A Galician Corpus for Misogyny Detection Online

Lucía M. Álvarez-Crespo
Universidade da Coruña
A Coruña (Spain)
lucia.maria.alvarez.crespo@udc.es

Laura M. Castro-Souto
Universidade da Coruña
A Coruña (Spain)
lcastro@udc.es

Abstract

Social networks are virtual spaces where millions of people share ideas, opinions, and experiences. However, this broad social interaction also exposes negative and harmful behaviors, such as harassment and misogyny. Misogyny, particularly, is a worrying phenomenon that perpetuates gender inequality and undermines the dignity and rights of women.

In this context, Natural Language Processing (NLP) emerges as a promising tool to analyze and understand the discourse of social networks. However, most of NLP research, sentiment analysis, and hate speech, focuses on languages such as English and, to a lesser extent, Spanish. This implies that other languages in general, and minority languages, such as Galician, in particular, are beyond the scope of this research, and that extrapolation of results and techniques is not explored.

This work describes the development process of a Galician corpus for the detection of misogyny online. The results are made available to the research community to facilitate further analysis by third-parties interested in studying this same subject.

1 Introduction

Social networks are not only the online spaces where most human communication occurs nowadays, but also the ones where both men and women suffer the highest levels of harassment. According to a study by the Pew Research Center[1], approximately four in ten Americans have experienced online harassment (Vogels, 2021). This study revealed significant differences regarding gender, showing that women are more likely than men to report cases of harassment, both sexual (16% versus 5%) and of other kinds (13% versus 9%). As much as 33% of women under the age of 35 have ever suffered online sexual harassment, compared to 11% of men of the same age. Among adults victims of online harassment, nearly half of women (47%) believe their harassment was rooted in their being women, compared to 18% of harassed men who think likewise (cfg. Fig. 1). These data highlight the need to address and understand the issue of online harassment, especially with regard to women, in order to promote greater safety and well-being on digital platforms.

Misogyny, defined as hatred or prejudice against women, can manifest itself in a variety of ways, including social exclusion, discrimination, hostility, threats of violence, and sexual objectification. Online misogyny has been compared to witch hunting (Siapera, 2019), as it shows a similar function: to coerce women to prevent them from expressing themselves freely. This type of violence, affects especially those women in public roles, most prominently in politics, giving birth to the term VAWIP (Violence Against Women In Politics) (Union, 2018; Krook and Restrepo Sanín, 2020): they suffer sexist attacks motivated by both their gender and their public visibility.

NLP combines computational linguistics techniques, machine learning (ML), and data processing, to extract valuable information from large volumes of text (Kurdi, 2017). The application of these techniques to the study of misogyny in social networks allows for the identification of specific trends and manifestations of this phenomenon, which in turn can contribute to social awareness and the adoption of preventive measures. In particular, research on sentiment analysis has great potential for extracting critical information from opinions shared on social networks that can help identify hate speech and discrimination. These technologies have been applied in multiple text classification tasks, such as irony (Zhang et al., 2019) or hate speech detection (Corazza et al., 2020). If we consider misogyny a form of hate speech, then hate speech detectors should work

[1] https://www.pewresearch.org

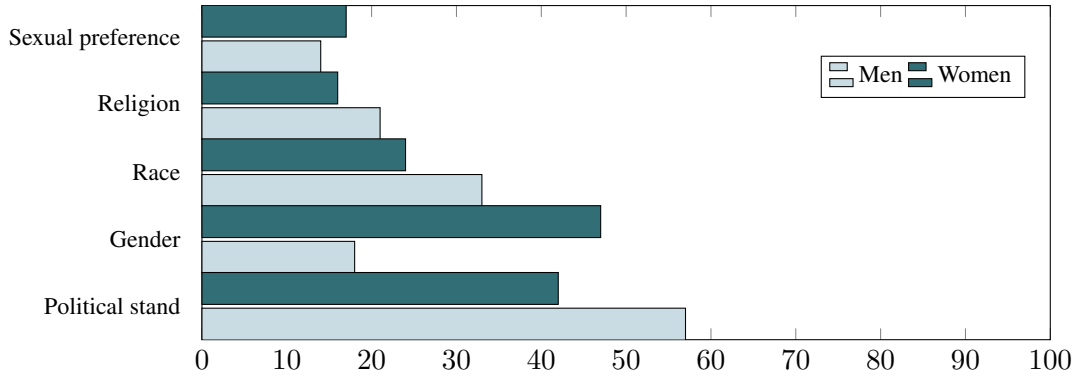

Figure 1: Reasons to which victims attribute the motivations of their online harassers (Vogels, 2021)

