Unveiling the Dark Side of Social Media: Developing the First Galician Corpus for Misogyny Detection on Twitter and Mastodon

Lucía Álvarez-Crespo, and Laura M. Castro

Models and Applications of Distributed Systems (MADS), Faculty of Computer Science, Universidade da Coruña, 15071 A Coruña, Spain
Correspondence: lucia.maria.alvarez.crespo@udc.es

DOI: *https://doi.org/10.17979/spudc.000024.14*

Abstract: This work aims to develop the first Galician corpus for the detection of misogyny on Twitter and Mastodon. We collect and analyze linguistic data in Galician on these social media platforms, identifying manifestations of misogyny in digital communication. The process involves data collection, text selection, and normalization, followed by thorough cleaning. We apply machine learning techniques to train accurate models for classifying the presence of misogyny. The resulting corpus facilitates analysis and study by research teams interested in misogyny in Galician. This scientific advancement contributes to the understanding and prevention of misogyny in the Galician-speaking community, promoting equality and respect in digital communication in Galician.

1 Introduction

Social media's rapid growth has exposed a rise in misogyny. Studies reveal a concerning prevalence of harassment, particularly of sexual nature, aimed at women, online (Emily A. Vogels, 2021). Despite this, natural language processing (NLP) research often overlooks minority languages like Galician, impeding our understanding and combating online misogyny in specific linguistic contexts.

Misogyny takes various forms online, including discrimination, threats, and sexual objectification (Fersini et al., 2018; Siapera, 2019). NLP emerges as a promising tool for understanding this discourse, but identifying subtle misogyny can be challenging due to cultural and contextual differences. However, automated misogyny detection is increasingly important in addressing this social issue.

This work introduces the GalMisoCorpus[1], a first approach of a Galician-language dataset, harvested from Twitter[2] and Mastodon, and develops and evaluates machine learning models for misogyny detection.

2 Related work

Detecting misogynistic discourse on social media is complex. This review explores innovative approaches in sentiment analysis for understanding abusive behavior.

[1] *https://github.com/luciamariaalvarezcrespo/GalMisoCorpus2023/*

[2] As of the summer of 2023, Twitter changed its name to 'X'. However, for the sake of readability and common usage, we will continue to use the term 'Twitter' throughout this text.

Research teams have been using sentiment analysis for some time, especially on Twitter. This technology helps understand public attitudes in various contexts worldwide. For example, it was used to study opinions on COVID-19 (Manguri et al., 2020). In Japan, computational tools analyzed misogynistic hate speech towards female politicians on Twitter, combining quantitative and qualitative methods (Fuchs and Schäfer, 2021).

Addressing online misogyny is crucial due to its prevalence. Automation can assist in rapidly identifying and combating abusive content. Challenges include distinguishing between active and passive misogyny. Tasks like IberEval's AMI and EvalIta's AMI approached misogyny as a binary classification problem (Fersini et al., 2018, 2020).

Mastodon, a decentralized network, lacks sentiment analysis research. Recent studies are emerging (Cerisara et al., 2018; Monachelis et al., 2022). However, the scarcity of academic literature hinders the understanding of misogyny in Mastodon communities.

This project aims to optimize a misogyny detection system in Galician on Twitter and Mastodon. The goal is creating accurate machine-learning models for identifying misogynistic messages. Limited literature exists in Galician on this topic, making this research a valuable contribution to understanding misogyny in Galician on these platforms.

3 Corpus

The data collection phase is crucial, especially for our dataset focused on Galician language texts mirroring common online language and topics, including misogyny. We opt for automated methods following a binary classification approach (Fersini et al., 2018; García-Díaz et al., 2021) for misogyny detection, requiring a dataset with Galician samples from both Twitter and Mastodon, classified as misogynistic or not.

We begin by acquiring a non-misogynistic class via toots from the Galician Mastodon instance. Using tailored data scraping tools, we retrieve toots in Galician from this instance, assured by stringent moderation guidelines. This allows us to select non-misogynistic toots without exhaustive content review. The Mastodon API facilitates this process, ensuring efficient and systematic data collection from May 2022 to February 2023.

Given the lack of prior work in Galician misogyny detection, we turn to the Spanish MisoCorpus-2020 (García-Díaz et al., 2021). This corpus, focused on Spanish misogyny, serves as a valuable resource. Despite being in Spanish, we leverage the CIXUG[3] translator to convert the texts to Galician, ensuring alignment with our study language. To circumvent translation issues, only tweets labeled as "Spanish from Spain" will be translated.

Merging data from the Galician Mastodon instance (representing a non-misogynistic class) with misogynistic tweets from Twitter, we create a balanced dataset. This enables accurate training and evaluation of our binary classification model. This dual approach ensures an effective distinction between misogynistic and non-misogynistic content in Galician on social media. This positions us to tackle the challenge of misogyny detection on digital platforms.

The resulting corpus comprises labeled toots and tweets, divided into two subsets. The `toots.csv`[4] contains a substantial 19,387 non-misogynistic samples. In contrast, `tweets.csv` is smaller, with 1,307 samples, as tweets lacking geolocation info in Spain were removed.

