli Studi di Torino
Torino, Italy

Giancarlo Ruffo
Dipartimento di Informatica
Università degli Studi di Torino
Torino, Italy

Paolo Rosso
PRHLT Research Center
Universitat Politècnica de València
València, Spain

November, 2019 - DOI:10.1016/j.datak.2019.101738

ABSTRACT

In the last decade, social media gained a very significant role in public debates, and despite the many intrinsic difficulties of analyzing data streaming from on-line platforms that are poisoned by bots, trolls, and low-quality information, it is undeniable that such data can still be used to test the public opinion and overall mood and to investigate how individuals communicate with each other. With the aim of analyzing the debate in Twitter on the 2016 referendum on the reform of the Italian Constitution, we created an Italian annotated corpus for stance detection for automatically estimating the stance of a relevant number of users. We take into account a diachronic perspective to shed lights on users' opinion dynamics. Furthermore, different types of social network communities, based on friendships, retweets, quotes, and replies were investigated, in order to analyze the communication among users with similar and divergent viewpoints. We observe particular aspects of users' behaviour. First, our analysis suggests that users tend to be less explicit in expressing their stances after the outcome of the vote; simultaneously, users who exhibit a high number of cross-stance relations tend to become less polarized or to adopt a more neutral style in the following phase of the debate. Second, despite social media networks are generally aggregated in homogeneous communities, we highlight that the structure of the network can strongly change when different types of social relations are considered. In particular, networks defined by means of reply-to messages exhibit inverse homophily by stance, and users use more often replies for expressing diverging opinions, instead of other forms of communication. Interestingly, we also observe that the political polarization increases forthcoming the election and decreases after the election day.

Keywords: Stance, Political Debates, Homophily, Twitter

1 Introduction

Social media have changed the information consumption and diffusion behavior, as in [32, 54], gaining a crucial role in the public debate about socio-political issues in both directions, as in [62]: from institutions and politicians to citizens (*top-down*), and conversely (*bottom-up*). Indeed, some political leaders make an extensive use of platforms like Twitter or Facebook to communicate with citizens, e.g., in [19, 15, 11], that, on the other hand, join in online discussions supporting or criticizing their political opinions. In this framework, social media provide a powerful

tool to test the public opinion mood and investigate how individuals are exposed to diverse viewpoints. Developing automated systems for a deep analysis of users' generated contents and interactions is becoming increasingly relevant, and recent works focused on detecting users' opinion towards a particular target, e.g., in [3, 40, 43, 64].

Recent studies suggest that web users tend to polarize their opinion and form partisan political communities, e.g., in [1, 13], following the *homophily* principle in [37, 41], according to which like-minded people are more likely to connect to each other. Despite the scientific debate is still open about the role of the architecture of these platforms in the formation of social groups, as in [7, 53], many scientists suggest that the presence of *echo chambers* (i.e., when users are exposed only to information from like-minded ones) and *filter bubbles* (i.e., when content is selected by algorithms according to the user's previous behaviors) reinforce people's pre-existing beliefs, filtering and censoring divergent viewpoints, as in [17, 65]. Moreover, Sunstein in [61] suggests that two persons, who only slightly disagree with each other, are likely to be even more opposed, after they have talked to each other, while democracies should be based on a conciliation among viewpoints.

In our previous studies, we analyzed two political debates on Twitter focusing on two events that have been considered as symptoms of a new nationalist populism in [26]: Trump's election and Brexit. In our fist analysis, we only focused on linguistic aspects with the aim to determine from the text of a tweet whether the author is in favor of the given target, i.e., Hillary Clinton, Donald Trump in [34]. Analyzing the political debate around the so called Brexit referendum in [36], we inspected the debate at user level by aggregating, in a diachronic perspective, the tweets posted by the same user in a single day. Information of existing follower/followee relations were analyzed and used to create network-based features that improved stance detection performance.

In this study we examine the political debate on Twitter about the Italian constitutional referendum held on December 4th, 2016 in Italy, adopting the machine learning model we obtained previously in the same scenario in [35]. Some political analysts tried to explain the result of the Italian constitutional referendum, similar to Brexit and Trump's election results, as a reaction of a sense of disbelief against the elites. The international pressure on this case arose when many influential actors expressed their support for (i.e., JP Morgan, Fitch, the Financial Times, the Wall Street Journal, and the 44th USA President Barack Obama etc.) or criticism of (i.e. the Economist) the constitutional reform as a way to stop the spreading of populism in western democracies. Referring to a national report about the Italian social situation in 2016 in [5, 4], a not negligible portion of Italian people use Facebook and get news from them: 38% of interviewed people declare to use Facebook as their main source of information, as opposed to the 60.6% and 22.4% of Italians that use TV and Radio news outlets. Although Twitter is not the dominant source of information for Italians, there is still a 4.8% of citizens that declare to use Twitter (10.6% among the young population) for such a purpose: apparently, studying opinion dynamics in data streaming out from such a platform is likely to return a signal of what is going on in the more general public debate.We are aware that a percentage of approximately 5% of citizens is not representative of the whole population's stances; in fact, we want to remark that our study is limited to how Italian users relate to Twitter when they expose their stances toward a given topic, and that these results are not necessarily generalizable to other communication networks commonly used. Our aim is to show that such techniques can at least complement traditional political polls, that are more accurate in terms of the selection of socio-demographic features of the sample, but that relies on a much smaller set of citizens that are directly interviewed by poll agencies. For example, it has been reported (source: `https://www.ilpost.it/2018/01/21/sondaggi-elezioni-2018/`) that during 2018 political elections in Italy, a number of telephone interviews varying from 259 to 850 has been placed to collect statistical data.

In order to investigate the online debate we used the dataset and the CONREF-STANCE-ITA annotated corpus described in [35]. Starting from these data, we already monitored, during our previous work, the interactions with other Twitter users (through followees, retweets, quotes, and reply-to information), providing a varied representation of the political debate on Twitter. We divided our dataset of tweets in four discrete temporal phases delimited by significant events occurred before the consultation period and immediately after election results were made publicly available, in order to analyze the dynamics of both users' stance (opinion towards the referendum) and social relations. In [35], we manually annotated the evolution of the stance for 242 users, creating a corpus for *Stance Detection* (henceforth SD), i.e. the task of automatically determining whether the author of a text is in favor, against, or neutral towards a given fact or target, as in [43]. As one of the main results of our previous work, an automatic classifier was trained on that corpus to automatically detect stance of other users not considered during learning.

The machine learning model presented in our previous work is capable to predict a huge number of users' stances with an accuracy comparable to the humans that participated to the manual annotation phase. Therefore, we can count on a set of tools that allows us to analyze the debate about this particular political event considering both users' stance and social media relations in a diachronic perspective. We observe social network communities focusing on opinion shifting and on the dynamics of the graph. We also investigate how the most commonly used types of relations between two users correlates with users' stance. The major findings of this work can be summarized as follows:

1. Constructing networks based on different types of communication interactions results in different network structures.
2. Users having different opinions are more likely to communicate by means of reply-to instead of retweets or quotes; in fact, the reply-to networks exhibit inverse homophily by stance.
3. Polarization by stance increases forthcoming the election and decreases immediately after the election day.
4. Users who exhibit a higher number of cross-stance relations in a given temporal phase of the debate tend to adopt a less polarized stance in the following phase.

The above mentioned contributions are in addition to the previous results we obtained in our study published in the Proceedings of NLDB 2018 in [35], that actually presents our machine learning model that classifies users' stance in this domain. In the study presented in this paper, the model has been extensively applied to automatically annotate previously unseen sequences of tweets; this allowed us to put new lights on the interplay between users' homophily, their stance dynamics, and their communication behavior.

The rest of the paper is structured as follows. We provide an overview of related work in Section 2, discussing some of the main aspects about stance detection and political debates in social media. Then, in Section 3 we describe our datasets and how we collected and pre-processed information from Twitter; after that, we focus on the creation of the CONREF-STANCE-ITA corpus in Section 4. Every data in the CONREF-STANCE-ITA corpus is then represented in terms of features vectors and used to train a supervised classification model of user's stance (Section 5). In addition to learn a stance detection model, we used collected and annotated data also to define four different types of networks (Section 6), that provide a useful tool to analyze the relationships between users exposing the same or a different stance. In fact, as described in Section 7, we used our machine learning model to automatically classify a larger set of user's stances to better understand the relationship between users' homophily and communication behavior in a temporal dimension. Conclusions and hints to future directions will be given at the end in Section 8.

2 Related Work

2.1 Political sentiment and stance detection

Techniques such as sentiment analysis (SA) and opinion mining are usually exploited to monitor people's mood, extracting information from users' generated contents in social media, as in [52, 48, 47, 58, 57, 56, 39, 6]. Other works focused on detecting users stance towards a particular target adopting several approaches, as in [3, 40, 72].

To our knowledge, the first shared task on SD in Twitter was held at SemEval-2016, i.e., Task 6 in [43], that is described as follows: "Given a tweet text and a target entity (person, organization, movement, policy, etc.), automatic natural language systems must determine whether the tweeter is in favor of the target, against the given target, or whether neither inference is likely". The chosen tasks were six commonly known topics in the United States, such as feminist movement or climate change. The majority of the teams that participated to the task exploited standard text classification features such as n-grams and word embedding vectors. Sentiment resources, such as EmoLex in [46], MPQA in [67], Hu&Liu in [29], and NRC Hashtag in [44], have also been used. The best result was obtained by a deep learning approach based on a recurrent neural network trained with embeddings of words and phrases initialized with the word2vec skip-gram method in [71].

Machine learning algorithms and deep learning approaches also appeared in a second shared task on gender and stance detection in Twitter held at IberEval-2017 in [64], where the given dataset referred to the political debate about "Independence of Catalonia" during the Catalan regional election that was held on September 2015. With regard to SD, participating teams exploited different kinds of features such as bag of words, bag of parts-of-speech, n-grams, word length, number of words, number of hashtags, number of words starting with capital letters, and so on. In this case, the best result was obtained by an SVM classifier exploiting three groups of features: *Stylistic* (bag of: n-grams, char-grams, part-of-speech labels, and lemmas), *Structural* (hashtags, mentions, uppercase characters, punctuation marks, and the length of the tweet), and *Contextual* (the language of each tweet and information coming from the URL in each tweet) ([33]).

2.2 Political debate on social media

Social media enabled researchers to investigate social networks analyzing new forms of interpersonal relations. Furthermore, the huge amount of user generated content allows to observe more easily social phenomena in a wide variety of disciplines compared with traditional survey data, as in [23, 24, 38, 66].

Despite social media potentially expose users to a larger range of different views, some studies suggested that the existence of echo chambers and filter bubbles mechanisms can have both positive and negative effects in on-line and off-line forms of political participation, as in [17, 65]. Sustein in [61] also discussed the phenomenon of group polarization drawing the attention of its implications for law and political theory. He suggests that two individuals, who only slightly disagree to each other, will tend to be even more opposed, after they have talked to each other. This phenomenon could also explain the emergence of extreme and radical tendencies in social media communities.

Lazarsfeld and Merto in [37] considered the role of homophily, observing that people tend to be connected with persons who have similar opinions, regardless of any differences in their status characteristics (i.e. gender, age, social status). For instance, Adamic et al. in [1] observed that blogs preferentially link to other blogs of the same political ideology. In the Twitter platform, there are evidences that users tend to retweet posts supporting the same political orientations, as in [13]. Analyzing an independent information and communication platform for Swiss politics in ([20], it has been possible to measure network polarization among politicians exploring the relation between ideology and social structures in on-line interactions.

