University of Trento
Department of Sociology and Social Research
Computational Social Science
A.Y. 2021/2022

Sentiment analysis of Italian political candidates during the 2022 election campaign

Emma Benedetti
223 996
emma.benedetti@studenti.unitn.it

***Abstract*—This research analyses the Twitter posts of the political candidates of the 2022 Italian election campaign. The goal is to detect the tone of voice used by the candidates, in order to understand the general sentiment behind the text posts, as well as checking for the presence of multiple topics among these posts. To do so, we first collected a total of 2.829 tweets among six political leaders in the span of two months. Then, we retrieved the more relevant text information through the Term Frequency-Inverse Document Frequency (TF-IDF) method. Finally, we analysed the most-recurring keywords through the topic-modelling, Latent Dirichlet Analysis model.**

***Keywords*—Twitter, Election Campaign, Sentiment Analysis, NLP, Topic Modelling, Latent Dirichlet Allocation**

I. Introduction

For anyone who wants to spread a message or start a campaign, social media platforms are the best way to reach as many people as possible, as quickly as possible. It comes to no surprise, then, that many political figures created accounts on these platforms to make their agenda known. The unknown variable, instead, is how they choose to do it.
The goal of this work is, therefore, to understand which communication style -whether one more neutral or one more targeted towards the followers of the politician- was preferred by the political candidates of the Italian 2022 election campaign. In particular, we will focus on the text posts published on the micro-blogging platform Twitter (here also referred to as *tweets*).
To answer the research question, we will first collect the tweets of the eight most relevant political leaders who campaigned in the past elections, namely:

- **Giorgia Meloni:** leader of the right-wing party *Brothers of Italy*. Her party, and the coalition it is part of, obtained the majority of votes. She was later appointed prime minister and tasked with the organization of the new government.
- **Matteo Salvini:** from the right-wing party *League*, Salvini is a prolific poster on social media. Also his party was part of the winning coalition.
- **Silvio Berlusconi:** leader of the right-wing party *Forza Italia*, third and last party in the winning coalition.
- **Enrico Letta:** leader of the most popular left-wing party, *Partito Democratico*, which competed alone in the past campaign.
- **Matteo Renzi:** leader of the left-wing party *Italia Viva*, he ran in coalition with *Azione*, the party of Carlo Calenda.
- **Carlo Calenda:** from the left-wing party *Azione*, which competed together with *Italia Viva*.
- **Luigi di Maio:** former leader of the *Five Star Movement*, di Maio ran with his new left-wing party *Impegno Civico*.
- **Giuseppe Conte:** former prime minister, he ran as leader of the *Five Star Movement*, which has an undefined political alignment.

Then, we will compute the polarity score of each leader to analyse the general sentiment of their text posts. Afterwards, we will investigate whether the leaders trying to promote their own agendas led to the creation of sub-topics discussed during the campaign, by applying the topic-modelling Latent Dirichlet Allocation model.
To ensure reproducibility, a dashboard containing the research process, as well as the scripts used for data collection and elaboration, can be found on the following GitHub repository: EmmaBenedetti/CampaignSentiment.
For a better understanding of this work, Section II provides an overview of current scholarship about this topic. Section III instead reports the procedure used for data collection, text pre-processing, and topic modelling, as well as the results obtained. Finally, in Section IV we will discuss the advantages and disadvantages of the methodology applied to answer the research question, as well as draw the conclusive remarks.

II. Literature Review

Social media platforms have established their role as virtual squares, places where users are free to share their thoughts and opinions. It comes to no surprise, then, that politicians created accounts on those platforms to reach out to their voters and advertise their campaigns. How relevant the tone of voice used during such interactions, however, is not yet clear (S. Stier, A. Bleier, H. Lietz, M. Strohmaier, 2018).
Previous literature concluded that the *modus operandi* of politicians on social media platforms is to share

similar messages across their platforms, while restricting communication with their followers (R. Gibson, A. Römmele, A. Williamson, 2014; A. O. Larsson, 2015). According to the rules of mass communication, candidates should adjust their messages by adopting a neutral tone, in order to reach out to an audience as wide as possible. However, especially in interactive social media like Twitter, the single user tends to be networked with other like-minded individuals in a personalised echo-chamber. As a consequence, politicians are exposed to a demographic with very specific political interests, and might feel the need to tailor their language to that specific demographic.