perfectly well by analyzing text containing misogynistic traces. However, in many cases, misogyny is presented in very subtle and obscure ways, so it may not be as easy to identify (Lundquist and Adams, 2023). In addition, cultural and context differences can complicate this identification work (McPherson, 2018). Still, automatic identification of misogyny is gaining relevance (Yin et al., 2023; Priyadharshini et al., 2022).

In order to contribute to closing the gap that non-English speakers suffer when it comes to technological advance, it is essential to address, in particular, the research and development of misogyny detection tools in other languages, especially those which are or have been minoritized, such as Galician. It is key to properly understand and address this phenomenon in each specific linguistic context, and to promote greater inclusion in the analysis of feelings and detection of hate speech research. This may require adapting and improving existing techniques, but also developing specific resources for these languages.

In this work, we address the task of detecting misogyny in texts from one of the most popular social networks, X (formerly known, and from now on referred to, as Twitter), as well as from its free alternative, Mastodon. These two platforms offer a wide space for social interaction and opinion expression, where anonymity is a prominent feature (Parlangeli et al., 2019), and in which very different moderation approaches are conducted. This is why they constitute very valuable complementary data sources for analysis. Our contributions are twofold:

- First, we have developed and made available under an open license what is, to the best of our knowledge, the first Galician corpus for the detection of misogyny. This corpus, consisting of a set of texts collected on social networks Twitter and Mastodon, constitutes a fundamental database for the training and evaluation of automatic learning models.
- Second, we have evaluated the corpus for detecting misogyny in texts from the 'Galisphere' using different ML algorithms and exploring different approaches to achieve high performance and accuracy.

The rest of the paper is structured as follows: Sect. 2 presents previous work on the matters relevant to our own; Sect. 3 explains in detail the process we followed to develop the corpus, which Sect. 4 describes, in turn. Next, Sect. 5 explains how the corpus was used to train different ML models, and their compared evaluation. Finally, we wrap up by summarizing our conclusions and future work lines on Sect. 6.

2 Related work

The detection of misogynistic discourse and offensive behavior in social media is a complex, multidimensional challenge. In recent years, many research teams have been working on sentiment analysis on social networks, especially in the context of Twitter (Manguri et al., 2020). Focusing on misogyny specifically, we find a multidimensional exploratory study on instances of misogynistic or sexist hate speech and abusive language aimed at political women in the context of Japan (Fuchs and Schäfer, 2021), and an analysis of court rulings in Portugal (Cantante, 2020).

The prevalence of misogynistic abuse on online networks, both due to its high volume and its persistence, presents challenges for both users and platform suppliers. For the latter, automated detection is interesting for expediting identification

Figure 2: Example of *passive* misogyny (LC, 2020), translated into Galician by the authors.

and combating of abusive content. (Hewitt et al., 2016) explores previous research on online misogyny, and presents an experiment that highlights the challenges of sentiment analysis to detect this phenomenon. The most notable of these is the differentiation between *active* and *passive* messages (cfg. Fig. 2), depending on whether or not they are addressed to a specific woman. This binary classification approach to the problem of misogyny detection is also present in (Fersini et al., 2018), who worked with a corpus in Spanish and another in English, both with messages labeled as active or passive, and in (Fersini et al., 2020), who worked with corpuses in English and in Italian.

Another challenge influencing misogyny detection in social media is the common use of informal language, which is not always properly registered in corpuses. (Lynn et al., 2019) used the Urban Dictionary to collect misogyny-related *slang* and studied how considering those influenced the performance of their models.

As we see, while relevant literature regarding misogyny identification does exist, it is most prominently performed within the context of the English language. We did find some research in Spanish (García-Díaz et al., 2021), but during the course of our own research we found none in Galician, and little in Portuguese: only a study of misogyny in written magazine texts (Santos et al., 2015), apart from the aforementioned analysis of bias in court rulings (Cantante, 2020). Research regarding sentiment analysis in Galician does exist, although often messages are translated from Galician into English to be able to apply already existing sentiment analysis techniques (Loureiro et al., 2022). This translation is not without issues, as it ignores the unique characteristics of the original language (in this case, Galician) that are lost in translation. For instance, when it comes to the detection and interpretation of misogyny, the loss of grammatical gender marks is absolutely crucial.