The relation between social media network structure and sentiment information extracted from posted contents has been explored consistently. Xu et al. in [69] introduced the concept of *sentiment community*, trying to maximize both the intra-connections among nodes and the sentiment polarities using movies' ratings collected from Flixster. Deitrick et al. in [14] combined sentiment analysis and community detection techniques by using Twitter's relations among users and sentiment classification of tweets. This ensures to perform community detection iteratively considering edge weights in a social network based on friend relations. Some preliminary results in [36] give a signal that a strong relation exists between user's stance and social media community the user belongs to. All these studies suggest that a strong correlation exists between the existence of internal connections within social communities and agreement on stance between individuals, when they are involved in polarized debates. It is likely that such signal can be exploited to detect stances more accurately.

2.3 Political debate in a diachronic perspective

Several works proposed methods to analyze the temporal dimension and to model the behavior of dynamical systems for prediction purposes, as in [8, 28]. Focusing on social relations, a way to represent a dynamic system is aggregating empirical data over time considering different granularities such as one day (in [2]), one week (in [49]), one month (in [59]), and several months in ([25]). Albeit the choice of the time window size is often dictated by the availability of data and this issue is often neglected in the literature, in [31] has been observed that the structural features of networks change considering different time intervals and therefore size of time-windows should be selected wisely.

Time evolution of polarization has been quantitatively analyzed in [20] on a daily basis showing that the polarization tends to increase significantly during election campaigns compared to other period. Recently, in [36] we also explored the time evolution of the stance toward BREXIT at user level by aggregating tweets posted by the same user over 24 hours time windows. Quite interestingly, stance may change after relevant events, a finding supported also by other researches in [42, 70].

3 Data Collection and Preprocessing

3.1 Data Collection

We collected a corpus of tweets about the Referendum held in Italy on December 4th, 2016, where citizens were asked to express their opinion towards a reform of the Italian Constitution. They could have voted *yes* if they wanted to approve the reform or *not* if they did not: 59.11% of voters rejected the reform (40.88% voted to approve it), causing the resignation of Matteo Renzi, the Prime Minister that proposed the reform. The data collection consisted of four steps:

1. About 900K tweets were collected between November 24th and December 7th through the Twitter's Stream API, using as keywords the following hastags: #referendumcostituzionale, #iovotosi, #iovotono[1]. From this set of seeds, we executed a one-step snowball sampling using retweets, quotes, replies, and followers. Steps 2-5 below describe what we have done for each type of observed interaction between users.
2. The original tweet from each retweet was recovered by identifying the tweet embedded within the JSON field *retweeted_status*. Then, we used the *statuses/retweets/:id* Twitter REST API to collect all the retweets for each original tweet. We used this data to create the retweets network described in Section 6.2.

[1]Translatable as #constitutionalreferendum, #Ivoteyes, #Ivoteno

3. We recovered the quoted tweet from each quote identifying the tweet embedded within the JSON field *quoted_status*. We used this data to create the quotes network described in Section 6.3.
4. We recovered complete conversations, as in [27], using recursively the Twitter REST API *statuses/show/:id* passing, as argument, the id specified in the field *in_reply_to_status_id* of each replied tweet.
5. We recovered the friend (i.e., followers and followees) list for each user using the *friends/ids* Twitter REST API. We used this data to create the quotes network described in Section 6.4.

Following these steps, we obtained a larger collection of tweets (more than 2M) where different types of interactions among users (friends, retweets, quotes, and replies) are considered.

The selection of the three seeding hashtags (i.e., #referendumcostituzionale, #iovotosi, #iovotono) is somehow arbitrary; nevertheless, we wanted to get the three most representative hashtags of a negative bias toward the vote (#iovotono), of a positive one (#iovotosi), and a more general hashtag describing the topic (#referendumcostituzionale). To detect which hashtag to use to such a purpose, we used the *Twita* corpus[2], in [10], [9], made of significant samples of italian tweets; focusing on two weeks before our observation period, we found the hashtags ranking that is displayed in Table 1.

[Table 1 about here.]

Observe that some other apparently related hashtags (e.g., #bastaunsi) were discarded because used mainly for a broader political discussion unrelated to the upcoming referendum. As a consequence, the first 900K tweets dataset was collected by means of the Twitter Stream API between November 24th and December 7th with #referendumcostituzionale, #iovotosi, and #iovotono hashtags as filters.

3.2 Diachronic Perspective

Our objective is to focus on the dynamics of users stance, looking also at the evolution of the social network topology while the discussions and the controversies between users take place. To extract information relevant to build significant graphs from tweets expressing different opinions regarding the referendum, we applied a methodology that we already adopted for the Brexit referendum in [36].

First of all, we identified four different events that, to our understanding, capitalized media attention during the monitoring period, and that also corresponded to spikes in the amount of generated tweets.

Then we defined four temporal phases, each delimited by an event and the following one. The range for three temporal phases out of four is equal to 3 days (72 hours). In the remaining temporal phase, the range is equal to 4 days. Nevertheless, we decided to define a fixed range of 3 days following every event to deal with comparable time spans, leaving out from our analysis all the tweets collected on Dec. 3th, that incidentally is the day before referendum day, when media are asked to observe election silence.

[Figure 1 about here.]

The four 72-hours temporal phases, and the tweets that have been generated by users during these periods, are displayed in Figure 1, and defined in terms of the following events:

- "The Economist" (EC): The newspaper *The Economist*[3] sided with the "no" campaign of the referendum (tweets retrieved between 2016-11-24 00:00 and 2016-11-26 23:59).
- "Demonstration" (DE): A demonstration[4] supporting the "no" campaign had been held in Rome exactly one week before the referendum (tweets retrieved between 2016-11-27 00:00 and 2016-11-29 23:59).
- "TV debates" (TD): The Italian Prime Minister, Matteo Renzi, who supported the "yes" campaign of the referendum, participated in two influential debates on TV (tweets retrieved between 2016-11-30 00:00 and 2016-12-02 23:59).
- "Referendum outcome" (RO): The phase includes the formalization of the referendum outcome, and the resignation of the Italian Prime Minister (tweets retrieved between 2016-12-4 00:00 and 2016-12-6 23:59).

[2]Download the corpus from here `http://valeriobasile.github.io/twita/about.html` or contact the author.

[3]"Why Italy should vote no in its referendum", `https://www.economist.com/leaders/2016/11/26/why-italy-should-vote-no-in-its-referendum`; printed article published: Nov. 24th, 2016.

[4]`http://www.ansa.it/sito/notizie/speciali/referendum/2016/11/26/referendum-domenica-movimenti-in-corteo-per-il-r7bb4b689-69c9-404e-b76c-f2612a79822c.html`

The four temporal phases consist in about 1M tweets posted by more than 100K users. The average number of tweets in each phase is about 250K, with a minimum number of tweets posted during EC phase (171,476) and a maximum posted during RO phase (324,464).

Observe that, as stated above, we left the tweets collected on Dec. 3th out of our analysis. We made this decision to define temporal phases with a fixed 72 hours range; moreover, we wanted to have as many tweets triplets as possible (see Section 4.1), and other choices of 24-hours and 48-hours ranges would have returned too few data to work on. We had some reluctance because Dec. 3th is just one day before the referendum day, and some discussions during election silence could have been of interest. However, we have to remind here that our main objective is to understand the dynamics of users' stance, also with the purpose to detect opinion shifts before and after some given events. Leaving the referendum day out gives us more data to detect stances both before and after the main event: for the purpose of this study, this is likely the best decision to take.

4 A New Corpus for Stance Detection

As mentioned before, we are interested in studying stance evolution; consequently, we enriched our corpus with an annotation schema previously defined in [45] that considers three labels: FAVOR, AGAINST, NONE. In order to engage correctly human annotators, we used the following annotation guidelines:

From reading the following tweets, which of the options below is most likely to be true about the tweeter's stance or outlook towards the reform subjected to the Italian Constitutional referendum?

1. **AGAINST**: We can infer from the tweet that the tweeter is against the reform.
2. **FAVOR**: We can infer from the tweet that the tweeter supports the reform.
3. **NONE**: We can infer from the tweet that the tweeter has a neutral stance towards the reform or there is no clue in the tweet to reveal the stance of the tweeter towards the reform.

Although some scholars can express some concerns on the NONE label that can be used to classify a clearly neutral as well as an unintelligible stance, we remind that much discussion has been carried on during the last years over this issue among scholars working on sentiment analysis and stance detection. For any other details and motivation about the schema that we adopted, we refer to [45].

4.1 Stance at user level

In line with our previous work in [36], we consider the stance at user level rather than the stance at tweet level, meaning that we infer the stance from multiple texts written by the same user rather than inferring the stance of a single anonymous text. Therefore we define a *triplet* as a set of three tweets written by the same user in a single temporal phase: a tweet, a retweet and a reply (see Table 8).

As we already said in Section 3.2, we could have done other choices regarding the time span of each temporal phase. We opted for 72-hours ranges instead of 24 or 48 because we wanted to maximize the number of triplets and users involved in our analysis, to investigate stance evolution and dynamics. In fact, with 24-hours ranges no users satisfy the above mentioned constraints; within 48-hours, 127 users satisfy the requirements; finally, for each of our four *72-hours temporal phases*, we found that *242 users* wrote twitter triplets, for a total of *968 triplets* to be analyzed. This is our "Users Sample", at the core of our annotated corpus.

[Table 2 about here.]

4.2 Manual annotation

Two native speakers provided two independent annotations for all the 968 triplets. When they did not agree, we used CrowdFlower [68], a crowd-sourcing resource that allows researchers to create HIT (Human Intelligent Tasks) to be assigned to human annotators. We considered 100 tweets as test question in order to evaluate the CrowdFlower annotators. We required that annotators were native Italian speakers living in Italy. The annotators have been evaluated over the test questions and only if their precision was above 80% they were included in the task. A further annotator

was required unless at least 60% of the previous annotators agreed on the stance of a given triplet. We required a maximum of 3 additional annotators, in addition to the 2 native speakers we contacted directly, to assign a label to ambiguous triplets. Therefore, each triplet was annotated by at least 2 annotators to a maximum of 5.

4.3 Agreement

We discharged triplets annotated by 5 annotators having less than 3 annotators in agreement on the same label. Luckily enough, only 5 triplets were discarded: the final gold standard we produced on Italian tweets related to the constitutional referendum has been called CONREF-STANCE-ITA and it consists of 963 triplets in total, each annotated with a label in (FAVOR, AGAINST, NONE) identifying users' stance. Inter-annotation agreement has been calculated as follows:

$$IAA = \frac{A_{agree}}{A_{tot}}$$

where IAA is the number of pairs of annotators who agree labeling the same triplet A_{agree} over the total possible number of pairs of annotators A_{tot} who labeled the same triplet. This normalized form of inter-annotator agreement has been proposed in [45] to overcome the problem of calculating agreement over a set of documents annotated by a varying number of annotators. The IAA calculated over all 968 triplets is 74.7%. Table 3 shows IAA calculated on each temporal phase.

[Table 3 about here.]

