A Topology-Based Approach for Predicting Toxic Outcomes on Twitter and YouTube

Gabriele Etta, Matteo Cinelli, Niccolò Di Marco, Michele Avalle, Alessandro Panconesi, Walter Quattrociocchi

***Abstract*—The benefits of an information ecosystem based on social media platforms came at the cost of the rise of several antisocial behaviours, including the use of toxic speech. To assess the aspects that concur with the formation of toxic conversations, we provide a multi-platform comparison on Twitter and YouTube between the 2022 Italian Political Elections, representing a potentially polarising topic, and the Italian Football League, a topic close to the country's popular culture. We first probe structural and conversational toxicity differences by analyzing 257K conversations (3.7M posts, 1M users) on both platforms. Then, we provide a machine learning approach that, by leveraging the previous features, identifies the presence of the following toxic comment in different stages of conversations. We show that football tends to exhibit lower toxicity levels than politics, with the latter producing more extended conversations that attract a broader audience and consequently fostering the polarization phenomenon. The implemented classifiers resulting from the conversation stage-based approach achieve state-of-the-art performances despite a restricted set of features. Furthermore, our cross-topic comparison shows that models trained on a divisive topic can be applied to other discussions without causing a degradation of their performance.**

***Index Terms*—social media, hate speech, information cascades, moderation**

I. Introduction

Social media platforms have reshaped how users inform themselves and participate in online discussions [1]–[3]. Indeed, the microblogging features and the decentralized scheme proposed by these platforms provided the opportunity to be involved in an unprecedented number of debates, with the result of promoting the emergence of new ideas [4] and becoming rapidly aware of a multitude of topics [5]. Despite the potential benefits, social media are also considered responsible for a number of issues, such as: fostering the spreading of online misinformation [6], the emergence of echo chambers [7], and increasing intolerance expressed in online debates [8]–[10]. These debates are characterized by several antisocial behaviours like cyberbullying [11], sexual harassment [12], trolling [13], and hate speech [14] which, in turn, contribute to the rise of individual and societal problems [15]–[17]. Therefore, to pursue the development of safer digital environments, it is crucial to identify early warnings of emergent toxicity and adequately moderate them. Many scholars have already faced this challenge with a mixture of approaches that ranged from the analysis of conversation cascades [9], [10], [18], [19] to Machine Learning (ML) [20]–[28]. A potential driver for the emergence of toxic speech may be the topic of discussion; another one is the platform on which the topic is debated. Yet, another one may be how the conversation evolved from the point of view of both structure and tone.

To explore these mechanisms, we provide a multi-platform comparison regarding the rise and the prevalence of toxic speech that results in developing a machine learning model able to predict the emergence of toxicity in a conversation. As a case study, we consider online debates on Twitter and YouTube around two topics of national interest: the Italian Political Elections held in 2022 and the 2022 Italian Football League.

In more detail, we first compare how toxicity evolved on the two topics and platforms, understanding which factors may have contributed to its prevalence. Then, we exploit conversation trees made up of comments to understand their structural properties using a set of cascade metrics. Finally, we provide a ML approach that, by leveraging the previous features, predicts the appearance of the following toxic comment in different stages of conversations.

From a conversational perspective, our findings suggest that a divisive topic, namely politics, tends to exhibit higher toxicity levels than a topic close to popular culture, such as football, producing more extended conversations that attract a broader audience. Lastly, the classifiers resulting from the stage-based approach achieve state-of-the-art performances despite a restricted set of features. Furthermore, our cross-topic comparison shows that models trained on a more toxic topic, namely political elections, can be generalized to other discussion arguments without causing a degradation of their performance. Our results suggest that both the cultural context and the conversation stage should be considered when developing tailored automatic moderation tools[1].

II. Related Work

A. Defining toxicity on social media

The definition of online toxicity and toxic behaviors has evolved over the years due to its interdisciplinary nature and the role of context. Prior work on this topic coined the term *hateful speech*, referring to any speech expressing

The work is supported by IRIS Infodemic Coalition (UK government, grant no. SCH-00001-3391), SERICS (PE00000014) under the NRRP MUR program funded by the European Union - NextGenerationEU, project CRESP from the Italian Ministry of Health under the program CCM 2022, PON project "Ricerca e Innovazione" 2014-2020 and project SEED n. SP122184858BEDB3.

Corresponding author: W. Quattrociocchi (email: quattrociocchi@di.uniroma1.it).

[1]Additional information for this paper can be found at https://osf.io/qdr7f

hatred by the author against a person or people based on their identity [29]. Similar definitions from the juridical literature defined hateful speech as any form of expression that can increase harassment towards individuals or groups due to some characteristics they share or affiliation [30]. A further advance in the definition of toxicity was made in recent years by the United Nations, which formalized the concept of hate speech as "*any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language regarding a person or a group based on who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factors*" [31]. More recently, researchers from Google Jigsaw, contextually with the introduction of their Perspective Application Programming Interface (API) [32], defined toxic content as any content characterized by "*rude, disrespectful, or unreasonable language that is likely to make someone leave a discussion*." [33].

B. Conversation cascades and toxicity dynamics

Conversation cascades are an instance of the so-called *information cascades* whose properties and insights have been observed for years [19]. Despite the prior knowledge, the problem of curating online conversations has attracted increasing interest due to the societal implications it has [23], [34], [35]. Prior research efforts on this topic investigated the topological structures of conversations [36], [37] and proposed new generative models [29], [38] for their reconstruction. From a social media perspective, scholars made an extensive effort to analyze conversations and their role in anti-social behaviors like harassment [12], [39], spreading of misinformation [9] and trolling [8], [13]. Moreover, it was found that users tend to concentrate their anti-social efforts on a small number of threads [40], providing no evidence for the presence of "pure haters" [35]. From a dynamic perspective, it was observed how discussions on YouTube tend to degenerate towards increasingly toxic exchanges of views [35]. Such exchanges, however, have been demonstrated not to nourish misinformation spreading on social media [9]. Finally, a stream of work investigated the predictive power of structural content and user features to identify toxic comments and anti-social behaviors [10], [29], [41], [42], achieving important results in the selection of features to employ in the automatic identification of toxic elements.

C. ML for toxicity identification

From a ML perspective, the non-trivial task of identifying the presence of toxicity in online conversations has collected an increasing interest due to its implications for society and the technical challenge it poses. Researchers achieved promising results by applying architectures that ranged from traditional classifiers [20]–[24] to deep learning approaches, including Recurrent Neural Networks (RNN) [25] and Natural Language Processing (NLP) [24], [26]–[28]. Along this path, in 2017, Google Jigsaw introduced Perspective API [32], [43], a ML system that detects toxicity of online comments [44]. Despite its initial criticism [45], [46], the API was employed by multiple research works [9], [10], [47]–[49], being recognized as a state-of-the-art tool in the context of online toxicity quantifying. Despite the extensive use of this model in research [9], [10], [47]–[49], its usage has been associated with various criticism from researchers. Indeed, it was shown how adversarial approaches can effectively reduce the toxicity associated with toxic content so that the system assigns significantly lower scores [45], [46]. Moreover, recent benchmarking tools [50] revealed pitfalls in toxicity recognizing Perspective API's capabilities on many categories, demonstrating instead how GPT-3 approaches may perform better. The reasons can be found in toxicity detection models' limited ability to contextualize conversations [51], [52]. Indeed, these models often struggle to incorporate factors other than text, such as the participant's personal characteristics, relationships and the overall tone of the conversation [51]. Consequently, what is considered toxic content can vary significantly among different groups, such as ethnicities or age groups [53], leading to potential biases. These biases may stem from the annotators' backgrounds and the datasets used for training, which might not adequately represent cultural heterogeneity. Additionally, subtle forms of toxic content, like sarcasm or memes that target specific groups, can be particularly challenging to detect. Therefore, recent advances in applying transformer-based models to identify toxicity show how specific feature combination strategies [54] and ensemble models [55] achieve promising results. Finally, researchers evaluated the ability of Generative Pre-trained Transformers (GPTs) to create synthetic datasets which can serve as input for deep learning architectures [56].

