Identification of Misogynistic Accounts on Twitter through Graph Convolutional Networks

Fontanella Lara[a], **Emiliano del Gobbo**[b], Alex Cucco[c]

[a] *G. d'Annunzio University, Chieti-Pescara, Italy*

[b] *University of Foggia, Italy*

[c] *Imperial College, London, UK*

ClaDAG 2023
11-13 September 2023

Outline

Automatic detection of producers of misogynistic online content

Misogyny: cultural attitude of hatred for females because they are female.

Online Misogyny: online content that conveys *hatred*, *aversion*, and *distrust*, and deep-seated *prejudices* against women.

Research Aim: ***Identification of producers of misogynistic online content.***

Integration of

- **textual data**
- **network relational data**

in a **collective classification** task

ICOMIC: Identifying and Countering Online Misogyny
Project funded by EU Next Generation, MUR-Fondo Promozione e Sviluppo-DM 737/2021
Funded by the European Union

Textual data collection

- ***Social media***: Twitter (*now rebranded X*)
- ***Temporal interval***: real time download from August 2022 to December 2022
- ***Software***: Socialgrabber `https://www.socialgrabber.net/`
- ***Keywords*** related to:
  - *Political Women*: Chiara Appendino, AnnaAscani, Lucia Azzolina, Anna Maria Bernini, Laura Boldrini, Giulia Bongiorno, Maria Elena Boschi, Mara Carfagna, Marta Cartabia, Susanna Ceccardi, Monica Cirinnà, Ilaria Cucchi, Paola De Micheli, Michela Di Biase, Federica Gasbarro, Mariastella Gelmini, Barbara Lezzi, Sanna Marin, Giorgia Meloni, Alessia Morani, Paola Nugnes, Virginia Raggi, Elly Schlein, Valentina Vezzali
  - *Female television personality and Influencers*: Caterina Balivo, Ilary Blasi, Giulia De Lellis, Elodie , Elisa Esposito, Chiara Ferragni, Michelle Hunziker, Vanessa Incontrada, Elisa Isoardi, Miriam Leone, Emma Marrone, Aurora Ramazzotti, Belén Rodríguez
  - *Female Journalists*: Lucia Annunziata, Bianca Berlinguer, Luisella Costamagna, Ilaria d'Amico, Veronica Gentile, Diletta Leotta

Relational data collection

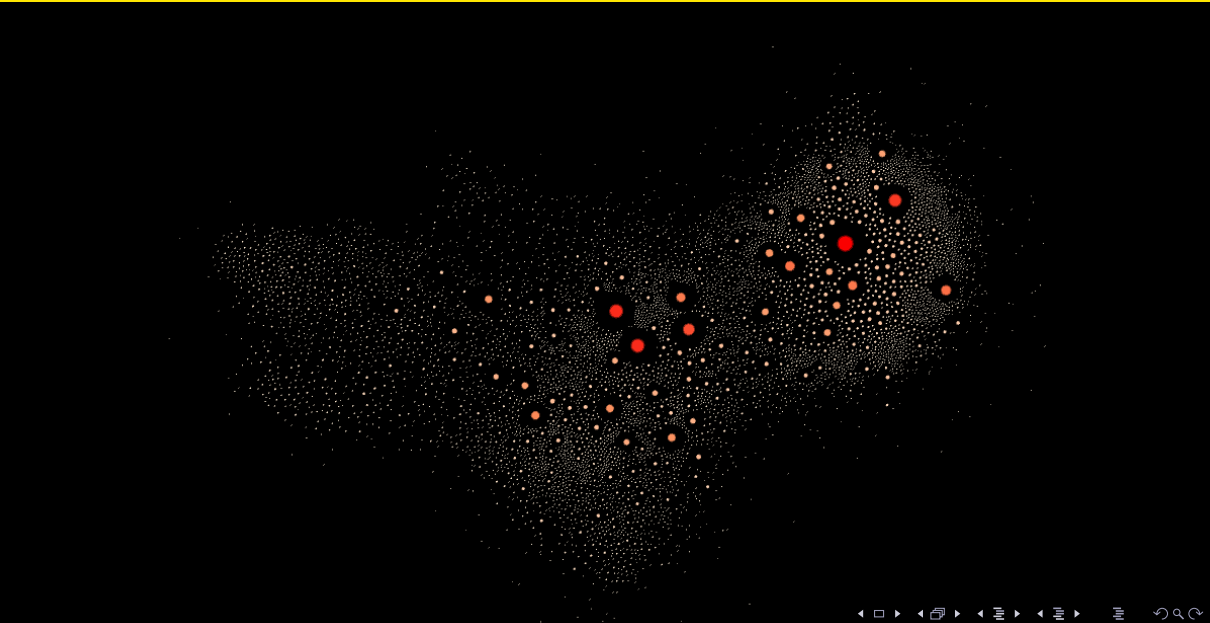
User selection rules and relational data collection:

- removal of accounts related to information providers (e.g., *newspapers, radio stations, television channels, and television programs, news aggregators*)
- removal of accounts with less than 5 tweets.
- removal of accounts with an outlier number of tweets.
- downsampling of accounts with tweets focused only on Giorgia Meloni
- removal of accounts no longer existing
- retrieval of friend/follower relations from Twitter using Socialgrabber

Number of accounts in the network: 7,354

Number of tweets: 82,807

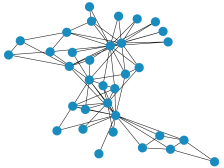
Twitter accounts' friend-follower network



Twitter accounts' textual data

Network representation and nodes' information

The structure of a network can be represented by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.



Different information can be associated to each node $v \in \mathcal{V}$:

- a set of *local features* $\mathbf{x}_v$, generally assumed known for the entire network;
- a *label* Y_v, which can be partially observed.
  - $\mathbf{Y}^{(l)}$: labelled subset
  - $\mathbf{Y}^{(u)}$: unlabelled subset.

Given an unlabelled node v, there are three distinct types of correlations that can be utilised to predict its label:

- the correlations between the label of v and the observed attributes $\mathbf{x}_v$;
- the correlations between the label of v and the observed attributes and observed labels of nodes in its neighbourhood;
- the correlations between the label of v and the unobserved labels of objects in its neighbourhood.

Collective classification

Collective classification is an important Statistical Relational Learning task:

- *related data instances are classified simultaneously as opposed to independently as in classical Machine Learning*

Collective classification: combined classification of a set of interlinked objects that exploits the attributes of an element in addition to observed attributes and labels and unobserved labels of neighbouring elements in order to predict its label [Sen et al., 2008].

Networked data: node collective classification can be considered a regular *transductive semi-supervised learning task* [van Engelen and Hoos, 2020].

- ***Semi-supervised learning*** exploits both labelled and unlabelled instances in the classification algorithm.
- ***Transductive semi-supervised learning algorithms***: given labelled data $\left(\mathbf{X}^{(l)}, \mathbf{Y}^{(l)}\right)$ and unlabelled data $\mathbf{X}^{(u)}$, provide exclusively predictions $\mathbf{Y}^{(u)}$ without producing a predictor that can operate over the entire input space.

Graph Convolutional Networks

Graph Convolutional Networks [GCNs, Kipf and Welling, 2016] allow to perform collective node classification exploiting all the types of correlation in the networked data.

GCNs extend the convolution operator to non-Euclidean data, by considering the non-Euclidean nature of the input data [Bronstein et al., 2017].

GCNs' goal: learn a function of signals/features on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ which may take as input:

- a feature description $\mathbf{x}_v$ for every node $v \in \mathcal{V}$ summarised in the feature matrix $\mathbf{X}$
- a representative description of the graph structure in the form of the adjacency matrix $\mathbf{A}$

and produces node-level output $\mathbf{Y}$

Every neural network layer can then be written as a non-linear function

$$\mathbf{H}^{(l+1)} = f(\mathbf{H}^{(l)}, \mathbf{A}, \mathbf{\Omega}_l)$$

$\mathbf{H}^{(0)} = \mathbf{X}$, $\mathbf{H}^{(L)} = \mathbf{Y}$, L: number of layers. $\mathbf{\Omega}_l$ is the weight matrix and f is the activation function.

