In: The 5th International Symposium, ALGOCLOUD 2019, Munich, Germany, 2019.

Developing a Cloud-based Algorithm for Analyzing the Polarization of Social Media Users

Loris Belcastro[1], Fabrizio Marozzo[1,2], Domenico Talia[1,2⋆], and Paolo Trunfio[1,2]

[1] DIMES, University of Calabria, Italy,
[lbelcastro, fmarozzo, talia, trunfio]@dimes.unical.it
[2] DtoK Lab Srl, Rende, Italy

Abstract. Social media analysis is a fast growing research area aimed at extracting useful information from social media. Several opinion mining techniques have been developed for capturing the mood of social media users related to a specific topic of interest. This paper shows how to use a cloud-based algorithm aimed at discovering the polarization of social media users in relation to political events characterized by the rivalry of different factions. The algorithm has been applied to a case study that analyzes the polarization of a large number of Twitter users during the 2016 Italian constitutional referendum. In particular, Twitter users have been classified and the results have been compared with the polls before voting and with the results obtained after the vote. The achieved results are very close to the real ones.

Keywords: Social Data analysis · Cloud computing · Big Data · User polarization · Sentiment analysis

1 Introduction

With the growth of utilization of social media, every day millions of people produce huge amount of digital data containing information about human dynamics, collective sentiments, and the behavior of groups of people. Such data, commonly referred as Big Data, overwhelms our ability to make use of it and extract useful information in reasonable time. Cloud computing systems provide elastic services, high performance and scalable data storage, which can be used as large-scale computing infrastructures for complex high-performance data mining applications. Combining Big Data analytics and machine learning techniques with scalable computing systems allows the production of new insights in a shorter time [5]. The analysis of such information is clearly highly valuable in science and business, since it is suitable for a wide range of applications: tourism agencies and municipalities can know the most important regions-of-interest visited by users [6], transport operators can reveal mobility insights in cities such as incident locations[12], business managers can understand the opinions of people on a topic, a product or an event of interest.

⋆ corresponding author: Domenico Talia - email: talia@dimes.unical.it

In this work we propose a new parallel and distributed algorithm for discovering the polarization of social media users in relation to a political event, which is characterized by the rivalry of different factions or parties. Examples of political events are:

- municipal elections, in which a faction supports a mayor candidate;
- political elections, in which a faction supports a party;
- presidential elections, in which a party (or a coalition of parties) supports a presidential candidate.

To deploy and run the designed algorithm on the Cloud, it has been written using ParSoDA (*Parallel Social Data Analytics*) [7], a Java library for building parallel social media analysis algorithms and simplifying the programming task necessary to implement these class of algorithms on parallel computing systems. To reach this goal, ParSoDA includes functions that are widely used for processing and analyzing data gathered from social media so as to find different types of information (e.g., user trajectories, user sentiments, topics trends).

The algorithm is designed to deal with Big Data. For this reason, it is based on the MapReduce model and can be executed in parallel on distributed systems, such as the HPC and Cloud platforms. The main benefit of using ParSoDA is that it was specifically designed to build Cloud-based data analysis applications. To this end, ParSoDA provides scalability mechanisms based on two of the most popular parallel processing frameworks (Hadoop[3] and Spark[4]), which are fundamental to provide satisfactory services as the amount of data to be managed grows.

To assess the accuracy of our algorithm, we present a case study application to extract the political polarization of Twitter users. In particular, the algorithm has been applied on a case study that analyzes the polarization of a large number of Twitter users during the 2016 Italian constitutional referendum. The obtained results are very close to the real ones and significantly more accurate than the average of the opinion polls, assessing the high accuracy and effectiveness of the proposed algorithm.

The paper is organized as follows: Section 2 discusses related work and compares other techniques with the one proposed here. Section 3 introduces the algorithm details and Section 4 discusses the case study on which the proposed algorithm has been used. Section 5 draws some conclusions.

2 Related work

Several researches are working on the design and implementation algorithms for measuring public opinion and predicting the polarization of social users according to political events.

Graham et al. [10] performed an hand-coded content analysis for understanding how British and Dutch parliamentary candidates used Twitter during the

[3] https://hadoop.apache.org/
[4] https://spark.apache.org/

2010 general elections. Anstead and O'Loughlin [2] analyzed the 2010 United Kingdom election and suggested the use of social media as a new way to understand public opinion. Gruzd and Roy [11] investigated the political polarization of social network users during the 2011 Canadian Federal Election by analyzing a sample of tweets posted by social media users that self-declared political views and affiliations.