In (Ortega et al., 2022), automatic translation between different languages was explored, including Galician. The research team proposed an approach that takes advantage of the proximity between Portuguese and Galician to automate translation. This technique involved transliteration, which is the action of transcribing the written terms of one language into the other word by word, in this case from Portuguese to Galician. In turn, (Fernández and Campos, 2011) proposed a semi-automatic methodology to generate resources for sentiment analysis in Galician, taking advantage of resources in Spanish and also using Portuguese as an intermediary language due to its proximity to Galician. These studies offer valuable strategies that, avoiding translation, manage to bypass the loss of important information.

In (Agerri et al., 2018) authors describe the development of NLP processing resources and tools for Galician, including manually annotated corpuses and specific NLP modules. However, these tools and information are not useful when it comes to analyzing social media messages. The fact that they use as data sources like Wikipedia or official government websites means that the language variant is formal, and does not necessarily reflect the informal language used online.

Last but not least, we must mention, regarding the Mastodon platform, admittedly much less popular than Twitter, that the research community has also started to study it (Cerisara et al., 2018; Monachelis et al., 2022).

3 Corpus development

After exploring the state of the art, we decided to develop a Galician corpus for misogyny detection. The development process consisted of several steps to obtain and prepare the necessary data, which we describe next.

3.1 Data collection

First, we proceed with the collection of relevant data from social networks. Data collection plays a critical role in the development of any corpus: in our case, we intended to obtain a large and diverse sample of online texts in Galician that reflect the language style and usual conversation subjects present on social media, including the presence of

misogynistic content. We adopt a binary classification approach, as observed in literature (Fersini et al., 2018, 2020; García-Díaz et al., 2021).

We started the process by obtaining a non-misogynistic class for our dataset by collecting *toots* from the Galician instance of Mastodon (via its public API). Mastodon's API allows access to public data, and retrieval of messages (*toots*) via HTTP requests. This automated approach simplifies the harvesting process, making it efficient and systematic. We are confident that we do not find misogynistic content when collecting these texts thanks to the strict moderation guidelines enforced by this instance administrators, which promote respectful and inclusive communication (Alcalde-Azpiazu, 2023). This allowed us to select messages for the non-misogynistic class of our dataset without the need to perform a comprehensive review of downloaded content. The temporal range used was May 2022 to March 2023.

Regarding the misogynistic class, we initially considered the possibility of obtaining samples from `https://masto.pt`, the Portuguese Mastodon instance, given the proximity between Galician and Portuguese, as well as the existence of transliteration tools that would allow us to convert texts in Portuguese to Galician. However, the analysis of `masto.pt`'s code of conduct revealed that this instance also explicitly prohibits misogynistic behavior, and that messages are moderated accordingly (Gameiro, 2023).

After careful consideration, we chose to make use of the Spanish dataset MisoCorpus-2020 (García-Díaz et al., 2021)[2], a Spanish corpus specifically focused on misogyny. This is a balanced corpus that contains representative examples of different types of misogynistic behavior extracted directly from the social network Twitter. Specifically, the corpus is classified into three interrelated subsets: (1) the first addresses violence against relevant women, providing specific samples of those behaviors; (2) the second refers to messages that harass women in Spanish from Spain and Spanish from Latin America, offering a comprehensive view of this problem in different linguistic contexts; (3) the third encompasses general traits related to misogyny, allowing us to study their manifestation in various forms. The latter subset was the one we chose as most useful to our objective. In this case, we did not use a temporal range, but rather collected all available samples from the original MisoCorpus. We will later on address the issue of sample size difference between the misogynistic and non-misogynistic classes.

[2] `https://pln.inf.um.es/corpora/misogyny/misocorpus-spanish-2020.rar`

3.2 Data translation

The next step was to automatically translate the selected Spanish messages from MisoCorpus, to Galician. For this task, we wanted to use the tools provided by Proxecto Nós (Vladu et al., 2022). The choice of the Nós Tradutor (Ortega et al., 2022) was motivated by its commitment to the promotion and use of Galician, as well as by its quality and accuracy.