It is difficult to understand whether one or the other language is more effective, due to the high fragmentation of the political communication field that has been proven difficult to solve. There are many factors preventing this to happen, such as the different fields where research on this topic is carried out -along with their approaches and theoretical assumptions- as well as a general lack of unifying theories on this matter (J. Strömbäck, S. Kiousis, 2014). Moreover, the majority of research on political communication uses data on the highly-atypical American Congress, while there is scarce cross-national research on this topic.

Therefore, it might be plausible that the communication style of candidates of political elections greatly varies according to the analysed country, the affiliation that voters have with the competing parties, or even the time of the election. For example, Stier et al. (2018) analysed the political candidates of the federal German election campaign of 2013 and concluded that a communication targeted towards a specific audience is more effective. They argue that, due to the interactivity of social media platforms, the posts created by the candidates should be strictly related to the preferred topics of their immediate communication network. However, party identification is significantly strong in Germany (K. Arzheimer, 2006).

Italian voters, on the other hand, are being more and more affected by party leaders and their charisma (D. Garzia, F. F. da Silva, A. De Angelis, 2021). One should also note that, as opposed to Germany, Italian parties tend to be tailored on the figure of the leader itself, to the point that the surname pf the leader is often included on the symbol or the party name. This union of party and leadership might bring the political candidates to use more moderate tones, in order to convince as many voters as possible to vote for their party.

III. Research Methodology

A. Data Collection

The search for an answer for the research question started with a collection of Twitter text posts from the most prominent political leaders in the Italian political framework.

We used Twitter APIs to scrape the user timeline of each leader. To avoid any possible echo-chamber effect, we collected only original tweets from the party leaders. Namely, we excluded any retweet from other users and any answering post. As for the time range, we set as delimiting dates July 21^{st}, 2022 -the day when the Italian President dissolved the previous Parliament- and September 25^{th}, 2022, i.e. the day of the election. Due to the nationality of both candidates and voters, most tweets were written in Italian, but we also removed tweets written in any other foreign language.

With these criteria, we managed to collect 2.829 tweets, unevenly distributed among the authors. For a clearer understanding of tweet allocation, please see table II.

The final dataset, after a first step of text pre-processing, presents the attributes shown in table I.

TABLE I
Description of the columns included in the dataset

Column Name	Data Type	Variable Description
Date	datetime	Day of publication of the tweet
Tweet ID	integer	Unique identifier of the tweet
Leader Name	string	Name and Surname of the tweet author
Followers	integer	Total of the leader's followers
Text	string	Original text of the tweet
Hashtags	list	List of hashtags included in the tweet
Polarity Score	float	Compound text polarity score
Tokenized Text	list	Tokenized tweet text, used for NLP

During this step, we also noticed the distribution of the tweets over the election campaign, shown in figure 1. As expected, the number of tweets rose over time, reaching its highest value two days before the election date. This increase is partially due to the subscription of Calenda to Twitter, depicted in gray. After Salvini, Calenda was the second steadiest and most prolific politician to post on the platform. On the other hand, the low number of tweets on the days before and of the election is to attribute to the mandatory electoral silence that must be held at this specific time (Italian Parliament, 1956).

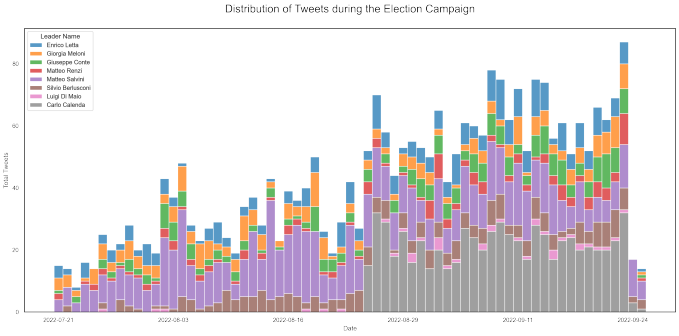

Fig. 1. Chronological distribution of tweets throughout the campaign, divided by leader.

B. Text Preprocessing

To provide an usable corpus for the analysis, it was necessary to pre-process the tweet dataset. Firstly, we removed all unicode characters and emojis from the texts. Then, we removed all Italian stopwords provided by the NLTK library (S. Bird, E. Klein, E. Loper, 2009), as well as additional stopwords still present in the texts but not in the library (such as city names, or numbers written in their literal

form). After that, by using the VADER-Sentiment library (C. Hutto, E. Gilbert, 2014), we were able to detect the polarity score of each tweet, whose average for each leader was reported in table II.
The polarity score is a relevant metric in sentiment analysis, used to quantify the general opinion expressed in a text in an interval between -1 and +1. According to the score sign, the text is classified as negative, neutral, or positive. Table II shows that every candidate assumed a generally positive tone of voice in their text posts, although two candidates leaned towards a more neutral tone.

