Drivers of hate speech in political conversations on Twitter: the case of the 2022 Italian General election

Francesco Pierri[1*]

[1]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Via Giuseppe Ponzio 34, Milan, Italy.

Corresponding author(s). E-mail(s): francesco.pierri@polimi.it;

Abstract

We study the prevalence and characteristics of toxic speech on Twitter in the run-up to the 2022 Italian General election. We analyzed over 8.5 M tweets shared by 450k unique users and employed a machine learning classifier to estimate the toxicity score of their messages. We found that supporters of different political coalitions exhibit varying levels of toxic speech, sending between 6-8% of toxic messages overall. Notably, Centre-Left politicians received more toxic messages on average, with the largest target of hate receiving over 15 000 toxic replies. We employed Generalized Linear Models to study factors that drive hate speech to political targets, finding that, importantly, politicians employing more abusive and harmful language are also more likely to attract more inflammatory speech. Our findings underscore the critical need for targeted interventions to address hate speech online, fostering healthier dialogue and safeguarding democratic discourse.

Keywords: hate speech, online social networks, toxicity, Twitter

1 Introduction

Online social media have become an essential component in modern democracies, allowing users to share and discuss content, and freely express their views [1]. During elections' season, they can play a pivotal role in shaping public opinion, amplifying political discourse, and mobilizing voters, as well as increase political polarization [2].

However, in recent years there have been alarming reports of a concerning surge in hate speech and harmful content, raising questions about moderating online discourse while ensuring rights of free expression [3, 4]. The COVID-19 pandemic has spurred an unprecedented *infodemic* of misinformation, conspiracy theories, and misleading content which has hindered effective public health communication efforts, and posed significant challenges to combating the spread of the virus [5–8]. It has also highlighted the need for collaborative efforts among authorities, platforms and scientists to safeguard the integrity of online conversations [9].

On online social media, *toxic speech* refers to any communication that is abusive, disrespectful, or intended to harm others, often through insults, harassment, or inflammatory remarks [10]. *Hate speech*, on the other hand, is a more specific form of toxic speech that targets individuals or groups based on characteristics such as race, religion, ethnicity, gender, sexual orientation, or disability [11]. In this paper, we use toxic speech and hate speech interchangeably as we examine harmful conversations related to politics. During political events platforms may become fertile grounds for inflammatory rhetoric and divisive discourse [12, 13]. This can foster polarization within societies and undermine the democratic process. Addressing this problem is challenging, as it requires a diverse approach that includes strong and transparent content moderation, higher user awareness, and collaboration between policymakers and social media platforms, protecting rights to free speech while mitigating the negative impact of hateful content [14]. The traditional approach for automatically identifying hateful speech on online social media is to build labeled datasets of such posts and then train and test machine learning classifiers [15]. One of the most widely adopted models is Perspective API, developed by Google and publicly available to content moderators, practitioners and researchers. It is currently employed by numerous online news websites and social platforms such as The New York Times and Reddit [16].

In this work, we analyze the prevalence of toxic and hateful interactions among Twitter users in the context of the 2022 Italian election, which took place on September 25th. The collapse of the Italian government of national unity led by Mario Draghi on July 21st, 2022, prompted the country's first snap election in the fall, characterized by a reduced number of seats in both the House of Chambers (from 630 to 400) and the Senate (from 315 to 200) following the 2020 Italian constitutional referendum. With a voter turnout reaching a record low of less than 64% of eligible voters, the right-wing coalition led by Giorgia Meloni secured a significant victory, capturing over 43% of the vote share. In contrast, the Centre-left coalition garnered around 25% of the vote, while the populist Movimento 5 Stelle fell short, obtaining less than 16%. The liberal and centrist Third Pole, inclusive of Matteo Renzi's party Italia Viva, emerged as the fourth largest group with nearly 8% of the vote share.

In this variegated political setting, where public opinion can be distributed across numerous parties, we formulate the following research questions:

RQ1 Are there differences in the hateful behaviour of coalitions' supporters?
RQ2 Are there political and gender differences in targets of hate?
RQ3 What are the determinants of toxic speech targeting politicians?

To address them, we leverage a large-scale dataset of over 8.5 M tweets related to the 2022 Italian General election and shared by 450 K unique users over 2 months. We estimate the political affiliation of users analyzing their retweeting behavior, and employ a state-of-the-art machine learning model to classify the toxicity of shared messages. We find that different coalitions' supporters exhibit a heteregenous level of toxic interactions with other users and politicians, with Movimento 5 Stelle (M5S) followers being the most toxic on average and those sending out the largest number of toxic replies, on average. We find that replies are generally more toxic than other kinds of tweets (original and quotes), and that suspenders users exhibit a larger usage of abusive language. We observe that Centre-Left and Right politicians are the most frequent targets of hate, and that female politicians generally receive a larger number of toxic replies compared to male politicians. Importantly, we find that politicians that employ more toxic language are also more likely to become targets of hate speech. Our results provide further insights on the interplay between political discourse and harmful online interactions, which can inform intervention strategies to counter and reduce the spread of hateful speech on social platforms.

The outline of this paper is the following: in Section 2, we provide an overview of the related literature and the position of our work; then, in Section 3 we describe the dataset and methods employed in our analysis; next, in Section 4 we provide results of our experiments to address the above research questions; finally, in Section 5 we sum up our contributions, draw implications of our work, highlight limitations and discuss future directions.

2 Related Work

Early work on abusive usage of social platforms focuses on detecting spammers, social bots, and state-backed trolls that eventually get suspended and removed by platforms, [17–20]. Previous research on the hateful behaviour of Twitter users has found that such individuals differ from other users in terms of their activity patterns, word usage and network structure [21].

Next, we overview recent contributions that analyze abusive and harmful online language on social media that are more related to the present work, and then we position our contributions. We refer the interested reader to literature reviews on the topic [13, 22–25], and to [26] for hate speech detection approaches for the Italian language.

[27] study abusive replies, through a dictionary-based approach, directed at tweets by UK Members of the Parliament (MP) during the lead-up to the 2015 and 2017 UK general elections. They provide evidence of a rising trend in online abuse towards UK politicians, proposing a structured framework that examines differential treatment based on gender and political affiliation. They also present a comparative analysis of abuse towards MPs who stand for re-election versus those who do not, aiming to understand its impact on political careers and representation more broadly.

Hua et al. [28] employ a combination of automated tools and qualitative coding techniques to offer a longitudinal characterization of Twitter users involved in what they dub "adversarial" interactions with political candidates leading up to the U.S.

midterm election in November 2018. Their dataset encompasses 1.2 million replies to over 700 candidates' tweets generated by 0.4 million unique Twitter users. Their primary focus is on examining replies to candidates' tweets, which are inherently visible to other users engaging with the candidates' posts on Twitter. Using Perspective API [14] to identify adversarial interactions, they show that such harmful interactions stem from users who interact with candidates infrequently, and that more than 35% of adversarial replies originate from just 10% of users who consistently post contentious content directed towards candidates.

In follow-up work, Hua et al. [29] characterize adversarial interactions directed towards political candidates by investigating candidate attributes–such as gender and affiliated party–that may influence the frequency of adversarial interactions they receive.

[30] create and curate a dataset of hate speech in the public discourse between citizens and UK MPs on Twitter over a two month period. Their investigation into hate speech targeting MPs reveals distinct patterns: MPs from ethnic minority backgrounds encounter higher levels of hate speech compared to their white counterparts, while male and female MPs experience similar rates of hate speech.

[31] also empirically analyze how the user base on Twitter responds to posts from members of the U.S. Congress, with the goal of understanding how personal characteristics of politicians, such as their party affiliation, gender, and ethnicity can drive hate speech. They find that, all else being equal, tweets are more likely to attract hate speech in replies if they are authored by people of color from the Democratic party, white Republicans, or women.

