ORIGINAL ARTICLE

CrossMark

Analyzing polarization of social media users and news sites during political campaigns

Fabrizio Marozzo[1] · **Alessandro Bessi**[2]

Abstract

Social media analysis is a fast growing research area aimed at extracting useful information from social networks. Recent years have seen a great interest from academic and business world in using social media to measure public opinion. This paper presents a methodology aimed at discovering the behavior of social network users and how news sites are used during political campaigns characterized by the rivalry of different factions. As a case study, we present an analysis on the constitutional referendum that was held in Italy on December 4, 2016. A first goal of the analysis was to study how Twitter users expressed their voting intentions about the referendum in the weeks before the voting day, so as to understand how the voting trends have evolved before the vote, e.g., if there have been changes in the voting intentions. According to our study, 48% of Twitter users were polarized toward *no*, 25% toward *yes*, and 27% had a *neutral* behavior. A second goal was to understand the effects of news sites on the referendum campaign. The analysis has shown that some news sites had a strong polarization toward *yes* (unita.tv, ilsole24ore.it and linkiesta.it), some others had a *neutral* position (lastampa.it, corriere.it, huffingtonpost.it and repubblica.it) and others were oriented toward *no* (ilfattoquotidiano.it, ilgiornale.it and beppegrillo.it).

Keywords Social media analysis · Public opinion · Online information · News sites · Users' polarization · Social networks · Political events

1 Introduction

In the last years, the production rate of digital data has increased exponentially, with a great contribution from social networks such as Facebook, Twitter, Qzone and Instagram. The large volumes of data generated and gathered by social media platforms can be used to extract valuable information regarding human dynamics and behaviors (Belcastro et al. 2017). Social media analysis is a fast growing research area aimed at extracting useful information from this big amount of data (Talia et al. 2015). For example, it is used for the analysis of collective sentiments (Pang and Lee 2008), for understanding the behavior of groups of people

(Cesario et al. 2015, 2016), and to improve the communication between companies and customers (Hanna et al. 2011).

Recently, there has been a great interest from academic and business world for using social media to measure public opinion (Anstead and O'Loughlin 2015). Several researchers have used social media data for predicting election results (Franch 2013), measuring how public opinion changes after important political debates (Elmer 2013) or studying the effects of social media during important recent historical events (e.g., Arab Spring Howard et al. 2011). Other researchers have examined the impact of social media spaces on news consumption (Hermida et al. 2012) and on how information spreads through social networks (Lerman and Ghosh 2010).

This paper presents a methodology aimed at discovering the behavior of social network users and how news sites are used during political campaigns characterized by the rivalry of different factions. The methodology is composed of five steps: (1) definition of the factions and collection of the keywords associated with a political event; (2) collection of all the posts generated by social network users containing one or more keywords defined at first

✉ Fabrizio Marozzo
   fmarozzo@dimes.unical.it

   Alessandro Bessi
   bessi@isi.edu

1   DIMES, University of Calabria, Via P. Bucci, 42C, 87036 Rende, Italy

2   Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA

step; (3) preprocessing of the posts and creation of the input dataset; (4) data analysis and mining; and (5) results visualization. On the one hand, the methodology allows to study the users' polarization before a political event, what arguments they used to support their voting intentions, and if such intentions change in the weeks preceding the vote. On the other hand, the methodology allows to study the effects of news sites on a political event, e.g., how many users used information from news sites to support their voting intentions and what news sites can be considered in favor, against or neutral to a given faction.

Unlike works in the literature that classify a post manually (Gruzd and Roy 2014) or with text mining techniques (Ceron et al. 2014; Burnap et al. 2016; Tumasjan et al. 2010), the methodology exploits keywords contained in a post to classify it in favor of a faction. In this way, a post is classified in favor of a faction only if it shows a clear voting indication for a such faction; otherwise, we consider the post as neutral. With regard to studying the polarization of news sites, different works in literature use a direct approach that analyzes the contents of articles published by such news sites to understand their political orientation (Wagner 2017; Dallmann et al. 2015). Our approach instead uses a novel approach that analyzes how users referred to these news sites in their posts for supporting their voting intentions. Other aspects of novelty of the methodology are some analyses we have proposed such as statistical significance of collected data, mobility flows and polarization prediction.

Although the methodology is able to analyze political events characterized by *n*-factions, in this paper we focus on a subset of political events distinguished by the rivalry of only two factions (i.e., *two-faction political events*). This subset includes salient political events, such as *referenda* that see the opposition of two factions (e.g., in favor of *yes* or *no*) or *ballots* (runoff voting) that see the opposition of two candidates competing for the final victory. These events are characterized by some interesting features that make them interesting to study: (1) People show a special attention and sensitivity to these events as they are very important for a nation; (2) people present a strong polarization in favor of one of the two factions, and this allows separating them in two distinct groups; (3) accurate analysis can be done since each user can choose only between two values.

As a case study, we present an analysis on the constitutional referendum that was held in Italy on December 4, 2016. The Italian voters were asked whether they approve a constitutional law that amends the Italian Constitution to reform the composition and powers of the Parliament of Italy. The main supporter of *yes* was the Democratic Party and its leader and Italian Prime Minister Matteo Renzi, whereas in favor of *no* were the main opposition parties and several citizen committees. The referendum saw a high voter turnout (approximately 65% of voters) and a clear victory of *no* (59% of the expressed preferences).

In the weeks before the referendum, we identified a number of keywords (i.e., hashtags) that were used in Twitter to publish *neutral* posts on the referendum, for supporting either *yes* or *no*. We collected 338,592 tweets (1,165,176 if we also consider retweets) that contained those hashtags from October 23 (5 weeks before the voting day) to December 3, 2016 (1 day before). The number of Twitter users under analysis is 50,717 (139,066 considering also those who published a retweet).

A first goal of the analysis was to study how Twitter users expressed their voting intentions about the referendum in the weeks before the voting day, so as to understand how the voting trends have evolved before the vote, e.g., if there have been changes in the voting intentions. According to our study, 48% of Twitter users were polarized toward *no*, 25% toward *yes*, and 27% had a *neutral* behavior. Regarding the change of opinion in the weeks preceding the vote, the majority of users categorized as supporters of *no* have never changed during the weeks preceding the vote, while a consistent part of the *neutral* users moved toward *no*. A second goal was to understand the effects of news sites on the referendum campaign. The 22% of tweets contained URLs to news related to the referendum. The analysis has shown that some news sites had a strong polarization toward *yes* (unita.tv, ilsole24ore.it and linkiesta.it), some others had a *neutral* position (lastampa.it, corriere.it, huffingtonpost.it and repubblica.it) and others toward *no* (ilfattoquotidiano.it, ilgiornale.it and beppegrillo.it).

The structure of the paper is as follows. Section 2 describes the methodology proposed in this paper. Sections 3 and 4 describe, respectively, how the methodology has been exploited on the Italian constitutional referendum and which results have been achieved. Section 5 discusses related work. Finally, Sect. 6 concludes the paper.

2 Methodology

Given a political event $\mathcal{P}$ to be analyzed, five are the main steps of the proposed methodology:

1. Definition of the factions F and collection of the keywords K associated with $\mathcal{P}$;
2. Collection of all the posts P generated by social network users containing one or more keywords in K;
3. Preprocessing of P and creation of the input dataset D;
4. Data analysis and mining of dataset D;
5. Results visualization.