Having access to both the trained models and the translator's website, but due to the lack of an API to the mentioned web service that allowed automating the translation process, we tried to use the models directly. Unfortunately, our system turned out to be incompatible with the OpenNMT tool (Klein et al., 2017), which was necessary to run the translation models. Specifically, the version of Torch (Paszke et al., 2017) that we could install on our system was not compatible with the one required by OpenNMT. This meant that we could not make use of the full OpenNT functionality due to said incompatibility between versions.

In search of alternatives, we resorted to a translator available at CIXUG (cixug22). This tool allowed the translation of text files (`.txt`), which was a good match for our needs. The only limitation we encountered was the inability to properly translate messages from Latin-American Spanish. As a solution, we decided to use only messages geolocalized in Spain, even if the counterpart was (another) significant reduction of the sample.

4 Corpus description

We now describe the dataset we produced, which we have named GalMisoCorpus2023.

4.1 Structure

The proposed dataset is a collection of messages in Galician collected from Twitter and Mastodon.gal. This dataset consists of two CSV files: the first, `toots.csv`, contains a sample of non-misogynistic messages obtained from Mastodon.gal; the second, `tweets.csv`, contains a sample of misogynistic messages obtained from Twitter. As explained before, messages on

`toots.csv` were selected to represent the Galician language used generally on Mastodon.gal, with no misogynistic content; in turn, messages on `tweets.csv` were collected using MisoCorpus-specific criteria, and then translated to Galician.

Both files have the same structure and contain the following columns:

- `id`: a unique identifier for each message in the dataset that is the same as the one assigned by the social network of origin.
- `language`: the language in which each message is written.
- `content`: the text or content of the message.

4.2 Size

The file `toots.csv` contains a large set of 19,387 samples. Since Mastodon.gal allows users to label their own messages with the language tag of their choosing, we find messages not only in Galician, but also (although to a much lesser extent) in Spanish, English, Asturian, Catalan, Italian, and Portuguese. During the thorough analysis of this dataset, we identified an interesting situation in relation to the language attribution: a substantial number of samples labeled as *Portuguese* were, in fact, written in the *lusist* or *reintegrationist* Galician variant (Collazo, 2014). Although arguably not the case with Galician and Portuguese, the "tagging freedom" implies that users can make mistakes when identifying the language of their *toots*, leading to discrepancies between the actual language of a sample and its assigned value in the `language` field. These discrepancies should be considered, when using this corpus.

The file `tweets.csv` contains a considerably smaller set, with a total of 1,307 samples. Despite the translations we performed in this class, it is important to note that not all samples are written in Galician either. Samples were also collected in Catalan, already in the original MisoCorpus, which have been preserved intact.

Admittedly, the proposed corpus is not balanced, as the percentage of misogynistic samples is approximately 6.74% of the total. This must be taken into account in the analysis and interpretation of results derived from its use.

4.3 License

Our corpus has been released[3] under a Mozilla license to encourage and facilitate further research.

[3] https://github.com/luciamariaalvarezcrespo/GalMisoCorpus2023

For the public distribution of the dataset we must oblige by Twitter's policies of use, and consequently the `content` field of the file `tweets.csv`, must be empty. Interested parties must, thus, use the `id` field to retrieve messages from Twitter directly. Fortunately, this restriction does not apply to the Mastodon dataset, since its policies do allow the distribution of the complete contents of *toots*.

Additionally, to protect the identity of users, we have ensured that the data provided in the files do not contain directly identifiable personal information, such as user names. By taking appropriate measures to ensure anonymity and privacy, we enable the (re)use of this data for research purposes without compromising the privacy or security of the involved individuals.

5 Corpus evaluation

Next, we present the validation of our corpus by using it with several ML models for evaluation. We discuss the training procedures, and the selection of appropriate metrics for its evaluation.

5.1 Data pre-processing

Prior to any training experiment, preprocessing of the data was performed. This step involves several key tasks that contribute to the quality and reliability of ML model training results, such as removal of irrelevant characters or symbols, removal of HMTL tags, removal of emojis, and other normalization techniques (i.e. lowercasing).