2.1 Position of our work

Our work is overall similar in spirit to [27–29, 31, 32], who specifically focus on toxic speech targeting politicians, while most of the other work described above focuses on studying detection techniques. These contributions, however, study toxic interactions between online users and politicians in bipolar political settings (US Democrats and Republicans, UK Conservatives and Labour).

In contrast to the bipolar political environments studied in existing research, such as the US with its Democrats and Republicans or the UK with its Conservatives and Labour, Italy's multi-party system introduces a higher level of complexity to political dynamics. In a bipolar system, political discourse is often framed within a clear binary opposition, which can simplify the nature of toxic speech and its targets. However, Italy's multi-party system, with its array of competing parties and coalitions, creates a more fragmented political landscape. This fragmentation means that political toxic and hate speech can be directed at a wider range of political entities and ideologies, each with its own distinct voter base and influence. Consequently, this diversity may result in a more complex pattern of toxic interactions, as the interplay of multiple parties can lead to varied and potentially more nuanced forms of political animosity. Understanding this complexity is crucial for analyzing how toxic speech manifests in a system where political allegiances and adversaries are not as clearly defined. To the best of our knowledge, we provide the following novel contributions to existing literature:

1. We provide the first longitudinal analysis of online toxicity in Italian political conversations on Twitter.
2. We show that replies are generally more toxic than normal tweets, across the entire political spectrum.
3. We find that users employing more toxic language are more likely to be suspended by Twitter.
4. We pinpoint the top targets of hate speech in the Italian political scenario during the 2022 Election.
5. We analyze the determinants of hate speech targeting politicians in the Italian multi-party system, with results similar to other countries.

3 Dataset and Methods

3.1 Twitter

We leverage the ITA-ELECTION-2022 dataset of tweets related to the 2022 Italian election provided by [33] and collected by using tweepy[1] Python library to query Twitter v1.1 Filter streaming API endpoint [2] in the period September 2nd, 2022 - October 20th, 2022. This includes election day, as they were held on September 25th, 2022. A snowball sampling approach was employed [34, 35] to generate a list of relevant keywords, starting with seed terms such as \elezioni2022" and \elezioni"[3]; the final list contains 62 keywords, and a sample is provided in Table 1. In addition, they set the API parameter lang=IT to retrieve only Italian tweets.

The total collection of tweets contains 8 671 855 tweets shared by 456 942 unique users, corresponding to 1 244 051 original tweets and quotes, 2 615 809 replies and 4 811 995 retweets. We show in Figure 1 a time series of the number of tweets collected on a daily basis, and the daily proportion of tweets that are toxic (toxicity > 0:5, see next). The latter time series shows a slightly decreasing trend towards the end of the period of analysis.

We also identified suspended accounts by leveraging the POST /2/compliance/jobs endpoint via twarc2 in February 2023.[4]. Twitter may suspend an account for various reasons, including promoting violence, glorifying crime, hate speech, spam, and impersonation. Further information can be found in Twitter's documentation.[5]

3.2 Italian political parties

The 2022 Italian general election was held on September 25, 2022, to elect members of the Chamber of Deputies and the Senate of the Republic. This election followed the resignation of Prime Minister Mario Draghi, which led to the dissolution of Parliament. The election resulted in a significant victory for the right-wing coalition led

[1] https://www.tweepy.org
[2] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/overview
[3] In Italian \elezioni" means elections.
[4] twarc-project.readthedocs.io/en/latest/twarc2_en_us/#compliance-job
[5] help.twitter.com/en/managing-your-account/suspended-twitter-accounts

by Giorgia Meloni. This outcome marked a shift towards more nationalist and conservative policies in Italian politics. Giorgia Meloni subsequently became Italy's first female Prime Minister, leading a government focused on issues such as immigration, economic reform, and national sovereignty. We also notice that, in August 2022, Carlo Calenda was initially set to form an alliance with the Partito Democratico (PD), but he later shifted course and forged a pact with Matteo Renzi's Italia Viva, ultimately leading to the creation of the Third Pole. In this work, we leverage a list of Twitter accounts matched to the following parties:

- Alleanza Verdi Sinistra : A left-wing coalition formed by the Green Party and leftist groups focusing on environmental issues, social justice, and progressive policies.
- Partito Democratico (PD) : A center-left political party, it is one of Italy's major parties, advocating for social democracy, progressive reform, and economic justice.
- Movimento 5 Stelle (M5S) : Founded by comedian Beppe Grillo, this party is known for its anti-establishment stance, emphasizing direct democracy, environmentalism, and political reform.
- Azione : A centrist political party that focuses on pragmatic and reformist policies, aiming to address Italy's economic and political challenges through moderate and innovative approaches.
- Italia Viva : A centrist to center-left party founded by former Prime Minister Matteo Renzi, emphasizing reforms, modernization, and a pro-European stance.
- Lega: Originally known as Lega Nord, this right-wing party focuses on regional autonomy, nationalism, and conservative policies, and has evolved to a more nationalistic and populist stance under Matteo Salvini.
- Fratelli d'Italia (FdI) : A right-wing party with a focus on Italian nationalism, conservative values, and traditionalism, led by Giorgia Meloni.
- Forza Italia : Founded by former Prime Minister Silvio Berlusconi, this center-right party advocates for liberal economic policies, privatization, and a pro-business stance.
- Noi Moderati : A centrist party with a focus on moderate and pragmatic policies, aiming to bridge gaps between different political ideologies and promote consensus-driven governance.

3.3 Political affiliation

To estimate the political affiliation of Twitter users, we resort to the list of Twitter handles for 471 elected members in the Senate and Chamber of Deputies, corresponding to the four major coalitions, available in the ITA-ELECTION-2022 dataset:

- Centre-Left (\Alleanza Verdi Sinistra" and\Partito Democratico"): 113 accounts.
- M5S (\Movimento 5 Stelle"): 58 accounts.
- Third Pole (\Azione" and \Italia Viva"): 29 accounts.
- Right (\Lega", \Fratelli d'Italia", \Forza Italia" and \Noi Moderati"): 273 accounts.

We notice that M5S should be considered as a populist party, whereas the Third Pole is a liberal and centrist group.

Next, for each user in our dataset that retweeted a politician at least once (92 143), we compute the proportion of retweets to politicians in each coalition. We choose this approach as we do not have access to the follower/following information, which could also be employed to estimate the leaning [36], and because retweets usually represent an explicit form of endorsement on Twitter, with only a negligible of accounts (usually journalists) including the statement \RT != endorsement" in their pro le [37]. Instead, quotes and replies are usually employed to signal a strong negative reaction or disagreement with the original post, a mechanism which is denoted as \ratioing" usually employed to communicate with ideological opponents across polarized environments [38].

In total, we identify 522 646 retweets in our dataset (10.86%) of posts authored by politicians. We then assign to such users the label of the coalition that they retweeted the most, obtaining the breakdown of Table 2, similar to [39]. Overall, these users account for 1 565 176 tweets (excluding retweets) in our dataset. We remark that this procedure is equivalent to a label propagation [40] process run on the network of users retweeting politicians, assigning to each node the majority of the labels of its neighbors (i.e., politicians).

3.4 Toxicity detection

We classify the toxicity of tweets posted by users in our dataset by leveraging the Perspective API[6], a Large Language Model based classi er provided by the Jigsaw unit of Google [41], which is widely used within academia and for content moderation on news and social media platforms [16]. We remark that we are using the models-is, meaning that we do not carry out ne-tuning nor we aim to evaluate its accuracy performance.

Perspective API de nes a toxic message as a text that uses \rude, disrespectful, or unreasonable language that is likely to make someone leave a discussion" [41]. The API is multilingual (it supports Italian language) and returns several di erent toxicity scores, namely Severe toxicity, Insult, Profanity, Identity attack, Threat , and Sexually explicit We focus on the attribute \Toxicity", i.e., the estimated probability of a comment being perceived as toxic, as it is the most general and widely employed in the literature. The score ranges in the value [0, 1], and commonly a threshold of 0.5-0.7 is used to de ne a text as toxic [29, 42].