2.1 Definition of the factions *F* and collection of the keywords *K* associated with $\mathcal{P}$

The political event $\mathcal{P}$ is characterized by the rivalry of different factions $F = \{f_1, f_2, \ldots, f_n\}$. Examples of political events and relative factions are: (1) municipal election, in which a faction supports a mayor candidate; (2) political election, in which a faction supports a party; (3) presidential election, in which a faction supports a presidential candidate. In this step, we collect the main keywords K used by social network users to write posts associated with $\mathcal{P}$. The keywords can be divided in different subsets, e.g., $K = K_{\text{neutral}} \cup K_{f1} \cup \cdots \cup K_{fn}$ as described below:

– The general keywords that can be associated with $\mathcal{P}$ but cannot be associated with any factions in F (i.e., are *neutral*) are assigned to K_{neutral}.
– For each faction $f_i \in F$, K_{fi} contains the keywords used to support f_i.

In this paper, we focus on a subset of political events characterized by the rivalry of only two factions $F = \{f_1, f_2\}$. Examples of *two-faction events* are: (1) *referendum*, in which a faction supports a position (e.g., in favor of *yes* or *no*); (2) *ballot* (or runoff voting), in which a faction is one of the two candidates competing for the final victory. For these events, the keywords are divided in three subsets, $K = K_{\text{neutral}} \cup K_{f1} \cup K_{f2}$, where K_{neutral} contains the *neutral* keywords, K_{f1} and K_{f2} are the keywords associated, respectively, with f_1 and f_2.

2.2 Collection of all the posts *P* generated by social network users containing one or more keywords in *K*

Through the API provided by social networks, we download all the posts containing one or more *keywords* in *K*. The posts are not collected in real time, but downloaded a given time after their publication (e.g., 24 h). In this way, we are able to get some statistics related to the popularity of a post. For example: (1) number of *shares*, which indicates how many users shared a post with their friends; (2) number of *likes*, which indicates how many users found a post useful. Each collected post has at least one key in *K*, but may have also other keywords (*co-keywords*) that are useful to understand the arguments used to support the voting intentions.

2.3 Preprocessing of *P* and creation of the input dataset *D*

The goal of this phase is to preprocess the posts in *P* to make them ready for the subsequent analysis. Specifically, after preprocessing each post $p \in P$ is structured as a tuple

Table 1 Classification of a post by analyzing its keywords in an *n*-factions event

K_{neutral}	K_{f1}	$\ldots$	K_{fn}	Class
–	X	–	–	f_1
X	X	–	–	f_1
–	–	–	X	f_n
X	–	–	X	f_n
X	–	–	–	Neutral
–	X	–	X	Neutral
X	X	–	X	Neutral

$\langle$*user, text, timestamp, keywords, statistics, URLs, domains, class*$\rangle$ where

– *user* is the identification of the user who published *p*;
– *text* is the text of the post;
– *timestamp* is the timestamp indicating when *p* was published;
– *statistics* contains some statistic data about *p*;
– *keywords* contains the keywords of *p*;
– *URLs* contains all the URLs present in *p*;
– *domains* contains, for each $u \in$*URLs*, the corresponding domain names;
– *class* is a label that indicates how a post is classified.

The following operations are performed to preprocess the *keywords*, *URLs* and *domains* fields: (1) All the keywords are transformed to be lowercase and without accented letters (e.g., IOVOTOSI or iovotosí → iovotosi); (2) all the short URLs are changed into the corresponding long URLs (for example larep.it → repubblica.it); (3) all the alias domains are changed into a single domain (e.g., beppegrillo.it and beppegrillo.com → beppegrillo.it).

The *class* label is computed by analyzing the keywords of a post. A post may be labeled as one of the classes *neutral*, $\{f_1, f_2, \ldots, f_n\}$. Considering the keywords $\{K_{\text{neutral}}, K_{f1}, \ldots, K_{fn}\}$, Table 1 reports how a post *p* is associated with one of the factions $f_1, \ldots, f_n$ or classified as *neutral*. A post is classified as f_i if it contains at least one keywords in K_{fi}, possibly some keywords in K_{neutral}, but no one in other factions $\{K_{f1}, \ldots, K_{fi-1}, K_{fi+1}, \ldots, K_{fn}\}$. A post is categorized as *neutral* if has keywords in K_{neutral} and/or keywords in two or many factions $\{K_{f1}, \ldots, K_{fn}\}$. Although there are other approaches in the literature for classifying a post (e.g., manually or with text mining techniques), through our approach, a post is classified in favor of a given faction by analyzing the keywords contained in it, i.e., only if it shows a clear voting indication for that faction; otherwise, we consider the post as *neutral*.

In the case of two-faction events, a post may be labeled as one of the three classes $\{neutral, f_1, f_2\}$. Considering the keywords $\{K_{\text{neutral}}, K_{f1}, K_{f2}\}$, Table 2 reports how a post *p* is associated with one of the two factions f_1, f_2 or classified

Table 2 Classification of a post by analyzing its keywords in a two-faction event

K_{neutral}	K_{f1}	K_{f2}	Class
–	X	–	f_1
X	X	–	f_1
–	–	X	f_2
X	–	X	f_2
X	–	–	Neutral
–	X	X	Neutral
X	X	X	Neutral

as *neutral*. A post is classified as f_1 if it contains at least one keywords in K_{f1} and possibly some keywords in K_{neutral}. Similarly, a post is classified as f_2 if it contains at least one keywords in K_{f2} and possibly some keywords in K_{neutral}. A post is categorized as *neutral* if it has keywords in K_{neutral} and/or keywords in all the two factions $\{K_{f1}, K_{f2}\}$.

2.4 Data analysis and mining of dataset *D*

After having built the input dataset *D*, it is analyzed through algorithms and techniques for discovering the polarization of social network users and news sites during political campaigns characterized by the rivalry of different factions. In particular, the main goals of this step are as follows.

1. *Analysis of aggregate data D* is analyzed to derive statistics about data and to discover the main arguments used by the different factions whose posts are present in *P*.
2. *Statistical significance of collected data* The goal is to assess the significance of *D*.
3. *Temporal analysis* The goal is to analyze how the number of posts supporting the different factions vary along time.
4. *Polarization of users* Collected data are analyzed to discover how users are polarized toward the different factions.
5. *Mobility flows* The evolution of users' polarization is studied in the weeks preceding the political event.
6. *Polarization prediction* The goal is to predict the polarization of users before the political event.
7. *Polarization of news sites* Collected data are analyzed to discover how news site is polarized toward the different factions.

2.5 Results visualization

Results visualization is performed by the creation of info-graphics aimed at presenting the results in a way that is easy to understand to the general public, without providing complex statistical details that may be hard to understand to the intended audience. The graphic project is grounded on some of the most acknowledged and ever-working principles underpinning a 'good' info-graphic piece. In particular, we follow three main design guidelines: (1) preferring a visual representation of the quantitative information to the written one; (2) minimizing the cognitive efforts necessary to decoding each system of signs; and (3) structuring the whole proposed elements into graphic hierarchies (Cesario et al. 2016).