When performing the data pre-processing, we follow the same procedure used in MisoCorpus (García-Díaz et al., 2021) from which we extract our misogynistic samples. We add one additional previous step, and we then apply the preprocessing pipeline to both our data classes. This facilitates the comparison with previous contributions that make use of the MisoCorpus, and ensures coherence and consistency. The steps are:

1. Removal of emojis (*Mastodon messages*)
2. Lowercasing
3. Removal of empty lines and HTML tags
4. Removal of hashtags and mentions
5. Fixing typos (*not performed*)
6. Removal of repeated characters

We added the first step because it was required for the samples from Mastodon.gal. Although emojis do contain relevant information, their interpretation and analysis requires specific tools and, given their absence from the MisoCorpus samples, we chose to remove them to maintain concordance and comparability between the two data classes. Given that emoji removal may result in an empty message, we made sure we eliminated those and preserved only samples with textual content.

The second step involved converting all text samples to lowercase, with the goal of unifying the way words are written.

The third step was the removal of blank lines, which do not contain any textual content. Since empty lines do not contribute to the analysis, they can be omitted, resulting in more coherent and compact texts. URLs were also removed.

The fourth step was the elimination of hashtags by removing the special character (#) and keeping the word (so that `#feminist` becomes `feminist`). In this step we also remove mentions to other accounts and/or users (character `@`). Mentions are deleted with the aim of removing, as already mentioned, direct references to specific users.

Even if listed here for completeness, step 5 was actually not performed: no spell correction was applied to messages in Galician. Spell correction is a complex task that requires specific tools. Given the reality of the limited resources available for text processing in Galician, we preferred not to modify the data in this regard. However, it is important to take this limitation into account when analyzing and interpreting the results derived from this preprocessed corpus.

Last, we proceed to eliminate characters and symbols that are repeated within text messages. This step materializes the fact that, in many cases, repetition does not provide relevant information to textual analysis, while it may adversely affect later stages of the process. Thus, by removing repeated symbols, we seek to reduce noise and ensure a cleaner and more concise representation of the textual content of the samples.

The resulting pre-processed dataset is also publicly available in the aforementioned repository (cfg. Sect. 4.3), under the same license.

5.1.1 Word embeddings

We now address the process of generating *sentence embeddings* from the preprocessed texts. Sentence embeddings are representations that capture semantic and contextual information of texts, and constitute very relevant elements in their analysis and comparison.

Sentence embeddings are composed of *word embeddings*, which are dense representations of words within a high-dimensional space, creating clusters of words that are semantically similar. Sentence embeddings can be represented as an *average of word embeddings* in the text. Sentence embeddings behave similarly to word embeddings, as they share the same main properties (Arora et al., 2017). In our work, we apply the Galician FastText model (Joulin et al., 2016), which contains pre-trained word embeddings from Common Crawl and Wikipedia.

However, it is important to note that, unlike in the original study (García-Díaz et al., 2021), the extraction of linguistic features was not possible in our case. The tool they use, UMUTextStats (García-Díaz et al., 2022), gives detailed linguistic information about texts (i.e. word counting, letter frequency, etc.) only for Spanish. Due to the lack of equivalent tools for Galician, we could not extract linguistic features from our texts. Consequently, we miss a valuable source of information about specific aspects of the language that could influence the detection of misogynistic messages. Linguistic features include elements such as grammatical structure, the use of certain words or expressions, and characteristics inherent to the language. These aspects are important to fully understand the content of texts and to detect subtleties or nuances that may reveal misogynistic content. Without this, we can be missing opportunities to identify misogynistic messages that are expressed in Galician in particular ways.

5.2 Training experiments

We now explore several ML algorithms, specifically Random Forest (Breiman, 2001), Support Vector Machine (Vapnik, 1999) and Linear Support Vector Machine (Cortes and Vapnik, 1995), for the task of misogyny identification in Galician social network messages.

First, we train the models with our cleaned-up, unbalanced dataset. We use the Scikit-Learn (Pedregosa et al., 2011) and (1) for RF we maintain the library's default values for the hyperparameters, following the example of (García-Díaz et al., 2021); (2) for SVM we use a polynomial kernel and C=1, again following on the footsteps of (García-Díaz et al., 2021); (3) for LSVM we

apply an L1 penalty and squared hinge loss, once more as in (García-Díaz et al., 2021).