We ran the API in November 2022 to estimate the toxicity of all tweets in our dataset (except retweets) obtaining scores for 99.28% of all tweets. Missing results correspond to tweets that were not recognized as Italian language by the API{we remark that the original dataset contains only Italian tweets which were collected specifying the attribute lang=it in the Streaming API [33]. We show examples of replies to politicians with their estimated toxicity score in Figure 2, while we provide in Figure 3 the distribution of toxicity scores for tweets in our dataset. We notice that it is a heavy-tailed distribution with most of the mass lying near the origin.

[6] https://www.perspectiveapi.com/

4 Results

4.1 Hateful behavior of coalitions' supporters

To answer (RQ1) , i.e., whether there are di erences in the hateful behavior of coalitions' supporters, we rst look at the overall distribution of toxicity for messages posted by di erent coalitions' supporters. As shown in panel A of Figure 4, there are signi cant di erences in the distributions of the coalitions (Kruskal-Wallis test $P < 0.001$) but they are very small (Median values are all 0.11) and most likely driven by outliers. We also investigate whetherreply tweets are more toxic than other tweets. As shown in panelB of Figure 4, replies are, on average, much more toxic than the original tweets and quotes, for all coalitions (Kruskal-Wallis $P < 0.001$ for all pairwise distributions).

We then observe larger signi cant di erences between coalitions when computing the average toxicity of users, as shown in panelA of Figure 5 (Kruskal-Wallis test $P < 0.001$). M5S supporters are the most toxic on average (Median = 0.17) and Right supporters are the least toxic (Median = 0.12), with Centre-Left and Third Pole supporters exhibiting similar values (Median = 0.13). We also analyze the average toxicity of politicians, which is available only for 293 accounts that tweeted at least once, and panelB of the same gure shows that they are much less toxic than their supporters: Median values are 0.04 for all coalitions except Right politicians, who are the least toxic on average (Median = 0.02).

Next, we focus on tweets de ned astoxic, i.e., with a toxicity score above 0.5. As shown in panelA of Figure 6, we observe that M5S and Right supporters shared overall a larger proportion of toxic tweets (8%) compared to Centre-Left (6.4%) and Third Pole (6.0%). Panel B of the same gure shows instead the average number of toxic tweets shared by each user, which ranges between 2-4 toxic comments, with M5S having the larger value.

Lastly, we look at the distribution of users that were suspended (as of February 2023) across the coalitions, and their average toxicity score. As shown in panel(A) of Figure 7, the Right coalition has the largest proportion of suspended users (3%) compared to the others (< 1%). For what concerns the average toxicity of suspended versus active users (as of February 2023), we notice that the former group is generally more toxic than the latter for all coalitions except for Right supporters.

Findings and Remarks: We showed that Twitter users a liated with di erent coalitions' of the Italian political spectrum exhibit small but signi cant di erences in the levels of online toxic speech. Replies, in particular, tend to be particular more toxic than other kinds of tweets. Users a liated with the M5S were those exhibiting the highest toxicity, compared to other coalitions'. Overall, a small but non-negligible proportion (6-8%) of all political messages shared by Italian Twitter users with a clear political a liation was employed abusive language. Moreover, users that eventually got suspended by the platform were more likely to employ abusive language.

4.2 Targets of hate

To answer (RQ2) , i.e., whether there are political and gender di erences in targets of hate, we focus on replies, i.e., responses to another person's post. According to Twitter's guidelines[7], when two users reply to one another, only other relevant users, such as their followers, will see the reply in their timeline. Speci cally, we analyze 732 765 replies received by 297 distinct politicians.

We rst compute the distribution of the average toxicity of replies received by each politician in the coalitions. As shown in panel A of Figure 8, there are signi cant di erences across distributions (Kruskal-Wallis $P < 0.001$), with Centre-Left (Median = 0.22) and M5S (Median = 0.20) politicians receiving replies that are slightly more toxic on average, compared to Third Pole (Median = 0.18) and Right (Median = 0.19).

Panel B of the same gure shows, instead, the distribution of the number of toxic replies received by each candidate, i.e., replies with a toxicity score above 0.5. We can observe very di erent patterns, with Third Pole (Median = 58) and Centre-Left (Median = 44) politicians receiving much more toxic replies than M5S (Median = 9) and Right ones (Median = 9).

We next perform the same analysis disentangling the gender of di erent candidates. As shown in panel (A) of Figure 9, only M5S and Third Pole exhibit di erences in the toxicity of the average reply received by male and female politicians. Panel (B) , instead, shows that the number of toxic replies is signi cantly larger for female than male politicians only for Centre-Left and Right, while the converse applies for M5S and Third Pole (Kruskal-Wallis $P < 0.001$ for all pairwise distributions).

Lastly, we show in Figure 10 the Top-15 politicians that received the largest number of toxic replies (the 1st and the 15th in this ranking received respectively 18 483 and 480 toxic replies. We can observe that all coalitions' leaders appear among the main targets of hate speech (Enrico Letta, Matteo Salvini, Carlo Calenda, Giuseppe Conte and Giorgia Meloni), and that half of the politicians in this ranking are female.

Findings and Remarks: We highlighted that politicians across the political spectrum receive replies with heteregenous levels of toxicity. In particular, Centre-Left and Third Pole politicians receive a larger number of toxic replies compared to other coalitions. We also observed that only Centre-Left and Right female politicians receive many more toxic replies than male ones, while the viceversa occurs for M5S and Third Pole. Finally, we saw that the top targets of hate include leaders of the di erent coalitions but also many female politicians.

4.3 Determinants of toxicity

To address (RQ3) , i.e., identify the determinants of toxic speech targeting politicians, we employ multiple generalized linear regression models (GLM) to analyze which factors drive toxicity of interactions between regular users and politicians. Following previous literature [29, 31], we consider the following set of control variables that might explain toxicity:

- Categorical variables encoded with dummy variables: Coalition (reference: Centre-Left), Gender (Male = 1), Twitter Veri cation (Veri ed = 1).

[7] https://help.twitter.com/en/using-x/mentions-and-replies

- Continuous variables: Number of Followers (log scale), Number of Tweets Shared (log scale), Average Toxicity of Tweets Shared.

In the first model, our dependent variable y is the probability of a reply received by politicians to be toxic: $y = 1(t > 0.5)$ where t = toxicity score. The number of observations is $N = 732506$. We fit a Logit (or Logistic) model, which is more suitable than a Linear one for binary variables (pseudo $R^2 = 0.006$ vs $R^2 = 0.004$). As shown in panel A of Figure 11, all variables are statistically significant ($P < 0.001$) but they have different magnitudes and signs. For what concerns political affiliation, Centre-Left politicians receive the most toxic replies (cf. odds ratio below 1 for other coalitions), with Third Pole (odds ratio = 0.61) politicians receiving the least toxic replies[8]. Male politicians (odds ratio = 1.31) with a verified account (odds ratio = 1.23) also tend to receive replies that are more likely to be toxic, on average. The number of followers and shared tweets have small yet significant opposite effects (resp. odds ratio = 1.03 and odds ratio = 0.95). Lastly, the stronger effect is given by the average toxicity generated by the candidate receiving a reply, with odds ratio = 8.15.

In the second model, our dependent variable y is the number of replies received by each politician that are toxic, i.e., with $t > 0.5$. We apply the transformation $y' = \log(y + 1)$ to account for politicians that received 0 toxic replies. The number of observations is $N = 254$. We fit both a Negative Binomial and a Poisson model, which are more suitable than a Linear one as the variable represents counts, and report results for the latter which exhibits a higher fit (pseudo $R^2 = 0.23$ vs $R^2 = 0.29$). As shown in panel B of Figure 11, differently from the previous regression not all variables are statistically significant ($P < 0.001$). For what concerns political affiliation, Centre-Left politicians still receive a larger amount of toxic replies (cf. negative sign of other coalitions), with Third Pole (odds ratio = 0.72) politicians receiving the least amount of toxic replies. Male politicians (odds ratio = 0.81) receive a smaller number of toxic replies compared to female politicians. Those with a verified account (odds ratio = 1.22) receive more toxic replies. The number of followers and shared tweets have small yet significant effects (resp. odds ratio = 1.22 and odds ratio = 1.25). Lastly, also in this case an extremely strong effect is given by the average toxicity generated by candidates, with odds ratio = 246.6.

