How to analyze hate networks in social media with machine learning and statistical physics

Jara Juana Bermejo-Vega*

Departamento de Electromagnetismo y Física de la Materia,

Avenida de la Fuente Nueva, 18071 Granada,

Universidad de Granada, Granada, Spain and

Institute Carlos I for Theoretical and Computational Physics,

Campus Universitario Fuentenueva,

Calle Dr. Severo Ochoa, 18071, Granada, Spain.

(Dated: June 30, 2023)

Social networks are used millions of users to read news and acquire information. Despite being popular hubs for news consumption, it is know that they can be used for the spread of misinformation and potentially harmful content [1, 2]. In recent years, the usage of social media to spread hateful content of transphobic nature has been reported in several countries such as the US, UK and Spain in the form of, e.g., misinformation about trans laws or medical treatments for trans people. Several of these trends make use of specific hashtags with political intent (e.g., attacking the Spanish "trans law"), which makes them easy to identify and data-hoard.

In this talk, we apply machine learning methods and tools from statistical physics to analyze how hateful content of the social network Twitter to disseminate transphobic hatred. We employ Python libraries such as T-Hoarder and twarc2 to data-mine data from atypical trends that make heavy use of hashtags. We use supervised machine learning tools combined with qualitative human verification to identify giant connected components of accounts displaying inauthentic behavior. Specifically, we employ data analysis tools (ATLAS algorithm, Gephi), as well as statistical physics methods (entropy based models [1]) to analyze the diffusion of hateful content in complex network. Equipped with these tools, we characterize echo chambers in Twitter Spain that distribute hateful transphobic content. Our results can be used to characterize inauthentic behavior and common narratives used to spread disinformation that harms the transgender community in Twitter. This is useful for the development of public strategies to protect a marginalized communities from cyberbullying in social media.

Part of our results have been presented in the 2022 Python Conference PyConES, University of Granada [3]. Our analysis based on statistical physics entropy-based model is part of the Bachelor Thesis of Laura Cuellar (July, 2023, University of Granada), submitted as a separate poster submission.

---

[1] Becatti, C., Caldarelli, G., Lambiotte, R. et al. Extracting significant signal of news consumption from social networks: the case of Twitter in Italian political elections. Palgrave Commun 5, 91 (2019). https://doi.org/10.1057/s41599-019-0300-3

* jbermejovega@go.ugr.es

[2] Ferrara, Emilio, Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election (June 30, 2017). http://dx.doi.org/10.2139/ssrn.2995809

[3] J. Bernejo Vega, Análisis de red del discurso de odio queerfóbico en Twitter, Contributed Talk at PyConES (September 30th-Oct 2, 2022) Conference, University of Granada, Online Talk https://www.youtube.com/watch?v=nC58DAXGhgw