TABLE II
DISTRIBUTION OF TWEETS AND AVERAGE POLARITY SCORE

Leader Name	Total Tweets	Polarity Score
Giorgia Meloni	258	0.175
Matteo Salvini	821	0.218
Silvio Berlusconi	250	0.266
Enrico Letta	282	0.186
Matteo Renzi	135	0.261
Carlo Calenda	615	0.076
Luigi di Maio	35	0.061
Giuseppe Conte	194	0.201

C. Keyword Finding with TF-IDF

For the extraction of the most relevant terms in the tweets, we relied on the Term Frequency-Inverse Document Frequency (TF-IDF) method (K. S. Jones, 1972).
TF-IDF works by assigning a score to each term in the corpus document, which increases proportionally by the ratio the term appears in each document of the corpus (TF) and is counterbalanced by the total number of documents in the whole corpus (IDF). The final score corresponds to the weight of the considered term: the higher it is, the more meaningful the term is in the corpus.
Assigning a TF-IDF score to each term has the additional advantage of converting the texts into vectors of numbers. This way, we are able to quantify the relevance of a word within the corpus without giving any weight to the stop-words included in the texts (H. C. Wu, R. W. P. Luk, K. F. Wong, K. L. Kwok, 2008).
By applying TF-IDF vectorization, we were able to find 431 keywords inside the corpus. However, this is a gross estimation of the actually relevant keywords inside of the corpus, since it includes also keywords that appeared rarely. As such, for the following analysis we will consider only the keywords that have a minimum of five occurrences in the corpus. After this step, we derived 48 significant keywords.

D. Keyword Centrality

To make a prediction of which will be the more important keywords in the tweet semantic network, we studied the application of two centrality measures, namely degree centrality and betweenness centrality (S. P. Borgatti, 2005).
Degree centrality can be defined as the number of edges in- or outgoing (or, in case of undirected graphs, connected to) a node. In other words, it shows how many connections can be found for each keyword. Betweenness centrality, on the other hand, reflects the importance of a node over the graph over the amount of paths flowing through that node.
In this work, we considered only keywords with a minimum degree centrality of 5, while the central nodes in figure 2 show the keywords with the highest betweenness centrality.

E. Spectral Clustering

To test whether political leaders preferred to promote their own agendas and varied messages during the campaign, we will use a clustering technique called *spectral clustering* (A. Ng, M. Jordan, Y. Weiss, 2001) to converge the keywords into multiple distinct topics.
Spectral clustering relies on the eigenvalues of the similarity matrix of the data, in our case the relevant keywords, to execute a dimensionality reduction task before grouping the variables according to the most relevant features. Both the silhouette coefficient (D.-T. Dinh, T. Fujinami, V.-N. Huynh, 2019) and the low number of keywords were used to evaluate the total number and consistency of the clusters. Eventually, we settled on two clusters, which included respectively 33 and 14 keywords, and reported a silhouette coefficient of 0.107.
Finding only two clusters leads us to the conclusion that, during the campaign, the candidates chose not to tailor the tone of voice to a specific audience. With such a low silhouette coefficient, we are not expecting a significant distinction between the clusters, further confirming the goal of the candidates to reach a wider audience by keeping a more neutral tone.

F. Semantic Networking

After finding the more important keywords and discovering two topic categories, we are ready to display the semantic network graph.
Semantic networks are useful, graphical representations of a topic's main concepts, namely the network nodes, and the relationships among them, shown as edges connecting two or more nodes. In text analytics, and in this work in particular, they are particularly useful for representing the relationships among keywords (P. Drieger, 2013).
The general structure of a semantic network hints to the significance of each concept within the corpus (Wu et al., 2008), as well as its relations with other concepts. For example, hubs of highly-connected nodes show both the similarity of the concepts and their importance inside the network.
Figure 2 shows the semantic network related to this work. We created an undirected graph, where the previously-found 48 keywords assume the role of nodes. Each edge then represents the co-occurrences between the keywords, where the weight increases proportionally to the count of times the connection appears in a tweet.
The colors of the graph nodes represent the cluster the keywords are part of. As anticipated by the spectral coefficient, here we see a very sparse network, without distinctive hubs. Three nodes in particular (*votafdi*, *elezionipolitiche*, and