The highest IAA was achieved during TD phase (86.2%), when the TV debate was held; on the contrary, the lowest value of IAA was achieved in RO final phase, in particular two days after the elections (63.4%). This is a signal that users no longer express stances that are as clearly polar as before. Possible explanations of this phenomenon is that the Twitter conversation or the overall political context has changed, and other topics are being discussed. It could also be the case that the political climate is no longer as polarized, and that users are concealing their stances over this subject; in fact, in Section 5 we will describe the results of our network homophily analysis where we also found a signal of a reduced level of political polarization during the RO final phase.

4.4 Label distribution

In Table 4 we show the distribution of labels over our temporal phases.

[Table 4 about here.]

We can observe that percentage of triplets labeled as AGAINST is higher than the rest of labels, accordingly the referendum outcome (40.88% vote "yes", 59.12% vote "no") [18].

A further point concerns the frequency of the label NONE over the temporal phases. As we can see, the distribution of this label constantly increases from phase EC to phase RO. We also explored if users' stance changes over time. We found that 66.8% of the users were labeled with the same stance in all three intervals (55.0% AGAINST, 10.9% FAVOR, 0.8% NONE). For what concerns users that change stance across different time intervals, about 12% of them varies annotated stance in the last phase (10% AGAINST $\rightarrow$ NONE; 2.5% FAVOR $\rightarrow$ NONE). The number of users that take a clear side rapidly decreases when the result of the referendum is finally reported from the main news media outlets; accordingly, the number of users annotated with the NONE stance, that have been slowly increasing before the result, almost doubled from 11.6% to 22.3%. We observed a similar tend during the debate of the so called BREXIT referendum, namely the United Kingdom European Union membership referendum that took place in 2016 in [36].

5 Learning a Stance Detection Model

We aim to automatically estimate the stance to annotate all the users in our dataset. This could support us with new and progressively more accurate tools to understand how stance is distributed across users, how it changes over time, and how interactions between users are associated to the stance. First, we propose and validate a machine learning supervised approach using linear SVM. Then, predicting also the stance of users without annotation, we will be able to analyze the evolution of stance polarity and the political debate using a diachronic perspective (see Section 7).

5.1 Automatic Stance Detection

We recall from Section 4 that our corpus CONREF-STANCE-ITA is made of 968 triplets of tweets (i.e., one tweet, one retweet, and one reply) from the same user. Each triplet has been manually annotated with a stance that can have one

of three values: AGAINST, FAVOR, and NONE. If stance is the target we want to predict using a machine learning model, we have to represent each triplet in terms of a features vector. These features are defined below:

- **Bag of Hashtags** (*BoH*): hashtags considered as terms to build a vector with binary representation. For example, with reference to the triplet shown in Table 8, we can extract the following hastags: "#ioDicoNo", "#bastaunSi", "#ioDicoNo", "#IoVotoNo', and "#vergognaPD".
- **Bag of Mentions** (*BoM*): mentions considered as terms to build a vector with binary representation. Again, with reference to the triplet of Table 8, we have the following mentions: "@fattoquotidiano", "@ComitatoDelNo", "@GiorgiaMeloni", "@MatteoRenzi", "@angelinascanu", "@AntonellaGramig", and "@Rainbowit66".
- **Bag of HashtagsPlus** (*BoHplus*): tokens (the longest words found in an Italian dictionary) extracted from the hashtags considered as terms for building a vector with binary representation. The Italian dictionary was created with the most common words extracted from Wikipedia's Italian pages. Even if we did not perform tokenization/lemmization of words extracted from tweets, we paid a particular attention the verb *to vote*: if the hashtag contains an inflection of this verb we consider the lemma as token. This feature, for instance, extracts from the hashtags #IoVotoNO (#IVoteNo) and #iohovotatono (#IVotedNo) the tokens "votare" and "no". As a consequence, in our example, our tokens are: "io", "dico", "no", "basta", "un", "si", "votare", "vergogna", and "pd".
- **Bag of Mention Plus** (*BoMplus*): tokens extracted from the name of the mentioned users considered as terms for building a vector with binary representation. Names have been extracted from the *User Object name* field of the mentioned user, and tokens are the result of the *name* splitting using the space as separator. In our example, we have tokens "il", "fatto", "quotidiano" extracted from mention "@fattoquotidiano", "giorgia" and "meloni" from "@giorgiameloni", and so on. Please, observe that this is not just a tokenization of the mention string: for instance, from mention @meb (the official Twitter account of Maria Elena Boschi, who held at the time the position of the Minister for Constitutional Reforms) the tokens extracted were retrieved directly using Twitter APIs and they were: "maria", 'elena" and "boschi".

Moreover, we consider also other two features that use only the text of the replied tweet (adding a prefix to differentiate the tokens):

- **Bag of Hashtags for Replies** (*BoHplusreply*): same as *BoHplus*, but using the text contained in the replied tweet. In our example, in the text after the TO indented field in Table 8, we extracted: "basta", "un", "si", "io", and "votare".
- **Bag of Mentions for Replies** (*BoMplusreply*): same as *BoMplus*, but using the text contained in the replied tweet. In our example, in the replied tweet there are no mentions, so this feature is empty.

We performed a five-cross validation on our training set to learn a SVM model. To evaluate the performance of our model, we computed two macro-average of the F_{micro} metrics i.e. F_{avg} and $F_{avg_{AF}}$. The first one computes the average among f-AGAINST, f-FAVOR, and f-NONE F_{micro} metrics. The second one, proposed in both SemEval-2016 Task 6 and IberEval-2017 SD tasks ([43] and [64]), computes the average between f-AGAINST and f-FAVOR F_{micro} metrics. We compared our results with two baselines such as: unigrams, bigrams and trigrams Bag of Words using SVM (*BoW*) and Majority Class (*MClass*). The combination of *BoHplus*, *BoMplus*, and *BoHplusreply* achieved the highest results ($F_{avg} = 0.76$ and $F_{avg_{AF}} = 0.85$). Both the F_{avg} and $F_{avg_{AF}}$ (see Table 5) change in time consistently with IAA (already shown in Table 3).

[Table 5 about here.]

Interesting, removing *BoHplusreply*, F_{avg} decrease to 0.69 and $F_{avg_{AF}}$ decrease to 0.83. Therefore, Table 6 shows the F1-score, precision and recall achieved for each class.

[Table 6 about here.]

The model achieved very high values of $Precision$ for both AGAINST and FAVOR classes, whereas the class NONE achieved the highest $Recall$. For the sake of completeness, we also report F_{avg} and $F_{avg_{AF}}$ obtained by SVM trained with one of each proposed feature compared with highest result and baselines as shown in Figure 2.

[Figure 2 about here.]

We can observe that the feature *BoHplus* achieved an high $F_{avg_{AF}}$, but a relative low $F_{avg_{AF}}$. Furthermore, the feature *BoHplusreply* achieved high values for both $F_{avg_{AF}}$ and $F_{avg_{AF}}$ metrics, but still significantly lower than the highest result.

In the NLDB 2018 paper where we presented our stance classification model, for each triplet in the corpus we added three network-based attributes in the feature vector. These attributes have discrete values that represent the community the user belong to and that were found with the Louvain community detection algorithm in [12] in three of our four networks: retweet, quote, and reply-to networks (see Section 6). Although the addition of these three network-based features improved consistently the performance of our predictions, we decided not to use them to learn the model we need here and that is applied in the analysis presented in Section 7. In fact, our model is used here to automatically annotate the stance of the users in order to study homophily evolution over time in our networks. Since homophily is considered one of the moving forces that causes the emergence of communities in networks, it is important to remove any bias in the classifier that could lead us to self-fulfilling predictions. If you want more information about the exploitation of network-based features for automatic stance detection, please refer to our previous paper in [35].

6 Debates as Networks

Network Analysis provides other useful tools to represent and analyze relations among objects and has applications in several fields including physics, computer science and sociology ([50]). A given complex system is simplified in terms of a network (or a graph), where individuals are nodes (or vertices) and the interactions between them are links (or edges). A network is *weighted* when each link is labeled with a numerical value, that in some domain may reflect the number of times an interaction has taken place between two connected nodes. We used this representation to study four types of networks where links between nodes i and j are created when one of the following types of interaction or relationship between Twitter users is taken into account: i is follower of j; j retweeted a tweet of i; j quoted a tweet of i; i replied to j.

6.1 Follower Networks

We begin with the definition of the network defined by the *follower* relationships among users. Intuitively, we created the social network based on relations explicitly created by users, meaning that a directed link (i, j) exists if i, the follower, *follows* j, the followee. In Twitter "to follow" refers to the specific relationship in which a user i subscribes to another user's feeds, namely user j. It is important to recall that Twitter allows asymmetric relations between users: user x can follow user y, though y is not required to follow back user x; therefore the network is directed.

We recreated the static graph of followers after the collection: this means that we do not know which relations have been formed before, during or after the four temporal phases. Even if these changes could be of interest for our study, recovering the whole follower and followee lists is a time consuming procedure (because of time and space limitations of the *friends/ids* Twitter REST API). Most importantly, Twitter allows to collect the current user's friend list and it does not allow to retrieve any information about changes happened in the past. Therefore, for these reasons, we did not collect data on this specific network evolution during each temporal phase.

Anyway, we gathered the followers list of a subset of 2,671 users. We focused on users belonging to our "Users Sample" and on users who replied to at least once to one of these users. The obtained graph consists in 1,383,740 nodes connected by 5,039,152 edges. 89,928 edges exist among users belonging to "Users Sample".

6.2 Retweet Networks

From the original 900K tweets we collected using the Twitter Stream API (see Section 3.1), we wanted to define a retweets-based network.

Because of intrinsic limitations of the Twitter Stream API, many retweets are not returned. So we adopted a one-step snowball sampling process to retrieve all the relevant retweets that have been created during the observation period. First of all, we removed duplicates out of the 900K tweets. Of the remaining 649,306 tweets, the majority (72.95%) has never been retweeted. For each tweet that has been retweeted at least once, we used the *statuses/retweets/:id* Twitter REST API to collect all the retweets. At the end of the process, we had a set R of relevant retweets, s.t. $|R| = 881,975$.

We used R to create a directed graph for each temporal phase, and also a fifth graph representing users and their retweets for the whole period. Nodes of our *Retweet Network* are users whose tweet $t \in R$ was a retweet or was retweeted. Hence, we have a direct link (i, j) if user j retweeted a tweet of user i. Please, observe that this network

is weighted: if j retweets a tweet of i more than once, no new link is added, but a counter w_{ij} is just incremented accordingly.

Table 7 shows the number of nodes and links of the networks created for each temporal phase. Both the number of the users (nodes) involved in the debate and the number of retweets (links) increase until referendum day.

[Table 7 about here.]

6.3 Quote Networks

The *Quote Networks* have been created similarly of the Retweet Networks, but focusing on quotes instead of retweets. A *quote* is a retweet in which the user adds their own comment. As before, we created a directed weighted graph for each temporal phase, and also a union graph representing the total period of observation. Table 8 shows the number of nodes and edges for each temporal phase. The number of users involved in the debate increases until referendum day; quite interestingly, the number of quoted tweets decreases slightly after the referendum outcome.

[Table 8 about here.]

6.4 Reply-To Networks

Finally, we created four replies-based networks (and also a fifth, that is the union of the others). In particular, an edge (i, j) between two users exists if user j replied to (RT) user i during a given temporal phase. Hence, for each temporal phase we have a weighted directed graph.