III. Preliminaries and definitions

A. Data Collection

We collect social media data concerning the 2022 Italian Political Elections and Football League. The first topic, Italian Elections, is known for being a polarising topic, especially in the case of the 2022 Italian Elections, where a strongly conservative party participated and won the elections, nourishing phenomena like echo chambers and polarization [57] and, eventually, offline disorders. Instead, the motivation for choosing the Italian Football League as a proxy for Italian popular culture is twofold. From a relevance perspective, football in Italy has the highest number of teams, thus a large geographical and media coverage, and it receives the highest number of public investments among all Italian sports [58]. From a toxicity perspective, we chose football due to its ability to spark anti-social behaviours, including tumults and brutal acts of violence [59], [60], which have the potential to be correlated with division and anti-social behaviors online.
The collection of posts and comments was performed on Twitter and YouTube to compare two regulated environments that rely on different media types, namely the text messages for Twitter and the videos for YouTube. The analysis includes all posts published from $25/08/2022$ to $25/12/2022$ with the corresponding comments. This period was suitable to to capture both the social media debate around the Italian electoral campaign and the Italian Football League. Indeed, the chosen time window captured the election day that was

on September 25, 2022, the following debate between the political parties involved once the winners were announced and the first stages of the Football League that started on the $21/08/2022$.

For the Football topic, we look for all posts containing at least one hashtag that refers to the Italian *Serie A* League team names and their slogans. Then, for each obtained post, we collect all the corresponding comments. The same approach was applied to the Elections topic, with the difference in the search hashtags that refer to political parties, exponents and general terms used by newspapers.

On Twitter, the data collection was performed by using the Twitter API for Academic Research [61], producing a total of 3.6M posts for both topics, published by 300K users, and 8.2M Italian comments, identified by using Google's Compact Language Detector 3 (CLD3), from 550K users (see table I for further details). On YouTube, instead, posts with their comments were collected using the YouTube Data API [62], resulting in a dataset of 87K posts for both topics published by 10K channels, which produced 2.6M Italian comments, again identified with CLD3, from 381K users commenting (see table I for further details).

B. Toxicity Labelling

In the current paper, we refer to toxic content using the definition provided by Google Jigsaw, which identifies as toxic any content that is "*rude, disrespectful, or unreasonable language likely to make someone leave a discussion*" [33]. Consistently with the authors of this definition, the toxicity content classification is based on Google Jigsaw Perspective API [32]. Such API uses a ML model [44] to provide a score ranging from 0 to 1, indicating the probability that a reader would perceive the comment as toxic [63]. To define an appropriate threshold, we draw from the existing literature [9], [10], [63], indicating that any content with a toxicity score ≥ 0.6 is considered toxic. To assess the validity of this threshold, we also performed content classification with a threshold of 0.5 and 0.7. Among all topics and platforms, the 0.6 threshold provided the best tradeoff between the percentage of classified elements and the size of the resulting dataset to employ for the training of toxicity classifiers.

By applying Perspective API, we quantify the toxicity of the 98.6% of the total number of posts and comments in the dataset (see table I for further details). The remaining 1.4% comprises all those contents for which the model failed to produce a toxicity score. This scenario may happen with texts containing only emojis, special characters or lexical elements for which the API did not quantify their toxicity [44].

C. Conversation Cascade Reconstruction

We model a conversation cascade as a directed tree graph $T = (V, E)$, where $V = \{1, \ldots, n\}$ represents the set of nodes and $E = \{1, \ldots, m\}$ the set of links. Each node $v \in V$ can be either an original post that started the conversation, representing the tree's root, or a comment. On both platforms, the tree's root is characterized by an identifier (ID) that uniquely defines the conversation, shared by other nodes through the *conversation_id* attribute on Twitter and by the *video_id* on YouTube. The edges $e \in E$ instead represent the act of replying that links a node v_j to a node v_i, with $j > i$. For instance, the edge $e_1 = (v_1, v_2)$ means that the comment made by node v_2 replied to the node v_1, which can be another comment or the root.

We implement the following procedure to reconstruct the conversation trees on each social media platform. On Twitter, we start from the root node and iterate on its children whose parent, represented by the *in_reply_to_id* attribute, corresponds to the root ID. For each identified node, we recursively look at their children with the same rationale until we reach all the tree leaves. The same procedure is applied on YouTube. However, in case of sub-conversations starting from a comment node v_i, YouTube will always indicate as v_i the parent of these nodes, despite the fact they may have replied to a child node v_j. Such limitations may prevent the algorithm from reconstructing the actual cascade structure. To overcome this problem, we apply a heuristic to reconstruct the tree by looking at the latest comment posted by the user mentioned in a message (referring to its username indicated by *@Username*). If no username is found in the text, we indicate as the parent of the comment the root of the tree, i.e., the original post. Otherwise, we assign as the parent of the comment the ID of the most recent comment node posted by the user identified by its username in the sub-conversation. Finally, we label the nodes on both platforms based on the toxicity score of the element, as described in section III-B. The resulting structure from this process is represented in fig. 1.

D. Cascade metrics

To provide a comparison between cascades, we define two categories of metrics. The first one called *structural metrics*, refers to those features that only depend on the number of nodes and links in a graph and their toxicity score. The second, named *conversational metrics*, refers to additional information that is not strictly related to the topology of each conversation tree.

1) Structural Metrics:

a) Tree Size: We define as Tree Size the number of nodes in a tree graph, denoted as $n = |V|$, where $|\cdot|$ is the cardinality of the set V. Conversation trees with high tree size values indicate a participated discussion. We assume a user can post multiple replies and interact with different users within the conversation.

b) Depth: The Depth $D(T)$ is the distance d of the deepest node in the conversation, which also coincides with the tree's diameter, i.e., the longest shortest path between the root node and any other node in the graph. The depth can be expressed as

$$D(T) = max(d_{rj}) \forall j, j \neq r, \tag{1}$$

where r is the root node. The deeper a conversation tree is, the more direct exchanges happen in the discussion.

Social	Topic	Posts	Users Posting	Comments	Users Commenting	Percentage of labelled elements	Percentage of toxic elements
YouTube	Football	52 023	5 431	1 296 837	193 907	99.8	2.4
YouTube	Elections	35 479	5 087	1 393 369	187 791	99.6	5.2
Twitter	Football	1 404 010	120 407	1 780 583	235 385	98.6	2
Twitter	Elections	2 258 988	183 252	6 426 742	331 310	99.6	3.4
Total		3 750 500	314 177	10 897 531	948 393	99.4	3.2

TABLE I: Data breakdown Twitter and YouTube data about Italian Football League and Elections.