We've shared the compiled corpus with the community, though Twitter guidelines only allow sharing tweet IDs to respect content creators' rights online[5]. Also, the GalMisoCorpus is released under an open license to promote its use and the advancement of this type of research.

[3] *https://tradutor.cixug.gal/*

[4] *https://github.com/luciamariaalvarezcrespo/GalMisoCorpus2023/*

[5] *https://github.com/luciamariaalvarezcrespo/GalMisoCorpus2023/*

4 Evaluation

The model development process included data preprocessing, algorithm selection, and performance evaluation. Preprocessing, following the pipeline of García-Díaz et al. (2021), involved: (1) Converting text to lowercase, (2) Removing HTML tags and blank lines, (3) Eliminating mentions and hashtags, (4) Spelling error correction (not done due to limited resources), (5) Reducing continuously repeated symbols (see figure 1).

Figure 1: Proposed pipeline.

After preprocessing, fastText (Joulin et al., 2016) was used to generate sentence embeddings. Algorithms like Random Forest (RF), Support Vector Machine (SVM), and Linear Support Vector Machine (LSVM) were employed. SVMs used manual hyperparameters (polynomial kernel and C=1), RF retained default settings, and LSVM employed L1 penalty and squared hinge loss. Data was split 70-30 for training and testing, facilitating model evaluation and generalization testing. The evaluation used 10-fold cross-validation, dividing data into 10 parts for 10 iterations, measuring model performance with F1-score, ideal for imbalanced datasets. For evaluation, a Bag of Words (BoW) text representation was applied, with TF-IDF for unigram relevance calculation. The chi-squared method selected the most relevant unigrams for distinguishing misogynistic and non-misogynistic texts.

In summary, the pipeline included text conversion, relevance calculation, feature selection, and classification. This approach aimed to accurately identify misogynistic content in Galician on Twitter and Mastodon.

5 Results

The results observed in table 1 reveal that machine learning models, including Random Forest (RF), Support Vector Machine (SVM), and Linear Support Vector Machine (LSVM), exhibit very similar performance in this iteration. The F1-score, a metric that balances precision and recall, is high for all three models, approximately 0.90. This indicates that they all achieve a good balance between accurately classifying positive cases and finding all positive cases. Precision is high for all three models, with values above 0.86, indicating a minimization of false positives. Recall, which assesses the ability to find all positive cases, is also high, with values around 0.93. Precision and recall align with the accuracy metric, which is approximately 0.93 for all three models, indicating a high proportion of correct predictions overall.

	RF	**SVM**	**LSVM**
F1-score	0.9038	0.9101	0.8975
Precision	0.9390	0.9428	0.8664
Recall	0.9348	0.9391	0.9308
Accuracy	0.9348	0.9391	0.9308

Table 1: Metrics

Among the three models, SVM stands out as the best choice in the first iteration due to its higher F1-score, approximately 0.9101. It also has high precision (approximately 0.9428) and

recall (approximately 0.9391). This means that the SVM model has an excellent ability to correctly classify positive cases, minimize false positives, and find the most positive cases in the dataset. Considering these factors, SVM emerges as the strongest choice due to its combination of a high F1-score, high precision, and high recall.

In summary, the results of the first iteration are promising and indicate that the models perform well in the task of classifying misogyny in the Galician language corpus.

6 Discussion

Despite sentiment analysis' popularity, few solutions focus on misogyny detection, especially in minority languages like Galician. Detecting misogyny on platforms like Twitter has the potential to foster respectful online communities and protect human rights. The results, with high precision, recall, and F1-scores, show promise in identifying misogynistic content in Galician on Twitter and Mastodon.

Bibliography

C. Cerisara, S. Jafaritazehjani, A. Oluokun, and H. Le. Multi-task dialog act and sentiment recognition on mastodon. *arXiv preprint arXiv:1807.05013*, 2018.

Emily A. Vogels. The state of online harassment. *https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/*, 2021. [Online; accessed 13-September-2023].

E. Fersini, P. Rosso, and M. Anzovino. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228, 2018.

E. Fersini, D. Nozza, and P. Rosso. Ami @ evalita2020: Automatic misogyny identification. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, pages 21–28, 2020. doi: 10.4000/books.aaccademia.6764.

T. Fuchs and F. Schäfer. Normalizing misogyny: hate speech and verbal abuse of female politicians on japanese twitter. In *Japan forum*, volume 33, pages 553–579. Taylor & Francis, 2021.

J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, and R. Valencia-García. Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114:506–518, 2021.

A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

K. H. Manguri, R. N. Ramadhan, and P. R. M. Amin. Twitter sentiment analysis on worldwide covid-19 outbreaks. *Kurdistan Journal of Applied Research*, pages 54–65, 2020.

P. Monachelis, P. Kasnesis, L. Toumanidis, C. Patrikakis, and P. Papadopoulos. Evaluation and visualization of trustworthiness in social media – eunomia's approach. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 217–222, 2022.

E. Siapera. *Online Misogyny as Witch Hunt: Primitive Accumulation in the Age of Techno-capitalism*, pages 21–43. Springer International Publishing, Cham, 2019. ISBN 978-3-319-96226-9.