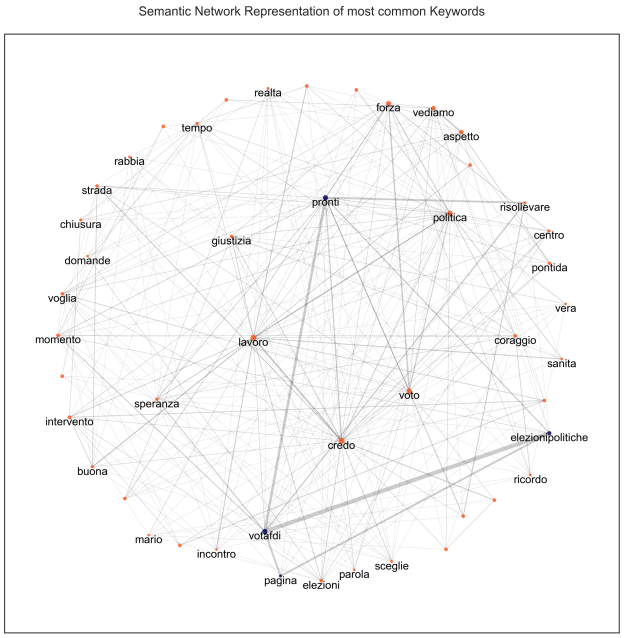

Fig. 2. Semantic network representation of the previously-extracted keywords.

pronti) appear to be strongly connected. Since two out of three are hashtags, it should be inferred that these keywords were the slogan of one of the analysed political leaders.
In figure 2 we can also see that the majority of keywords hold a positive connotation (like *speranza*, *momento*, or *pronti*), but there are also a few negative keywords (like *rabbia* or *chiusura*). Therefore, we can hypothesize that the leaders introduced themselves as the ready solution to the problems that afflict the country.
Finally, we notice that the keywords *lavoro* and *giustizia* (labor and justice) show a high betweenness centrality, hinting to their recurrence in the tweets and consequently to their importance in the political agendas. Another recurring keyword of the campaign is *sanità* (healthcare), although it presents a lower importance in the network.

G. Latent Dirichlet Allocation

The last step of our analysis consists of employing the Latent Dirichlet Allocation (LDA) model for topic modelling (D. M. Blei, A. Y. Ng, M. I. Jordan, 2003).
LDA is a generative probabilistic model that efficiently complements the TF-IDF scheme described in section III-C. In fact, while TF-IDF successfully identifies the more important terms in a corpus, it does not inform on the inter- or intra-document connections among such terms.
LDA, on the other hand, considers the documents in the corpus as '*random mixtures over latent topics*' (Blei et al., 2003). In other words, LDA assigns a topic to a document, based on the words included in that same document. It does so by parsing the document through a hierarchical, three-level Bayesian model, where the words in the document are given a probability to belong to a specific topic. Then, the model assigns the document to the topic, according to the words in the former with the highest probability to belong to the latter. We applied LDA on the keywords found through TF-IDF (see section III-C). Due to the low number of keywords, we processed only unigrams, namely co-occurrences of single words. For the sake of coherence with the number of clusters, we set only two topics to detect.
The resulting division of the keywords is represented in figure 3, while the topics can be interpreted as follows:

- **Topic 1** includes all keywords that encourage the public to vote at the elections. The most frequent examples of this topic are keywords like *voto*, *politica*, and *forza*.
- **Topic 2** instead includes all the keywords related to the political agendas and what they include. The main keyword of this topic is *lavoro*, but it includes also encouraging terms like *risollevare* and *incontro*.

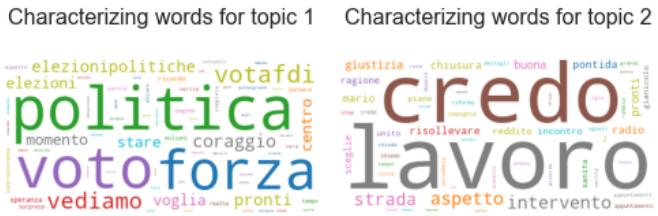

Fig. 3. Wordcloud of the characterizing keywords of each LDA topic.