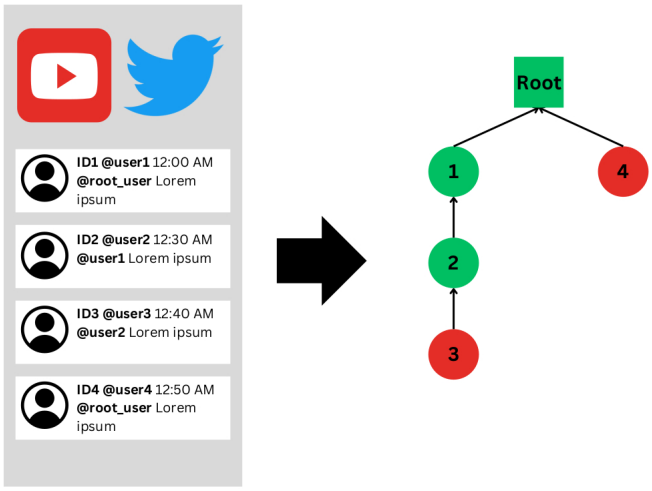


Fig. 1: Graphical representation of a conversation tree on YouTube and Twitter. The root node representing the post is a square, while the children nodes (comments) are represented as circles, with the number representing the comment ID. The nodes' colours represent the toxicity category assigned from their text. A node in green represents content whose text was identified by Perspective API with a toxicity score < 0.6, whilst a red node identifies an element with a toxicity score ≥ 0.6. Finally, grey nodes represent all those contents for which the API could not quantify their toxicity.

c) Wiener Index: The Wiener Index $W(T)$ measures the structural complexity of the conversation tree T and its potential virality [64]. It is the average shortest path between each pair of nodes i, j. In the case of a directed tree, the Wiener Index can be defined as

$$W(T) = \frac{2}{n(n-1)} \sum_i \sum_{j>i} d_{ij}, \tag{2}$$

where $\frac{2}{n(n-1)}$ is a normalization factor to account for all paths among couples of nodes. The Wiener index ranges between $[1, \infty)$ and, in general, it is minimized for broadcast structures and maximized for low branching structures [64]. In a conversation tree, a lower Wiener Index indicates that comments are more reachable from each other compared to comments in a tree with a higher Wiener Index.

d) Toxicity Ratio: The Toxicity Ratio $TR(T)$ is the average number of toxic comments in the conversation tree T, considering the number of toxic replies out of the total number in the conversation. The toxicity ratio can be defined as

$$TR(T) = \frac{|\tau|}{|V|}, \tag{3}$$

where $\tau = \{v \in V \mid s(v) \geq 0.6\} \subseteq V$, $s(v)$ is the toxicity score of the comment $v \in V$ and 0.6 is the toxicity threshold value. The higher the ratio, the more toxic the discussion is. The rationale behind this measure is to quantify the toxicity of the conversation up to the moment a comment node takes place. Therefore, it does not represent the toxicity of the parent comment but, instead, it describes a conversational state up to comment $c \in T$.

e) Average Toxicity Distance: The Average Toxicity Distance $TD(T)$ is the average normalized distance of toxic comments from the root $r \in V$, defined as

$$TD(T) = \frac{1}{|\tau|} \sum_{j \in \tau} \frac{d_{rj}}{D(T)}. \tag{4}$$

$TD(T)$ is bounded in $(0, 1]$, and low values of this quantity imply that toxic comments are, on average, located close to the root.

f) Assortativity: The assortativity coefficient r measures the extent to which similar nodes tend to be connected with each other [65]. It is defined in the $[-1, 1]$ range: values close to -1 indicate disassortativity (i.e., nodes with different features tend to be interconnected less than expected at random), whilst values close to 1 indicate assortativity (i.e., nodes with similar features tend to be interconnected more than expected at random). A value close to 0 means the distribution of node features is close to random. We consider as node feature their toxicity label, and to compute the assortativity coefficient, we ignore the direction of the links, obtaining the following equation:

$$r(T) = \frac{\sum_{ij}(a_{ij} - \frac{k_i k_j}{2m}) x_i x_j}{\sum_{ij}(a_{ij} x_i^2 - \frac{k_i k_j}{2m} x_i x_j)}, \tag{5}$$

where a_{ij} are elements of the adjacency matrix $A = (a_{ij})_{i,j \in V}$ in which $a_{ij} = 1$ ($a_{ij} = 0$) indicates the presence (absence) of a link between nodes i and j; $k_i = \sum_{j=1}^{n} a_{ij}$ is the node degree, and x_i is the feature assigned to node i.

2) Conversational Metrics:

a) Average Comment Intertime: To quantify the average time, in seconds, lasting from the appearance of a comment and its successor in a conversation, we introduce a measure called Avg. Comment Intertime $CI(T)$. Given tree graph T, it is defined as

$$CI(T) = \frac{1}{n-1} \sum_{e \in E} \Delta t(e), \tag{6}$$

where $\Delta t(e) = t(w) - t(v)$ represents the difference between the timestamps associated to the nodes w and v, with $e = (w, v) \in E$. The rationale behind this measure is to assess whether heated discussions tend to have shorter waiting times between responses rather than discussions that do not present toxicity traits.

b) Number of Unique Users: The number of unique users $U(T)$ is the number of distinct users appearing in a post by posting or commenting, which is lower or equal to the Tree Size $TS(T)$, then $U(T) \leq n$.

c) Root Toxicity: To account for the influence that the text of the initial post can have on the conversation, we assign a toxicity label to the root of each tree T, as described in section III-B.

E. Permutation Test

To assess differences in the distribution of cascade metrics between different topics, we perform permutation tests whose algorithm is described in algorithm 1. For each metric, we consider the two distributions X_{ele} and Y_{foot} relative to the *Elections* and *Football* topics, keeping track of which population an observation is taken from. We begin by computing the test statistic m, defined as the absolute difference value between the mean of X_{ele} and Y_{foot}. Then, we unify the cascade distributions of two topics into a new one, called Z, and we shuffle the labels of the measures, obtaining Z^*, a set containing the same observations but (possibly) with different labels. Such operation allows us to perform the permutation tests by extracting the two shuffled distributions, i.e., X^*_{ele} and Y^*_{foot} based on their labels in Z^* and performing the absolute difference for their mean m^*. We repeat the procedure 1000 times and, as a result, we compute the probability that the test statistics m^*, observed in our null model, is higher (in absolute value) than m. We decide to use the permutation test since it can reduce the effects of imbalances in the sample sizes that may interfere with other tests, such as the Kolmogorov-Smirnov (KS) test.

Algorithm 1 Permutation test algorithm to assess statistical differences in the cascade metrics of two topics.

Input: Two topic metric distributions X_{ele} and Y_{foot}, where each measure posses a label identifying its provenience
Parameter: N, number of permutations
Output: p, the p-value resulting from the permutation test

$c = 0$
$N = 1000$
Calculate the test statistic $m = |\overline{X_{ele}} - \overline{Y_{foot}}|$
$Z = X_{ele} \cup Y_{foot}$ (maintaining the label of each observation)
let $i = 1$
while $i \leq N$ **do**
    Z^* = shuffle the labels of observation in Z
    Extract ${X_{ele}}^*$ and ${Y_{foot}}^*$ from Z^* according to their label in Z^*
    $m^* = |\overline{{X_{ele}}^*} - \overline{{Y_{foot}}^*}|$
    if $m^* \geq m$ **then**
        $c = c + 1$
    end if
    $i = i + 1$
end while
$p = \frac{c}{N}$
return p

F. Toxicity comment prediction in a conversation

Content moderation algorithms play a crucial role in the maintenance of online ecosystems. On the one hand, they must promptly limit the diffusion of harmful content. At the same time, too much limitation can prevent the emergence of vibrant discussions, impacting freedom of speech. Recent approaches to designing effective moderation [10], [29], [41], [42] tools focused on structural aspects of the conversations without effectively considering the relationship between the topic discussed and the community involved. To address this gap, we propose a ML approach that differs from the current literature for two main reasons. First, we aim to provide a minimal yet effective feature set based on previously computed cascade metrics. Second, since it is known that structural feature importance is subjected to decaying as the tree size grows [19], we implement 4 different classifiers, each trained with comments belonging to specifc stages of a conversation. In terms of toxicity, we hypothesise such a solution will capture its evolution in the different stages of a conversation.