Displaying quantitative information by visual means instead of just using numeric symbols—or at least a combination of the two approaches—has been proven extremely useful in providing a kind of sensory evidence to the inherent abstraction of numbers, because this allows everybody to instantly grasp similarities and differences among values. In fact, basic visual metaphors (e.g., the largest is the greatest, the thickest is the highest) enable more natural ways of understanding and relating sets of quantities (Tufte 198).

3 Case of study: Italian constitutional referendum, 2016

We applied the methodology described in the previous section to the constitutional referendum that was held in Italy on December 4, 2016. The Italian voters were asked whether they approve a constitutional law that amends the Italian Constitution to reform the composition and powers of the Parliament of Italy, as well as the division of powers between the State, regions, and administrative entities.[1] The main supporter of the referendum (i.e., in favor of *yes*) was the Democratic Party (in Italian Partito Democratico, PD) and its leader and Italian Prime Minister Matteo Renzi, on the other hand, in favor of *no* the main opposition parties (e.g., Movimento 5 Stelle, Forza Italia) and different citizen committees. The referendum saw a high voter turnout (approximately 65% of voters) and a majority of the votes opposed to the reform (i.e., voting *no*), which exceeded 59% of the expressed preferences. A political effect of the referendum's result was the resignation of the Italian prime minister.

The political event under analysis $\mathcal{P}$ is a two-faction event $F = \{yes, no\}$. We collected the main keywords K used as hashtags in tweets related to $\mathcal{P}$. Such keywords have been grouped as follows:

– $K_{\text{neutral}} = \{$#referendumcostituzionale, #siono, #riformacostituzionale, #referendum, #4dicembre, #referendum4dicembre$\}$
– $K_{\text{yes}} = \{$#bastaunsi, #iovotosi,#italiachedicesi, #iodicosi, #leragionidelsi$\}$
– $K_{\text{no}} = \{$#iovotono, #iodicono, #bastaunno, #famiglieperilno, #leragionidelno$\}$

Table 3 Examples of tweets on the Italian constitutional referendum

Text	Timestamp	Keywords	URLs	Class
Why is important to be well informed on #Referendum-Costituzionale	25 Oct 2016 08:00:00	#Referendum Costituzionale	youtube.com/…	Neutral
#IoVotoNO: all the reasons to vote against this reform	15 Nov 2016, 09:00:00	#iovotono	ilfattoquotidiano.it/… ilgiornale.it/…	No
Now, wait the results! #referendum4dicembre #iovotoNO #democrazia	3 Dec 2016 10:00:00	#referendum 4dicembre #iovotono #democrazia	–	No

Table 4 Statistics about collected tweets

Filter	#Tweets	#Retweets	#Favorites	Total
None	338,592	826,584	987,010	2,152,186
Contains co-hashtags	146,687	449,198	518,088	1,113,973
Contains URL	74,973	139,417	148,888	363,278

Given the keywords $K = K_{neutral} \cup K_{yes} \cup K_{no}$, we collected 338,592 tweets containing at least one of these keywords posted from October 23 (5 weeks before the voting day) to December 3, 2016 (1 day before). The tweets were not collected in real time, but with a delay of 24 h after their publications so as to capture: (1) the number of *retweet*, which indicates how many users shared a tweet with their friends; (2) the number of *favorites*, which indicates how many users found a tweet useful.

Collected tweets were preprocessed as described in Sect. 2.3. For instance, Table 3 shows three tweets published by a user u_i before the voting day (translated in English for the Reader's convenience). For each tweet, the main fields have been reported in the table.

In the first tweet, u_i expresses the importance of going to vote by using a *neutral* hashtag (#ReferendumCostituzionale) and including a Youtube URL. This tweet is classified as *neutral*. In the second tweet, u_i shows his/her dissatisfaction with the reform by using a hashtag supporting *no* (#iovotono) and two news sites for motivating his/her voting intention. It is classified as in favor of *no*. The third tweet contains a *neutral* hashtag (#referendum4dicembre), a hashtag supporting *no* (#iovotono) and a co-hashtag (#democrazia). This tweet is classified as in favor of *no*.

4 Analysis and results

4.1 Analysis of aggregate data

Table 4 reports some statistics about the tweets collected: 338,592 are tweets, 826,584 are retweets and 987,010 are favorites. Filtering the data, we discovered that 43% of tweets contain co-hashtags (e.g., #democrazia, #renzi) and 22% contain URLs. Co-hashtags are useful to understand
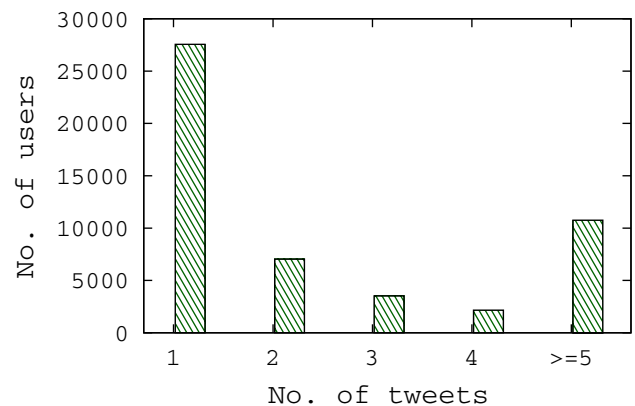
Fig. 1 No. of tweets posted by users

the arguments used in favor of one or another position. The URLs allows understanding what news site were used by users to support their voting intentions. The number of users under analysis is 50,717 (139,066 considering also the retweets). Figure 1 shows that more than half (54%) of the users published only one tweet on the referendum, 14% two tweets, 7% three, 4% four and 21% five or more tweets.

Table 5 reports some statistics about the main hashtags used for collecting tweets, grouped in *yes*, *neutral* and *no*. Next to each hashtag, the number of tweets, retweets and favorites containing such hashtags are reported. The percentage of tweets published with *yes* or *neutral* hashtags are similar (respectively, 23 and 24%) and are both half of those in favor of *no* (53%). We also studied how users have used these hashtags to write their tweets: 88% of tweets contain only one or more hashtags of a group (*yes*, *neutral* or *no*), 11% of tweets contain hashtags of two groups (yes/neutral, no/neutral or yes/no), and 1% of tweets contain hashtags of all the groups (yes/neutral/no).

Table 6 shows the main ten co-hashtags used by Twitter users, divided into *yes*, *neutral* and *no*. We note that, in many cases, users who supported *yes* did it by posting tweets reporting Prime Minister's statements (e.g., #matteorisponde), the opportunity to improve the political system (e.g., #avanti) or information propagated by opponents (e.g., #lebufaledelno). On the other hand, users supporting *no* posted tweets reporting positions from the political

Table 5 Main hashtags related to yes, neutral and no

Hashtag	#Tweets	#Retweets	#Favorites	Total
#bastaunsi	37,268	94,730	133,774	265,773
#iovotosi	38,373	64,419	95,479	198,273
[All hashtags supporting yes]	76,257	161,306	231,875	469,445
#referendumcostituzionale	36,283	61,940	68,967	167,191
#siono	14,678	28,958	44,460	88,096
#riformacostituzionale	12,233	29,232	30,248	71,715
#referendum	9727	26,440	27,241	63,409
#4dicembre	7028	24,715	29,889	61,633
[All neutral hashtags]	81,764	175,123	205,157	462,050
#iovotono	152,638	379,988	430,268	962,895
#iodicono	26,574	107,669	117,233	251,476
[All hashtags supporting no]	180,562	490,147	549,972	1,220,684

Table 6 Main co-hashtags related to yes, neutral and no tweets

Category	Co-hashtags
Yes	#renzi, #sivainpiazza, #matteorisponde, #leopolda7, #avanti, #midiconoche, #m5s, #matteorenzi, #pd, #lebufaledelno
Neutral	#agcom, #serracchiani, #renzi, #pd, #mafiacapitale, #mafia, #accozzagliachi, #bufale, #bastapocochechevo, #themancettacandidate
No	#renzi, #salvini, #m5s, #movimentonesti, #trenotour, #costituzione, #nonrubo, #pd, #renziacasa, #deluca

opposition (e.g., #m5s, #salvini), willing to leave the constitution as it is (e.g., #costituzione) or hoping to send the prime minister home (e.g., #renziacasa). The *neutral* co-hashtags highlight topics that were treated during the referendum campaign.