We apply a usual 70-30 division of the corpus (Vrigazova, 2021), meaning we use 70% of the corpus samples for training and 30% for testing. We also apply a 10-fold cross validation, where we divide the whole dataset into 10 parts (folds) and iterate 10 times, using a different fold as test set each time, and the rest as training data. As comparison metric, we use F1-score instead of accuracy because F1-score combines precision and recall. This is an especially relevant combination in the presence of unbalanced classes, as it is our case, since it takes into account both false positives and false negatives.

We evaluate our models using a BoW (bag of words) text-representation model, a very common text representation technique in NLP. In this technique, each message is treated as an unordered set of words without considering any grammatical information. This representation model is simple and yields good results in NLP tasks (Cámara et al., 2011), although we must consider that it requires a lot of resources, both time and memory.

To calculate the percentage of unigrams (individual words) in documents we calculate the Term Frequency-Inverse Document Frequency (TF-IDF) to measure the relevance of each feature within the corpus, using the frequency of the normalized term to avoid bias with common unigrams. Our reference research (García-Díaz et al., 2021) does not specify which feature selection algorithm they use to filter the most discriminatory unigrams, so we use the Chi-square (χ^2) method, as a sensible choice for feature selection for text classification tasks (Mohd A Mesleh, 2007). This method is based on a homonimous statistical test, which helps us measure the relationship between categorical variables. In our case, we consider each unigram as a categorical variable, and we want to determine which are the most relevant unigrams for the classification between misogynistic and non-misogynistic texts. By applying the χ^2 method, we can calculate a score of importance for each unigram relative to the target variable, which is the classification as misogynistic or non-misogynistic. Unigrams that have a higher χ^2 score are considered more relevant and have a greater influence on the classification between the two types of texts.

In short, our experiment procedure can be summarized as follows:

1. Convert text to a BoW representation.

2. Calculate the importance of each unigram in documents using TF-IDF with the frequency of the standardized term.

3. Use the χ^2 scoring function to perform a selection of attributes.

4. Apply each of the previously proposed classifiers (RF, SVM and LSVM) with their respective selected hyperparameters.

In a second iteration of our experiments, we apply random subsampling (RUS) (Japkowicz and Stephen, 2002) to treat our data unbalancing. We follow the same training procedure we have just described, but we apply the RUS technique to our majority class data (non-misogynistic samples), in order to reduce its size and balance the distribution of both classes in the dataset. In particular, we randomly remove samples from the majority class until the ratio is the same.

The results are presented in Tab. 1 and Tab. 2, which reveal two different scenarios. In both, the three ML models exhibit very similar performance.

	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: Model Metrics (first iteration)

Table 1 shows a very promising scenario, where we see that 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. In this scenario, SVM emerges as the strongest choice due to its combination of a high F1-score, high precision, and high recall.

	RF	SVM	LSVM
F1-score	0.5118	0.4484	0.3226
Precision	0.5425	0.4766	0.2404
Recall	0.5375	0.4736	0.4903
Accuracy	0.5375	0.4736	0.4903

Table 2: Model metrics (second iteration –w/RUS–)

However, Table 2 depicts a different image. We can see that the values for F1-score, precision, recall and accuracy are quite low in general. This indicates that the models are not showing good performance in the detection task at hand. The F1-score is especially low for all three models, with values ranging from 0.3226 to 0.5118. This indicates that models are having difficulty achieving a balance between accuracy and the ability to find positive cases in data. Accuracy is also low, with values ranging from 0.2404 to 0.5425. This means that models are returning many false positives when classifying cases. The recall value, which represents the ability to find positive cases, is also low, with values ranging from 0.4736 to 0.5375. This implies that models are letting many positive cases go undetected. Finally, accuracy is also low, with values ranging from 0.4736 to 0.5375. This indicates that models are not making correct predictions in general.

Our conclusion is that the application of the RUS technique led to a significant loss of misogynous class-related information. In other words, the subsampling affected the ability of models to correctly identify the cases of misogyny, resulting in unsatisfactory overall performance. In this sense, it is important to consider other approaches to treat unbalanced data, such as minority class oversampling or the use of ML algorithms designed to directly treat class imbalance. Other strategies to improve model performance, such as hyperparameter optimization could also be explored.

6 Conclusions and future work

Despite the great popularity of sentiment analysis, few research is focused on detection of misogyny, and even less on minority languages, such as Galician. The impact of research focused on toxic language detection is potentially huge, both in number of online interactions and in terms of mental health benefits: fighting discrimination, promoting a more respectful online community and fostering a safe and inclusive environment for all users deserves more attention in this research field.