1) Dataset creation:

a) Computing cascade metrics at comment-level: We begin the dataset creation procedure by reconstructing, for each topic and platform, the conversation cascades as described in section III-C. During the reconstruction, we filter out all those conversations with less than one comment to ensure the existence of at least a pair of toxic/non-toxic comments. Next, we compute the evolution of the features described in section III-D at the insertion time of each comment.

b) Creating a dataset for the toxicity prediction task: In ML tasks involving cascades, it is mandatory to account for the decaying importance of their features as the size grows [18], [19], [66]. If not, the predictions produced by models trained on these data may be biased from the tree's current state. From a structural perspective, previous results [67] showed how logarithmic binning enhances differences in the evolution of structural measures concerning the cascade size. Given the following motivations, we apply a dataset creation strategy that performs a logarithmic binning on the cascade size. Indeed, each unfolded conversation is split into four intervals, i.e., $(1, 10)$, $(10, 100]$, $(100, 1000]$, $(1000, 10000]$, according to the position assigned to a comment by entering in the conversation (comment index). This approach allows the creation of subsets that describe the different stages at which a conversation evolves, potentially helping the emergence of topological or conversational dynamics.

To optimize the separation between toxic and non-toxic elements, on each subset, we retain only those comments with a toxicity score provided by Perspective API less than 0.2, representing elements with a low presence of toxic language and greater or equal to 0.6, representing the toxic elements.

For each conversation in a subset, we create a pair of comments that include a toxic/non-toxic element until all toxic comments have a unique counterpart. However, to account for all those toxic comments without a counterpart, we randomly assign them a non-toxic element chosen from the subset in the exam. Then, we extract the features of both comments from all pairs, obtaining a cascade snapshot from a structural and conversational perspective when a toxic and non-toxic comment in the different conversations is posted. Finally, we end the dataset creation by performing an 80/20 split to obtain the train and test sets for the model training and testing phase.

2) Model training: To predict the occurrence of a toxic comment in a conversation, we implement an ensemble approach that consists of four ML sub-models, each specialized for a specific conversation stage as described in section III-F1. We train these models on a set of structural and conversational features, defined in section III-D, to capture the different aspects that can bring to the production of toxic content in a conversation. We implement several ML-supervised models to identify the consistency of results and the most suitable model for this task, namely Logistic Regression (LR) models, Random Forests (RF), Decision Trees (DT), AdaBoost (AB), Support Vector Machines (SVM) and Gradient Boosted Regression Trees (GBRT). For each model, we tune its hyperparameters through a 10-fold CV. The best model is refitted on the entire training set based on its accuracy score. For each dataset interval, we choose the best model with the highest F1 score, considering the Accuracy score in the case of a draw.

To estimate the predictive power of singular features, we proceed as follows. We first compute the F1 score s obtained by fitting the model m on the original dataset X. Next, we randomly shuffle its values for each feature $j \in [1, P]$ of the dataset, where P is the total number of features. For every shuffle $k \in [1, 10]$, we fit the model m on the dataset $\tilde{X}_{j,k}$ with the j-th column shuffled, obtaining a new score $s_{k,j}$. The importance of the feature i_j is defined as

$$i_j = s - \frac{1}{10}\sum_{k}^{10} s_{k,j}. \tag{7}$$

IV. Results

A. Toxicity Evolution

We begin the analysis by comparing the toxicity evolution for the Italian Football League, representing a topic close to the Italian popular culture, and the 2022 Italian Political Elections, representing a potentially polarising topic. Fig.2a represents the average toxicity scores observed for each topic and social media platform during the analysis period. We observe that conversations about Italian Elections display higher toxicity levels than those about Italian Football. Indeed, on Twitter, Elections conversations produce an average daily toxicity score of 0.18 compared to the 0.09 for Football. The same behavior is found on YouTube, where the Elections topic attracts more toxicity than Football, with an average score 0.22 against the 0.13 of its counterpart. This result complies with the toxicity labelling results described in table I in which, on both social media, Elections contents have the highest percentage of toxic elements and is in line with previous studies [10], [68] reporting a low, but still problematic, prevalence of toxic speech in online social media. We statistically assess this result by applying the KS test on both topic distributions for each social, obtaining a p-value $p < 0.05$ for both cases. Ultimately, we provide the first evidence of how the topic of Football produces conversations characterized by a lower presence of toxic language compared to political Elections. The lower prevalence of toxic speech in the football debate, which is instead usually associated to events of hate and violence both offline and online, represents a counterintuitive aspect emerging from the data.

Next, we quantify the rate at which toxicity evolved during the analysis period, assessing whether events, such as the Italian Elections voting day on September 25th, 2022, may produce an effect on the toxicity of the corresponding debate. To achieve this goal, we estimate the evolution of toxicity on each topic through Ordinary Least Squares (OLS) regression models, defined as $\text{Toxicity}_t = \beta_0 + \beta_1 \text{Date}_t$.

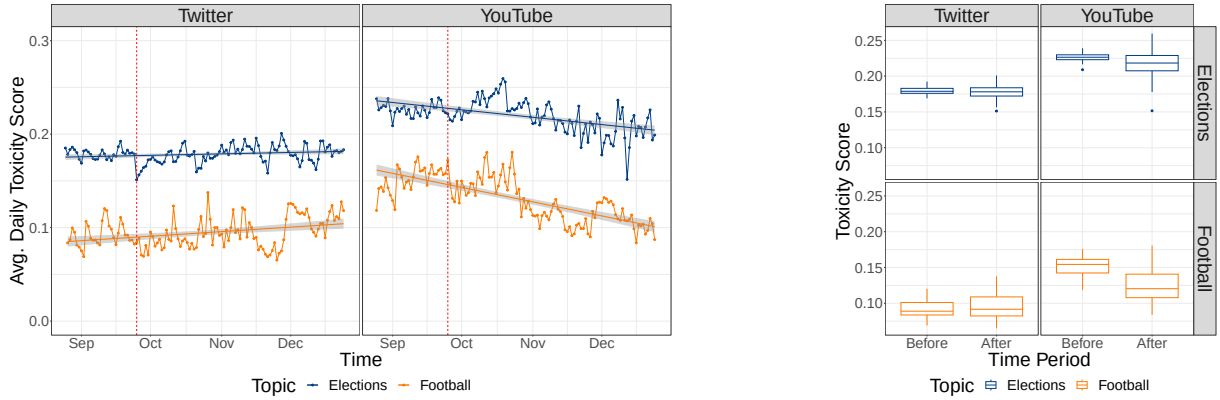
Results from the fitting procedure show that YouTube is characterized by a decreasing trend of the toxicity scores for both topics ($\beta_1 = -2.59\times10^{-4}$ Elections and $\beta_1 = -5\times10^{-4}$ for Football), whilst Twitter presents a stationary trend for the Elections topic ($\beta_1 = 5.06 \times 10^{-5}$) and an increasing one for Football ($\beta_1 = 1.59 \times 10^{-4}$).