4.2 Statistical significance of collected data

The goal is to assess the statistical significance of the input dataset. Specifically, we studied whether the Twitter users captured in our analysis were actual voters of the referendum, i.e., whether they were Italian citizens able to vote (at least 18 years old).

From the metadata present in the tweets used in our analysis, we extracted aggregate information on the language used to write them and on the location of users who wrote them. Specifically, from the tweet metadata we analyzed the *lang* field,[2] which is a language identifier corresponding to the machine-detected language of the Tweet text (e.g., "en" for English, "it" for Italian, "und" if no language could be detected). In addition, from the user metadata we analyzed the *location* field,[3] which indicates the user-defined location for the accounts profile (e.g., San Francisco, CA).

By analyzing the metadata described above, we can say that:

– All the tweets under analysis have the *lang* field equals to "it" (Italian). The Italian language is mainly used by Italians who reside in Italy (60 million) or abroad (about 4 million). Italian is used as first language[4] only by a small part of Swiss (about 640,000 people), and a very small part of Croats and Slovenes (about 22.000 people).
– 98% of users who have defined the location in their profile live in Italy.

To further show the statistical value of user locations, in Table 7 we compared the number of Twitter users captured in our analysis with the total number of citizens grouped by Italian regions. There is a strong correlation (Pearson coefficient 0.9) between these sets of data. Similar results are obtained by comparing the number of users and the total number of citizen grouped by Italian cities. Also in this case, as shown in Table 8 there is a very strong correlation between these sets of data (Pearson coefficient 0.96).

These statistics give us strong indications about the users analyzed in our case study: It is highly likely that they are voters of the referendum that is adult Italians citizen.

---

[2] Twitter API, https://dev.twitter.com/overview/api/tweets.

[3] Twitter API, https://dev.twitter.com/overview/api/users.

[4] Italian language, https://it.wikipedia.org/wiki/Lingua_italiana.

Table 7 Comparison of the number of users and the total number of citizens grouped by region

Region	No. of users	No. of citizen
Lazio	4169	5,893,935
Lombardy	4129	10,014,304
Campania	1739	5,840,219
Tuscany	1628	3,743,370
Emilia-Romagna	1621	4,447,419
Sicily	1431	5,055,838
Veneto	1331	4,907,284
Piedmont	1186	4,394,580
Apulia	1174	4,066,819
Sardinia	675	1,654,587
Liguria	671	1,565,566
Calabria	565	1,966,819
Friuli-Venezia G.	449	1,218,068
Marches	396	1,539,316
Abruzzo	380	1,322,585
Umbria	310	889,817
Trentino-S. Tyrol	255	1,061,318
Basilicata	202	571,133
Aosta Valley	76	126,732
Molise	73	310,685

Table 8 Comparison of the number of users and the total number of citizens grouped by cities (only 20 of the major Italian cities)

City	No. of users	No. of citizen
Rome	3499	2,874,529
Milan	2221	1,353,467
Naples	747	969,456
Turin	548	885,651
Florence	486	382,346
Bologna	452	388,567
Palermo	348	672,398
Genoa	313	582,870
Bari	215	323,503
Catania	198	312,895
Cagliari	188	154,194
Padua	185	209,475
Venice	177	261,496
Verona	172	257,815
Bergamo	168	120,358
Brescia	159	196,205
Modena	127	184,642
Trieste	125	203,974
Udine	123	99,245
Salerno	119	134,857

Regarding the last point, statistics show that 96% of Italian Twitter users are adults.[5]

4.3 Temporal Analysis

Figure 2 shows the time series of the number of tweets published during the 5 weeks preceding the referendum. The tweets in the figure are classified as supporting *yes* (solid blue line), *neutral* (black dashed line), or *no* (solid red line). A fourth time series on all the tweets is represented as a solid black line.

All four time series have a similar growing trends (Pearson coefficients of the *yes*, *neutral* and *no* series versus the all series range from 0.87 to 0.97) and show some peaks in the following dates:

– 29th October: It was the day after a major television confrontation between Matteo Renzi in favor of *yes* and the former PM Ciriaco De Mita in favor of *no*;[6]
– 12th and 23rd November: Debates and discussions in different cities of Italy in favor of *yes* or *no*;
– 2nd December: The last day to make propaganda before the election silence day (3rd December).

We observe that, during the whole observation period, tweets supporting *no* were more than those supporting *yes* or *neutral*. Statistically, every day the number of tweets supporting *neutral* or *yes* is similar, and they both are half of the tweets supporting *no*.

Figure 3 shows the number of tweets aggregated by week day. The interest on referendum increases from Monday to Friday and then decreases during the weekend.

4.4 Polarization of users

Polarization of a user $\rho_u \in [-1, 1]$ is defined as

$$\rho_u = 2 \times \frac{|yes_u|}{|yes_u| + |no_u|} - 1,$$

where $|yes_u|$ and $|no_u|$ represent, respectively, the number of tweets published by u classified as *yes* and *no* (Bessi et al. 2015). A value of ρ_u close to 1 means that user u tends to be polarized toward *yes*, while when ρ_u is close to -1 it means that user u is polarized toward *no*. In all the analyses of our paper, we focused on users who showed a strong polarization toward a given faction. For this reason, we chose a
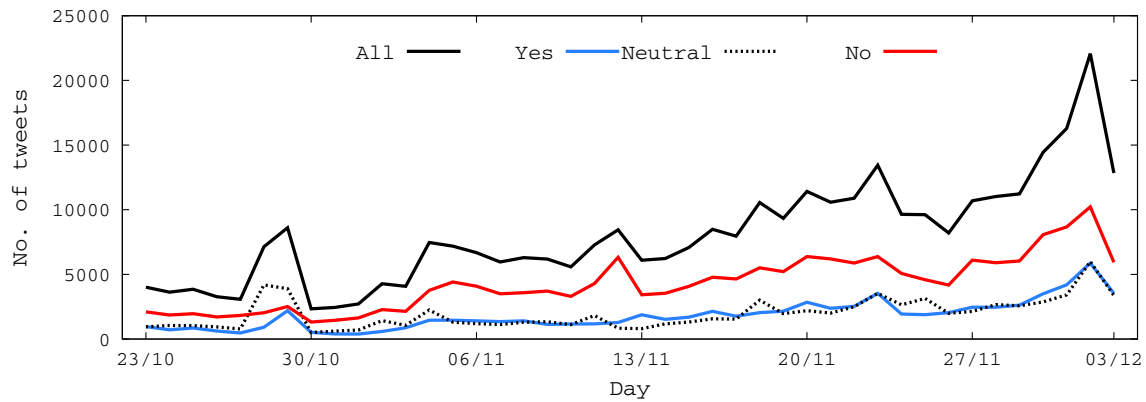
Fig. 2 Time series of tweets published from 23rd October to 3rd December 2016
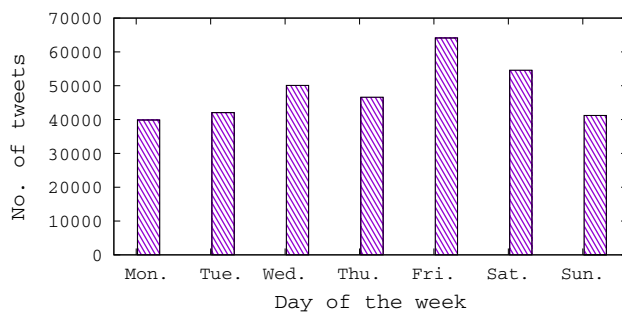
Fig. 3 No. of tweets per week day