Findings and Remarks: we employed two GLMs to estimate the probability for candidates of receiving a toxic reply, and the number of toxic replies received, using a set of features that could explain these variables. We found that party affiliation and gender were both significant drivers of hate speech targeting politicians. Importantly, we observed that politicians employing more abusive language are more likely to receive hate speech messages.

4.4 Hate speech among coalitions' supporters

As an additional analysis, we examine how frequently supporters of different coalitions engage in toxic interactions. To measure this, we calculate the proportion of toxic

[8] We remark that the Logistic regression coefficients correspond to the logarithm of the odds of a reply being toxic.

replies (toxicity score > 0:5) exchanged between users from di erent political coalitions, including interactions within the same party and excluding tweets directed to or from politicians. Figure 12 shows that, on average, out-group interactions tend to be more toxic (o -diagonal values in the range [5:36; 11:84%) than in-group ones (values in the range [4:80; 5:56], with M5S supporters being the most toxic when replying to Third Pole supporters (11:84%) or the Right coalition (9:22%). The second most toxic group appears to be Right coalition supporters, aligning with previous ndings. Even within the same coalition, there is a small yet non-negligible amount of toxic in-group exchanges, signaling that political discussions can often become a breeding ground for harmful communication even among users sharing the same ideology. Findings and Remarks: we computed the proportion of toxic interactions between users a liated with di erent political groups, nding that out-group interactions are more toxic than in-group exchanges, and that M5S and Right supporters are the most hateful in their out-group interactions.

5 Discussion

5.1 Contributions

We studied the prevalence of toxic and hateful speech in political conversations on Twitter, in the run up to the 2022 Italian General election.

We found that supporters of di erent political coalitions exhibit di erent levels of toxic speech, which is conveyed more through replies than other kinds of tweets. This might be due to the fact that replies allow users to directly engage with contentious or polarizing content, which can lead to more aggressive and hostile language. Additionally, replies tend to be spontaneous reactions that may lack the same level of thoughtfulness. In contrast, original tweets are typically crafted to express ideas, share information, or initiate discussions, making them more deliberate and controlled. Reply tweets, being reactive and often written in the heat of the moment, tend to be more emotionally charged and less re ective. Additionally, the anonymity provided by online platforms can lead users to express themselves more assertively, which may further contribute to the prevalence of aggressive language in these interactions. Replying to politicians' tweets ts within targeted political participation, which refers to direct interactions with politicians. Expressive participation, on the other hand, involves conveying political opinions in a more general manner, such as through yard signs or posting undirected tweets [43].

We observed that M5S and Right followers were on average the most hateful users, with approx. 8% of shared tweets that were toxic, and a median number of 3-4 toxic replies per user. Supporters of M5S often exhibit higher levels of toxicity due to the party's strong anti-establishment stance and populist rhetoric, which fosters animosity towards traditional political gures and institutions, leading to more aggressive and confrontational behavior [44]. Additionally, the high engagement of M5S supporters on social media platforms, where anonymity and immediacy can intensify hostile interactions, might further contribute to the elevated levels of toxicity [45]. Similarly, the adoption of populist and nationalist rhetoric by some factions within the Right

can also exacerbate toxic discourse [46]. Users who eventually got suspended by Twitter exhibited a higher usage of abusive language across all coalitions except for the Right, which however exhibits the largest percentage of suspended users (3%).

We also uncovered a variation in hate speech targets, with Centre-Left and Third Pole politicians receiving a higher volume of toxic replies, on average, compared to other coalitions. This is likely because these politicians were targeted by M5S and Right supporters, who display the highest rates of abusive behavior. Centre-Left politicians have historically been targeted by hate speech from M5S supporters and politicians, re ecting a long-standing pattern of animosity and political con ict between these groups. This trend underscores the contentious relationship and frequent hostility that Centre-Left gures encounter from the M5S faction, often resulting in a disproportionate amount of negative and abusive rhetoric directed at them. Besides, Matteo Renzi and Carlo Calenda, both members of the Third Pole, are particularly controversial gures, which likely contributes to the signi cant volume of toxic speech they attract.

Female politicians received more toxic replies only in the case of Centre-Left and Right coalitions, while the converse applies to the other coalitions { most likely because of the smaller number of female politicians with an associated Twitter account. The increased visibility on social media of female politicians can amplify hostile interactions and lead to a greater volume of toxic responses. Nevertheless, the dynamics and nature of online interactions may vary depending on the coalition's political landscape and the presence of female politicians in prominent positions.

We delved into an analysis of factors that could explain di erent levels of toxic speech received by politicians, nding that political a liation is a signi cant variable: Third Pole politicians were the least likely to receive toxic replies, while Centre-Left ones exhibit the highest probability; Male politicians received on average replies that are more toxic, but Female politicians received more toxic replies overall; politicians with a Veri ed account, more active on Twitter and with a larger number of followers were more likely to be targets of hate speech. Most importantly, politicians that generated on average more toxic conversations were also more likely to receive hateful speech and more toxic replies.

Lastly, when considering the frequency of toxic interactions, we nd that cross-party exchanges are signi cantly more toxic than within-party interactions, with M5S and Right supporters being the most toxic political groups. This highlights the heightened potential for harmful communication in polarized political discussions.

5.2 Implications

The implications of our ndings are multifaceted.

We observed that some dynamics of online hate speech in Italy's multi-party system align with ndings from bipolar political environments, such as those in the US and the UK. Consistent with results from [27, 28, 32], we observe that more prominent candidates, i.e., those with a larger follower base, are more likely to be targeted by hate speech. Additionally, our analysis con rms that female candidates receive a higher proportion of hate speech on average. Furthermore, [31] reports that candidates who share more negative content tend to attract more hate, which aligns with our nding

that candidates using more toxic language experience a higher volume of harmful interactions. Similarly, [27, 32] identi es that right-leaning users are more active in generating hate speech, which is consistent with our results. However, some analyses of regression coe cients from bipolar political systems, such as Hua et al. [29]'s nding that Democrats were more frequently targeted, are less straightforward to translate to Italy's multi-party context.

To address the spread of toxic speech in political discourse on social media, one practical solution is the development of automated moderation systems using machine learning algorithms that can ag or limit the visibility of toxic replies in real-time. This would reduce the reach of harmful content while minimizing manual intervention. Partnering with social media platforms to re ne these systems, along with increasing transparency in how harmful speech is managed, could enhance public trust in moderation e orts. Additionally, platforms could adopt a tiered approach to intervention|ranging from warnings for minor infractions to temporary bans for repeat o enders|based on the severity of the toxicity. Nevertheless, our analysis shows that Twitter did not consistently enforce its policies against toxic speech, which may contribute to the ongoing prevalence of harmful content on the platform. Despite the existence of moderation guidelines, the e ectiveness of their application appears inconsistent, potentially allowing toxic behavior to persist and negatively impact the discourse. This highlights the need for more robust and uniform enforcement practices to e ectively mitigate the spread of harmful content and promote a healthier online environment.