In terms of differences found in coincidence of the voting day, fig. 2b reports a toxicity decrease of -21.98% on Football and -3.57% on YouTube, whilst on Twitter we note an increase of 3.03% for Football and a decrease of -0.42% for Elections. However, by conducting a KS test on the sample concerning the pre and post-event periods, we observed that the only significant change in toxicity happened on YouTube with p-value < 0.05 for both topics against the 0.22 and 0.35 in the case of Twitter.

To conclude our analysis, we look at the possible factors explaining the observed toxicity trends. To do so, we compute the Pearson correlation coefficient between the toxicity score and a set of measures related to the volume of content and the user's behaviour, namely the number of posts (*Posts*), comments (*Comments*), the number of users commenting (*Users Commenting*) and the comment they produce (*User Comments*). From the results reported in table II, we observe that, on YouTube, the evolution of toxicity is positively linked with all the measures taken into account, identifying the role of the content volume in the production of online hate for both topics. More specifically, the positive correlation between the number of comments and users involved provides evidence of how online toxicity is closely associated with the length of discussions - represented by the number of comments - and with the commenting activity of users - represented by the number of user comments. On Twitter, toxicity in Football conversations appears to be linked to the number of posts generated about the topic, without being influenced by the commenting perspective. For the Elections topic instead, results confirm what was observed on YouTube, i.e., toxicity has a strict direct relationship with the commenting activity.

Ultimately, we provide evidence of how a popular culture topic

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2024.3398219



(a) (b)

Fig. 2: Left panel:Average daily toxicity score reported on Twitter (left) and YouTube (right). The straight horizontal lines represent the linear fit performed on each trend. The red vertical line represents the date of the voting day for the Italian Elections (September 25, 2022). Right panel: toxicity score distributions for each social media and topic before and after the date concerning the Italian Elections voting.

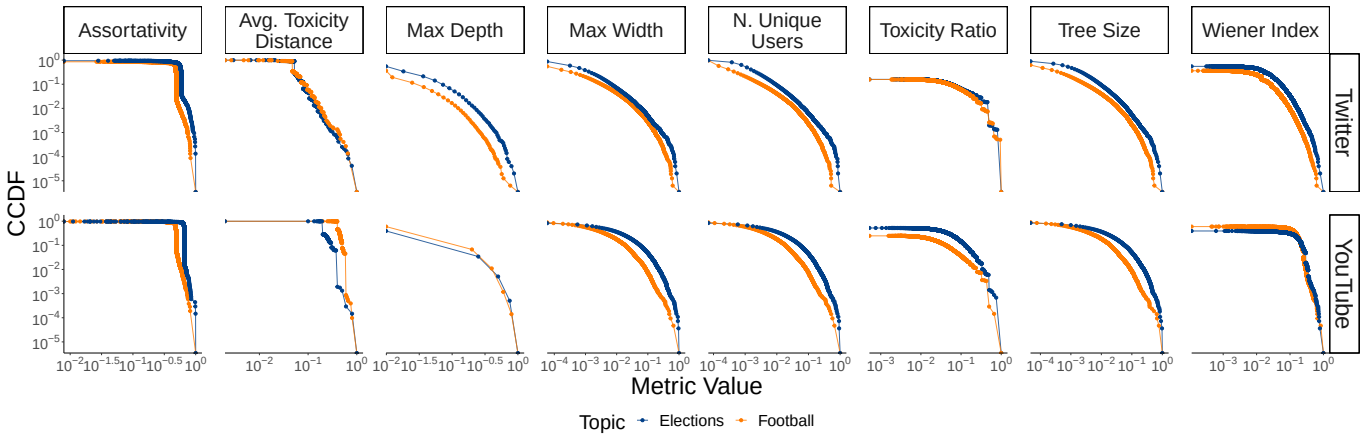


Fig. 3: CCDFs of the standardized cascade metrics for Twitter (top) and YouTube (bottom).

like football tends to attract less hate than those inherently divisive, such as political elections. From a social media perspective, we observe how the topic may be a discriminant in the evolution of toxicity,even in unexpected cases such as football.

Twitter				
Topic	**Posts**	**Comments**	**Users Commenting**	**User Comments**
Football	0.53	0.01	-0.03	-0.04
Elections	0.04	0.38	0.39	0.28
YouTube				
Topic	**Posts**	**Comments**	**Users Commenting**	**User Comments**
Football	0.40	0.61	0.56	0.77
Elections	0.34	0.40	0.50	0.56

TABLE II: Pearson correlation scores between average daily metrics concerning the toxicity scores, the number of posts and comments as well as the number of users commenting and how many times they commented daily.

B. *Structural analysis*

We continue our comparison by investigating how the structure of conversations diverges according to their topic and platform. We first compute a set of structural metrics, described in section III-D; then, we assess the statistical validity of the obtained distributions using a permutation test, described in section III-E, with a Bonferroni correction to account for multiple comparisons, considering p-values less than 0.00625 $(0.05/8)$ as significant. Fig.3 reports the Complementary Cumulative Distribution Functions (CCDFs) computed on the previous cascade metrics for both topics and social media. We observe how, on Twitter, the Elections topic tends to attract bigger (*Tree Size*), wider (*Max Width*) and deeper (*Max Depth*) conversations. From a content perspective instead, Elections conversations are more likely to carry more toxic tweets (*Toxicity Ratio*) than those from Football. Conversely, users participanting in Football conversations have a higher chance to find a toxic comment earlier than in the Elections ones (*Avg. Toxicity Distance*). The p-value of the statistical tests evidences how *Max Width* and *Number*

of unique users are the only metrics on Twitter having no differences regardless of the topic.

C. Predicting the following toxic comment in a conversation

As a final step, we predict the probability that the following comment in a conversation is toxic. Our results show that GBRT models achieve the highest performance on most configurations, whose results are reported in fig. 4a. We report results containing the $(1, 10]$ interval for the sake of completeness, but we do not include them in the discussion of the results. The reason is that newborn conversations with few comments may not have established proper conversational dynamics yet, therefore not representing an adequate asset for toxicity predictors. The F1 scores reported for the Elections topic range between $[0.72, 0.78]$ on Twitter and $[0.70, 0.76]$ on YouTube. For Football instead, F1 scores range between $[0.79, 0.84]$ on Twitter and $[0.77, 0.84]$ on YouTube. Next, we create a baseline by training each model on datasets obtained by unifying all intervals for each topic-platform combination. The resulting metrics unveil how, in all configurations, the $(10, 100]$ interval produces greater or equal F1 scores than the baseline, providing evidence of how accounting for the different stages of a conversation may produce models with better performance and, therefore, with the ability to keep digital ecosystems safer.