Fig. 4 Probability density function of the users' polarization

Fig. 5 Complementary cumulative distribution function of the number of tweets published by users polarized toward yes and toward no

polarization, and another one polarized toward *no*. Specifically, the 48% of users under analysis have a strong polarization toward *no*, 25% toward *yes*, and 27% are *neutral*.

Figure 5 illustrates production patterns of polarized users. In particular, the figure shows the complementary cumulative distribution function (CCDF) of the number of tweets published by users polarized toward *yes* and toward *no*. Both curves point out very similar production patterns between users polarized toward *yes* and users polarized toward *no* (Lievrouw et al. 2014). The number of tweets posted by a user does not depend on its polarization: There are a similar number of users who have published at least x tweets among users polarized both toward *no* and toward *yes*.

4.5 Mobility flows

Figure 6 represents the evolution of users' polarization in the 5 weeks preceding the referendum. To study the mobility flows of users, we restricted our analysis on users who have

high threshold (0.9) to select users with strong polarization in favor of *yes* or *no*. Specifically, we consider users with $\rho_u > 0.9$ as polarized toward *yes*, users with $\rho_u < -0.9$ as polarized toward *no*, otherwise *neutral*. Figure 4 shows the probability density function of the users' polarization. We observe a trimodal distribution, indicating that a group of users are polarized toward *yes*, another one has a *neutral*
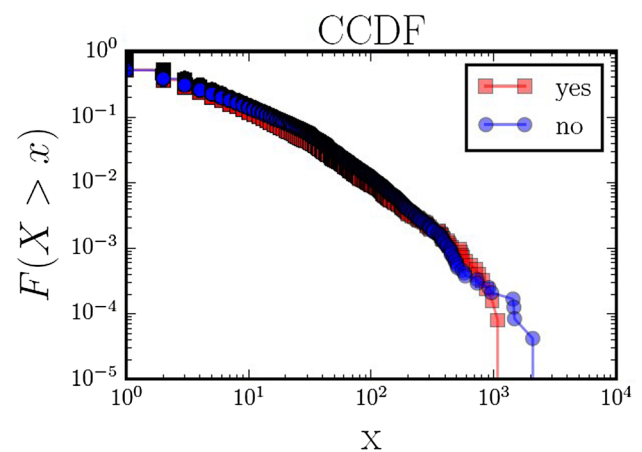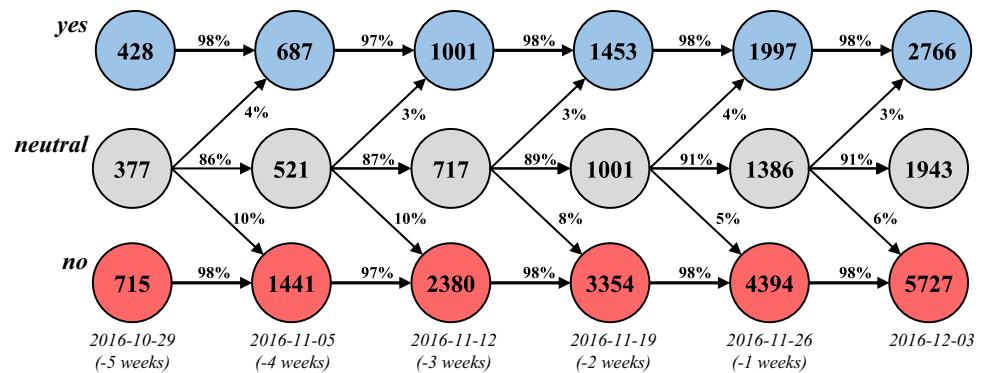
Fig. 6 Evolution of users' polarization in the 5 weeks preceding the referendum day: Users polarized in favor of yes (blue circles), in favor of no (red circles), and neutral (gray circles)

published at least 5 tweets (i.e., 10,436 users). The figure shows how vary the number of users polarized toward *yes* (blue circles), the number of users polarized toward *no* (red circles), and the number of *neutral* users (gray circles) in the 5 weeks preceding the referendum. Arrows in the figure show the percentage of users who after 1 week are polarized as in the previous week and the percentage of *neutral* users who move toward *yes* or *no*. We do not report the moving from *yes* toward *no* (and viceversa) because they are low numbers (less than 3%). Notice that the number of users under analysis increase from week to week (from 1520 to 10,436), because by collecting new tweets we are able to categorize new users.

We observe that, over the 5 weeks preceding the referendum, the vast majority of users polarized toward *yes* and *no* tend to maintain their polarization. The biggest changes occur only among users categorized as *neutral*: 10% of neutral users moves toward the *yes* and 20% toward the *no*.

We can conclude that almost all users polarized toward *no* have not changed position during the weeks preceding the vote, and one fifth of the *neutral* users moved toward *no*. Also supporters of the *yes* were very compact, while a lower number of *neutral* users have moved to *yes*.

4.6 Polarization prediction

The goal of this section is to predict the polarization of users before the referendum day. Different machine learning techniques has been studied to evaluate their appropriateness in the considered domain. Among those, some classification algorithms have been tested and the Random Forest (RF) (Breiman 2001) algorithm was selected as it achieved the best performance in terms of accuracy and recall, with limited model building time. Other research works exploited RF for social media analysis due to its high level of accuracy (e.g., see Zhang et al. 2011; Gokulakrishnan et al. 2012; Monti et al. 2013), (Kwon et al. 2013).

Random Forests have been trained for predicting the polarization that a user will have before the voting day, by using information available n weeks before the referendum,

where n varies from 5 to 1. Specifically, we trained five Random Forest models (one for each value of n), each of them trained from this information:

– The input is composed by aggregate information contained in tweets posted by a user at least n weeks before the referendum. This information is: (1) number of tweets containing *yes* hashtags, (2) number of tweets containing *no* hashtags, (3) number of tweets containing *neutral* hashtags, (4) total number of tweets and (5) number of hashtags used.
– The class is a label that indicates the final polarization of a user (*yes*, *no* or *neutral*) calculated by our methodology using all the information contained in all the tweets posted by a user (i.e., it does not depend on n).