Our research also underscores the need for collaborative policy development between governments, tech companies, and civil society to address online toxicity more holistically. This could include public awareness campaigns aimed at promoting civil discourse or legislation focused on ensuring accountability for toxic behavior while safeguarding free speech. Furthermore, fostering an online environment where users are encouraged to self-moderate through reporting mechanisms could empower communities to curb harmful speech without heavy-handed oversight from platforms. The European Digital Services Act (DSA) represents a signi cant regulatory e ort aimed at enhancing the accountability and transparency of online platforms across Europe. Implemented to address the challenges posed by digital services, the DSA establishes clear rules for content moderation, requiring platforms to take more proactive measures against illegal content and misinformation. It mandates greater transparency in how content is moderated, and promotes user empowerment by enforcing clearer mechanisms for reporting and appealing decisions. By setting higher standards for platform responsibility and user protection, the DSA seeks to create a safer and more equitable digital environment, ensuring that online platforms uphold democratic values and protect users from harmful content.

Additionally, increasing digital literacy among users is a vital step toward fostering a healthier online space. By promoting digital education initiatives in schools and through public programs, we can equip users with the skills to critically assess content, recognize misinformation, and avoid participating in or spreading toxic speech. However, implementing digital literacy programs at scale presents challenges such as funding, curriculum development, and reaching at-risk populations, which will

need to be addressed through multi-stakeholder partnerships involving educators, policymakers, and technology companies.

Finally, our findings contribute to broader discussions on balancing content moderation with free speech rights. We acknowledge that this balance is complex and context-dependent. Carefully crafted legal frameworks, drawing on input from diverse stakeholders, can help ensure moderation practices are both effective and aligned with democratic values. For example, clear guidelines on what constitutes harmful speech, combined with transparent and appealable moderation decisions, could prevent overreach while still addressing toxicity effectively. By considering ethical and legal challenges, we can develop a framework that promotes a respectful online environment without compromising the right to free expression.

5.3 Limitations

Our work comes with some limitations.

First, we estimate the toxicity of tweets using a black-box machine learning model that might not be completely accurate [16], and we did not perform a manual evaluation of its performance. This would be particularly challenging given that the classifier is based on a Large Language Model that estimates the probability for a text to be toxic. However, it is the most widely employed tool by the research community for analyzing the integrity of online social platforms and understanding the spread of hateful speech, being cited over 1400 times on Google Scholar.

Next, we do not filter out for automated and inauthentic accounts that might be present in our data sample, as this is not a trivial task and there is no ground truth information [47]. We do perform an analysis of suspended users, although we cannot discern the actual reasons behind the platform's intervention.

Also, we use retweeting as a proxy for endorsement to estimate the political affiliation of users, and this might not correspond to their actual leaning. Nevertheless, research indicates that instead quotes and replies are often used to express a strong negative reaction or disagreement with the original post (\ratioing") in politically polarized environments [38].

Moreover, in our analyses of determinants of toxicity we do not take into account the content of tweets shared by politicians that receive toxic replies, which might play a relevant role in attracting abusive and harmful language, and we leave it for future analyses.

The Logistic model employed in the regression analysis yields a low predictive power{ despite showing than R^2 values higher than a Linear model{ which is likely due to the very large number of observation ($N = 732506$), and this may limit the generalizability of the conclusions drawn from the results.

Furthermore, we did not distinguish cases when original tweets convey hateful speech and its replies are positive and supporting the hate expressed by the original poster.

Other potential biases in the methodology could arise from the over-representation of certain coalitions due to retweet behavior, where parties with more active or larger follower bases may dominate the conversation. This can skew the results, making

it appear as though these parties have a disproportionate in uence or popularity compared to others.

Finally, Twitter's user-base is not fully representative of the Italian general demographics[9] and it is also not the most used platform[10], therefore our result might not generalize to the actual population.

5.4 Future work

Future avenues of research should consider performing comprehensive comparative analyses of toxic speech across di erent social media platforms, including but not limited to Facebook or Reddit, highlighting similarities and di erences in how o ensive and derogatory language manifests across online ecosystems. Additionally, future investigations could study di erent topics of online conversation to understand whether the political landscape is a stronger catalyzer of abusive and harmful language. Lastly, harnessing novel NLP technologies such as Large Language Models might allow researchers to gain deeper insights into the underlying mechanisms driving negative interactions. This, in turn, would allow more e ective communication strategies aimed at mitigating the proliferation of hate speech in digital environments.

List of abbreviations

- API = Application Programming Interface
- GLM = Generalized Linear Model
- NLP = Natural Language Processing

Declarations

Availability of data and material

A list of tweet IDs is publicly available [11] in the repository associated to the resource paper [33]. In the past, these could be used to retrieve tweet objects, except for removed or protected tweets, by querying Twitter APIs that were freely available to researchers and practitioners. As of February 2024, we acknowledge that a paid endpoint is required to (re)collect large-scale datasets of tweets, and we encourage authors interested in analyzing this data to contact the corresponding author. The repository also contains the list of Twitter accounts associated with elected representatives of the Italian Senate and Chamber of Deputies described in the next section. Our data analysis adheres to ethical research standards by focusing solely on accounts associated with Italian politicians, without attempting to identify anonymized users. We consider this approach reasonable given the signi cance of political accountability and transparency.

[9] https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/
[10] https://transparency.twitter.com/dsa-transparency-report.html
[11] https://github.com/frapierri/ita-election-2022

Competing interests

The author declares he is an Associate Editor of EPJ Data Science.

Funding

This work was supported in part by the European Union (NextGenerationEU project PNRR-PE-AI FAIR) and the Italian Ministry of Education (PRIN PNRR grant CODE prot. P2022AKRZ9 and PRIN grant DEMON prot. 2022BAXSPY).

Author contributions

F.P. formulated the research questions, analyzed the data, ran the experiments and wrote the manuscript.

Acknowledgments

During the preparation of this work the author(s) used OpenAI chatGPT in order to proof-check the grammar of some paragraphs and re ne the language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

[1] Fujiwara, T., Mu ller, K., Schwarz, C.: The e ect of social media on elections: Evidence from the united states. Journal of the European Economic Association, 058 (2023) https://doi.org/10.1093/jeea/jvad058

[2] Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., Starnini, M.: The echo chamber e ect on social media. Proceedings of the National Academy of Sciences 118(9), 2023301118 (2021) https://doi.org/10.1073/pnas.2023301118

[3] Jhaver, S., Boylston, C., Yang, D., Bruckman, A.: Evaluating the e ectiveness of deplatforming as a moderation strategy on twitter. Proceedings of the ACM on Human-Computer Interaction 5(CSCW2), 1{30 (2021) https://doi.org/10.1145/3479525

[4] Terizi, C., Chatzakou, D., Pitoura, E., Tsaparas, P., Kourtellis, N.: Modeling aggression propagation on social media. Online Social Networks and Media 24, 100137 (2021) https://doi.org/10.1016/j.osnem.2021.100137

[5] Ferrara, E., Cresci, S., Luceri, L.: Misinformation, manipulation, and abuse on social media in the era of covid-19. Journal of Computational Social Science 3, 271{277 (2020) https://doi.org/10.1007/s42001-020-00094-5

[6] Gallotti, R., Valle, F., Castaldo, N., Sacco, P., De Domenico, M.: Assessing the risks of `infodemics' in response to covid-19 epidemics. Nature human behaviour 4(12), 1285{1293 (2020) https://doi.org/10.1038/s41562-020-00994-6

[7] Shahi, G.K., Dirkson, A., Majchrzak, T.A.: An exploratory study of covid-19 misinformation on twitter. Online social networks and media 22, 100104 (2021) https://doi.org/10.1016/j.osnem.2020.100104

[8] Trevisan, M., Vassio, L., Giordano, D.: Debate on online social networks at the time of covid-19: An italian case study. Online Social Networks and Media23, 100136 (2021) https://doi.org/10.1016/j.osnem.2021.100136

[9] Horta Ribeiro, M., Hosseinmardi, H., West, R., Watts, D.J.: Deplatforming did not decrease parler users' activity on fringe social media. PNAS nexus2(3), 035 (2023) https://doi.org/10.1093/pnasnexus/pgad035

[10] Pradel, F., Zilinsky, J., Kosmidis, S., Theocharis, Y.: Toxic speech and limited demand for content moderation on social media. American Political Science Review, 1{18 (2024) https://doi.org/10.1017/S000305542300134X