Next, we investigate the generalizing power of models concerning the topics they were trained from. To do so, we perform a cross-topic evaluation for each social media: each stage model is trained on one topic and tested on its counterpart. fig. 4b displays the result of this comparison, where we observe a twofold scenario. On YouTube, training on Football data and testing against the Elections test set decreased F1 score by an average of 7%. The same result is observed even by training on Elections data and testing against the Football test set, with an average decrease of F1 score equal to 9%. Conversely, on Twitter, we observe a twofold effect. Whilst training on the Football comments and testing on Elections produced an average reduction of the F1 score equal to 8%, the opposite scenario produced an average increase of the metric equal to 8%. Such a result indicates that Football, whose conversations are less toxic and participated, cannot generalize toxicity dynamics occurring in more toxic topic like Elections, resulting in a drop in performance. Instead, the models trained on cascades with a more articulated structure, like the Elections ones, tend to better generalize unknown observations in their feature space, achieving higher performance on a cross-topic benchmark. Finally, we assess the importance of each employed metric in this prediction task by measuring how the $F1$ would be impacted if a feature is removed. Results displayed in fig. 5 show, as expected, that the toxicity ratio (*Toxicity Ratio*) is the most significant feature for predicting the toxicity of a comment, leading to an average reduction of 22% in the F1 score on both platforms, followed by the average toxicity distance (*Avg. Toxicity Distance*) (2%) and the assortativity (*Assortativity*) (1%). This result describes how combining cascade features with domain-specific information can be relevant in predicting harmful content.

V. Conclusion

In this paper, we proposed a Twitter and YouTube comparison between the Italian football championship and the 2022 Italian general elections. We first assessed their differences in toxicity evolution, understanding which factors induce changes in the prevalence of toxic speech. Then, we compared conversations from a topological perspective by employing a set of structural metrics typical of cascades. Finally, we employed a ML approach, which, by creating four sub-models accounting for the different stages of a conversation, predicted the presence of the following toxic comment in a conversation. Our findings provide a counterintuitive example of how football, a topic close to popular culture that is usually associated with episodes of extreme hate and violence, tends to exhibit lower toxicity levels than politics, a potentially divisive topic. This comparison also sheds light on a trend towards affective polarisation, which implies increased negativity towards the members of the opposing political parties [69], [70] at a national level. From a structural perspective, conversations from the Elections are broader, more toxic and involve more users. Moreover, the classifiers resulting from the stage-based approach achieved state-of-the-art results despite a minimal set of features, with models from early stages of conversations performing as well as those trained on the entire datasets. Our findings could be employed to support human moderators by providing a warning signal related to conversations that display a higher likelihood of generating toxic exchanges.

Despite positive aspects such as the multi-platform and multi-topic nature of our study, it presents some limitations. The first limitation relates to only one language - Italian - in the conversations. Our results may also suffer from the lack of deleted content despite our data collection being performed with a short delay (a few days at most) concerning the posting time.

In future works, we aim to generalize this approach by extending the number of topics chosen and the list of platforms, including unmoderated social media platforms. Finally, to advance the quality of predictions, we also aim to define newer structural and conversational metrics to include in our models.

References

[1] J. Brown, A. J. Broderick, and N. Lee, "Word of mouth communication within online communities: Conceptualizing the online social network," *Journal of interactive marketing*, vol. 21, no. 3, pp. 2–20, 2007.

[2] R. Kahn and D. Kellner, "New media and internet activism: from the 'battle of seattle'to blogging," *New media & society*, vol. 6, no. 1, pp. 87–95, 2004.

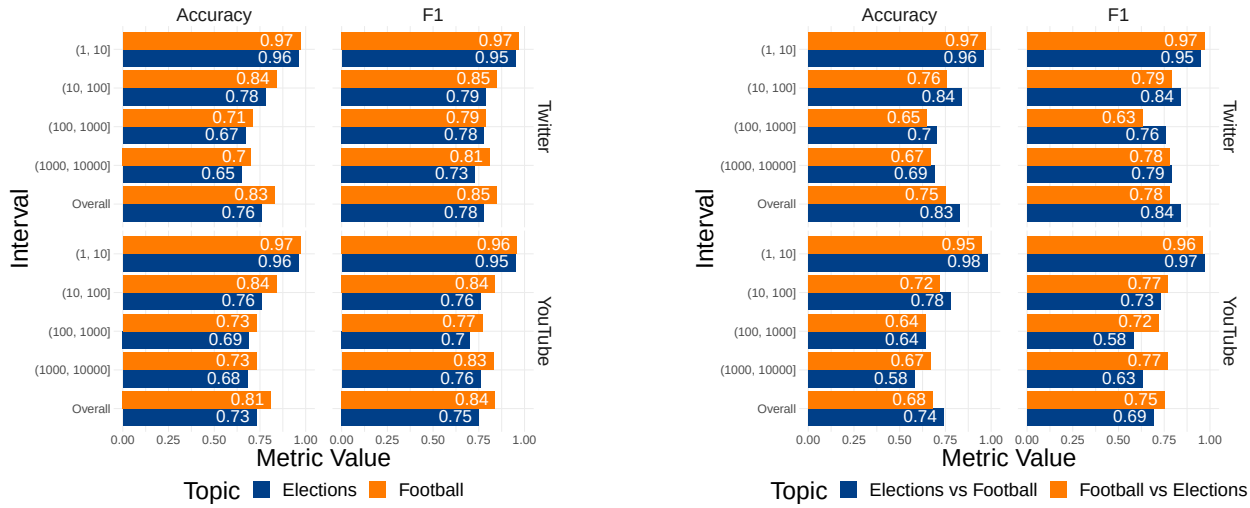
[3] W. Quattrociocchi, G. Caldarelli, and A. Scala, "Opinion dynamics on interacting networks: media competition and social influence," *Scientific reports*, vol. 4, no. 1, p. 4938, 2014.

[4] J. I. Criado, R. Sandoval-Almazan, and J. R. Gil-Garcia, "Government innovation through social media," pp. 319–326, 2013.

[5] G. Etta, E. Sangiorgio, N. Di Marco, M. Avalle, A. Scala, M. Cinelli, and W. Quattrociocchi, "Characterizing engagement dynamics across topics on facebook," *Plos one*, vol. 18, no. 6, p. e0286150, 2023.

[6] M. Del Vicario, G. Vivaldo, A. Bessi, F. Zollo, A. Scala, G. Caldarelli, and W. Quattrociocchi, "Echo chambers: Emotional contagion and group polarization on facebook," *Scientific reports*, vol. 6, no. 1, p. 37825, 2016.

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2024.3398219



(a) (b)

Fig. 4: Left panel: prediction results of the GBRT model trained on intervals from each social media and topic. Right: Prediction results from a cross-topic comparison on each social media. We observe how performing out-of-topic prediction reduces prediction scores.

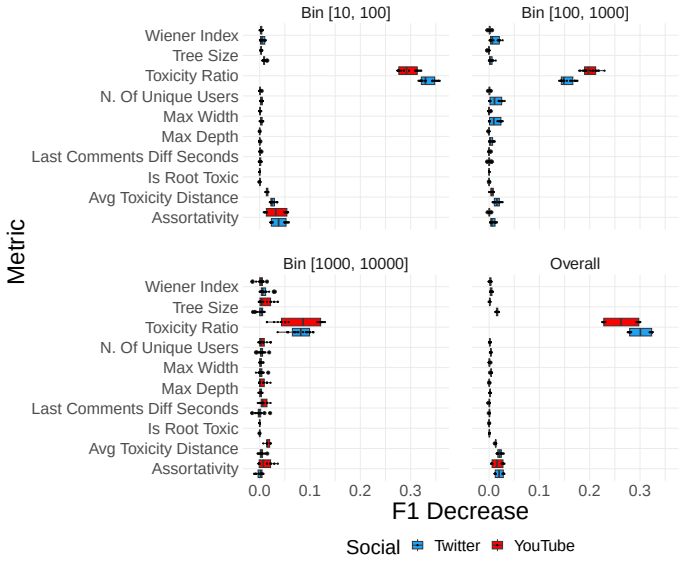


Fig. 5: Representation of the importance of the features employed in the model, quantified by the average drop in F1 score corresponding to removing a specific feature.