The specific models differ in how $f(\cdot, \cdot, \cdot)$ is chosen and parameterised and by the number of layers.

GCNs in Text Analysis

Given two sets of nodes (i.e., *users* and *words*), it is possible to apply a **two layer GCN** to classify a specific type of node (e.g. *users*) based on the entire graph structure [Yao et al., 2019].
The two layer structure allows message passing among nodes that are at maximum two steps away.

Implemented GCN:

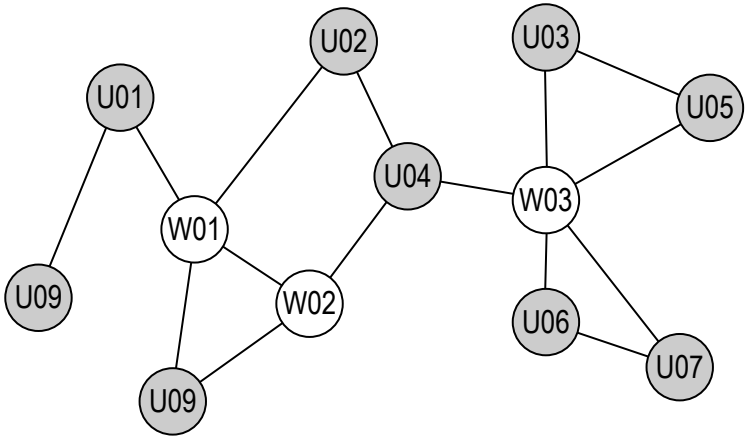
$$\mathbf{Z} = \text{softmax}(\tilde{\mathbf{A}}\,\text{ReLU}(\tilde{\mathbf{A}}\mathbf{X}\Omega_0)\Omega_1) = \text{softmax}(\tilde{\mathbf{A}}\,\text{ReLU}(\mathbf{B})\Omega_1) = \text{softmax}(\mathbf{C})$$

- $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$: normalised adjacency matrix.
- Ω_0 and Ω_1: trained weights of the first and second layer of the network.
- $\text{ReLU}(b_{ij}) = \max(0, b_{ij})$: activation function for the first layer.
- $z_{i,j} = \text{softmax}(\boldsymbol{c}_i)_j = \frac{\exp(c_{ij})}{\sum_k \exp(c_{ik})},\ j = 1, \ldots, K$: activation function for the second layer.

Each layer, after the training of the weights, represents the documents and words embeddings.

Loss function: cross-entropy error **over all labelled documents**.
Optimization function: Adam.

Representation of the network with words and users nodes

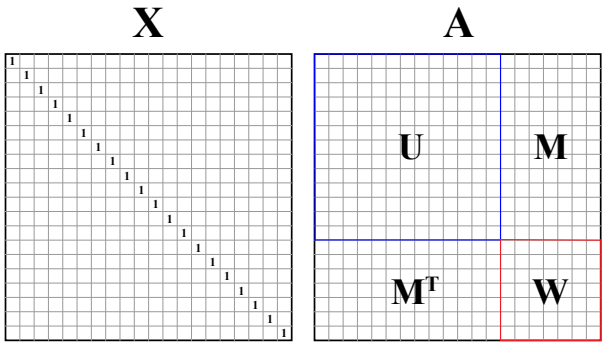


Adjacency structure and feature matrix

X = feature matrix

A = adjacency matrix

- **U** = users' friend/follower adjacency matrix
- **W** = co-occurrence terms matrix
- **M** = users/terms TF-IDF matrix



