Quantifying Topological Differences in Online Conversations

social media; hate speech; online conversations; reply cascades; toxicity prediction;

Extended Abstract

The rise of social media platforms has unequivocally affected how people express themselves online. Within these platforms, users have the potential to bring discussions to unprecedented scale and efficacy, interacting with a large number of peers in a short time. Despite the benefits brought by this new way of communicating, the amplified potential exposure of users to hate speech and toxic conversations is an issue that has become crucial in designing and maintaining healthy and safe online environments. Indeed, the advances in the study of online conversations have unveiled meaningful connections between toxicity and the way discussions are structured. A recent work [1] showed that the conversation's structural characteristics could predict whether the next reply posted by a specific user will be toxic or not. In the context of online hate and misinformation, it was observed [2] that online debates tend to degenerate towards increasingly toxic exchanges of views when interventions are not applied. Despite these advances, the majority of works about hate speech and toxic conversations mainly focus on political-related topics [3, 1, 4, 5] and are frequently limited to the analysis of a single social media [3, 1, 2, 6, 5]. Therefore, it is crucial to provide a quantitative assessment of the structural dynamics revolving around online conversations by comparing multiple social media and heterogeneous topics. To address this gap, in this work we provide a comparative analysis of the online discussions regarding two different topics on two of the significant regulated social media platforms to study their differences from a topological and behavioural perspective. To achieve this goal, we collected a total of $\sim 300K$ conversations from Twitter and YouTube, including $\sim 5M$ pieces of content from $\sim 500K$ users, during a period that ranged from August to November 2022. To effectively quantify the differences in how people communicate online, we selected content from two unrelated topics - the 2022 Italian Elections and Italian Football - using the latter as a benchmark. To measure the toxicity of each comment in a conversation, we employed Google's Perspective API [7] due to its effectiveness and extensive use in recent literature [1]. The API provides a continuous score between 0 (non-toxic) and 1 (highly toxic) that quantifies the toxicity of a piece of text. We modelled the discussion originating from each post as a directed conversation tree, defined as a graph $G = (V,E)$, where $V = \{1,\ldots,n\}$ represents the set of nodes and $E = \{1,\ldots,m\}$ the set of links. From these trees, we quantified the topological aspects of online conversations by computing a set of structural metrics for each topic on both social media. We then assessed the presence of statistical differences between the two topics by performing a Kolmogorov-Smirnoff test on the previously computed metrics. Finally, we test the effectiveness of these metrics in the task of predicting whether the following reply in a tree will be toxic. To further extend the recent results from the literature [1], we propose a set of logistic regression models that perform the prediction on specific tree size intervals, namely $[1,10]$, $(10,100]$, $(100,1000]$, $(1000,10000]$.

Our results provide evidence of different conversational behaviours related to the discussed topics. Indeed, from the Complementary Cumulative Distribution Functions (CCDFs) of the structural metrics represented in Fig. 1, we observe that discussions concerning a polarizing topic like the Italian Elections tend, compared to general topics like Italian Football, to attract a more significant number of users who produce comments with a higher toxicity score, mainly located at a lower distance from the root of the conversation tree. Such a tendency is reflected by the observed assortativity of the conversation trees, which indicates the extent to which comments with similar toxicity scores tend to be connected. From a classification perspective, our models predict whether the following comment will be toxic with an accuracy of 61.6% (AUC: 70.4%), providing comparable results with other models in the same tasks despite a restricted amount of features.

Our work provides further advances in understanding the dynamics that regulate users' toxic behaviours and, therefore, in the emergence of toxic online conversations. Indeed, understanding that specific topics may trigger toxicity can be crucial in defining social media regulations and policies tailored to a specific topic or event of interest. From a prediction perspective, achieving satisfactory results with a set of models trained for specific tree size ranges and with a minimal number of features has several implications. First, it confirms how properly framing the problem - in our case, by differentiating conversation trees due to their size - can provide further advancements in developing regulation systems. Second, it confirms the importance of studying collective dynamics on social media to find meaningful features that can improve the performance of the current tools that regulate the circulation of harmful content in online ecosystems.

References

[1] Martin Saveski, Brandon Roy, and Deb Roy. The structure of toxic conversations on twitter. In *Proceedings of the Web Conference 2021*, WWW '21, page 1086–1097, New York, NY, USA, 2021. Association for Computing Machinery.

[2] Matteo Cinelli, Andraž Pelicon, Igor Mozetič, Walter Quattrociocchi, Petra Kralj Novak, and Fabiana Zollo. Dynamics of online hate and misinformation. *Scientific reports*, 11(1):22083, 2021.

[3] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 world wide web conference*, pages 913–922, 2018.

[4] Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. ETHOS: a multi-label hate speech detection dataset. *Complex &amp Intelligent Systems*, 8(6):4663–4678, jan 2022.

[5] Bojan Evkoski, Andraž Pelicon, Igor Mozetič, Nikola Ljubešić, and Petra Kralj Novak. Retweet communities reveal the main sources of hate speech. *PLOS ONE*, 17(3):1–22, 03 2022.

[6] Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. In *Proceedings of the Web Conference 2021*, WWW '21, page 1134–1145, New York, NY, USA, 2021. Association for Computing Machinery.

[7] Google Jigsaw. Perspective api.

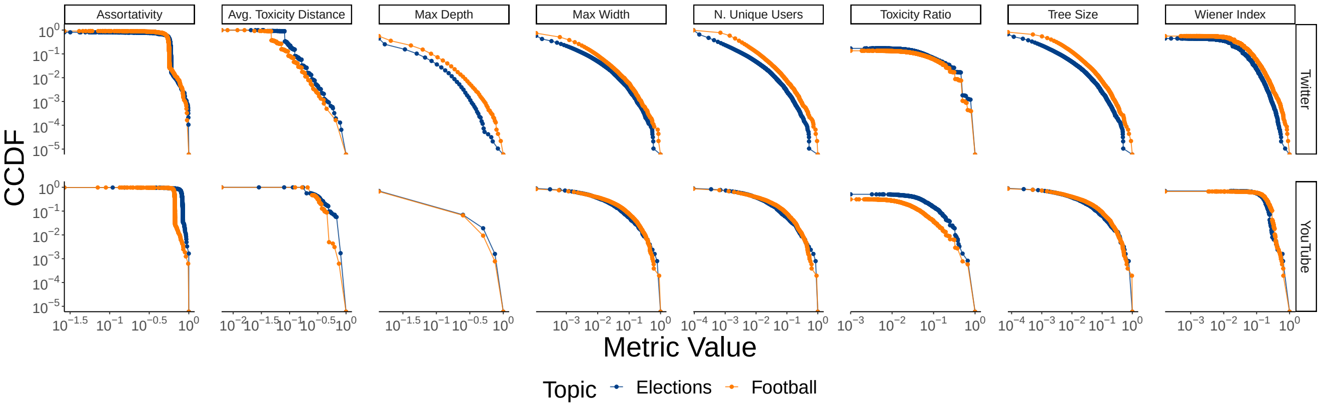

Figure 1: CCDFs of the structural metrics in the conversation trees for Twitter (upper panel) and YouTube (lower panel) computed on conversations regarding Italian elections and football.