[7] M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, "The echo chamber effect on social media," *Proceedings of the National Academy of Sciences*, vol. 118, no. 9, p. e2023301118, 2021.

[8] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Anyone can become a troll: Causes of trolling behavior in online discussions," in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 2017, pp. 1217–1230.

[9] A. Quattrociocchi, G. Etta, M. Avalle, M. Cinelli, and W. Quattrociocchi, "Reliability of news and toxicity in twitter conversations," in *Social Informatics*, F. Hopfgartner, K. Jaidka, P. Mayr, J. Jose, and J. Breitsohl, Eds. Cham: Springer International Publishing, 2022, pp. 245–256.

[10] M. Saveski, B. Roy, and D. Roy, "The structure of toxic conversations on twitter," in *Proceedings of the Web Conference 2021*, 2021, pp. 1086–1097.

[11] C. E. Robertson, N. Pröllochs, K. Schwarzenegger, P. Pärnamets, J. J. Van Bavel, and S. Feuerriegel, "Negativity drives online news consumption," *Nature Human Behaviour*, pp. 1–11, 2023.

[12] A. G. Chowdhury, R. Sawhney, R. Shah, and D. Mahata, "# youtoo? detection of personal recollections of sexual harassment on social media," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2527–2537.

[13] J. Hannan, "Trolling ourselves to death? social media and post-truth politics," *European Journal of Communication*, vol. 33, no. 2, pp. 214–226, 2018.

[14] A. Matamoros-Fernández and J. Farkas, "Racism, hate speech, and social media: A systematic review and critique," *Television & New Media*, vol. 22, no. 2, pp. 205–224, 2021.

[15] M. C. Parent, T. D. Gobble, and A. Rochlen, "Social media behavior, toxic masculinity, and depression." *Psychology of Men & Masculinities*, vol. 20, no. 3, p. 277, 2019.

[16] K. Saha, E. Chandrasekharan, and M. De Choudhury, "Prevalence and psychological effects of hateful speech in online college communities," in *Proceedings of the 10th ACM conference on web science*, 2019, pp. 255–264.

[17] C. M. Valensise, M. Cinelli, and W. Quattrociocchi, "The dynamics of online polarization," *arXiv preprint arXiv:2205.15958*, 2022.

[18] R. Rotabi, K. Kamath, J. Kleinberg, and A. Sharma, "Cascades: A view from audience," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 587–596.

[19] F. Zhou, X. Xu, G. Trajcevski, and K. Zhang, "A survey of information cascade analysis: Models, predictions, and recent advances," *ACM Comput. Surv.*, vol. 54, no. 2, mar 2021. [Online]. Available: https://doi.org/10.1145/3433000

[20] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 27, no. 1, 2013, pp. 1621–1622.

[21] P. Burnap and M. L. Williams, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," *Policy & internet*, vol. 7, no. 2, pp. 223–242, 2015.

[22] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 29–30.

[23] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, 2017, pp. 512–515.

[24] S. Malmasi and M. Zampieri, "Challenges in discriminating profanity from hate speech," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 30, no. 2, pp. 187–202, 2018.

[25] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on Twitter," in *Proceedings of the NAACL Student Research Workshop*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 88–93. [Online]. Available: https://aclanthology.org/N16-2013

[26] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1391–1399.

[27] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th international conference on World Wide Web companion*, 2017, pp. 759–760.

[28] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)," *arXiv preprint arXiv:1903.08983*, 2019.

[29] V. Gómez, H. J. Kappen, and A. Kaltenbrunner, "Modeling the structure and evolution of discussion cascades," in *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, 2011, pp. 181–190.

[30] M. Herz and P. Molnár, *The content and context of hate speech: Rethinking regulation and responses*. Cambridge University Press, 2012.

[31] A. Guterres *et al.*, "United nations strategy and plan of action on hate speech," *Taken from: https://www. un. org/en/genocideprevention/documents/UN*, no. 20Strategy, 2019.

[32] G. Jigsaw, "Perspective api," 2022. [Online]. Available: https://perspectiveapi.com/

[33] Jigsaw, "Toxicity," *The Current*, no. 3, 2023. [Online]. Available: https://jigsaw.google.com/the-current/toxicity/

[34] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech," *Proc. ACM Hum.-Comput. Interact.*, vol. 1, no. CSCW, dec 2017. [Online]. Available: https://doi.org/10.1145/3134666

[35] M. Cinelli, A. Pelicon, I. Mozetič, W. Quattrociocchi, P. K. Novak, and F. Zollo, "Dynamics of online hate and misinformation," *Scientific reports*, vol. 11, no. 1, p. 22083, 2021.

[36] S. Gonzalez-Bailon, A. Kaltenbrunner, and R. E. Banchs, "The structure of political discussion networks: A model for the analysis of online deliberation," *Journal of Information Technology*, vol. 25, no. 2, pp. 230–243, 2010. [Online]. Available: https://doi.org/10.1057/jit.2010.2

[37] L. Backstrom, J. Kleinberg, L. Lee, and C. Danescu-Niculescu-Mizil, "Characterizing and curating conversation threads: expansion, focus, volume, re-entry," in *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013, pp. 13–22.

[38] R. Kumar, M. Mahdian, and M. McGlohon, "Dynamics of conversations," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 553–562. [Online]. Available: https://doi.org/10.1145/1835804.1835875

[39] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," *Proceedings of the Content Analysis in the WEB*, vol. 2, no. 0, pp. 1–7, 2009.

[40] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities," in *Proceedings of the international aaai conference on web and social media*, vol. 9, no. 1, 2015, pp. 61–70.

[41] S. Levy, R. E. Kraut, J. A. Yu, K. M. Altenburger, and Y.-C. Wang, "Understanding conflicts in online conversations," in *Proceedings of the ACM Web Conference 2022*, ser. WWW '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 2592–2602. [Online]. Available: https://doi.org/10.1145/3485447.3512131

[42] M. Raghavendra, K. Sharma, S. Kumar *et al.*, "Signed link representation in continuous-time dynamic signed networks," *arXiv preprint arXiv:2207.03408*, 2022.

[43] A. Lees, V. Q. Tran, Y. Tay, J. Sorensen, J. Gupta, D. Metzler, and L. Vasserman, "A new generation of perspective api: Efficient multilingual character-level transformers," *arXiv preprint arXiv:2202.11176*, 2022.

[44] G. Jigsaw, "Perspective api - attributes & languages," 2022. [Online]. Available: https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages

[45] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, "Deceiving google's perspective api built for detecting toxic comments," *arXiv preprint arXiv:1702.08138*, 2017.

[46] E. Jain, S. Brown, J. Chen, E. Neaton, M. Baidas, Z. Dong, H. Gu, and N. S. Artan, "Adversarial text generation for google's perspective api," in *2018 international conference on computational science and computational intelligence (CSCI)*. IEEE, 2018, pp. 1136–1141.

[47] G. Russo, L. Verginer, M. H. Ribeiro, and G. Casiraghi, "Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning," *arXiv preprint arXiv:2209.09803*, 2022.