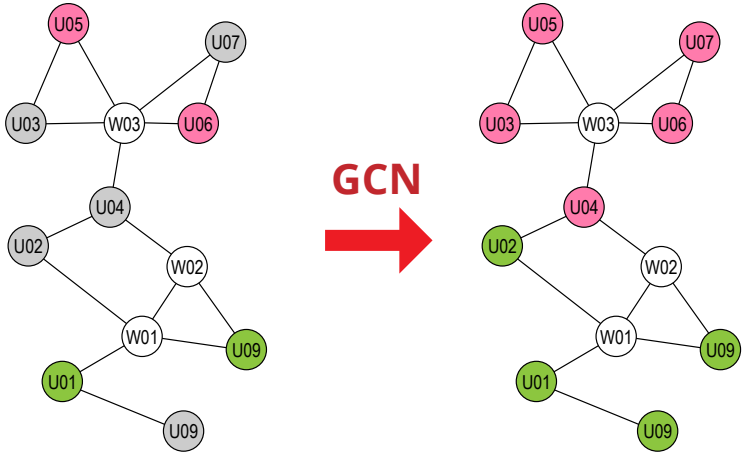
Adjacency matrix construction

- Users' adjacency matrix **U** $\rightarrow$ **Relational data**: we assume an indirect graph for the users' friend/follower relations.

- Co-occurence matrix **W** and users/terms TF-IDF matrix **M** $\rightarrow$ **Textual data**:
  - all tweets of a given account were collapsed into a single document;
  - corpus vocabulary: bag of words approach based on lexical and lexicon-based features.

  The vocabulary contains
  - all the **functional words** (i.e., pronouns, prepositions, conjunctions)
  - **non specific domain terms** selected considering two-by-two association scores (*keyness*) to compare the word distributions across the different search domains and remove the domain specific terms
  - word extracted from a **misogynistic tailored lexical dictionary** developed in the ICOMIC project.

Semi-supervised classification with GCN



Training data

A subset of accounts was selected based on the following criteria:

- Node centrality measures: a well-spread sample on the network likely includes nodes with the highest degree and betweenness indexes [del Gobbo, 2021]
- Distribution of tweets by the women chosen for corpus construction
- Level of offensiveness: the revised Hurtlex dictionary [Tontodimamma et al., 2022] was used to derive an offensiveness score at the producer level