[11] Mondal, M., Silva, L.A., Benevenuto, F.: A measurement study of hate speech in social media. In: Proceedings of the 28th ACM Conference on Hypertext and Social Media, pp. 85{94 (2017). https://doi.org/10.1145/3078714.3078723

[12] Yin, W., Zubiaga, A.: Hidden behind the obvious: Misleading keywords and implicitly abusive language on social media. Online Social Networks and Media 30, 100210 (2022) https://doi.org/10.1016/j.osnem.2022.100210

[13] Mladenovic, M., Osmjanski, V., Stankovic, S.V.: Cyber-aggression, cyberbullying, and cyber-grooming: a survey and research challenges. ACM Computing Surveys (CSUR) 54(1), 1{42 (2021) https://doi.org/10.1145/3424246

[14] Sheth, A., Shalin, V.L., Kursuncu, U.: De ning and detecting toxicity on social media: context and knowledge are key. Neurocomputing490, 312{318 (2022) https://doi.org/10.1016/j.neucom.2021.11.095

[15] Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L.A., Anderson, K., Kohli, P., Coppin, B., Huang, P.-S.: Challenges in detoxifying language models. In: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 2447{2469 (2021). https://doi.org/10.18653/v1/2021. ndings-emnlp. 210

[16] Nogara, G., Pierri, F., Cresci, S., Luceri, L., Törnberg, P., Giordano, S.: Toxic bias: Perspective api misreads german as more toxic. ICWSM (2025)

[17] Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: Proceedings of the 26th Annual Computer Security Applications Conference,

pp. 1{9 (2010). https://doi.org/10.1145/1920261.1920263

[18] Yang, C., Harkreader, R., Zhang, J., Shin, S., Gu, G.: Analyzing spammers' social networks for fun and pro t: a case study of cyber criminal ecosystem on twitter. In: Proceedings of the 21st International Conference on World Wide Web, pp. 71{80 (2012). https://doi.org/10.1145/2187836.2187847

[19] Ferrara, E.: The history of digital spam. Communications of the ACM 62(8), 82{91 (2019) https://doi.org/10.1145/3299768

[20] Addawood, A., Badawy, A., Lerman, K., Ferrara, E.: Linguistic cues to deception: Identifying political trolls on social media. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, pp. 15{25 (2019). https://doi.org/10.1609/icwsm.v13i01.3205

[21] Ribeiro, M., Calais, P., Santos, Y., Almeida, V., Meira Jr, W.: Characterizing and detecting hateful users on twitter. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 12 (2018). https://doi.org/10.1609/icwsm.v12i1.15057

[22] Ejaz, N., Razi, F., Choudhury, S.: Towards comprehensive cyberbullying detection: A dataset incorporating aggressive texts, repetition, peerness, and intent to harm. Computers in Human Behavior 153, 108123 (2024) https://doi.org/10.1016/j.chb.2023.108123

[23] Yi, P., Zubiaga, A.: Session-based cyberbullying detection in social media: A survey. Online Social Networks and Media36, 100250 (2023) https://doi.org/10.1016/j.osnem.2023.100250

[24] Pamungkas, E.W., Basile, V., Patti, V.: Towards multidomain and multilingual abusive language detection: a survey. Personal and Ubiquitous Computing27(1), 17{43 (2023) https://doi.org/10.1007/s00779-021-01609-1

[25] Thomas, K., Akhawe, D., Bailey, M., Boneh, D., Bursztein, E., Consolvo, S., Dell, N., Durumeric, Z., Kelley, P.G., Kumar, D., et al.: Sok: Hate, harassment, and the changing landscape of online abuse. In: 2021 IEEE Symposium on Security and Privacy (SP), pp. 247{267 (2021). https://doi.org/10.1109/SP40001.2021.00028 . IEEE

[26] Sanguinetti, M., Comandini, G., Di Nuovo, E., Frenda, S., Stranisci, M., Bosco, C., Caselli, T., Patti, V., Russo, I.: Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task. Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2020)

[27] Gorrell, G., Bakir, M.E., Roberts, I., Greenwood, M.A., Bontcheva, K.: Which politicians receive abuse? four factors illuminated in the uk general election 2019. EPJ Data Science 9(1), 18 (2020) https://doi.org/10.1140/epjds/

s13688-020-00236-9

[28] Hua, Y., Naaman, M., Ristenpart, T.: Characterizing twitter users who engage in adversarial interactions against political candidates. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2020). https://doi.org/10.1145/3313831.3376548

[29] Hua, Y., Ristenpart, T., Naaman, M.: Towards measuring adversarial twitter interactions against candidates in the us midterm elections. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, pp. 272–282 (2020). https://doi.org/10.1609/icwsm.v14i1.7298

[30] Agarwal, P., Hawkins, O., Amaxopoulou, M., Dempsey, N., Sastry, N., Wood, E.: Hate speech in political discourse: A case study of uk mps on twitter. In: Proceedings of the 32nd ACM Conference on Hypertext and Social Media, pp. 5–16 (2021). https://doi.org/10.1145/3465336.3475113

[31] Solovev, K., Pröllochs, N.: Hate speech in the political discourse on social media: Disparities across parties, gender, and ethnicity. In: Proceedings of the ACM Web Conference 2022, pp. 3656–3661 (2022). https://doi.org/10.1145/3485447.351226

[32] Gorrell, G., Greenwood, M., Roberts, I., Maynard, D., Bontcheva, K.: Twits, twats and twaddle: Trends in online abuse towards uk politicians. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 12 (2018). https://doi.org/10.1609/icwsm.v12i1.15070

[33] Pierri, F., Liu, G., Ceri, S.: Ita-election-2022: A multi-platform dataset of social media conversations around the 2022 italian general election. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. CIKM '23, pp. 5386–5390. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3583780.3615121

[34] Yang, K.-C., Hui, P.-M., Menczer, F.: Bot electioneering volume: Visualizing social bot activity during elections. In: Companion Proceedings of The 2019 World Wide Web Conference, pp. 214–217 (2019). https://doi.org/10.1145/3308560.3316499

[35] Conover, M.D., Gonçalves, B., Flammini, A., Menczer, F.: Partisan asymmetries in online political activity. EPJ Data science **1**(1), 1–19 (2012) https://doi.org/10.1140/epjds6

[36] Stamatelatos, G., Gyftopoulos, S., Drosatos, G., Efraimidis, P.S.: Revealing the political affinity of online entities through their twitter followers. Information Processing & Management **57**(2), 102172 (2020) https://doi.org/10.1016/j.ipm.2019.102172

[37] Metaxas, P., Mustafaraj, E., Wong, K., Zeng, L., O'Keefe, M., Finn, S.: What do

retweets indicate? results from user survey and meta-review of research. Proceedings of the International AAAI Conference on Web and Social Media (ICWSM) (2015) https://doi.org/10.1609/icwsm.v9i1.14661

[38] Guerra, P., Nalon, R., Assunçao, R., Meira Jr, W.: Antagonism also flows through retweets: The impact of out-of-context quotes in opinion polarization analysis. In: Proceedings of the International Aaai Conference on Web and Social Media, vol. 11, pp. 536–539 (2017). https://doi.org/10.1609/icwsm.v11i1.14971

[39] Barberá, P., Jost, J.T., Nagler, J., Tucker, J.A., Bonneau, R.: Tweeting from left to right: Is online political communication more than an echo chamber? Psychological science **26**(10), 1531–1542 (2015) https://doi.org/10.1177/0956797615594620

[40] Sharma, K., Ferrara, E., Liu, Y.: Characterizing online engagement with disinformation and conspiracies in the 2020 us presidential election. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, pp. 908–919 (2022). https://doi.org/10.1609/icwsm.v16i1.19345