The set of RT tweets that we used to build our *Reply-To Networks* has been collected as follows. From the original set of 900K tweets (see Section 3.1) we extracted 81,321 RT tweets using the *statuses/show/:id* Twitter REST API. However, a RT tweet may be itself a RT tweet; therefore, we recovered the whole conversations recursively extracting RT tweets. At the end of the process, we had a set of 103,559 tweets. Table 9 shows the number of nodes and edges of the reply-to networks we created for each temporal phase. The number of users involved in the debate as well as the number of RT tweets increase before the referendum day; apparently users start to leave the conversations on this topic after the referendum results.

[Table 9 about here.]

6.5 Networks and Stance

After we created up to five networks for each type of interaction between users, we could analyze a preliminary signal of correlation of labels annotating the stance of the users at the endpoints of every link. First of all, we focused on links established between users annotated with the same stance, hereinafter referred to as *within-stance percentage* (see Table 10). For the sake of simplicity, here we report our analysis focused only on AGAINST or FAVOR labels and on our "Users sample" core dataset. Since a user can retweet, quote or reply more than once, for these kind of interactions we considered both unweighted and weighted network representations.

[Table 10 about here.]

Some observations for every type of networks we built follows here.

Followers

Since the follower/followee relation has a binary nature, links are easily represented by means of directions and no weights. Relations between users belonging to our "Users Sample" (excluding temporarily nodes annotated with the NONE label) are 16,224, and their distributions maintain comparable sizes distributed over the four temporal phases (4,461, 4,405, 4,316, and 3,042 respectively). In Table 10 we immediately observe that there is an overwhelming majority of relations between users having the same stance (92.5% in the union graph). Actually this is not surprising, because selection biases based on common interests, as in [30], or similar age and country of residence, as in [32], may have an important role in link formation. Furthermore the within-stance percentage increases after EC phase.

Retweets

The considered 3,099 reply-to interactions are respectively distributed over the four temporal phases as it follows: 749, 885, 989, and 476. As we can see in Table 10, the users usually retweet only messages tweeted by users having

the same stance (within-stance percentage is about 99%), without any significant differences between unweighted and weighted graph. We can notice that the within-stance percentage slightly decreases to about 97% in the last RO.

Quotes

If we restrict our analysis only to users that clearly exposed their stance (AGAINST or FAVOR), then we have 717 interactions based on quotes in the full period of observation (respectively 183, 179, 247, and 108 in each temporal phase). There are no significant differences in time, but the within-stance percentage is a little bit higher for weighted graphs (in particular from 94.8% to 97.6% overall), returning a signal that users have a tendency to quote tweets authored by already quoted users.

Replies

Focusing on reply-to interactions between users in our "Users Sample", the number of links drops down to 662, that are distributed over temporal phases as it follows: 172, 173, 207, and 110. However, in this case the within-stance percentage changes comparing unweighted and weighted graphs (in particular from 81.9% to 77.3% in the union network), and rates are generally lower than in previous types of networks.

6.6 Discussion

All the networks exhibit a high value of links among users whose tweets have been annotated with the same stance; in particular the retweet and the quote networks have within-stance percentage very close to 100%. This suggests that, in the context of this debate, Twitter users basically retweet almost exclusively content they agree with. In particular the percentages of quotes are a little bit lower than the ones for retweets: this can happen because quotes may also be used to negatively comment political opponents' posts, as already observed in [25]. Small differences between unweighted and weighted quote networks indicate that users quote even more than once users they tend to agree with.

Interestingly, we have a different behavior in reply-to networks, where there are approximately 20% of *cross-stance edges* (edges between two users with different stance) and this percentage is even higher if we consider the weighted network. This means that, even if users mainly reply to those they agree with, conflicting points of views in the political debate are more likely to be observed with this communication form than with interactions along followers/followees links, retweets, and quotes.

7 Modeling a Polarized Debate

In order to predict the stance of as many as possible users that were involved in this political debate and that we collected (see Section 3.1), we selected all the users who wrote at least one tweet, one retweet and one reply in at least one of the considered temporal phases. Excluding users belonging to the original "Users Sample" that were already manually annotated, we found other 6,465 triplets written by 4,731 different user. Using the model described in Section 5.1, we automatically annotated the stance of 4,731 different users who were active in at least one temporal phase. Figure 3 shows the label distribution in each temporal phase of both manually and automatically annotated triplets.

[Figure 3 about here.]

We have to observe that, although manually and automatically annotated datasets cannot be easily compared to each other because of a different composition and size, they have a similar label distribution. Nevertheless, the automatic classifier apparently tends to amplify the signal over label NONE, especially in the last period RO. This can be due to a reduced accuracy of the classifier in that period (see Table 5), or it can also be the case that less engaged users, that are likely to have been excluded for a manual annotation, have a higher tendency to show their stance less clearly after referendum outcome. However, this increased difficulty of detecting stance in this last period is subject to a broader discussion later in the paper.

Network analysis described in this section is based on four different graphs built with users as nodes, no matter if they were annotated manually or automatically. Therefore, actual total numbers and percentages are reported in Table 11.

[Table 11 about here.]

After that a large number of users have been automatically annotated, we explored the structure of our follower, retweet, quote and reply-to networks. We visualized these networks using the *force atlas* layout[5], hiding users without annotation. The annotated users have been colored depending on the manually or automatically annotated stance: green for FAVOR, red for AGAINST, blue for NONE. For each network, we also included a chord diagram to better represent the amount of links both within and cross clusters (see Figures 4, 7, 8, and 9).

First, we explore if our graphs exhibit *homophily* according to stance, meaning that we want to check if users with the same opinion tend to be more connected each other. Let us consider the subnetwork of just FAVOR and AGAINST users. To do this, let A be the fraction of all users annotated as AGAINST and F the fraction of all users annotated as FAVOR. Considering a given edge in any of our four networks, if we randomly assign label AGAINST to the first end of the edge with probability A, and label FAVOR to the other end of the edge with probability F, and vice-versa, than we can have a cross-stance edge with probability $2AF$. Then, applying the *homophily test* proposed in [16], we can just check if the fraction of cross-stance AGAINST-FAVOR edges (CE_{AF}) is significantly less than $2AF$. In such a case, we could conclude that there is a signal of homophily. We can generalize the test including in our observation nodes labelled as NONE. In this case, the probability of a random cross-stance edge is $2(AF+AN+NF)$ (where N is the fraction of all users annotated as NONE). The *homophily test* can be formulated as: "if the fraction of cross-stance edges (CE_{AFN}) is significantly less than $2(AF + AN + NF)$ then there is a signal of homophily".

Second, we use *modularity* Q_{AFN} in order to observe the evolution of the polarization between AGAINST, FAVOR and NONE labelled communities during the four temporal phases. Indeed, modularity Q is a network metric that provides a measure of the level of connection among the groups of nodes characterized by different features, or modules [50]. We compute modularity Q as it follows:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{ij} \tag{1}$$

where m is the total number of edges in the network, A_{ij} the element (i, j) of the adjacency matrix of the network ($A_{ij} = 1$ if there is a link between vertices i and j), k_i and k_j are respectively the degrees of nodes i and j. The Kronecker δ_{ij} is 1 if users i and j belong to the same group (i.e. are annotated with the same stance, considering AGAINST, FAVOR and NONE labels) and 0 otherwise (Q_{AFN}). Since we do not really know the opinion of NONE users, we also computed the modularity of the networks considering the subnetworks induced by AGAINST and FAVOR users (Q_{AF}). For the purpose of our analysis, just recall that a value of $Q = 0$ should represent a network with a number of within-community edges not higher than a null (or random) model. Values higher than 0 indicate a deviation from randomness. Q approaching to 1 indicates strong community structure (even if values higher then 0.7 are very rare, as explained in [51]).

Since we observed that users labeled as NONE increased in the last phase, we finally explored the likelihood for users to change their stances from AGAINST or FAVOR to NONE in function of the level of cross-stance edges in the previous phase. First, we computed the fraction of cross-stance edges ρ_i for each user i in phase t. Then we measured, for each value of ρ, the fraction of users (with the same value of ρ in the temporal phase t) that change their stance from AGAINST or FAVOR to NONE in phase $t + 1$.

7.1 Follower Networks

We start with the analysis of the follower networks. Figure 4 shows the evolution of the friends-based networks along the four temporal phases. Please notice that in this case the graph structure is the same for each diagram because we retrieved the list of followers only once after the referendum (see Section 6.1). However the stance of the single users may change.

[Figure 4 about here.]

We see segregated colored clusters within the network. As shown also in Table 11, although the number of users annotated as NONE increases over time, the percentage remains quite stable for the first three phases, but increases dramatically in the RO phase (after the referendum). The stance variation seems to affect users belonging to both AGAINST and FAVOR clusters. This is even clearer in the chord diagram of RO phase: users whose stances are now labeled as NONE have a significantly higher number of connections with users of the other groups, but also among themselves.

The follower networks do not exhibit homophily by stance: considering the mean and the standard deviation over the four temporal phases, we have a fraction of CE_{AFN} equal to $0.372 \pm \sigma 0.055$ that is slightly higher than $2(AF + AN +$

[5] As provided within Gephi, the network analysis and visualization framework we used here (downloaded from [22]).

$NF)$ 0.344$\pm\sigma$0.038. This means that we have almost a number of cross-stance edge that we could expect in a random network with the same characteristics.

However, if we consider each phase separately, we can observe a diverging trend in phase RO. Indeed a quite strong *inverse* homophily by stance emerges among the three clusters: the fraction of CE_{AFN} of 0.467 is significantly higher than $2(AF + AN + NF)$ 0.280, as also showed in Fig 5(a). This could be observed also in the chord diagram: the connections between users in different groups are visibly higher than connections among the same group in the last phase because of the the big proportion of users re-labeled as NONE.

[Figure 5 about here.]

Interestingly, results for the subnetwork induced by FAVOR and AGAINST users reveal a strong homophily by stance: the rate of CE_{AF} of 0.092$\pm\sigma$0.006 is significantly lower then $2AF$ 0.324$\pm\sigma$0.013. No significant differences appear considering the four temporal phases, meaning that users with a clearly polar stance (FAVOR or AGAINST) tend not to follow each other.

Finally, we analyzed the polarization among the three clusters computing modularity Q_{AFN} for each temporal phase. As showed in Fig. 6(a), the value changes during the debate starting from the minimum measured value of $Q_{AFN} = 0.096$ after the election outcome (RO phase) up to a maximum of $Q_{AFN} = 0.164$ and $Q_{AFN} = 0.160$ respectively on the DE and TD phases. This reveals a signal of polarization that however is mitigated just after the election results. If we calculate modularity for the subnetworks made of AGAINST and FAVOR users, the levels of polarization is higher.

[Figure 6 about here.]

7.2 Retweet Networks

Figure 7 shows the evolution of the retweet network along the four temporal phases. In this case, both the network structure and the users' stance may change.

[Figure 7 about here.]

As already observed and shown in Table 11, the number of users labelled as NONE increases in the last phase. Quite interestingly, the users affected by this phenomenon are likely those located in the middle of the retweet network, connected with both AGAINST and FAVOR clusters as the chord diagram suggests as well.