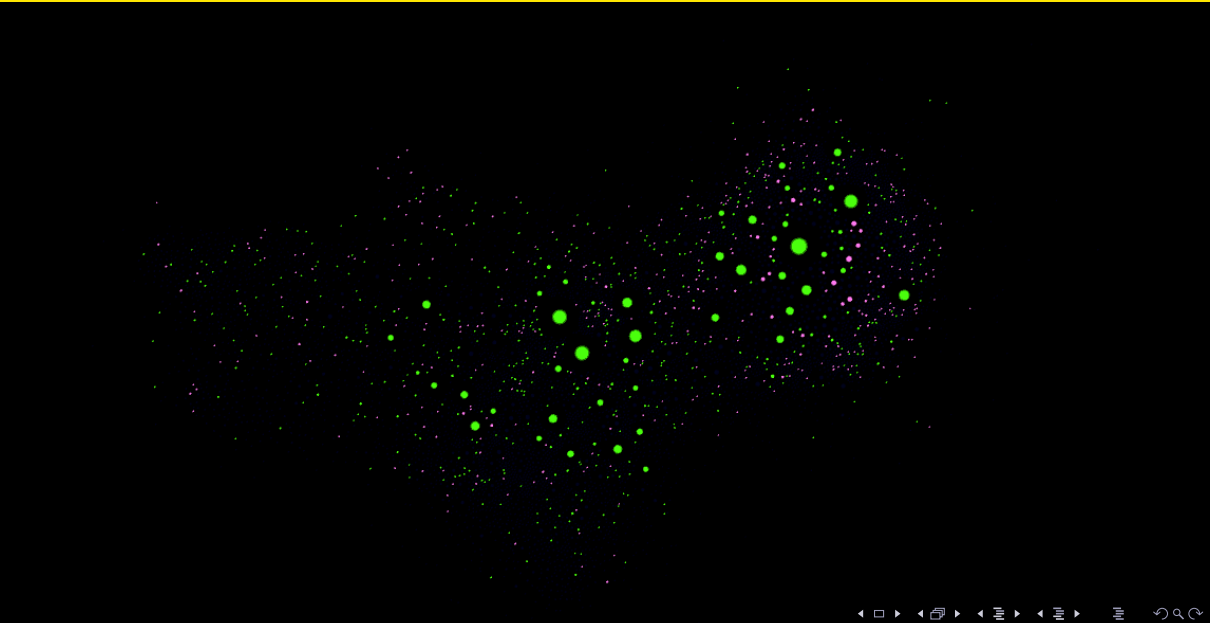
Number of sampled accounts: 937

During their internship program, students were trained on the concept of misogyny and issues related to annotation.

They analysed in detail the textual content shared by the sampled users and manually classified them using a binary coding schema

Annotated accounts: ***55.4% non-misogynistic; 44.6% misogynistic***.

Annotated Accounts on the Network



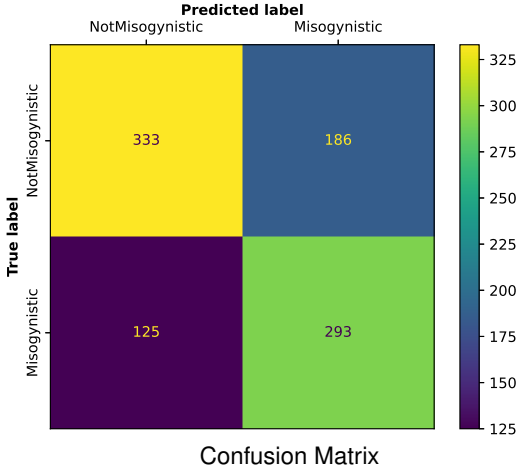
GCN 20-folds Cross-Validation

The Confusion Matrix shows the performances results of our model.

The table shows the model evaluation metrics comparing the results to other variants: using the same model, but using only the users' relationships as adjacency ($\mathbf{A} = \mathbf{U}$), or reducing the number of layers.

GCN Model	Acc.	$F1_M$	$F1_{NM}$
Users+Words | 2 Layers	0.668	0.682	0.653
Users | 2 Layers	0.630	0.684	0.552
Users | 1 Layer	0.590	0.450	0.673

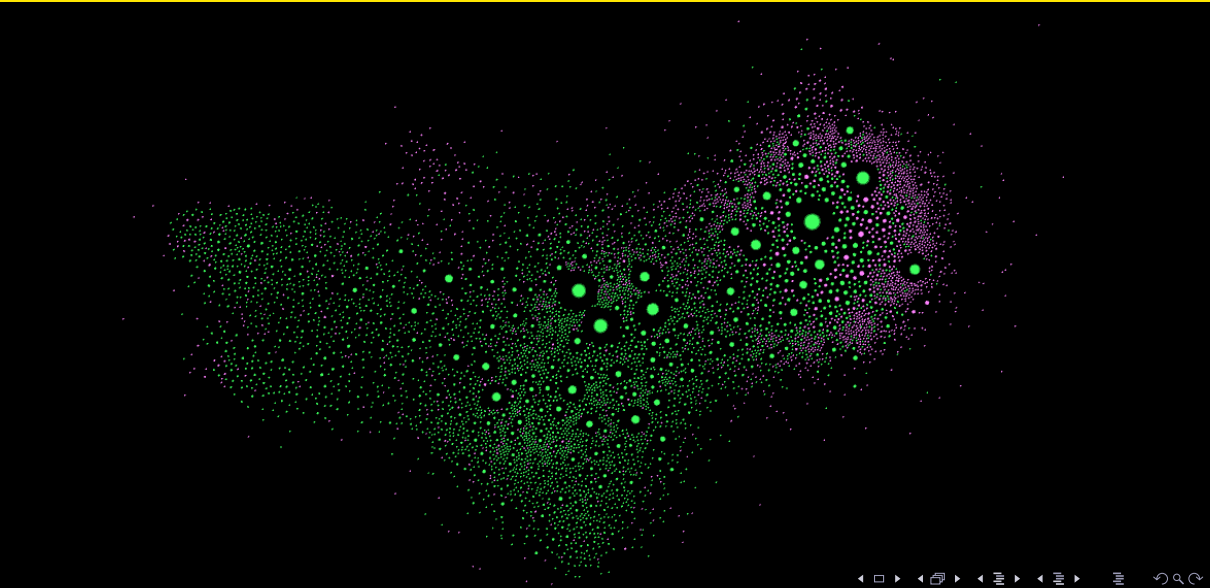
M and NM stand respectivily for Misogynistic and Not Misogynistic.