[41] Lees, A., Tran, V.Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., Vasserman, L.: A new generation of Perspective API: Efficient multilingual character-level transformers. In: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'22), pp. 3197–3207 (2022). https://doi.org/10.1145/3534678.3539147

[42] Saveski, M., Roy, B., Roy, D.: The structure of toxic conversations on twitter. In: Proceedings of the Web Conference 2021, pp. 1086–1097 (2021). https://doi.org/10.1145/3442381.3449861

[43] An, Z., Joseph, K.: An analysis of replies to trump's tweets. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 15, pp. 49–60 (2021). https://doi.org/10.1609/icwsm.v15i1.18040

[44] Serafini, L., Gottlob, A., Pierri, F., Ieva, F., Ceri, S.: Risk narratives on immigration during the covid-19 crisis in italy: A comparative analysis of facebook posts published by politicians and by news media. Journal of Communication Inquiry, 01968599231167939 (2023) https://doi.org/10.1177/01968599231167939

[45] Combei, C.R., Giannetti, D.: The immigration issue on twitter political communication. italy 2018-2019. Comunicazione politica **21**(2), 231–263 (2020) https://doi.org/10.3270/97905

[46] Bracciale, R., Martella, A.: Define the populist political communication style: the case of italian political leaders on twitter. Information, communication & society **20**(9), 1310–1329 (2017) https://doi.org/10.1080/1369118X.2017.1328522

[47] Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social

bots. Communications of the ACM **59**(7), 96–104 (2016) https://doi.org/10.1145/2818717

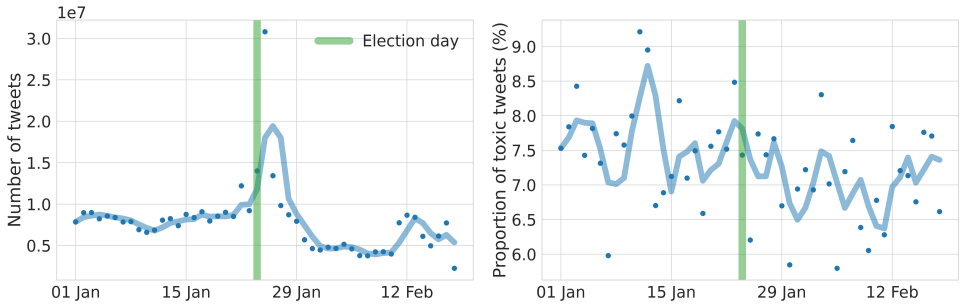

Fig. 1 Time series of the number of tweets in the analyzed dataset (left) and daily proportion of tweets that are classified as toxic (right, toxicity > 0.5).

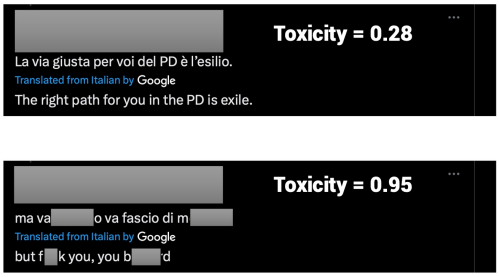

Fig. 2 Examples of tweets in our dataset and their toxicity score. We hide the username and censor extreme language.

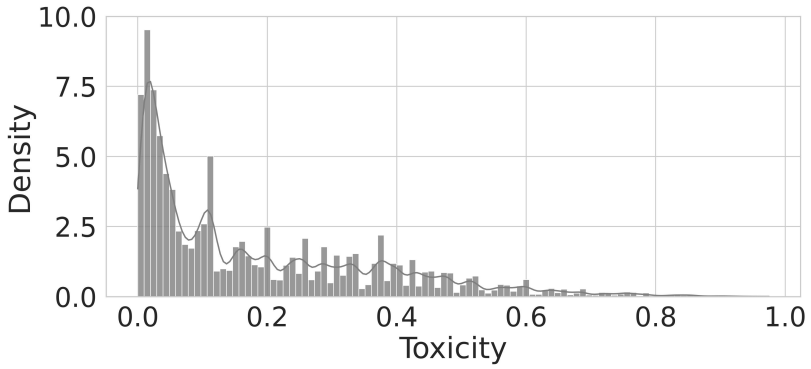

Fig. 3 Distribution of the toxicity scores of tweets in our dataset. The histogram is computed on 100 equal-width bins. Summary statistics: Mean = 0.18, 25th = 0.29, 50th = 0.11, 75th = 0.29, Max = 0.98.

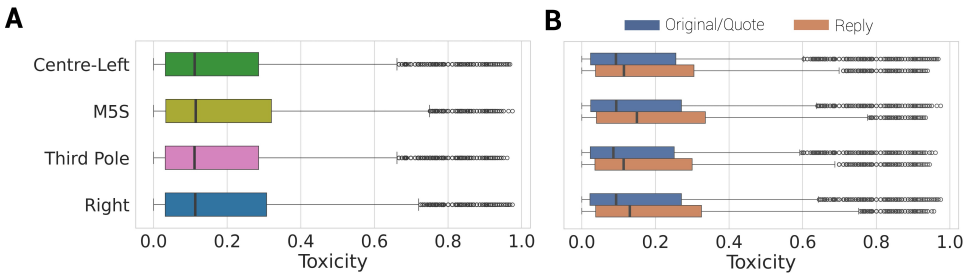

Fig. 4 Toxicity of tweets shared by coalitions' supporters. (A) Distribution of toxicity for tweets shared by users of different coalitions. Median value is 0.11 for all distributions. **(B)** Distribution of toxicity for tweets shared by users of different coalitions separating original/quote tweets from replies. Median values are: Centre-Left Original/Quote = 0.09, Reply = 0.11; M5S Original/Quote = 0.09, Reply = 0.15; Third Pole Original/Quote = 0.08, Reply = 0.11; Right Original/Quote = 0.09, Reply = 0.13. Distributions are statistically different according to Kruskal-Wallis test ($P < 0.001$).

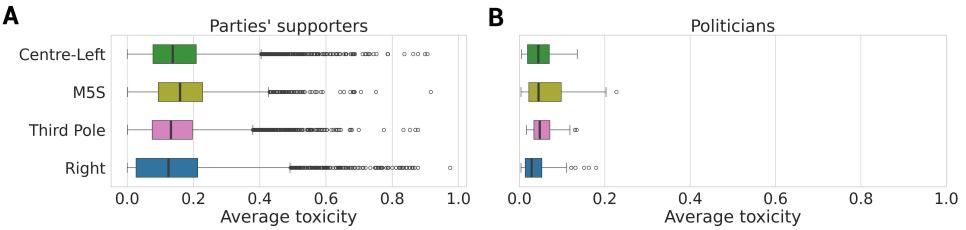

Fig. 5 Average toxicity of coalitions' supporters and politicians computed from the tweets they shared. (A) Average toxicity of coalitions' supporters. Median values are: Centre-Left = 0.14; M5S = 0.16; Third Pole = 0.13; Right = 0.12. Distributions are statistically different according to Kruskal-Wallis test ($P < 0.001$). **(B)** Average toxicity of coalitions' politicians. Median value is 0.04 for all coalitions except Right (0.02).

elezioni	partito democratico	berlusconi
renzi	movimento 5 stelle	salvini
calenda	di maio	politiche2022
meloni	elezioni2022	conte

Table 1 A sample of Italian language keywords related to the 2022 election that were used to collect the `ITA-ELECTION-2022` dataset.

