IDENTIFICATION OF MISOGYNISTIC ACCOUNTS ON TWITTER THROUGH GRAPH CONVOLUTIONAL NETWORKS

Lara Fontanella [1] and Emiliano del Gobbo [2] and Alex Cucco [3]

[1] G. d'Annunzio University, Chieti-Pescara, Italy, (e-mail: lara.fontanella@unich.it)

[2] University of Foggia, Italy, (e-mail: emiliano.delgobbo@unifg.it)

[3] National Heart and Lung Institute, Imperial College, London, UK, (e-mail: a.cucco20@imperial.ac.uk)

ABSTRACT: Misogyny is the hatred, dislike, and mistrust towards women simply because of their gender, accompanied by ingrained prejudice against them. Our study focuses on producers of misogyny on social media platforms, specifically examining content shared in Italian on Twitter. Using a substantial collection of Italian tweets, we analyse textual and relational data from the friend/follower network to classify Twitter accounts based on a binary misogyny scheme. We employ Graph Convolutional Networks to achieve this.

KEYWORDS: misogyny, textual data, relational data, Graph Convolutional Networks

1 Introduction

Cyberspace is often misused to spread offensive and abusive content. Women are among the most targeted groups for online abusive content (Amnesty International Italia, 2022). Hate speech against women is strongly linked to misogyny, which is the cultural attitude of hatred towards females simply because they are female. In our research, we focus on identifying producers of misogynistic content shared in Italian on Twitter. Specifically, we tackle an automatic classification task by utilising textual-based features extracted from the shared content, as well as relational data derived from the network of relationships between Twitter accounts. In hate speech research, studies have focused on automatically detecting abusive online content, while more recently, attention has shifted towards examining the behaviour and relationships of individuals who spread abusive comments on mainstream platforms. Only a few studies have adopted a network modelling approach (Chatzakou *et al.*, 2017; Mishra *et al.*, 2018). These studies have integrated graph-based features from the pro-

ducers' network into a classification model along with textual data to enhance the classification performance. However, as far as we know, networked data has not been utilized to identify misogynistic producers of online content.

2 Materials and methods

2.1 Textual and relational data

To build the textual corpus, we downloaded Italian tweets containing keywords, mentions, and hashtags related to approximately fifty politically-active women, feminists, journalists, influencers, and female television personalities. Tweets were downloaded in real time from August to December 2022, and the downloaded dataset contains 1,002,226 tweets, associated with 204,095 accounts. We filtered out accounts that no longer existed, information providers (e.g., newspapers, radio stations, television channels and programs, news aggregators), and accounts with less than 5 tweets. To ensure a less biased composition of the retrieved network, we down-sampled the accounts with tweets focusing only on Giorgia Meloni. This was necessary because approximately 75% of the total number of tweets mentioned her, which was a result of the electoral campaign and her subsequent role as Prime Minister. The final dataset includes 82,807 tweets from 7,371 accounts, and the friend/follower relations among these accounts were retrieved.
We manually annotated a subset of 942 accounts using a misogyny binary scheme. To select these accounts, we considered node centrality measures to ensure a well-spread sample on the network that included nodes with the highest degree and betweenness indexes. We also considered the distribution of tweets by the women included in the corpus construction to ensure a larger variability in the textual content and higher domain coverage. Finally, we used the revised Hurtlex dictionary (Tontodimamma *et al.*, 2023) to compute an offensiveness score at the producer level. Out of the annotated accounts, 44.6% were flagged as misogynistic.

2.2 Collective classification and Graph Convolutional Networks

Given a network, represented through a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ the set of edges, different information can be associated with each node $v \in \mathcal{V}$. In particular, we might have a set of local features $\mathbf{x}_v$, generally assumed known for the entire network, and a label y_v, which can be observed only on a node subset. In this setting, a collective classification pro-

cedure allows to jointly predict the unobserved labels considering the attributes of the nodes to be predicted in addition to the observed attributes and labels and unobserved labels of neighbouring elements. GCNs (Kipf & Welling, 2017) are a type of neural network that can perform collective node classification by learning a function f that maps a feature description $\mathbf{x}_v$ and the graph structure, represented by an adjacency matrix $\mathbf{A}$, to a node-level output $\mathbf{y}^{(U)}$, where $U \subset \mathcal{V}$ is the unlabelled nodes subset. By jointly considering the feature descriptions and the graph structure, GCNs can improve classification accuracy compared to traditional machine learning models that only use node features.
In our analysis, we utilised a binary scheme for the misogyny classification task. We employed users' textual data to extract node local features, while relational data were derived from the friend/follower users' network.