To fine-tune the model, we performed a grid search over the parameters' space and we found that the best results are provided by a Random Forest using the entropy criterion and 128 estimators.

Figure 7 shows the classification performance achieved by RF models at different times. In particular, we show micro- and macro-averaging (Van Asch 2013) of the area under the curve (AUC) computed for the model trained with information available at different times. Results are averaged over 10 Monte Carlo cross-validation iterations and indicate that the information available 5 weeks before the referendum day provide a classification performance of 0.849 ± 0.006 (micro-AUC) and 0.83 ± 0.006 (macro AUC). Such a classification performance increases with the amount of information available, reaching the value of 0.962 ± 0.002 (micro-AUC) and 0.949 ± 0.001 (micro-AUC) 1 week before the referendum day.

4.7 Polarization of news sites

Table 9 reports some statistics about tweets containing URLs from the main Italian news sites. Almost 3/4 of such tweets contain URLs from five of the major news sites: beppegrillo.it (36%), ilfattoquotidiano (17%), repubblica.

Fig. 7 User polarization prediction achieved by a Random Forest model using information posted by users from 5 to 1 week before the referendum day
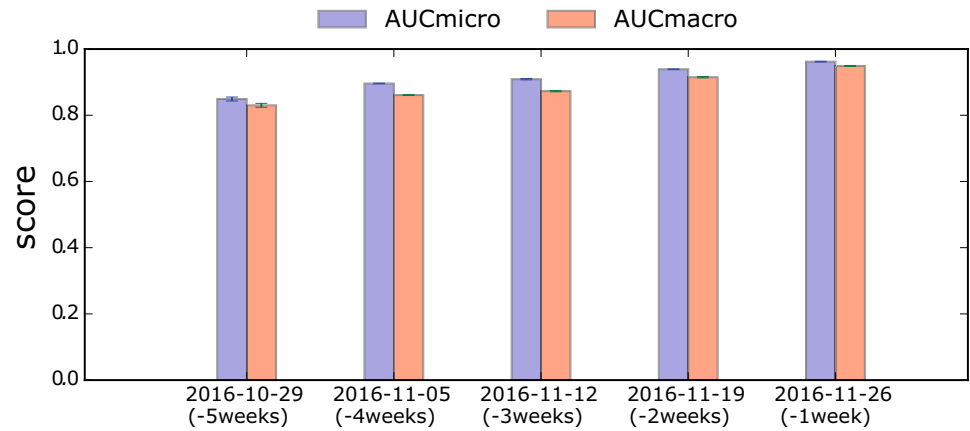

Table 9 Top 15 news sites used by Twitter users during the referendum campaign

Hashtag	#Tweets	#Retweets	#Favorites	Total
beppegrillo.it	4244	12,990	13,575	30,810
ilfattoquotidiano.it	2027	8935	7495	18,457
repubblica.it	1468	2537	2571	6576
huffingtonpost.it	957	3150	2763	6870
corriere.it	558	1083	1235	2876
unita.tv	509	1992	2716	5218
ilgiornale.it	482	873	764	2120
ansa.it	269	668	606	1543
ilsole24ore.it	268	386	349	1004
formiche.net	216	243	204	664
movimento5stelle.it	206	709	593	1508
lastampa.it	189	438	526	1153
possibile.com	173	804	627	1604
linkiesta.it	173	428	358	959
affaritaliani.it	143	606	541	1290

it (12%), huffingtonpost.it (8%) and corriere.it (5%). Since we have registered a greater presence of tweets supporting *no*, the popularity of news sites was been affected if the magazine has written articles close to the positions of *no*.

Given a news site *s*, we compute its polarization as follows

$$\rho_s = 2 \times \frac{|yes_s|}{|yes_s| + |no_s|} - 1,$$

where $|yes_s|$ and $|no_s|$ represent, respectively, the number of tweets classified as *yes* and *no* that contain a URL linking to the news site *s*. Figure 8 shows the polarization of the main Italian news sites for each category (*yes*, *neutral* and *no*). The figure highlights that some journals had a strong polarization toward *yes* (unita.tv, ilsole24ore.it and linkiesta.it), some others had a *neutral* position (lastampa.it, corriere.
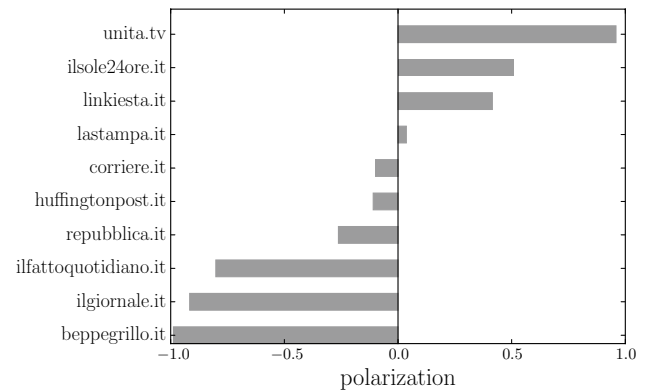

**Fig. 8** Polarization of the main Italian news sites for each category (yes, neutral and no)

it, huffingtonpost.it and repubblica.it) and others toward *no* (ilfattoquotidiano.it, ilgiornale.it and beppegrillo.it). This result can be explained in two ways: news sites that for editorial choices have supported the campaign of *yes* or *no*, or readers of a certain news site that for political reasons are closer to a certain position.

Figure 9 shows the evolution of the polarization of four representative news sites over the 5 weeks preceding the referendum day. The figure clearly indicates that the polarization of news sites does not show relevant changes over time.

5 Related work

In recent years, the use of social media for measuring public opinion has become one of the hot topics in social network research (Anstead and O'Loughlin 2015). In particular, two are the main areas of research related to this paper: (1) the use of social media to measure public opinion and predict election results and (2) the impact of social media on news consumption and on how information spreads through social

Fig. 9 Time series of the polarization of four representative Italian news sites
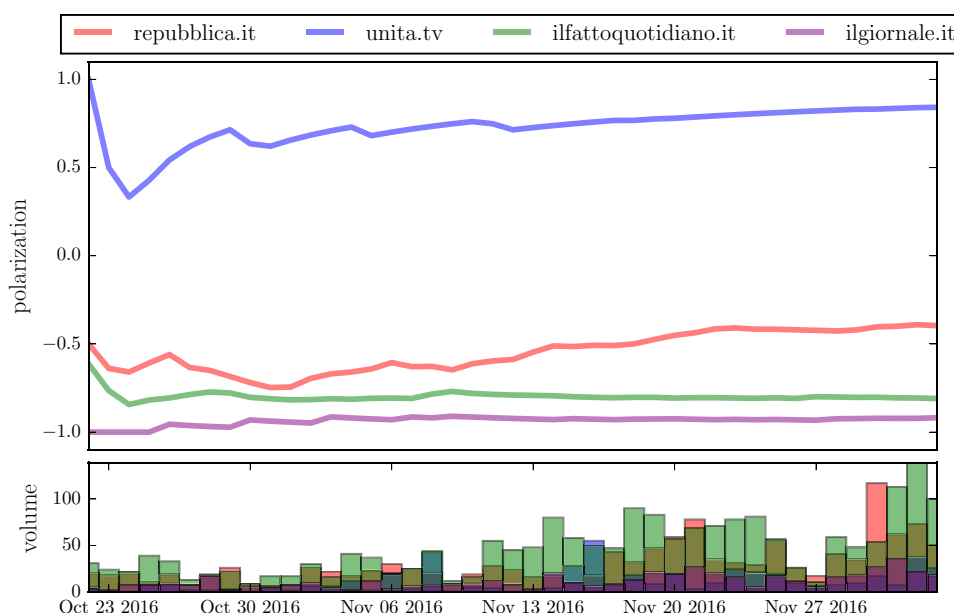
networks. For each of these research areas, the main related work has been described.