Figure legends

Table legends

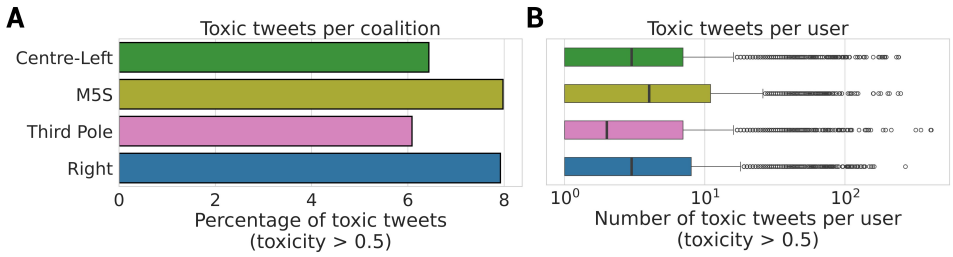

Fig. 6 Toxic tweets shared by coalitions' supporters. (A) Proportion of tweets shared by each coalition that are toxic. Values are: Centre-Left = 6.43, M5S = 7.97, Third Pole = 6.08, Right = 7.92. **(B)** Number of toxic tweets shared by coalitions' supporters. Median values are: Centre-Left = 3, M5S = 4, Third Pole = 2, Right = 3. Distributions are statistically different according to Kruskal-Wallis test ($P < 0.001$).

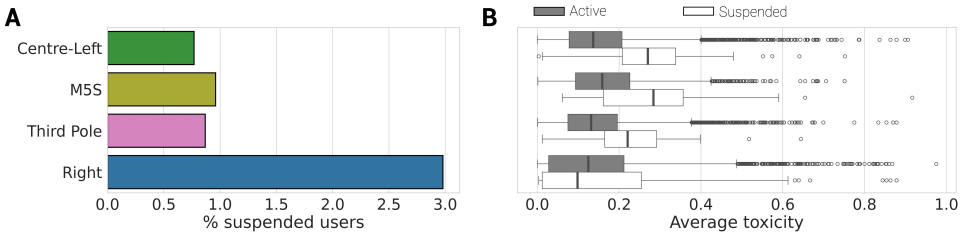

Fig. 7 Suspended users across coalitions. (A) Percentage of suspended users across coalitions. Values are: Centre-Left = 0.76, M5S = 0.95, Third Pole = 0.86, Right = 2.96. **(B)** Distribution of average user toxicity across coalitions. Median values are: Centre-Left Active = 0.13, Suspended = 0.27; M5S Active = 0.15, Suspended = 0.28; Third Pole Active = 0.13, Suspended = 0.22; Right Active = 0.12, Suspended = 0.09. Distributions of Active versus Suspended users are pairwise statistically different according to Kruskal-Wallis test ($P < 0.001$).

Coalition	Users	Tweets (excl. retweets)
Centre-Left	15173	440174
M5S	5307	251693
Right	58262	450945
Third Pole	13401	422364
Total	92143	1565176

Table 2 Number of users with political affiliation estimated in each coalition, and the number of tweets they shared.

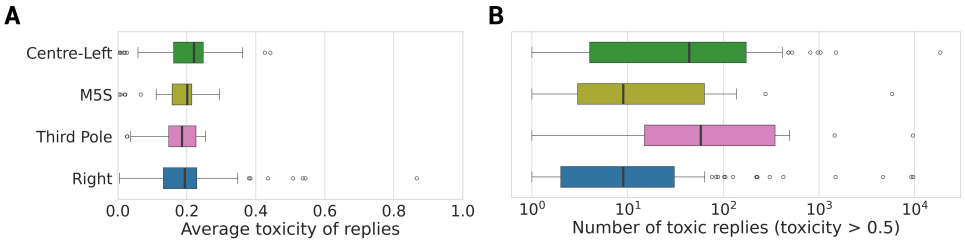

Fig. 8 Toxicity of replies received by politicians. (A) Average toxicity of replies received by politicians. Median values are: Centre-Left = 0.20, M5S = 0.22, Third Pole = 0.18, Right = 0.19. Distributions are statistically different according to a Kruskal-Wallis test ($P < 0.001$). **(B)** Number of toxic replies (toxicity > 0.5) received by politicians. Median values are: Centre-Left = 44, M5S = 9, Third Pole = 58, Right = 9. Distributions are statistically different according to a Kruskal-Wallis test ($P < 0.001$).

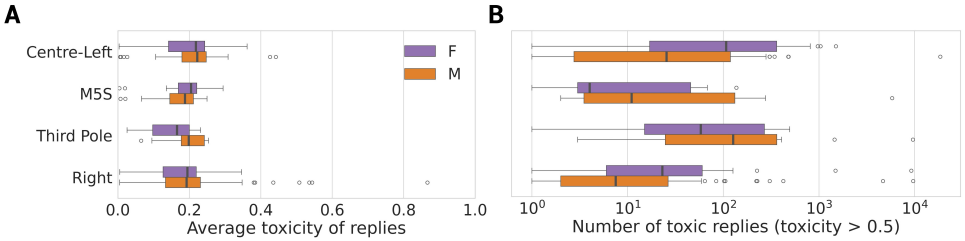

Fig. 9 Toxicity of replies received by politicians, by gender. (A) Average toxicity of replies received by politicians by gender. Median values are: Centre-Left M = 0.22, F = 0.21; M5S M = 0.18, F = 0.20; Third Pole M = 0.19, F = 0.16; Right M = 0.19, F = 0.19. Distributions are statistically different according to a Kruskal-Wallis test ($P < 0.001$). **(B)** Number of toxic replies (toxicity > 0.5) received by politicians. Median values are: Centre-Left = 44, M5S = 9, Third Pole = 58, Right = 9. Distributions are statistically different according to a Kruskal-Wallis test ($P < 0.001$).

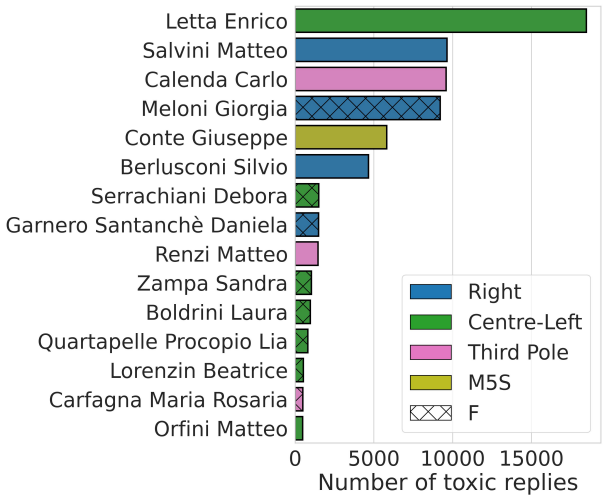

Fig. 10 Top 15 politicians ranked by the total number of toxic replies (toxicity > 0:5) received.

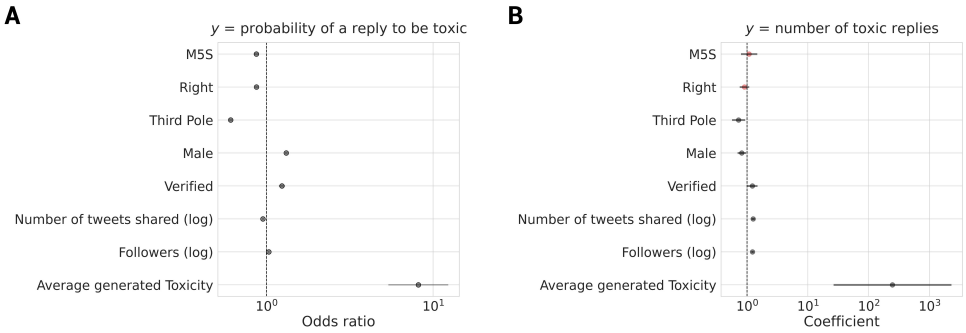

Fig. 11 Odds ratio values of the GLMs. (A) Odds ratio values of the Logistic Regression model fitted to predict the probability of a reply received by a politician to be toxic. All coefficients are statistically significant at = 0:05. Error bars represent the 95% C.I. but they might be not visible because of very small values. **(B)** Odds ratio values of the Poisson Regression model fitted to predict the number of toxic replies received by politicians. All coefficients are statistically significant at = 0:05 except for those in red (M5S and Right). Error bars represent the 95% C.I. but they might be not visible because of very small values.