Unveiling the Interplay between Hate and Misinfomation

Matteo Cinelli[1], Andraž Pelicon[2,3] Igor Mozetič[2], Walter Quattrociocchi[4], Petra Kralj Novak[2], and Fabiana Zollo[1]

[1] Ca' Foscari University of Venice, Italy
[2] Jozef Stefan Institute, Ljubljana, Slovenia
[3] Jozef Stefan International Postgraduate School, Ljubljana, Slovenia
[4] Sapienza University of Rome, Italy

1 Introduction

Public debates on social media platforms are often heated and polarised [1,3]. Back in the 90s, Mike Godwin coined a theorem, today known as Godwin's law, stating that "As an online discussion grows longer, the probability of a comparison involving Nazis or Hitler approaches to one". More recently, with the advent of social media, an increasing number of people is reporting exposure to online hate speech [8], leading institutions and online platforms to investigate possible solutions and countermeasures [4]. To prevent and counter the spread of hate speech online, for example, the European Commission agreed with Facebook, Microsoft, Twitter, YouTube, Instagram, Snapchat, Dailymotion, Jeuxvideo.com, and TikTok a "Code of conduct on countering illegal hate speech online". In general, the detection and contrast of hate speech is complicated since there are still ambiguities in the very definition of it, with academic and relevant stakeholders providing their own interpretation [8], including social media companies such as Facebook, Twitter, and YouTube. Here we intend hate speech as *episodes in which a speaker/user threatens, indulges, desires, or calls for physical violence against a target (e.g., minorities) or calls, denies or glorifies war crimes and crimes against humanity*. In other words, the notion of hate speech that we employ involves calls for violence against a target, in agreement with the literature and the regulators [8]. Furthermore, we look at inappropriate (e.g. profanity) and offensive language (e.g. dehumanisation, offensive remarks), which is not illegal, but deteriorates public discourse and can lead to a more radicalised society.

2 Results

We analyse a corpus of more than one million comments on Italian YouTube videos related to COVID-19 to unveil the dynamics and trends of online hate. First, we manually annotate a large corpus of YouTube comments for hate speech and fine-tune a hate speech deep learning model to accurately detect it. Then, we apply the model to the entire corpus, aiming to characterise the behaviour of users producing hate, and shed light on the (possible) relationship between the consumption of misinformation and usage of hate and toxic language. While there is a large body of literature about community-level hate speech [5], less is known about the behavioural features of users using hate speech on mainstream social media platforms, with few recent exceptions for Twitter [7] and Gab [6]. We distinguish YouTube channels into two categories: questionable, i.e., channels likely to disseminate misinformation, and reliable. This categorisation is in line with previous studies on the spreading of misinformation [2], and builds on a list of misinformation sources provided by the Italian Communications Regulatory Authority (AGCOM). Furthermore, to our knowledge, the relationship between online hate and misinformation is yet to be explored. Our results show that hate speech on YouTube is slightly more present than on other social media platforms [9] and that there are no significant differences between the proportions of hate speech detected in comments on videos from questionable and reliable channels (see Figure 1a in which we report proportion of comment types for the whole dataset that is consistent with the cases of questionable and reliable channels taken singularly). Interestingly, we do not find evidence of "pure haters", intended as active users posting exclusively hateful comments (see Figure 1b). Still, we note that users skewed towards reliable channels use on average a more toxic language –i.e. inappropriate, offensive, or violent– than their counterpart (see Figure 1c). Finally, we find that the overall toxicity of the discussion increases with its length measured in terms of the number of comments (see Figure 1d). In other words, online debates tend to degenerate towards increasingly toxic exchanges of views, in line with Godwin's law.

Acknowledgements

The authors acknowledge financial support from the Slovenian Research Agency (research core funding no. P2-103), and the European Unions Rights, Equality and Citizenship Programme under Grant Agreement no. 875263.