The network exhibits a quite strong signal of homophily considering AGAINST, FAVOR, and NONE clusters as showed in Fig. 5(b); in fact, the fraction of CE_{AFN} 0.243$\pm$0.086σ is significantly less than $2(AF + AN + NF)$ 0.344 $\pm$0.038σ. Again, as observed also in the follower network, an inverse trend appears in phase RO. Moreover, in Fig. 5(b) we can also see that the subnetwork of AGAINST and FAVOR clusters exhibits a strong homophily by stance: the fraction of CE_{AF} 0.032$\pm$0.006σ is significantly less than $2AF$ 0.324 $\pm$0.013σ.

The retweet networks appear to be highly segregated between supporters and critics of the reform. We computed the modularities Q_{AFN} and Q_{AF} for each temporal phase as shown in Fig. 6(b). The values change during the debate starting from the minimum measured value of Q_{AFN} 0.167 on EC phase at a maximum of Q_{AFN} 0.232 forthcoming the election on the DE phase, revealing an increasingly polarized debate. No relevant difference is observed considering Q_{AF} values. Observe that inverse homophily and lower values of modularity in the last phase suggest that users from different groups increased their cross groups interactions, but this phenomenon affects this network to a lesser extent compared to the follower one. This might be due to the fact that also communications among NONE users grow in the last phase, as it can be seen in the respective chord diagram in Fig.7.

7.3 Quote Networks

Figure 8 shows the evolution of the quote network along the four temporal phases. Both the network structure and the users' stance are subject to change.

[Figure 8 about here.]

As observed for the retweet network, the visualizations suggest that users that changed their stance to NONE during the last phase of the debate are more likely endpoints of links crossing clusters representing different viewpoints. This intuition will be confirmed also quantitatively later in Section 7.5.

The three clusters exhibit a very light signal of homophily by stance: the fraction of CE_{AFN} $0.31 \pm 0.038\sigma$ is slightly smaller than $2(AF + AN + NF)$ $0.344 \pm 0.038\sigma$. However, as in the follower and retweet networks, an inverse homophily signal emerges in phase RO. Nertheless, as showed in Fig. 5(c), a strong homophily signal is observed if we consider the subnetwork of AGAINST and FAVOR clusters: the fraction of CE_{AF} 0.106 ± 0.036 is significantly less then $2AF$ 0.324 ± 0.013.

The values of modularity change during the debate: Q_{AFN} and Q_{AF} are very similar in EC, DE, and TD phases, revealing a positive signal of polarization, while they diverge in the last phase RO (see Fig. 6(c)). In particular, there is an increasing level of polarization considering the three clusters (AGAINST, FAVOR, and NONE) and a decreasing level of polarization considering just AGAINST and FAVOR clusters in the phase RO. Observe also that in the first three phases the modularity values are smaller than the respective ones in the retweet networks.

7.4 Reply-To Networks

Figure 9 shows the evolution of the reply-to network along the four temporal phases. Both the network structure and the users' stance are subject to change.

[Figure 9 about here.]

Differently from the other three kind of networks we analyzed earlier, every snapshot of the reply-to network at different temporal phases exhibits a signal of inverse homophily by stance (see Figure 5(d)); in fact, the fraction of CE_{AFN} $0.443 \pm 0.053\sigma$ is significantly higher than $2(AF + AN + NF)$ $0.344 \pm 0.038\sigma$. Nevertheless, the subnetworks formed by AGAINST and FAVOR clusters do not exhibit homophily by stance: the fraction of CE_{AF} $0.321 \pm 0.052\sigma$ is comparable to $2(AF)$ $0.324 \pm 0.013\sigma$. Furthermore, Figure 5(d) shows that the homophily values significantly change during the four temporal phases.

In this case, we do not observe striking divisions between users exposing different stances as with the other networks. We computed the modularities Q_{AFN} and Q_{AF} for each temporal phase. The values change during the debate from a minimum measured value of Q_{AFN} and Q_{AF} (respectively 0.057 and 0.024) in DE phase to a maximum of Q_{AFN} and Q_{AF} (respectively 0.166 and 0.113) in the RO phase. Figure 6(d) reveals a lower polarization compared to that observed in the other networks. This can also be seen in the chord diagrams in Figure 9, in which we observe a considerable number of links among different groups. However, after the second phase, the polarization levels increase, meaning that cross-stance connections decrease, and this is also evident in the chord diagrams (specially the last one).

Inverse homophily and low modularity suggest that reply-to is the preferred interaction mode that Twitter users adopted for discussing about the Italian referendum from different view points.

7.5 Users' stances trends

We observed that users labeled with the NONE stance increase dramatically after the referendum outcome (i.e., the RO phase). Therefore, we aim to investigate if the tendency of users to change towards NONE is correlated with the fraction of cross-stance edges observed in the previous phases. Note that this does not mean necessarily that users changed their opinion after referendum result, but that they do not expose clearly their stance or that their opinion is expressed in a less polar way. However, we want to check if the probability for a user to change stance to NONE increases with their fraction of cross-stance links.

We computed the fraction of cross-stance edges for each user in phases EC, DE, and TD. Then, we computed the fraction of users that change stance from label AGAINST or FAVOR to label NONE respectively in the following phases DE, TD, and RO. Figure 10 shows the relation between the fraction of cross-stance edges and the likeliness to change from AGAINST or FAVOR to NONE for each network type (friends, retweets, quotes, and replies networks). Dashed lines are linear polynomials that interpolate the discrete set of known data points. The percentage of users that changes from AGAINST or FAVOR to NONE is not negligible (about 16%), as also observed in Section 4.4 for the *ConRef-STANCE-ita* manually annotated corpus.

[Figure 10 about here.]

This result suggests that users with more heterogeneous connections are also more likely to change their stance to NONE, i.e., their style is now less polarized and more neutral, or they started to doubt about their own vote.

In reply-to networks, we already observed that the users tend to create a higher number of heterogeneous connections compared to the other networks. In addition to this, in Figure 10(d) we observe smaller probabilities to change

stance and a smaller dependence on the number of cross-stance connections, compared to the other networks based on different kind of interactions.

Apparently, reply-to is the preferred message that users adopt to interact with other users expressing a different view; in fact we observed inverse homophily by stance for this kind of networks. This does not necessarily mean that users at the end points of a cross-stance link adopt a more neutral and less polarized style, or that they conceal their view points: the probability to change stance to NONE increases very slowly in presence of conversation with users expressing heterogeneous opinions.

8 Conclusions

In this work we created a manually annotated Italian corpus for addressing Stance Detection from a diachronic perspective, and then we used a machine learning model to annotate automatically other users' stance. Our aim here was to give a contribution to better understand the interplay between communication networks structure and how users express polar stances over time. We observed that, in this particular domain (the Twitter debate about the 2016 Italian Constitutional Referendum), an increasing fraction of users tend to express themselves in a more neutral way, specially after the referendum results. Indeed, a fraction of users previously labeled with a clear polar stance (FAVOR or AGAINST), are labeled as NONE in a following phase of the debate, suggesting that users' stances are less explicit, therefore the (human and mechanical) annotators are no longer able to accurately infer their opinion.

The investigation of network structures led to the observation that users are generally aggregated in homogeneous communities, except for the reply-to network. This is reasonable since users having different opinions often tend to discuss using replies, as in [21].

Our data analysis shows that the network structures based on followers, retweets, and quotes exhibit a signal of homophily by stance among supporters and critics of the reform, suggesting that users tend to connect more likely to others that express the same opinion. However, an inverse homophily by stance emerges in the last phase of the debate for all of these types of networks; in other words, in the last phase there are more connections among users labeled with different stances. Nevertheless, the snapshots of the reply-to network taken at different temporal intervals show an inverse homophily by stance, suggesting that "reply to" communication has its distinct role w.r.t. other Twitter interactions; moreover, this signal of inverse homophily proves that there are much more cross-stance links than in the other networks and more than expected if compared with a random null model. This implies that if we want to investigate on how conversations between users expressing diverging opinions take place on Twitter, then reply-to interactions are more likely to return valuable data for such a purpose. Also the modularity values reveal quite high levels of polarization in follower, retweet, and quote networks and an increasing polarization appears in the replies-based network forthcoming the elections and after the outcome: apparently, maintained discussions between users with different opinions just augmented distances instead of reducing them.

Finally, since the number of NONE labeled users increases dramatically immediately after the referendum outcome, we explored the relation between the level of diversity in the neighbourhood of FAVOR and AGAINST users (fraction of cross-stance edges) and the likelihood to be labeled as NONE in the next phase. In addition to that, it must be remarked that stance classification's performance decreases in the last phase of the debate, confirming that users tend to adopt a different communication style that makes their opinion less evident or more difficult to be detected. Therefore, the results suggest that users who exhibit a higher fraction of cross-stance connections at a given phase of the debate tend to express their stance less clearly in the following phase.

In related works on political debates on social media, as in [1, 13, 36] to cite just a few, some signal has been found suggesting that online political debate tends to structurally polarize users expressing different viewpoints. Here, we also observed that the type of chosen instance of communication is also very important, since different networks, built on different kind of interactions between users, can exhibit several levels of polarization and homophily. In particular, the reply-to graph is quite mixed, revealing a higher number of connections between users in groups expressing clearly opposite stances. This does not necessarily imply a tendency to change opinion more frequently and we need to investigate further in this direction to understand if this is due to the nature of the interaction itself (how the reply-to is used on Twitter) or if the communication between users with different viewpoints leads them to conceal their opinions.

Another important line of research that can put more lights on understanding opinion shifts dynamics and how polarized communities are formed is to better investigate the role of different users inside the given networks. It is well known that many fake accounts, e.g., in [60], try to manipulate online political debates and that diffusion of fake news can depend on the level of segregation in social and communication communities, e.g., in [63]. More empirical evi-

dence to support such intuitions is needed, and research conducted on corpora like CONREF-STANCE-ITA can help to better understand such dynamics.

Availability of data and material

Code, annotated corpus and edges list for each network type are available at: `https://www.researchgate.net/publication/324517807_Annotated_Corpus_for_Stance_Detection_-_Italian_Constitutional_Referendum_2016`.

Funding

The work of Viviana Patti and Giancarlo Ruffo was partially funded by the Fondazione CRT under research project the Hate Speech and Social Media (2016.0688), and the "Progetto di Ateneo/CSP 2016" under research project "Immigrants, Hate and Prejudice in Social Media" (S1618_L2_BOSC_01). The work of Paolo Rosso was partially funded by the Spanish MICINN under the research project "MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech" (PGC2018-096212-B-C31).

Acknowledgments

We would like to thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions. Moreover, we want to thank the editors of this special issue of DKE, namely Max Silberztein, Elisabeth Métais, Farid Meziane, Elena Kornyshova, Faten Atigui, for pre-selecting our paper previously published in the Proceedings of NLDB'18.

References

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 u.s. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05, pp. 36–43. Association for Computing Machinery, New York, NY, USA, 2005. 10.1145/1134271.1134277.

[2] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In *Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing*, STOC '00, pp. 171–180. Association for Computing Machinery, New York, NY, USA, 2000. 10.1145/335305.335326.

[3] P. Anand, M. Walker, R. Abbott, J. E. F. Tree, R. Bowmani, and M. Minor. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pp. 1–9. Association for Computational Linguistics, Stroudsburg, PA, USA, 2011.