The main objective of this work was to develop a first corpus for the detection of misogynistic social media messages in Galician language. The corpus, that we named GalMisoCorpus2023, is available both in its original and in processed form under an open license (galmisocorpus23). As a second objective, we built a classification system based on ML algorithms to automatically identify misogynistic messages, to demonstrate the usefulness of the GalMisoCorpus2023. This system went through several iterations, being evaluated and compared using different metrics. The results show promising performance in the detection of misogynistic messages in Galician online messages. Models of the first iteration, especially SVM, achieved high values of precision, recall and F1-score, indicating an adequate ability to correctly identify and classify misogynistic messages. However, a second iteration in which we tried to balance the two corpus classes (mysoginistic and non-mysognistic messages) showed much worse results, leaving open doors for further work.

We could expand the dataset used for training, as a larger amount and variety of messages could further improve system performance. This would require the collection and labeling of more data in Galician, to enrich and diversify the training set. One way of achieving this would be requesting access to the moderated *toots* in the Mastodon.gal instance. This would eliminate the need for translation, and thus constitute a valuable source of information, provided that moderated *toots* are preserved and available.

A different way of expanding the dataset would be the application of oversampling techniques. Oversampling is a technique used to address class imbalance in training data, that goes in the opposite direction of undersampling, the one we used in this work and which yielded unsatisfactory results. The application of oversampling techniques could have a different outcome.

Another important line of future work we would like to explore is the development of lexicons or models that support emojis. Emojis are elements that are widely used in social media and can convey specific emotions, attitudes, or feelings, and as such are surely important also in the identification of misogynistic or offensive content.

Finally, we aim to extend our experimentation to some Deep NLP models, like the multilingual base models provided by the HuggingFace project (Wolf et al., 2020). We also would like to

explore resources that might help overcome the deficiencies of sentiment analysis-based approaches in detecting offensive content based on genderedness (Dinan et al., 2020), which could result in an enriched corpus with pragmatic annotations.

References

Rodrigo Agerri, Xavier Gómez Guinovart, German Rigau, and Miguel Anxo Solla Portela. 2018. Developing new linguistic resources and tools for the galician language. In *International Conference on Language Resources and Evaluation*.

Rafael Alcalde-Azpiazu. 2023. Acerca de - mastodon.gal.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Eugenio Martínez Cámara, M Teresa Martín Valdivia, José M Perea Ortega, and L Alfonso Ureña López. 2011. Técnicas de clasificación de opiniones aplicadas a un corpus en español. *Procesamiento del Lenguaje Natural*, 47:163–170.

Inês Cantante. 2020. Deteção de bias num acórdão jurídico. *Redis: Revista de Estudos do Discurso*, 9:4378.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa Le. 2018. Multi-task dialog act and sentiment recognition on mastodon. *arXiv preprint arXiv:1807.05013*.

cixug22. 2022. Consorcio interuniversitario de galicia (cixug): Tradutor.

Silvia Duarte Collazo. 2014. O estándar galego: reintegracionismo vs. autonomismo. *Romanica Olomucensia*, 1:1–13.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Trans. Internet Technol.*, 20(2).

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Conference on Empirical Methods in Natural Language Processing*, pages 314–331. Association for Computational Linguistics.

Paulo Malvar Fernández and José Ramom Pichel Campos. 2011. Generación semiautomática de recursos de opinion mining para el gallego a partir del portugués y el español. In *Workshop on Iberian Cross-Language Natural Language Processing Tasks*.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. Ami @ evalita2020: Automatic misogyny identification. In *Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, pages 21–28.

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

Tamara Fuchs and Fabian Schäfer. 2021. Normalizing misogyny: hate speech and verbal abuse of female politicians on japanese twitter. *Japan Forum*, 33(4):553–579.

galmisocorpus23. 2023. Galimisocorpus 2023.

Hugo Gameiro. 2023. Sobre - mastodon (pt).

José Antonio García-Díaz, Pedro José Vivancos-Vicente, Angela Almela, and Rafael Valencia-García. 2022. Umutextstats: A linguistic feature extraction tool for spanish. In *Language Resources and Evaluation Conference*, pages 6035–6044.