Murphy et al. (2014) examined the potential impact of social media on public opinion research, as an important way for facilitating and/or replacing traditional survey research methods. The authors highlighted several problems related to this topic, for example, (1) not every member of the public uses social network platforms; (2) incomplete and not accurate information published by social users; (3) legal regulations about data collected. O'Connor et al. (2010) correlated Twitter data with several public opinion time series; Anstead and O'Loughlin (Anstead and O'Loughlin 2015), by analyzing the 2010 UK election, suggested the use of social media as a new way to understand public opinion. Others related work attempted to measure the publics evolving response to stimuli, examining both short-term events such as TV political debates (Elmer 201) or long-term events such as economic downturns (Gonzalez-Bailon et al. 2010). An indepth survey on this topic can be found in Jungherr (2016).

Hermida et al. (2012) have examined the impact of social media on news consumption, based on an online survey of 1600 Canadians. The study highlights that social networks are a significant source of news for Canadians: Two-fifths of users under analysis said that they received news from users who they follow, while a fifth got information from news organizations and individual journalists who they follow. Lerman and Ghosh (Lerman and Ghosh 2010) studied the spread of information on social networks and if their network structure affect how information is disseminated. Specifically, they extracted the active users and track how the interest in news spreads among them. Howard et al. (Howard et al. 2011) studied the effects of social media during the

Arab Spring.[7] By analyzing users posts from different social networks, the authors have reached three main conclusions: (1) Social media played a central role in guiding the political debates during such event; (2) a spike in online conversations often preceded major events in the real world; and (3) social media helped to accelerate the spreading of news and ideas in the world.

To highlight the level of novelty of the methodology we proposed, in the following we review some of the most related research works by discussing differences and similarities with our work. Ceron et al. (Ceron et al. 2014) used a text analysis approach (Hopkins and King 2010) for studying the voting intention of French Internet users in both the 2012 Presidential ballot and the subsequent legislative election. The authors mainly present the results of their analysis by comparing them with official data and predictions made by survey companies. Very few are the implementation details, e.g., it is not clear how the statistical value of data was assessed and how tweets and users were classified. Gruzd and Roy (2014) investigated the political polarization of social network users during the Canadian Federal Election, 2011. A sample of tweets posted by 1492 Twitter users were manually classified based on their self-declared political views and affiliations. The methodology we proposed allows to classify a user automatically by analyzing the posts he/she published—and the keywords he/she used—in the weeks preceding the vote. Nulty et al. (2016) surveyed the European landscape of social media using tweets originating from and referring to political actors during the 2014 European Parliament election campaign. With respect to

---

[7] https://en.wikipedia.org/wiki/Arab_Spring.

our paper, these authors do not present a methodology, but only a hashtag analysis per languages, political parties and candidates. Burnap et al. (Burnap et al. 2016) used Twitter data to forecast the outcome of UK General Election, 2015. The authors applied an automated sentiment analysis tool for classifying tweets. Differently from this methodology, we classified user posts—and consequently users who wrote these posts—by taking advantage of the keywords related to the political event under analysis. Similar considerations can be made for Tumasjan et al. (2010) that analyzed about 100.000 political tweets on 2009 German federal election using a text analysis software. Kagan et al. (2015) exploited Twitter data for predicting the electoral results of 2013 Pakistan and 2014 Indian elections. The authors studied how the support for a candidate (or opposition to a candidate) was spreading through Internet. In fact, the diffusion model proposed classifies a user by taking into account also the percentage of his/her neighbors (i.e., friends) that have expressed a positive/negative opinion on a candidate/faction. With respect to this work, our methodology evaluates only the content of posts published by a user, but it could be extended considering the opinion of friends of such user. Wagner (Wagner 2017) studied the 2014 Scottish independence referendum for understanding how local newspapers supported the campaign of the referendum. Specifically, the author has analyzed the political position of two local Scottish newspapers (i.e., The Courier and Evening Telegraph), by counting how many stories were neutral, in favor of, and opposed to Scottish independence. With respect to our work, this is a traditional approach that analyzes the textual content of articles published by a news site. We proposed to evaluate the political position of a news site by analyzing how users referred to such news site for supporting their voting intentions. Similar considerations can be made for Dallmann et al. (2015) that analyzed the behavior of four leading German online newspapers over a timespan of four years.

In summary, this paper presents a methodology aimed at analyzing the polarization of social network users and news sites during political campaigns characterized by the rivalry of two factions (e.g., referenda and ballots). Unlike works in literature that classify a post manually (Gruzd and Roy 2014) or with text mining techniques (Ceron et al. 2014; Burnap et al. 2016; Tumasjan et al. 2010), our methodology exploits keywords (e.g., hashtags) contained in a post to classify it in favor of a faction. In this way, a post is classified in favor of a faction only if it shows a clear voting indication for a such faction, otherwise we consider the post as neutral. With regard to studying the polarization of news sites, different works in literature use a direct approach that analyzes the contents of articles published by such news sites to understand their political orientation (Wagner 2017; Dallmann et al. 2015). Our approach instead uses a novel approach that analyzes how users referred to these news sites

in their posts for supporting their voting intentions. Other aspects of novelty of the methodology are some analyses we have proposed:

– *Statistical significance of collected data* to study the statistical significance of data used in our analysis. It gives strong indications about the users and if they are voters of the political event under analysis.
– *Mobility flows* to analyze the evolution of users' polarization in the weeks preceding the political event. It allows to study if users maintained the same polarization or if they changed their opinion.
– *Polarization prediction* to predict the polarization of users before the political event. This allows understanding with what precision the polarization of a user can be predict, using information available some weeks before the vote.

The whole methodology and all its analysis have been applied to a real application case such as the Italian constitutional referendum, 2016. We studied the behavior of 50,717 Twitter users by analyzing the 338,592 tweets posted on the referendum by them in the 5 weeks preceding the vote. The results demonstrate the applicability of our methodology in discovering the behavior of social network users and how news sites are used during political campaigns.

6 Conclusions

Social media analysis is an important research area aimed at extracting useful information from the big amount of data gathered from social networks. Recent years have seen a great interest from academic and business world in using social media to measure public opinion.

This paper presents a methodology aimed at analyzing the polarization of social network users and news sites during political campaigns characterized by the rivalry of different factions. On the one hand, the methodology allows to study the users' polarization before a political event, what arguments they used to support their voting intentions, and if such intentions change in the weeks preceding the vote. On the other hand, the methodology permits to analyze the effects of news sites on important political events, that is, how many users used information from news sites and what news sites can be considered in favor, against or neutral to a given faction.