[48] C. S. Czymara, S. Dochow-Sondershaus, L. G. Drouhot, M. Simsek, and C. Spörlein, "Catalyst of hate? ethnic insulting on youtube in the aftermath of terror attacks in france, germany and the united kingdom 2014–2017," *Journal of Ethnic and Migration Studies*, vol. 49, no. 2, pp. 535–553, 2023.

[49] M. Avalle, N. Di Marco, G. Etta, E. Sangiorgio, S. Alipour, A. Bonetti, L. Alvisi, A. Scala, A. Baronchelli, M. Cinelli *et al.*, "Persistent interaction patterns across social media platforms and over time," *Nature*, pp. 1–8, 2024.

[50] L. Rosenblatt, L. Piedras, and J. Wilkins, "Critical perspectives: A benchmark revealing pitfalls in perspectiveapi," in *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, 2022, pp. 15–24.

[51] A. Sheth, V. L. Shalin, and U. Kursuncu, "Defining and detecting toxicity on social media: context and knowledge are key," *Neurocomputing*, vol. 490, pp. 312–318, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231221018087

[52] W. Yin and A. Zubiaga, "Hidden behind the obvious: Misleading keywords and implicitly abusive language on social media," *Online Social Networks and Media*, vol. 30, p. 100210, 2022.

[53] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. A. Smith, "Annotators with attitudes: How annotator beliefs and identities bias toxic language detection," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 5884–5906.

[54] J. A. García-Díaz, S. M. Jiménez-Zafra, M. A. García-Cumbreras, and R. Valencia-García, "Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers," *Complex & Intelligent Systems*, vol. 9, no. 3, pp. 2893–2914, 2023.

[55] P. K. Roy, S. Bhawal, and C. N. Subalalitha, "Hate speech and offensive language detection in dravidian languages using deep ensemble framework," *Computer Speech & Language*, vol. 75, p. 101386, 2022.

[56] T. Wullach, A. Adler, and E. Minkov, "Towards hate speech detection at large via deep generative modeling," *IEEE Internet Computing*, vol. 25, no. 2, pp. 48–57, 2021.

[57] M. Scott, "Digital bridge: Italian election — midterm polarization — android fallout," *Politico*, 2022. [Online]. Available: https://www.politico.eu/newsletter/digital-bridge/italian-election-midterm-polarization-android-fallout/

[58] B. IFIS, "Osservatorio sullo sport system italiano," 2022.

[59] T. Jones, "Beyond the violence, the shocking power the ultras wield over italian football," *The Guardian*, 2018.

[60] N. Squires, "Italian riot police clash with football fans on the rampage in naples," *The Telegraph*, 2023.

[61] Twitter, "Twitter api for academic research," 2023. [Online]. Available: https://developer.twitter.com/en/products/twitter-api/academic-research

[62] YouTube, "Youtube data api," 2023. [Online]. Available: https://developers.google.com/youtube/v3

[63] Jigsaw, "About the scores," 2023. [Online]. Available: https://developers.perspectiveapi.com/s/about-the-api-score?language=en_US

[64] S. Goel, A. Anderson, J. Hofman, and D. J. Watts, "The structural virality of online diffusion," *Management Science*, vol. 62, no. 1, pp. 180–196, 2016.

[65] M. E. Newman, "Mixing patterns in networks," *Physical review E*, vol. 67, no. 2, p. 026126, 2003.

[66] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *Proceedings of the 13th international conference on World Wide Web*, 2004, pp. 491–501.

[67] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?" in *Proceedings of the 23rd international conference on World wide web*, 2014, pp. 925–936.

[68] J. W. Kim, A. Guess, B. Nyhan, and J. Reifler, "The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity," *Journal of Communication*, vol. 71, no. 6, pp. 922–946, 2021.
[69] J. C. Rogowski and J. L. Sutherland, "How ideology fuels affective polarization," *Political behavior*, vol. 38, pp. 485–508, 2016.
[70] M. Wagner, "Affective polarization in multiparty systems," *Electoral Studies*, vol. 69, p. 102199, 2021.

Gabriele Etta Gabriele Etta is a Ph.D. student from Sapienza University of Rome, Italy, working on Data Driven Modeling of Social Dynamics. He received his MSc degree in Data Science from the University of Padua, Italy, in 2020. His research interests include complex networks, information diffusion, and computational social science.

Matteo Cinelli Matteo Cinelli is a post-doc researcher at Ca' Foscari, University of Venice and associate researcher at ISC-CNR. His background is in Management Engineering and he obtained a PhD in Enterprise Engineering from the University of Rome "Tor Vergata". His research interests include network science, computational social science and big data.

Niccolò Di Marco Niccolò Di Marco is a Post-Doctoral Researcher at Sapienza University of Rome, Rome, Italy. He received the Ph.D. degree in Mathematics from the University of Florence, Italy, in 2023.

His background is in discrete mathematics and graph theory. His research interests include network science, complex systems, computational social science and graph theory.

Michele Avalle Michele Avalle received a MSc degree in Physics from Sapienza University of Rome in 2011, and a Ph.D degree in Physics in 2015 at University College London, working on quantum information and computation theory. He worked in the field of science communication for several years. He is currently post-doc researcher at Sapienza University of Rome at the Center of Data Science and Complexity for Society.

Alessandro Panconesi Alessandro Panconesi is Full Professor in the Computer Science Department at Sapienza University of Rome. He received his Ph.D. in Computer Science from Cornell University and a Laurea in Mathematics cum Laude from Sapienza University of Rome. He held research and teaching positions at esteemed institutions such as CWI Amsterdam, Freie and Humboldt Universität in Berlin, BRICS (University of Åhrus), Ca' Foscari University of Venice, and University of Bologna. He received prestigious prizes such as the Edsger W. Dijkstra Prize in 2019, Google Focused Awards in 2014 and 2018, Google Faculty Research Awards in 2010 and 2012, an IBM Faculty Award in 2009.

Walter Quattrociocchi Walter Quattrociocchi is Full Professor at Sapienza University of Rome where he leads the Center of Data Science and Complexity for Society (CDCS. His research interests include data science, network science, cognitive science, and data-driven modeling of dynamic processes in complex networks. His activity focuses on the data-driven modeling of social dynamics such as (mis)information spreading and the emergence of collective phenomena. Dr Quattrociocchi has published extensively in peer reviewed conferences and journals including PNAS. The results of his research in misinformation spreading have informed the Global Risk Report 2016 and 2017 of the World Economic Forum and have been covered extensively by international media including Scientific American, New Scientist, The Economist, The Guardian, New York Times, Washington Post, Bloomberg, Fortune, Poynter and The Atlantic). He published two books: "Misinformation. Guida alla società dell'informazione e della credulità" (Franco Angeli) and "Liberi di Crederci. Informazione, Internet e Post Verità" with Codice Edizioni for the dissemination of his results.

In 2017 Dr Quattrociocchi was the coordinator of the round table on Fake News and the role of Universities and Research to contrast fake news chaired by the President of Italy's Chamber of Deputies Mrs Laura Boldrini. Since 2018 he is Scientific Advisor of the Italian Communication Authority (AGCOM) and currently Member of the Task Force to contrast Hate Speech nomianted by the Minister of Innovation. Professor Quattrociocchi is regularly invited for keynote speeches and guest lectures at major academic and other organizations, having presented among others at CERN, European Commission, the University of Cambridge, Network Science Institute, Global Security Forum etc.