Marozzo and Bessi [13] presented a methodology aimed at discovering the behavior of social media users and how news sites are used during political campaigns characterized by the rivalry of different factions. The idea behind this technique is to use the keywords inside a tweet to classify it by calculating the degree of polarity. Ceron et al. [8] proposed a text analysis methodology for studying the voting intention of French Internet users during the 2012 Presidential ballot and the subsequent legislative election, comparing their results with the predictions made by survey companies. El Alaoui et al. [9] proposed an adaptive sentiment analysis approach for extracting user opinions about political events. Their approach classifies the posts by exploiting a series of word polarity dictionaries built from a selected set of hashtags related to a political event of interest. Oikonomou et al. [14] used a Naïve Bayes classifier for estimating the winning candidate of USA presidential elections in three US states (i.e., Florida, Ohio and North Carolina). Ahmed et al. [1] compared three different volumetric and sentiment analysis methods in order to predict the outcome of the elections from Twitter posts in three Asian countries: Malaysia, India, and Pakistan. Olorunnimbe et al. [15] presented an incremental learning method, based on a multiple independent Naïve Baye models for predicting the political orientation of users over time.

Our algorithm analyzes the tags used by social media users for supporting their voting intentions. As an important aspect of the analysis process, we evaluated the statistical significance of collected data, which gives strong indications about the users and if they are voters of the political event under analysis. The algorithm has been applied to a real case study: the 2016 Italian constitutional referendum. We studied the behavior of about 50,000 Twitter users by analyzing more than 300,000 tweets posted on the referendum by them in the five weeks preceding the vote. The achieved results are very close to the real ones and significantly more accurate than the average of the opinion polls, assessing the high accuracy and effectiveness of the proposed algorithm.

3 Algorithm details

As mentioned in Section 1, this work proposes a new algorithm for estimating the polarization of social media users during political events. Given a political event $\mathcal{E}$, a set of the factions F, and a set the keywords K associated to $\mathcal{E}$, the proposed algorithm consists of the following steps (see Figure 1):

- *Data collection*: during this step all tweets that contain one or more keywords in K are gathered from Twitter[5] through public API.

[5] https://developer.twitter.com/

- *Data preprocessing*: at this step several operations are done for cleaning data, including removal of duplicates and tweets without tags, normalization of texts;
- *Tweet polarization*: during this step, each tweet is assigned to a specific faction f by considering the polarization of the tags it contains.
- *User polarization*: for each social media user u, a heuristic is used to calculate a score v_u, which represents the polarization of the user u towards each faction under analysis.
- *Result visualization*: at this step, the polarization scores are exploited for creating info-graphics that presents the results in a way that is easy to understand to the general public.

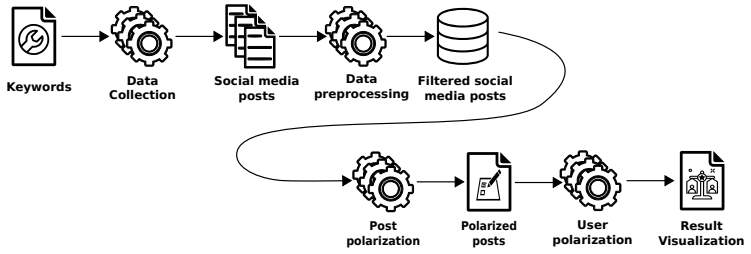

Fig. 1. Main steps of the proposed algorithm

3.1 Definition of keywords K

A political event $\mathcal{E}$ is characterized by the rivalry of different factions $F = \{f_0, f_1, ..., f_n\}$. The algorithm requires a set of the main keywords K used by social media users to write tweets associated to $\mathcal{E}$. Following the same approach used in [3], such keywords can be divided in *neutral* or in *favor* of a specific faction, i.e., $K = K^{\bigcirc} \cup K_F^{\oplus}$. Specifically:

- $K^{\bigcirc}$ contains all the keywords that can be associated to $\mathcal{E}$, but not to any faction in F.
- $K_F^{\oplus} = K_{f_0}^{\oplus} \cup ... \cup K_{f_n}^{\oplus}$, where $K_{f_i}^{\oplus}$ contains the keywords used by social media users for supporting $f_i \in F$.