Being a dataset of tweets of an electoral campaign, it was expected to see a topic like Topic 1. However, we were unable to separate the keywords in Topic 2 in different subtopics, one for each leader.

IV. Conclusion and Final Considerations

In this work, we analysed the tweets written by political candidates of the Italian 2022 election campaign. The goal was to study the general sentiment of the tweets, as well as understanding whether the candidates tried to reach a wider or a more targeted audience. To do so, we first computed the average polarity score for each leader. Then we extracted, clustered, and modelled the most frequent keywords into topics, with the help of Latent Dirichlet Allocation technique. This analysis led to the detection of two main topics very similar to each other and brought us to the conclusion that, in general, leaders prefer to reach out to a wider audience. Therefore, they maintain an overall encouraging tone of voice, and tend to promote topics of their agendas closer to the interests of the majority of voters, like labor or healthcare. Our work has brought results in line with those of other literature regarding social network analysis, but is not exempt of limitations related to the methodology that may have polluted the final result. Firstly, by collecting data only from Twitter, it is possible that we might have obtained results tailored for this platform. Being a micro-blogging platform, Twitter requires its users to write at most 280 characters. The necessity to write shorter texts might have led the candidates

to prefer neutrality and brevity in their texts, and to rely on media to spread their agendas. Since we considered only text posts, it is possible that we lost relevant information at the data collection stage.

Secondly, the small, unbalanced dataset led the analysis to skew the results towards the leaders that wrote more tweets. Table II shows that the three candidates of the right-wing coalition wrote in total 1.329 tweets, i.e. 45% of the total data. Consequently, the analysis was highly biased towards the campaign topics closer to the right-wing parties.

Another issue linked to the methodology is the highly vague general context at the basis of this work. By taking only the tweets -and the keywords- related to an election campaign, it was difficult for the techniques in Sections III-E - III-G to differentiate among more than two main topics. This issue as well hints to high bias in our analysis.

Despite these issues, this methodology allowed us to extract the more significant keywords and was more than sufficient to provide an answer to the research question. We believe that collecting data from more sources, like other social media platforms, might be an effective solution for further improvement of our analysis.

References

Arzheimer Kai. 'Dead men walking?'Party identification in Germany, 1977–2002 // Electoral Studies. 2006. 25, 4. 791–807.

Bird Steven, Klein Ewan, Loper Edward. Natural language processing with Python: analyzing text with the natural language toolkit. 2009.

Blei David M, Ng Andrew Y, Jordan Michael I. Latent dirichlet allocation // Journal of machine Learning research. 2003. 3, Jan. 993–1022.

Borgatti Stephen P. Centrality and network flow // Social networks. 2005. 27, 1. 55–71.

Dinh Duy-Tai, Fujinami Tsutomu, Huynh Van-Nam. Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient // International Symposium on Knowledge and Systems Sciences. 2019. 1–17.

Drieger Philipp. Semantic network analysis as a method for visual text analytics // Procedia-social and behavioral sciences. 2013. 79. 4–17.

Garzia Diego, Silva Frederico Ferreira da, De Angelis Andrea. Leaders without Partisans. 2021.

Gibson Rachel, Römmele Andrea, Williamson Andy. Chasing the digital wave: International perspectives on the growth of online campaigning. 2014. 123–129.

Hutto Clayton, Gilbert Eric. Vader: A parsimonious rule-based model for sentiment analysis of social media text // Proceedings of the international AAAI conference on web and social media. 8, 1. 2014. 216–225.

Italian Parliament . Act no. 212/1956. 1956. http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:1956-04-04;212!vig=2022-12-17.

Jones Karen Sparck. A statistical interpretation of term specificity and its application in retrieval // Journal of documentation. 1972.

Larsson Anders Olof. Green light for interaction: Party use of social media during the 2014 Swedish election year // First Monday. 2015.

Ng Andrew, Jordan Michael, Weiss Yair. On spectral clustering: Analysis and an algorithm // Advances in neural information processing systems. 2001. 14.

Stier Sebastian, Bleier Arnim, Lietz Haiko, Strohmaier Markus. Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter // Political communication. 2018. 35, 1. 50–74.

Strategic Political Communication in Election Campaigns. // . 06 2014. 109–128.

Wu Ho Chung, Luk Robert Wing Pong, Wong Kam Fai, Kwok Kui Lam. Interpreting tf-idf term weights as making relevance decisions // ACM Transactions on Information Systems (TOIS). 2008. 26, 3. 1–37.