References

1. Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, 2005.
2. Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. *Scientific Reports*, 10(1):1–10, 2020.
3. Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320, 2016.
4. Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. *Countering online hate speech*. Unesco Publishing, 2015.
5. Nicola F Johnson, R Leahy, N Johnson Restrepo, Nicolas Velasquez, Ming Zheng, P Manrique, P Devkota, and Stefan Wuchty. Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, 573(7773):261–265, 2019.
6. Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. Hate begets hate: A temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–24, 2020.

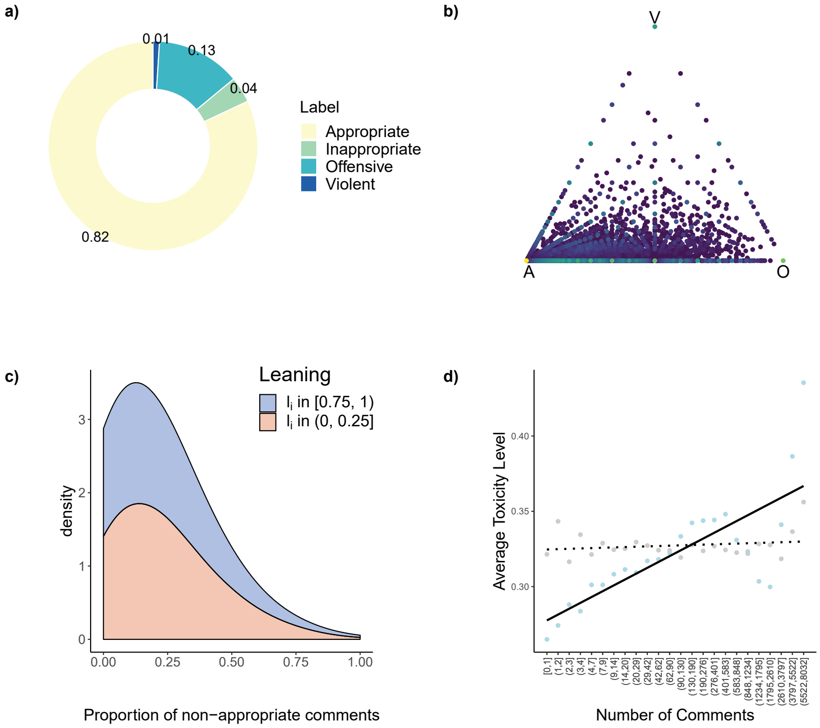

Fig. 1. Panel a: Proportion of the four hate speech labels, as provided by the machine learning model, in the whole dataset. Panel b: Users balance between different comment types, brighter dots indicate a higher density of users. Comments labelled as inappropriate (I) are eliminated. Each dot is mapped into the triangle using barycentric coordinates, and the more a user is close to a vertex of the triangle the more her/his commenting activity is based on that type of comments. Panel c: distribution of non-appropriate comments for users displaying a remarkable tendency to comment under videos posted by questionable ($l_j \in$[0.75,1)) and reliable ($l_j \in$(0,0.25]) channels. The user leaning l_j is the share of comments posted by user j under videos produced by questionable channels. Users skewed towards reliable channels post, on average, a higher proportion of non-appropriate comments ($\sim$ 23%) than users skewed towards questionable channels ($\sim$ 17%). Panel d: Linear regression model for toxicity level of conversation versus number of comments (grouped in logarithmic bins). The toxicity level of a discussion is the average of the toxicity values over all the comments of the discussion after assigning to each comment a toxicity value going from one (Appropriate) to four (Violent).

7. Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. Characterizing and detecting hateful users on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.
8. Alexandra A Siegel. Online hate speech. *Social Media and Democracy*, page 56, 2019.
9. Alexandra A Siegel, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, Joshua A Tucker, et al. Trumping hate on twitter? online hate speech in the 2016 us election campaign and its aftermath. *Quarterly Journal of Political Science*, 16(1):71–104, 2021.