Usually, this preparation step requires a minimal knowledge of the domain, that means it could be easily automated. In fact, keywords used for supporting a specific faction usually match some fixed patterns, such as the form "#vote + (faction / candidate / yes / no)". In data gathered from Twitter, such patterns can be searched in hashtags or words.

3.2 Data preprocessing

During this step, the tweets collected are pre-processed for making them suitable for the analysis. In particular, they are filtered and modified so as to:

- remove duplicates and stopwords;
- normalize all the keywords by transforming them in lowercase and replacing accented characters with regular ones (e.g., IOVOTOSI or iovotosí → iovotosi);
- improve data representativeness by filtering out all tweets having a language different from the one spoken in the nation hosting the considered political event.

The following operations are performed in parallel on multiple computing nodes exploiting the data parallelism provided by the MapReduce programming model. Since the algorithm has been developed using the ParSoDA library, it can run both on a Hadoop and a Spark cluster. In particular, some performance evaluation experiments we run show that the Spark version of ParSoDA is able to greatly reduce the execution compared to the Hadoop version of the library [4].

3.3 Tweet polarization

At this step, each tweet is assigned to a specific faction by considering the polarization of the tags it contains. In particular, if a tweet t contains only keywords that are in favor of a specific faction f, then t is classified as in favor of f; otherwise, t is classified as *neutral*. Algorithm 1 shows the pseudo-code of the tweet polarization procedure.

ALGORITHM 1: Polarization of tweets.

Input : Set of tweets T, set of factions F, set of keywords K_F for the different factions
Output : Dictionary of ⟨tweet, faction⟩ D_T

```
for t ∈ T do
    v_f ← [];
    for i = 0; i < F.size; i++ do
        if contains(t, K_{f_i}) then
            v_f[i] = 1;
    if sum(v_f) = 1 then
        f ← argmax(v);
        D_T ← D_T ∪ ⟨t, f⟩;
return D_P
```

3.4 User polarization

Using the classified tweets obtained at the previous step, the algorithm exploits a heuristic for estimating the polarization of each social user. Specifically, in a

two-factions political event, characterized by the rivalry between the factions f_0 and f_1, the polarization of a user u is defined as:

$$v_u = 2 \times \frac{|f_0|}{|f_0| + |f_1|} - 1 \tag{1}$$

where $|f_0|$ and $|f_1|$ represent the number of tweets published by u that have been classified in favor of f_0 and f_1 respectively. A value of v_u close to 1 means that user u tends to be polarized towards the faction f_0, while when v_u is close to -1 the user is polarized towards f_1.

To obtain more robust results, the algorithm requires a threshold th, usually set to a high value (e.g., 0.9), to select users with strong polarization in favor of f_0 or f_1. Specifically, we consider users with $v_u > th$ as polarized towards $|f_1|$, users with $v_u < -th$ as polarized towards $|f_1|$, otherwise *neutral*. The pseudo-code of the user polarization procedure is shown in Algorithm 2.

ALGORITHM 2: Polarization of users.

Input : Dictionary of ⟨tweet, faction ⟩ D_T, threshold th, two factions f_0 and f_1
Output : Dictionary of ⟨user, faction⟩ D_U

```
D_F ← ∅;
for t ∈ D_T do
    u ← t.user;
    f ← t.faction;
    D_F(u, f) + +;
D_U ← ∅;
for u ∈ D_F.users do
    v_u = 2 × |D_F(u,f_0)| / (|D_F(u,f_0)|+|D_F(u,f_1)|) − 1;
    if v_u > th then
        D_U ← D_U ∪ ⟨u, f_0⟩;
    else if v_u < −th then
        D_U ← D_U ∪ ⟨u, f_1⟩;
return D_U
```

3.5 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. 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 [16].

4 Case study and results

The algorithm has been applied to a case study that analyzes the polarization of a large number of Twitter users during the 2016 Italian constitutional referendum. The referendum, focused on changing the second part of the constitution, was characterized by the rivalry of two factions: *yes* and *no*. The results of the referendum saw the victory of the *no*, with about 60% of the votes. We collected the main keywords used as hashtags in tweets related to the political event. We collected the main keywords K used as hashtags in tweets related to the political event under analysis. Such keywords have been grouped as follows:

- $K^{\bigcirc}$ = {#referendumcostituzionale, #siono, #riformacostituzionale, #referendum, #4dicembre, #referendum-4dicembre}
- $K^{\oplus}_{yes}$ = {#bastaunsi, #iovotosi, #italiachedicesi, #iodicosi, #leragionidelsi}
- $K^{\oplus}_{no}$ = {#iovotono, #iodicono, #bastaunno, #famiglieperilno, #leragionidelno}

4.1 Statistical significance of analyzed data

The goal of this section is to assess the statistical significance of the dataset used for the analysis. Specifically, we studied whether the Twitter users included in our analysis were actual voters of the referendum, i.e., whether they were Italian citizens aged at least 18 years old. We also extracted aggregate information on the language used to write a tweet (e.g., "it" for Italian or "und" if no language could be detected) and on the location of users who wrote it. In addition, from the user metadata we analyzed the location field, which indicates the user-defined location for the accounts profile (e.g., Rome, Italy). By analyzing the metadata described above, we can say that:

- All the tweets under analysis have been written in Italian. Such language is mainly used by Italians who reside in Italy (about 60 million) or abroad (about 4 million). Italian is used as first language 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.

We calculated that there is a strong correlation (Pearson coefficient 0.9) between the number of Twitter users included in our analysis and the total number of citizens grouped by Italian regions. Similar results are obtained by comparing the number of users and the total number of citizen grouped by Italian cities (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 political event under analysis.

4.2 Analysis results

In the last few weeks before the mandatory stop to the polls, the *no* clearly prevailed on the *yes* in the totality of the opinion polls, maintaining about 4% of advantage. Figure 2 shows the comparison among the results achieved by our algorithm, the real voting percentages, the average of opinion polls before voting, and the post-voting percentages estimated for users aged 18-49. Specifically, our analysis focuses on two opposing factions, those in favor of the constitutional reform (i.e., *yes*) and the opposites (i.e., *no*).

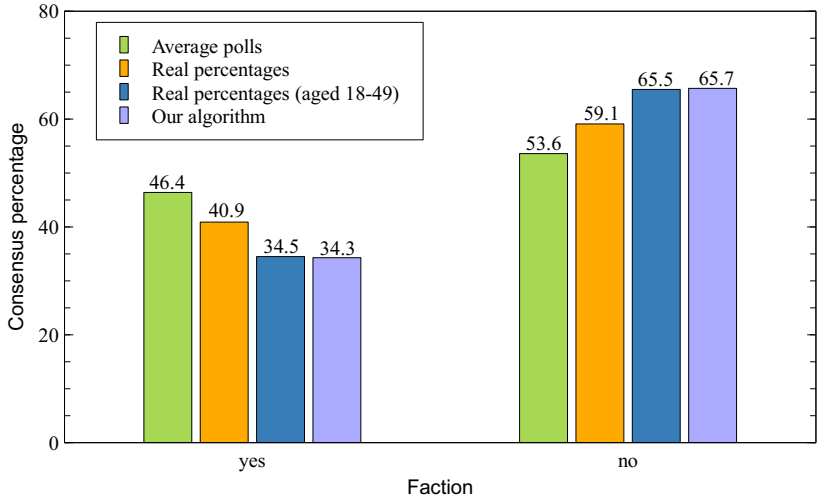

Fig. 2. Comparison between the obtained results, the real ones and the average of opinion polls.

The results achieved are very close to the real ones. This result assesses the high accuracy and effectiveness of the proposed approach. In particular, our algorithm estimated a consensus of 65.7% in favor of *no*, which is a slightly higher than the real one (59.1%), but really close to that estimated after the vote for users aged 18-49 (65.5%). Opinion polls underestimated the vote in favor of *no*, estimating only a percentage of about 53.6% for it. Differently from opinion polls, which tend to underestimate the results, our algorithm tends to overestimate them. This is most likely due to the Twitter data used for the analysis. As 75% of global Italian Twitter audiences were aged between 18 and 49 years, while only 14% of them are 50 or older[6]. An analysis carried out after the referendum[7] showed that the distribution of the vote by age was as follows:

[6] `https://datareportal.com/reports/digital-2019-q2-global-digital-statshot` (page 43)

[7] `https://www.youtrend.it/2016/12/09/referendum-costituzionale-tutti-numeri/`

– age 18-34: 64% *no*, 36% *yes*;
– age 35-49: 67% *no*, 33% *yes*;
– age 50-64: 57% *no*, 43% *yes*;
– age 65+: 51% *no*, 49% *yes*;

Since the majority of Italian Twitter users are aged between 18-49 (75% of audiences), our results strongly respect the distribution of the vote for these age groups. On the contrary, the polls are more cautious and generate more conservative estimates that tend to offset this gap.