[4] F. Angeli, ed. *50º rapporto sulla situazione sociale del paese 2016*. CENSIS, Rome, Italy, 2016. (in Italian).

[5] F. Angeli, ed. *Quattordicesimo Rapporto sulla comunicazione. I media e il nuovo immaginario collettivo*. CENSIS, Rome, Italy, 2017. (in Italian).

[6] Y. Arslan, D. Küçük, and A. Birturk. Twitter sentiment analysis experiments using word embeddings on datasets of various scales. In M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, and F. Meziane, eds., *Natural Language Processing and Information Systems*, pp. 40–47. Springer International Publishing, Cham, 2018.

[7] E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, Jun 2015. doi: 10.1126/science.aaa1160

[8] A.-L. Barabási, D. J. Watts, and M. Newman. *The structure and dynamics of networks*. Princeton University Press, Princeton, 2006.

[9] V. Basile, M. Lai, and M. Sanguinetti. Long-term social media data collection at the University of Turin. In E. Cabrio, A. Mazzei, and F. Tamburini, eds., *CLiC-it 2018 Italian Conference on Computational Linguistics*, number 2253 in CEUR Workshop Proceedings. Aachen, 2018.

[10] V. Basile and M. Nissim. Sentiment analysis on italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 100–107. Atlanta, 2013.

[11] M. A. Bekafigo and A. McBride. Who tweets about politics?: Political participation of twitter users during the 2011gubernatorial elections. *Social Science Computer Review*, 31(5):625–643, 2013. 10.1177/0894439313490405.

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

[13] M. Conover, J. Ratkiewicz, M. Francisco, B. Goncalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, ICWSM-11, pp. 89–96. AAAI Press, Menlo Park, CA, USA, 2011.

[14] W. Deitrick and W. Hu. Mutually enhancing community detection and sentiment analysis on twitter networks. *Journal of Data Analysis and Information Processing*, 1(3):19–29, July 2013. 10.4236/jdaip.2013.13004.

[15] G. Di Fraia and M. C. Missaglia. *The Use of Twitter In 2013 Italian Political Election*, pp. 63–77. Springer International Publishing, Cham, 2014.

[16] D. Easley and J. Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, Cambridge,UK, 2010.

[17] E. Elejalde, L. Ferres, and E. Herder. The nature of real and perceived bias in chilean media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, pp. 95–104. Association for Computing Machinery, New York, NY, USA, 2017. 10.1145/3078714.3078724.

[18] Italian constitutional referendum, 2016, 2018.

[19] D. Freelon. Tweeting to power: The social media revolution in american politics, by jason gainous and kevin m. wagner. *Political Communication*, 31(3):502–505, 2014. 10.1080/10584609.2014.923280.

[20] D. Garcia, A. Abisheva, S. Schweighofer, U. Serdült, and F. Schweitzer. Ideological and temporal components of network polarization in online political participatory media. *Policy & Internet*, 7(1):46–79, March 2015. 10.1002/poi3.82.

[21] K. Garimella, I. Weber, and M. De Choudhury. Quote rts on twitter: Usage of the new feature for political discourse. In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, pp. 200–204. ACM, New York, NY, USA, 2016. 10.1145/2908131.2908170.

[22] The open graph viz platform, 2018.

[23] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barábasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, Jun 2008. 10.1038/nature06958.

[24] B. Gonçalves, N. Perra, and A. Vespignani. Modeling users' activity on twitter networks: Validation of dunbar's number. *PloS one*, 6(8):e22656, August 2011.

[25] P. Guerra, R. Nalon, R. Assunção, and M. J. Wagner. Antagonism also flows through retweets: The impact of out-of-context quotes in opinion polarization analysis. In *Eleventh International AAAI Conference on Weblogs and Social Media*, ICWSM 2017, pp. 536–539. AAAI Press, Palo Alto, CA, USA, 2017.

[26] H. Gusterson. From brexit to trump: Anthropology and the rise of nationalist populism. *American Ethnologist*, 44(2):209–214, 2017. 10.1111/amet.12469.

[27] About conversations on twitter, 2018.

[28] P. Holme and J. Saramäki. Temporal networks. *Physics Reports*, 519(3):97 – 125, October 2012. https://doi.org/10.1016/j.physrep.2012.03.001.

[29] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pp. 168–177. Association for Computing Machinery, New York, NY, USA, 2004. 10.1145/1014052.1014073.

[30] J.-h. Kang and K. Lerman. Using lists to measure homophily on twitter. In *AAAI workshop on Intelligent Techniques for Web Personalization and Recommendation*, 2012.

[31] G. Krings, M. Karsai, S. Bernhardsson, V. D. Blondel, and J. Saramäki. Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Science*, 1(1):4, May 2012. doi: 10.1140/epjds4

[32] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 591–600. ACM, New York, NY, USA, 2010. 10.1145/1772690.1772751.

[33] M. Lai, A. T. Cignarella, and D. I. Hernández Farías. Itacos at ibereval2017: Detecting stance in catalan and spanish tweets. In Raquel et al. [55], pp. 185–192.

[34] M. Lai, D. I. Hernández Farías, V. Patti, and P. Rosso. Friends and enemies of clinton and trump: Using context for detecting stance in political tweets. In G. Sidorov and O. Herrera-Alcántara, eds., *Advances in Computational Intelligence: 15th Mexican International Conference on Artificial Intelligence, MICAI 2016, Cancún, Mexico, October 23–28, 2016, Proceedings, Part I*, pp. 155–168. Springer International Publishing, Cham, 2017. 10.1007/978-3-319-62434-1_13.

[35] M. Lai, V. Patti, G. Ruffo, and P. Rosso. Stance evolution and twitter interactions in an italian political debate. In M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, and F. Meziane, eds., *Natural Language Processing and Information Systems*, pp. 15–27. Springer International Publishing, Cham, 2018.

[36] M. Lai, M. Tambuscio, V. Patti, G. Ruffo, and P. Rosso. Extracting graph topological information and users' opinion. In G. J. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, and N. Ferro, eds., *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association*, CLEF 2017, pp. 112–118. Springer International Publishing, Cham, 2017. 10.1007/978-3-319-65813-1_10.

[37] P. F. Lazarsfeld and R. K. Merton. Friendship as a social process: A substantive and methodological analysis. In M. Berger, T. Abel, and C. Page, eds., *Freedom and Control in Modern Society*, pp. 18–66. Van Nostrand, New York, 1954.

[38] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. Life in the network: The coming age of computational social science. *Science*, 323(5915):721–723, Feb 2009. 10.1126/science.1167742.

[39] X. Li, Y. Rao, H. Xie, X. Liu, T.-L. Wong, and F. L. Wang. Social emotion classification based on noise-aware training. *Data & Knowledge Engineering*, 2017. https://doi.org/10.1016/j.datak.2017.07.008.

[40] W.-H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann. Which side are you on?: Identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pp. 109–116. Association for Computational Linguistics, Stroudsburg, PA, USA, 2006.

[41] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[42] E. Messina, E. Fersini, and J. Zammit-Lucia. All Atwitter about Brexit: Lessons for the Election Campaigns, 2017.

[43] S. M. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 31–41. Association for Computational Linguistics, Stroudsburg, PA, USA, June 2016.

[44] S. M. Mohammad, S. Kiritchenko, and X. Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 321–327. Association for Computational Linguistics, Stroudsburg, PA, USA, June 2013.

[45] S. M. Mohammad, P. Sobhani, and S. Kiritchenko. Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3):26:1–26:23, June 2017. 10.1145/3003433.

[46] S. M. Mohammad and P. D. Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, September 2013. 10.1111/j.1467-8640.2012.00460.x.

[47] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 1–18. Association for Computational Linguistics, Stroudsburg, PA, USA, June 2016.

[48] P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, and T. Wilson. Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the 7th International Workshop on Semantic Evaluation*, SemEval 2013, pp. 312–320. Association for Computational Linguistics, Stroudsburg, PA, USA, June 2013.

[49] A. A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi. On the structural properties of massive telecom call graphs: Findings and implications. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pp. 435–444. Association for Computing Machinery, New York, NY, USA, 2006. 10.1145/1183614.1183678.

[50] M. Newman. *Networks: An Introduction*. Oxford university press, Oxford, UK, 2010.

[51] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, Feb 2004. 10.1103/PhysRevE.69.026113.

[52] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, Jan. 2008. 10.1561/1500000011.

[53] E. Pariser. *The Filter Bubble: What the Internet is Hiding from You*. Penguin UK, London, Regno Unito, 2011.

[54] A. Perrin. Social media usage: 2005-2015. Technical report, Pew Research Center, 2015.

[55] M. Raquel, G. Julio, R. Paolo, M. Soto, and C.-d.-A. Jorge, eds. *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages*, number 1881 in CEUR Workshop Proceedings. Aachen, Germany, 2017.

[56] S. Rosenthal, N. Farra, and P. Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 502–518. Association for Computational Linguistics, Stroudsburg, PA, USA, August 2017.

[57] S. Rosenthal, P. Nakov, S. Kiritchenko, S. M. Mohammad, A. Ritter, and V. Stoyanov. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 451–463. Association for Computational Linguistics, Denver, Colorado, June 2015.

[58] S. Rosenthal, A. Ritter, P. Nakov, and V. Stoyanov. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 73–80. Association for Computational Linguistics and Dublin City University, Dublin, Ireland, August 2014.

[59] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove. Mobile call graphs: Beyond power-law and lognormal distributions. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pp. 596–604. Association for Computing Machinery, New York, NY, USA, 2008. 10.1145/1401890.1401963.

[60] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer. The spread of low-credibility content by social bots. *Nature Communications*, 9(1):4787, 2018.

[61] C. R. Sunstein. The law of group polarization. *Journal of Political Philosophy*, 10(2):175–195, December 2002. doi: 10.1111/1467-9760.00148

[62] C. R. Sunstein. *# Republic: Divided Democracy in the age of Social Media*. Princeton University Press, Princeton, NJ, USA, 2018.

[63] M. Tambuscio, D. F. M. Oliveira, G. L. Ciampaglia, and G. Ruffo. Network segregation in a model of misinformation and fact-checking. *Journal of Computational Social Science*, 1(2):261–275, September 2018. doi: 10.1007/s42001-018-0017-x

[64] M. Taulé, M. A. Martí, F. M. Rangel Pardo, P. Rosso, C. Bosco, and V. Patti. Overview of the task of stance and gender detection in tweets on catalan independence at ibereval 2017. In Raquel et al. [55], pp. 1–14.

[65] Y. Theocharis and W. Lowe. Does facebook increase political participation? evidence from a field experiment. *Information, Communication & Society*, 19(10):1465–1486, 2016. 10.1080/1369118X.2015.1119871.

[66] L. Weng, M. Karsai, N. Perra, F. Menczer, and A. Flammini. Attention on weak ties in social and communication networks. *ArXiv e-prints*, arxiv/1505.02399, 2015.

[67] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pp. 347–354. Association for Computational Linguistics, Stroudsburg, PA, USA, 2005. 10.3115/1220575.1220619.

[68] We make ai work in the real world, 2018.

[69] K. Xu, J. Li, and S. S. Liao. Sentiment community detection in social networks. In *Proceedings of the 2011 iConference*, iConference '11, pp. 804–805. Association for Computing Machinery, New York, NY, USA, 2011. 10.1145/1940761.1940913.