3 Preliminary results

For these preliminary results, the feature matrix was built through a bag of words approach, where functional words (i.e., pronouns, prepositions, conjunctions) and non specific domain terms, along with a misogynistic tailored lexical dictionary, were included in the document-term matrix. For the implementation of GCNs, we adopted the FastGCN algorithm (Chen *et al.*, 2018). In the classified network the misogynistic accounts amount to the 27.0% of the nodes. Figure 1 highlights some network characteristics of the misogynistic accounts: they are likely to be clustered, tend to follow more people, to be followed by less people, and to have less importance in the network structure.

4 Conclusion and future works

From our preliminary results, collective node classification performed through GCNs shows promising results regarding the prediction of misogynistic accounts. Our findings are in line with previous research on hater networks (Ribeiro *et al.*, 2018; Mathew *et al.*, 2019) that showed how hateful social media users are very densely connected and differ from normal ones in terms of their word usage and network structure. It also results from literature that haters are more likely to have a lower number of followers while following a larger number of accounts.
As future work, we will compare different GCN models using different types of embeddings to derive the feature matrix, and we will also explore masking techniques to ensure cross-domain comparison.

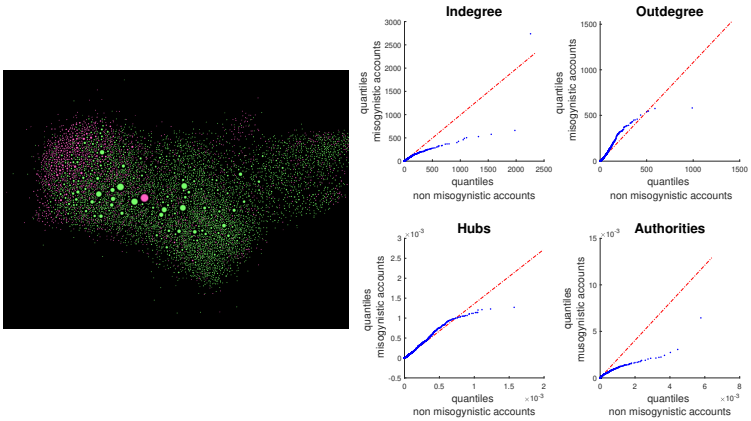

Figure 1. *Classified network - misogynistic accounts are depicted in pink - and centrality measures' qqplots*

Acknowledgements: This work was supported by EU Next Generation, MUR-Fondo Promozione e Sviluppo-DM 737/2021 [*ICOMIC: Identifying and Countering Online Misogyny*]

References

AMNESTY INTERNATIONAL ITALIA. 2022. *Odio in rete: italiani senza cittadinanza tra razzismo e xenofobia*. Tech. rept.

CHATZAKOU, D., KOURTELLIS, N., BLACKBURN, J., DE CRISTOFARO, E., STRINGHINI, G., & VAKALI, A. 2017. Mean Birds: Detecting Aggression and Bullying on Twitter. *In: WebSci '17: Proceedings of the 2017 ACM on Web Science Conference*.

CHEN, J., MA, T., & XIAO, C. 2018. FastGCN: Fast Learning with Graph Convolutional Networks via Importance Sampling. *In: 6th International Conference on Learning Representations, ICLR 2018*.

KIPF, T. N, & WELLING, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. *In: 5th International Conference on Learning Representations, ICLR 2017*.

MATHEW, B., DUTT, R., GOYAL, P., & MUKHERJEE, A. 2019. Spread of hate speech in online social media. *In: WebSci '19: Proceedings of the 10th ACM Conference on Web Science*.

MISHRA, P., DEL TREDICI, M., YANNAKOUDAKIS, H., & SHUTOVA, E. 2018. Author Profiling for Abuse Detection. *In: Proceedings of the 27th International Conference on Computational Linguistics*.

RIBEIRO, M., CALAIS, P., SANTOS, Y., ALMEIDA, V., & MEIRA JR, W. 2018. Characterizing and Detecting Hateful Users on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, **12**(1).

TONTODIMAMMA, A., FONTANELLA, L., ANZANI, S., & BASILE, V. 2023. An Italian lexical resource for incivility detection in online discourses. *Quality and Quantity*, **57**, 3019–3037.