The methodology has been validated with an important case study as the Italian constitutional referendum, 2016. According to our study, 48% of Twitter users were polarized toward *no*, 25% toward *yes*, and 27% had a neutral behavior. Regarding the change of opinion in the weeks preceding the vote, the majority of users categorized as supporters of

no or *yes* have never changed during the weeks preceding the vote, while a consistent part of the *neutral* users moved toward *no* (20%) and toward *yes* (10%). A second goal was to understand the effects of news sites on the referendum campaign. The analysis has shown that some news sites had a strong polarization toward *yes* (unita.tv, ilsole24ore.it and linkiesta.it), some others had a *neutral* position (lastampa.it, corriere.it, huffingtonpost.it and repubblica.it) and others were oriented toward *no* (ilfattoquotidiano.it, ilgiornale.it and beppegrillo.it). The polarization of news sites has remained almost unchanged in the weeks preceding the vote.

References

Anstead N, O'Loughlin B (2015) Social media analysis and public opinion: the 2010 UK general election. J Comput Mediat Commun 20(2):204–220. https://doi.org/10.1111/jcc4.12102

Belcastro L, Marozzo F, Talia D, Trunfio P (2017) Big data analysis on clouds. In: Sakr S, Zomaya A (eds) Handbook of big data technologies. Springer, Berlin, pp 101–142. https://doi.org/10.1007/978-3-319-49340-4_4 ISBN: 978-3-319-49339-8

Bessi A, Coletto M, Davidescu GA, Scala A, Caldarelli G, Quattrociocchi W (2015) Science versus conspiracy: collective narratives in the age of misinformation. PloS One 10(2):e0118093

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Burnap P, Gibson R, Sloan L, Southern R, Williams M (2016) 140 Characters to victory? Using twitter to predict the UK 2015 general election. Electoral Stud 41:230–233

Ceron A, Curini L, Iacus SM, Porro G (2014) Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizenspolitical preferences with an application to Italy and France. New Media Soc 16(2):340–358

Cesario E, Congedo C, Marozzo F, Riotta G, Spada A, Talia D, Trunfio P, Turri C (2015) Following soccer fans from geotagged tweets at fifa world cup 2014. In: Proceedings of the 2nd IEEE conference on spatial data mining and geographical knowledge services, pp 33–38. Fuzhou, China. ISBN 978-1-4799-7748-2

Cesario E, Iannazzo AR, Marozzo F, Morello F, Riotta G, Spada A, Talia D, Trunfio P (2016) Analyzing social media data to discover mobility patterns at expo 2015: methodology and results. In: The 2016 international conference on high performance computing and simulation (HPCS 2016), pp. 230–237. Innsbruck, Austria. ISBN: 978-1-5090-2088-1.

Dallmann A, Lemmerich F, Zoller D, Hotho A (2015) Media bias in German online newspapers. In: Proceedings of the 26th ACM conference on hypertext and social media, pp 133–137. ACM

Elmer G (2013) Live research: Twittering an election debate. New Media Soc 15(1):18–30. https://doi.org/10.1177/1461444812457328

Franch F (2013) 2010 UK election prediction with social media. J Inform Technol Polit 10(1):57–71. https://doi.org/10.1080/19331681.2012.705080

Gokulakrishnan B, Priyanthan P, Ragavan T, Prasath N, Perera A (2012) Opinion mining and sentiment analysis on a twitter data stream. In: 2012 International conference on advances in ICT for emerging regions (ICTer), pp 182–188. IEEE

Gonzalez-Bailon S, Banchs RE, Kaltenbrunner A (2010) Emotional reactions and the pulse of public opinion: measuring the impact of political events on the sentiment of online discussions. In: CoRR, abs/1009.4019

Gruzd A, Roy J (2014) Investigating political polarization on twitter: a Canadian perspective. Policy Internet 6(1):28–45

Hanna R, Rohm A, Crittenden VL (2011) Were all connected: the power of the social media ecosystem. Bus Horiz 54(3):265–273

Hermida A, Fletcher F, Korell D, Logan D (2012) Share, like, recommend. J Stud 13(5–6):815–824. https://doi.org/10.1080/1461670X.2012.664430

Hopkins DJ, King G (2010) A method of automated nonparametric content analysis for social science. Am J Polit Sci 54(1):229–247

Howard PN, Duffy A, Freelon D, Hussain MM, Mari W, Maziad M (2011) Opening closed regimes: what was the role of social media during the arab spring? SSRN. https://doi.org/10.2139/ssrn.2595096

Jungherr A (2016) Twitter use in election campaigns: a systematic literature review. J Inform Technol Polit 13(1):72–91. https://doi.org/10.1080/19331681.2015.1132401

Kagan V, Stevens A, Subrahmanian V (2015) Using twitter sentiment to forecast the 2013 Pakistani election and the 2014 Indian election. IEEE Intell Syst 30(1):2–5

Kwon S, Cha M, Jung K, Chen W, Wang Y (2013) Prominent features of rumor propagation in online social media. In: 2013 IEEE 13th international conference on data mining (ICDM), pp 1103–1108. IEEE

Lerman K, Ghosh R (2010) Information contagion: an empirical study of the spread of news on digg and Twitter social networks. In: CoRR, abs/1003.2664

Lievrouw L, Gillespie T, Boczkowski P, Foot K (2014) Materiality and media in communication and technology studies: an unfinished project. Media technologies: essays on communication, materiality, and society, pp 21–51

Monti C, Rozza A, Zappella G, Zignani M, Arvidsson A, Colleoni E (2013) Modelling political disaffection from twitter data. In: Proceedings of the second international workshop on issues of sentiment discovery and opinion mining, p 3. ACM

Murphy J, Link MW, Childs JH, Tesfaye CL, Dean E, Stern M, Pasek J, Cohen J, Callegaro M, Harwood P (2014) Social media in public opinion research executive summary of the aapor task force on emerging technologies in public opinion research. Public Opin Q 78(4):788. https://doi.org/10.1093/poq/nfu053

Nulty P, Theocharis Y, Popa SA, Parnet O, Benoit K (2016) Social media and political communication in the 2014 elections to the European parliament. Electoral Stud 44:429–444

O'Connor B, Balasubramanyan R, Routledge BR, Smith NA (2010) From tweets to polls: linking text sentiment to public opinion time series. ICWSM 11(122–129):1–2

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inform Retr 2(1–2):1–135. https://doi.org/10.1561/1500000011

Talia D, Trunfio P, Marozzo F (2015) Data analysis in the cloud. Elsevier, Amsterdam

Tufte ER (1986) The visual display of quantitative information. Graphics Press, Cheshire

Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with Twitter: what 140 characters reveal about political sentiment. ICWSM 10(1):178–185

Van Asch V (2013) Macro-and micro-averaged evaluation measures. Technical report

Wagner JP (2017) The media and national identity: local newspapers coverage of Scottish independence during the campaign of the 2014 Scottish independence referendum. In: Dealing with the local

Zhang K, Cheng Y, Xie Y, Honbo D, Agrawal A, Palsetia D, Lee K, Liao WK, Choudhary A (2011) SES: sentiment elicitation system for social media data. In: 2011 IEEE 11th international conference on data mining workshops (ICDMW), pp 129–136. IEEE