5 Conclusion

With the growth of social media, every day millions of people produce huge amount of digital data containing information about human dynamics, collective sentiments, and the behavior of group of people. In this work we presented a new parallel and distributed algorithm for discovering the polarization of social media users during political events, which are characterized by the rivalry of different factions or parties. The algorithm is based on the MapReduce model and can be executed in parallel on distributed systems, such as the Cloud, ensuring scalability as the amount of data to be analyzed grows. To validate the proposed algorithm, it has been applied to a real case study: the 2016 Italian constitutional referendum. The achieved results are very close to the real ones and are significantly more accurate than the average of the opinion polls, revealing the high accuracy and effectiveness of the proposed approach.

Acknowledgment

This work has been partially supported by the SMART Project, CUP J28C17000150006, funded by Regione Calabria (POR FESR-FSE 2014-2020) and by the ASPIDE Project funded by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 801091.

References

1. Saifuddin Ahmed, Kokil Jaidka, and Marko M Skoric. Tweets and votes: A four-country comparison of volumetric and sentiment analysis approaches. In *Tenth International AAAI Conference on Web and Social Media*, 2016.
2. Nick Anstead and Ben O'Loughlin. Social media analysis and public opinion: The 2010 UK general election. *Journal of Computer-Mediated Communication*, 20(2):204–220, 2014.
3. Loris Belcastro, Riccardo Cantini, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. Discovering political polarization on social media: A case study. In *The 15th International Conference on Semantics, Knowledge and Grids*, Guangzhou, China, 2019.

4. Loris Belcastro, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. Appraising Spark on large-scale social media analysis. In *Euro-Par Workshops*, Lecture Notes in Computer Science, pages 483–495, Santiago de Compostela, Spain, 28-29 August 2017. ISBN: 978-3-319-75178-8.
5. Loris Belcastro, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. Big data analysis on clouds. In Sherif Sakr and Albert Zomaya, editors, *Handbook of Big Data Technologies*, pages 101–142. Springer, December 2017. ISBN: 978-3-319-49339-8.
6. Loris Belcastro, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. G-RoI: Automatic region-of-interest detection driven by geotagged social media data. *ACM Transactions on Knowledge Discovery from Data*, 12(3), 2018.
7. Loris Belcastro, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. ParSoDA: High-level parallel programming for social data mining. *Social Network Analysis and Mining*, 9(1), 2019.
8. Andrea Ceron, Luigi Curini, Stefano M Iacus, and Giuseppe Porro. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New media & society*, 16(2):340–358, 2014.
9. Imane El Alaoui, Youssef Gahi, Rochdi Messoussi, Youness Chaabi, Alexis Todoskoff, and Abdessamad Kobi. A novel adaptable approach for sentiment analysis on big social data. *Journal of Big Data*, 5(1):12, 2018.
10. Todd Graham, Dan Jackson, and Marcel Broersma. New platform, old habits? Candidates' use of Twitter during the 2010 British and Dutch general election campaigns. *New media & society*, 18(5):765–783, 2016.
11. Anatoliy Gruzd and Jeffrey Roy. Investigating political polarization on Twitter: A canadian perspective. *Policy & Internet*, 6(1):28–45, 2014.
12. Ryong Lee, Shoko Wakamiya, and Kazutoshi Sumiya. Urban area characterization based on crowd behavioral lifelogs over twitter. *Personal and ubiquitous computing*, 17(4):605–620, 2013.
13. Fabrizio Marozzo and Alessandro Bessi. Analyzing polarization of social media users and news sites during political campaigns. *Social Network Analysis and Mining*, 8(1):1, 2018.
14. Lazaros Oikonomou and Christos Tjortjis. A method for predicting the winner of the USA presidential elections using data extracted from Twitter. In *2018 South-Eastern European Design Automation, Computer Engineering, Computer Networks and Society Media Conference (SEEDA_CECNSM)*, pages 1–8. IEEE, 2018.
15. Muhammed K Olorunnimbe and Herna L Viktor. Tweets as a vote: Exploring political sentiments on Twitter for opinion mining. In *International Symposium on Methodologies for Intelligent Systems*, pages 180–185. Springer, 2015.
16. Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA, 1986.