[70] S. Yardi and D. Boyd. Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology & Society*, 30(5):316–327, 2010. 10.1177/0270467610380011.

[71] G. Zarrella and A. Marsh. Mitre at semeval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval-2016, pp. 458–463. Association for Computational Linguistics, Stroudsburg, PA, USA, 2016. 10.18653/v1/S16-1074.

[72] C. Zirn and H. Stuckenschmidt. Multidimensional topic analysis in political texts. *Data & Knowledge Engineering*, 90:38 – 53, 2014. https://doi.org/10.1016/j.datak.2013.07.003.

List of Figures

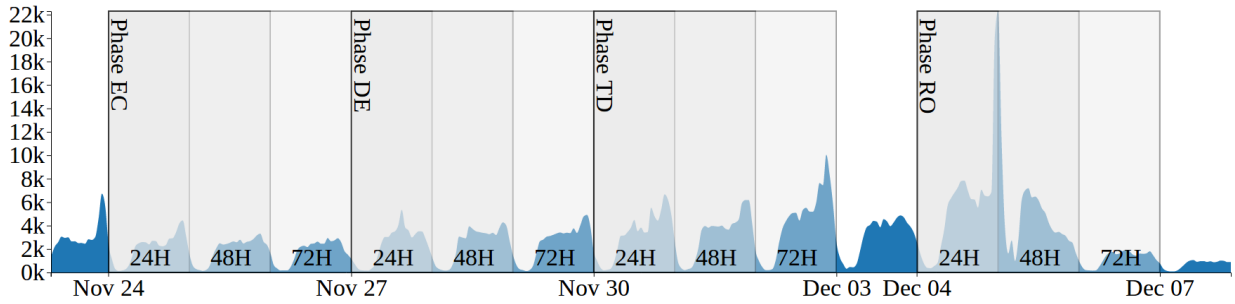

Figure 1: Hourly frequency of tweets posted in each considered temporal phase. Starting from the date the event happened, the figure also shows the enlargement of the considered temporal window.

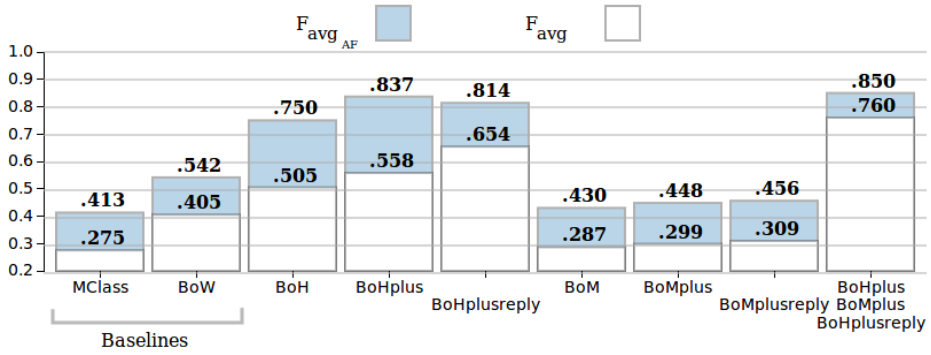

Figure 2: F_{avg} and $F_{avg_{AF}}$ obtained by SVM trained with each of the proposed features compared with the baselines and the best feature set result (*BoHplus*, *BoMplus*, and *BoHplusreply*).

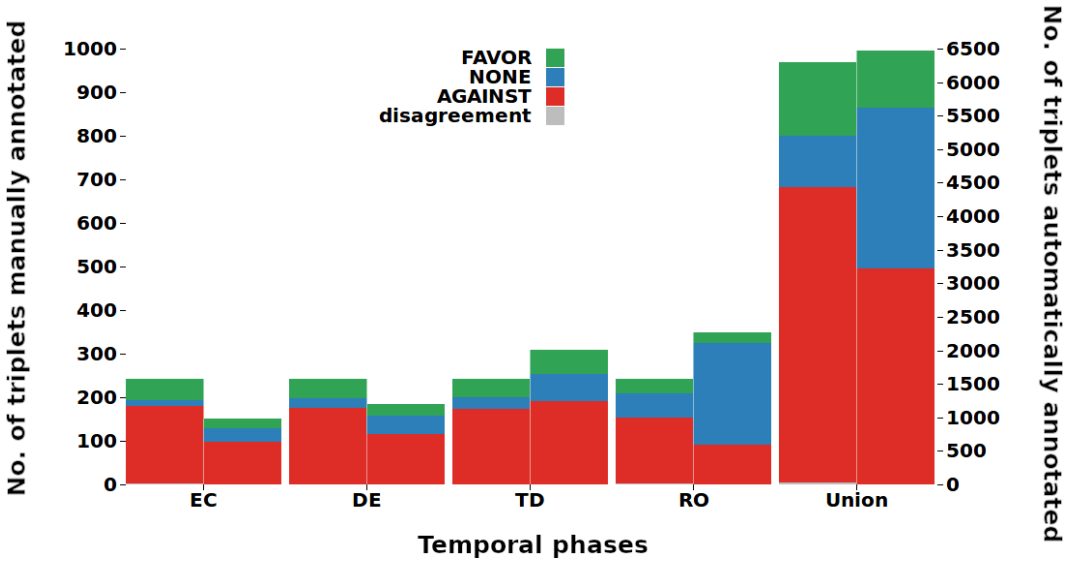

Figure 3: Distribution of 963 manually (bars on the left) and 6,465 automatically (bars on the right) annotated triplets over our given temporal phases RO, TD, DE, and EC. Datasets of manually and automatically annotated triplets have different sizes, hence scales on left and right y-axes have been re-scaled accordingly.

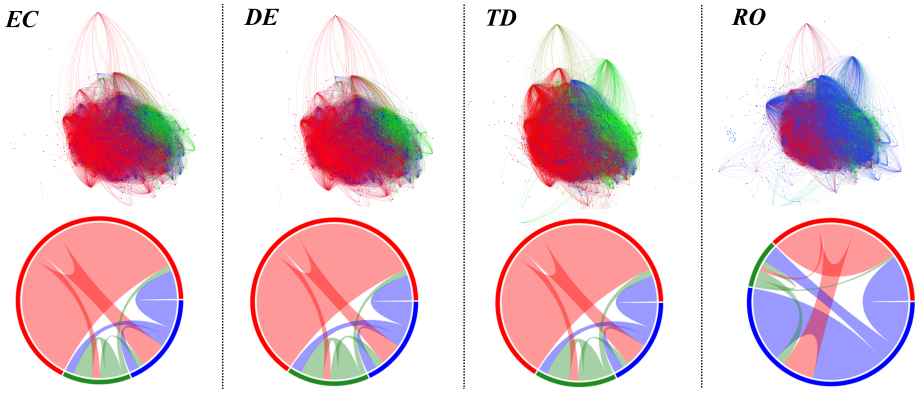

Figure 4: Follower Networks displayed using force atlas layout (above) and chord diagram (below) for each temporal phase.

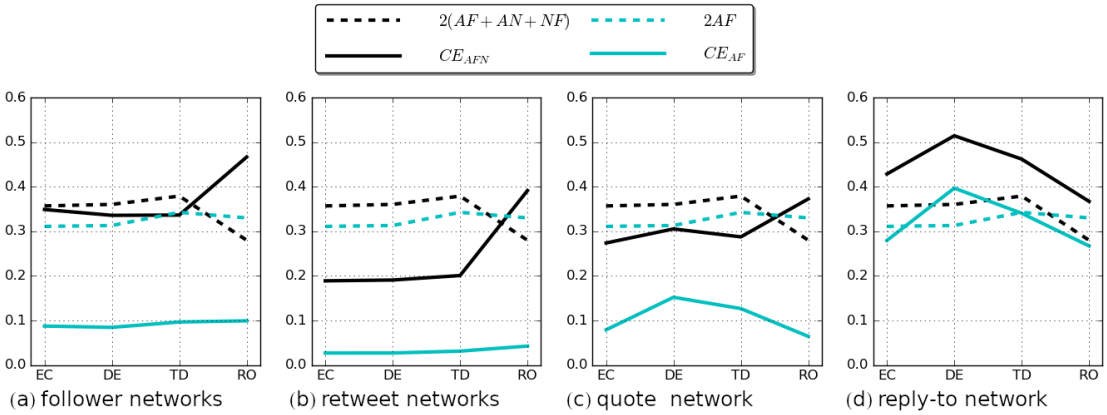

Figure 5: The homophily test according to stance for each temporal phase. We have homophily by stance if the fraction of cross-stance edges (CE) observed (solid lines CE_{AFN} and CE_{AF}) is significantly less than the probability that a cross-stance link is established in a null model (dashed lines $2(AF + AN + NF)$ and $2AF$).

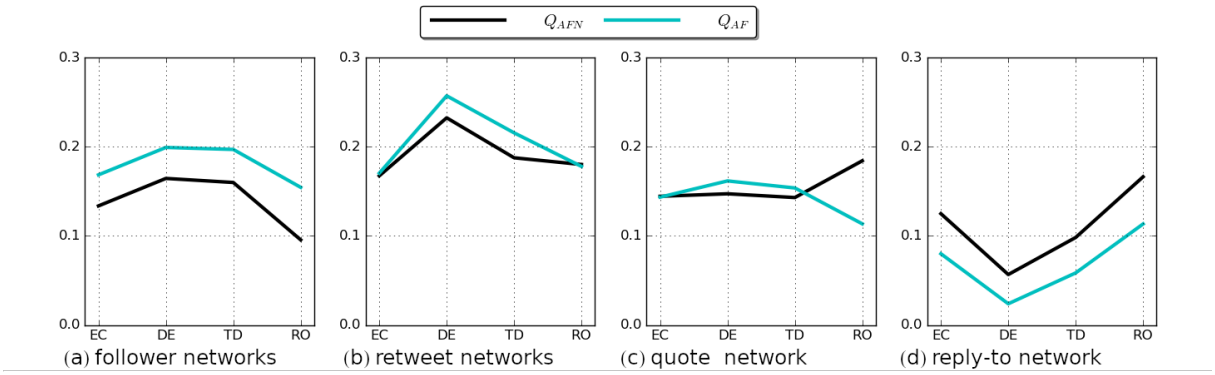

Figure 6: Evolution of modularity Q_{AFN} for all the networks at every phase; also modularity (Q_{AF}) is displayed for all the subnetworks formed by only AGAINST and FAVOR clusters.