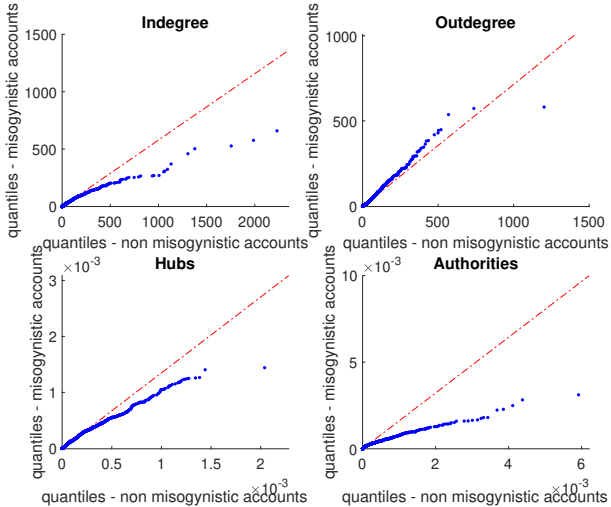
Confusion Matrix

Predicted misogynistic accounts on the Network

63.0% non-misogynistic; 37.0% misogynistic



Centrality measures on the friend directed network



Textual information comparison

misogynistic

non misogynistic

Conclusions and Future works

Preliminary results → collective node classification performed through GCNs shows promising results for the prediction of misogynistic accounts on the Network.

In line with previous research on hater networks, misogynistic Twitter accounts:

- are **clustered** [Johnson et al., 2019, Mathew et al., 2019]
- tend to **follow more people**, to **be followed by less people**, and to **be of less importance** in the network structure [Ribeiro et al., 2018].
- differ from other accounts in terms of their **word usage** [Ribeiro et al., 2018].

Future works:

- increase the number of annotated Twitter accounts to increase the performance of the classification
- include users' covariates (e.g., *total number of: friends/followers, posts, retweets, replies, mentions* [Klubička and Fernandez, 2018]) and words' covariates (e.g., *sentiment score*) in the feature matrix **X**
- compare the classification performance in a cross-domain study (e.g., train the model on the political woman textual data and predict on the entire network)

References

M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

E. del Gobbo. *Spatial and temporal aspects of online debates: a text mining approach*. PhD thesis, Scuola Superiore G. d'Annunzio, 2021.

N.F. Johnson, R. Leahy, N. Johnson Restrepo, N. Velasquez, M. Zheng, P. Manrique, P. Devkota, and S. Wuchty. Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, 573(7773):261 – 265, 2019.

T. N Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

F. Klubička and R. Fernandez. Examining a hate speech corpus for hate speech detection and popularity prediction. *arXiv preprint arXiv:1805.04661*, 2018.

B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182, 2019.

M. Ribeiro, P. Calais, Y. Santos, V. Almeida, and W. Meira Jr. Characterizing and detecting hateful users on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad. Collective Classification in Network Data. *AI Magazine*, 29:93–106, 2008.

A. Tontodimamma, L. Fontanella, S. Anzani, and V. Basile. An italian lexical resource for incivility detection in online discourses. *Quality and Quantity*, 2022.

J.E. van Engelen and H.H. Hoos. A survey on semi-supervised learning. *Mach. Learn.*, 109(1), 2020.

L. Yao, C. Mao, and Y. Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377, 2019.

Thank you for your attention!