José Antonio García-Díaz, Mar Cánovas-García, Ricardo Colomo-Palacios, and Rafael Valencia-García. 2021. Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114:506–518.

Sarah Hewitt, T. Tiropanis, and C. Bokhove. 2016. The problem of identifying misogynist language on twitter (and other online social spaces). In *ACM Conference on Web Science*, page 333335. Association for Computing Machinery.

Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Mona Lena Krook and Juliana Restrepo Sanín. 2020. The cost of doing politics? analyzing violence and harassment against female politicians. *Perspectives on Politics*, 18(3):740755.

Mohamed Zakaria Kurdi. 2017. *Natural language processing and computational linguistics 2: semantics, discourse and applications*, volume 2. John Wiley & Sons.

Lauren LC. 2020. Misogyny manifestation across all social media platforms.

Maria L. Loureiro, Maria Alló, and Pablo Coello. 2022. Hot in twitter: Assessing the emotional impacts of wildfires with sentiment analysis. *Ecological Economics*, 200:107502.

Caroline R. Lundquist and Sarah LaChance Adams. 2023. A continuum of women's agency under misogyny. *Hypatia*, 38(1):105113.

Theo Lynn, Patricia Takako Endo, Pierangelo Rosati, Ivanovitch Silva, Guto Leoni Santos, and Debbie Ging. 2019. A comparison of machine learning approaches for detecting misogynistic speech in urban dictionary. In *International Conference on Cyber Situational Awareness, Data Analytics And Assessment*, pages 1–8.

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

Rachel McPherson. 2018. *Variables Influencing Misogyny*. Ph.D. thesis, University of Central Florida.

Abdelwadood Mohd A Mesleh. 2007. Chi square feature extraction based svms arabic language text categorization system. *Journal of Computer Science*, 3(6):430–435.

Panagiotis Monachelis, Panagiotis Kasnesis, Lazaros Toumanidis, Charalampos Patrikakis, and Pericles Papadopoulos. 2022. Evaluation and visualization of trustworthiness in social media eunomia's approach. In *IEEE Annual Computers, Software, and Applications Conference*, pages 217–222.

John Ortega, Iria De-Dios-Flores, Pablo Gamallo, and José Campos. 2022. A neural machine translation system for spanish to galician through portuguese transliteration. In *Annual Conference of the Spanish Society for Natural Language Processing*.

Oronzo Parlangeli, Enrica Marchigiani, Margherita Bracci, Alison Margaret Duguid, Paola Palmitesta, and Patrizia Marti. 2019. Offensive acts and helping behavior on the internet: An analysis of the relationships between moral disengagement, empathy and use of social media in a sample of italian students. *Work*, 63(3):469–477.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298. Association for Computational Linguistics.

Anabela Santos, Carla Cerqueira, and Rosa Cabecinhas. 2015. Between the norm and the exception: gender asymmetries in portuguese newsmagazines. *Comunicação e Sociedade*, 27:457474.

Eugenia Siapera. 2019. *Online Misogyny as Witch Hunt: Primitive Accumulation in the Age of Technocapitalism*, pages 21–43. Springer International Publishing.

Inter-Parliamentary Union. 2018. Sexism, harassment and violence against women in parliaments in europe. Technical report, Inter-Parliamentary Union.

Vladimir Vapnik. 1999. *The nature of statistical learning theory*. Springer science & business media.

Adina Ioana Vladu, Iria de Dios-Flores, Carmen Magariños, John E Ortega, José Ramom, González González, Senén Barro, and Xosé Luis Regueira. 2022. Proxecto nós: Artificial intelligence at the service of the galician. In *Annual Conference of the Spanish Society for Natural Language Processing*.

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

Borislava Vrigazova. 2021. The proportion for splitting data into training and test set for the bootstrap in classification problems. *Business Systems Research: International Journal of the Society for Advancing Innovation and Research in Economy*, 12(1):228–242.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Wenjie Yin, Vibhor Agarwal, Aiqi Jiang, Arkaitz Zubiaga, and Nishanth Sastry. 2023. Annobert: Effectively representing multiple annotators label choices to improve hate speech detection. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):902–913.

Shiwei Zhang, Xiuzhen Zhang, Jeffrey Chan, and Paolo Rosso. 2019. Irony detection via sentiment-based transfer learning. *Information Processing & Management*, 56(5):1633–1644.