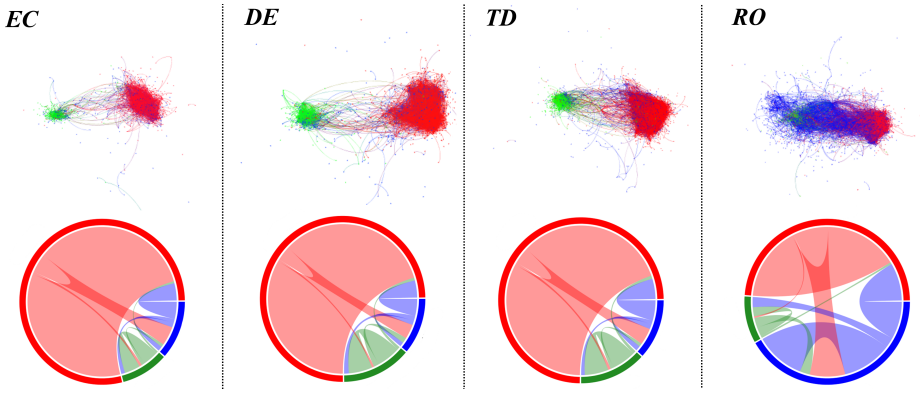

Figure 7: Retweet Networks displayed using force atlas layout (above) and chord diagram (below) for each temporal phase

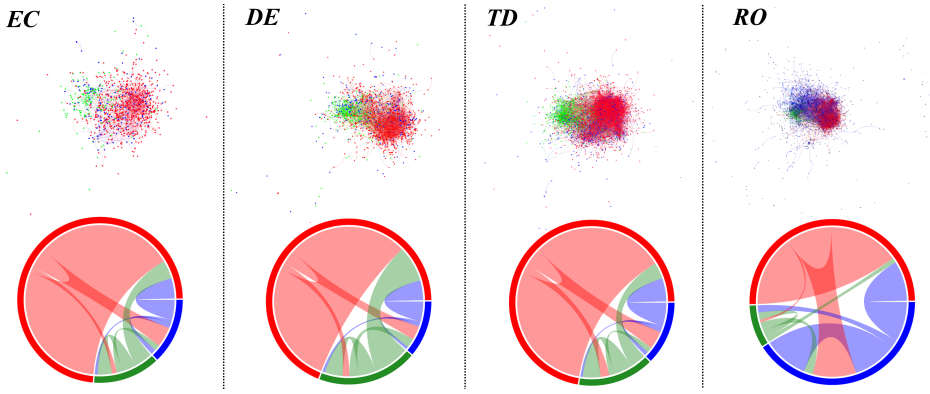

Figure 8: Quote Networks displayed using force atlas layout (above) and chord diagram (below) for each temporal phase

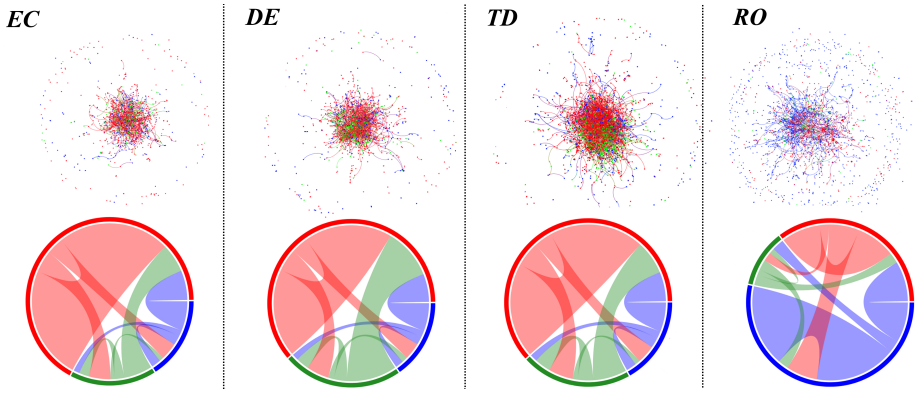

Figure 9: Reply-To Networks displayed using force atlas layout (above) and chord diagram (below) for each temporal phase

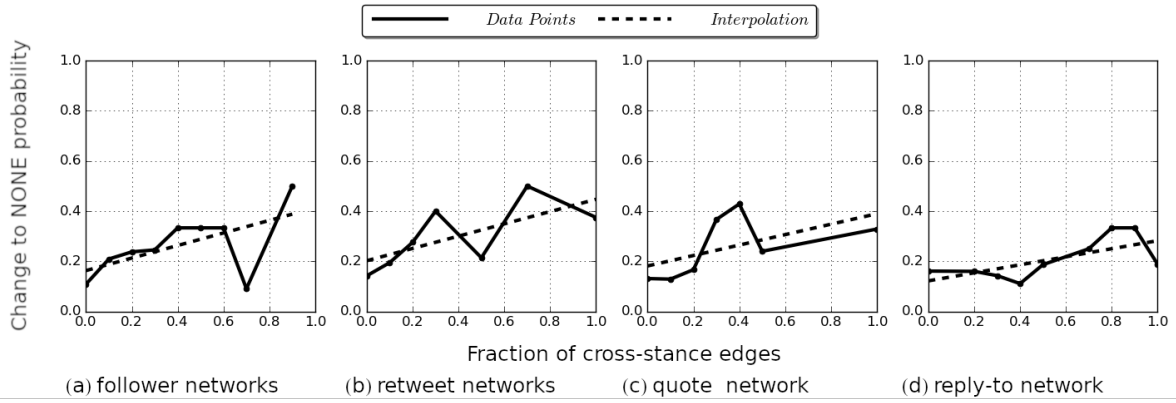

Figure 10: The likelihood to change from AGAINST or FAVOR to NONE in function of the fraction of cross-stance edges in the previous phase, for each type of network.

List of Tables

rank	no. of Tweets	#hashtags	rank	no. of Tweets	#hashtags
1	29741	Roma	16	8000	bastaunsì
2	29053	IoVotoNO	17	7764	PurposeTourBologna
3	24428	italia	18	7507	m5s
4	20812	Amici16	19	7396	MengoniLiveMilano
5	20338	XF10	20	7392	TeenWolf
6	18434	Milano	21	7018	PurposeBologna
7	17477	Renzi	22	6694	SerieA
8	14629	abilitatiTFA	23	6581	iovotosi
9	11425	TFAèConcorso	24	6458	assenzio
10	10845	news	25	6423	IoDicoNO
11	10161	doppiocanale	26	5700	hoBisogno
12	9966	BraccialettiRossi3	27	5662	AMAs
13	9891	MilanInter	28	5180	harrypotterelordinedellafenice
14	8786	ARIASJUSTINBIEBER	29	5156	ReferendumCostituzionale
15	8177	DeLuca	30	5144	referendum

Table 1: Ranking of the most frequently used hashtags in tweets from Nov. 11th to 23th 2016 according to *Twita*.

TWEET	• Travaglio: "Il 2 dicembre grande serata nostra Costituzione in diretta streaming" #ioDicoNo URL via @fattoquotidiano *(Travaglio: "The 4th December a great night for our Constitution in streaming live" #ISayNo URL through @fattoquotidiano)*
RETWEET	• RT @ComitatoDelNO: Brava @GiorgiaMeloni che ricorda a @matteorenzi di (provare a) dire la verità almeno 1 volta su 10! *(RT @NOCommittee: well done @GiorgiaMeloni who reminds to @matteorenzi to (try to) say the truth at least 1 time over 10!)*
REPLY	• @angelinascanu @AntonellaGramig @Rainbowit66 per la poltrona. La cosa più cara a voi del #bastaunSi #IoDicoNo #IoVotoNO #vergognaPD *(@angelinascanu @AntonellaGramig @Rainbowit66 for the chair. The most important thing for you of the #justaYES #ISayNo #IVoteNO #shamePD)*
↪ TO	○ Già dovrebbe spiegare...ma la risposta si conosce. Il 4 dicembre #bastaunSi #IoVotoSI URL *(He already should justify... but the answer is known. December, The 4th #justaYES #IVoteYES URL)*

Table 2: Triplet example: the user wrote a tweet, a retweet (RT), and a reply (black bullets). Please notice that the triplet also includes a tweet (white bullet) written by another user, embedded in the reply.

	EC	DE	TD	RO	UNION
IAA	78.6%	74.8%	**86.2%**	**63.4%**	74.6%

Table 3: IAA calculated for each of the four temporal phases, and also for the total period of observation.

LABEL	EC	DE	TD	RO	UNION
AGAINST	72.7%	72.7%	71.5%	62.8%	69.9%
FAVOR	19.8%	18.3%	16.9%	14.0%	17.2%
NONE	**6.2%**	**9.1%**	**11.6%**	**22.3%**	12.3%
disagreement	1.2%	0%	0%	0.8%	0.5%

Table 4: Label distribution for each of the four temporal phases, and also for the total period of observation.

	EC	DE	TD	RO	UNION
F_{avg}	$0.58_{\pm 0.03}$	$0.72_{\pm 0.07}$	$0.83_{\pm 0.08}$	$0.62_{\pm 0.11}$	$0.76_{\pm 0.03}$
$F_{avg_{AF}}$	$0.87_{\pm 0.04}$	$0.87_{\pm 0.05}$	$0.90_{\pm 0.04}$	$0.72_{\pm 0.12}$	$0.85_{\pm 0.02}$

Table 5: F_{avg} and $F_{avg_{AF}}$ achieved in the different temporal phases with the combination of BoHplus, BoMplus, and BoHplusreply features. Standard deviations calculated over the 5 folds of our cross validation analysis are shown for each measure.

	NONE	AGAINST	FAVOR
$Precision$	0.45	**0.96**	**0.94**
$Recall$	**0.89**	0.86	0.67
F_{micro}	0.60	0.91	0.79

Table 6: Scores achieved by SVM exploiting *BoHplus*, *BoMplus*, and *BoHplusreply*

	EC	DE	DT	RO	UNION
nodes	25,793	28,015	33,860	63,805	94,445
links	83,134	98,717	127,593	158,243	405,843

Table 7: Retweets-based graphs size for each of the four temporal phases, and also for the total period of observation. The fifth graph is an union of all the other weighted graphs, hence the number of nodes and links in the last column is not the sum of the others.

	EC	DE	DT	RO	UNION
nodes	6,907	7,577	9,599	14,919	24,976
links	13,574	15,665	22,479	21,977	69,240

Table 8: Quotes-based graphs size for each of the four temporal phases, and also for the total period of observation. The fifth graph is an union of all the other weighted graphs, hence the number of nodes and links in the last column is not the sum of the others.

	EC	DE	DT	RO	UNION
nodes	6,236	6,663	8,801	8,497	20,936
links	8,651	9,714	14,046	10,832	41,292

Table 9: Replies-based graphs size for each of the four temporal phases, and also for the total period of observation. The fifth graph is an union of all the other weighted graphs, hence the number of nodes and links in the last column is not the sum of the others.

	FOLLOWERS	RETWEETS		QUOTES		REPLIES	
	unweighted	*unweighted*	*weighted*	*unweighted*	*weighted*	*unweighted*	*weighted*
EC	90.0%	98.1%	98.9%	94.0%	96.9%	82.0%	71.9%
DE	93.9%	99.7%	99.8%	96.1%	97.9%	83.2%	81.0%
TD	93.0%	98.6%	99.4%	93.9%	97.7%	81.2%	78.9%
RO	93.6%	97.5%	97.6%	96.3%	97.9%	80.9%	77.1%
UNION	92.5%	98.6%	99.1%	94.8%	97.6%	81.9%	77.3%

Table 10: The within-stance percentage: rate of links between nodes in the "Users Sample" having the same stance (AGAINST or FAVOR)

LABEL	EC	DE	TD	RO	UNION
AGAINST	(809, 66.04%)	(933, 64.43%)	(1412, 62.73%)	(740, 29.55%)	(3894, 52.42%)
FAVOR	(193, 15.76%)	(225, 15.54%)	(397, 17.64%)	(195, 7.79%)	(1010, 13.60%)
NONE	(223, 18.20%)	(290, 20.03%)	(442, 19.64%)	(1569, 62.66%)	(2524, 33.98%)

Table 11: Label distribution for all the nodes for each of the four temporal phases, and also for the total period of observation.