Misogyny Detection

Isabella Cadisco

July 29, 2024

1 Introduction

Misogynistic speech, a form of hate speech against women, is a significant social problem with deeply rooted cultural implications. Addressing this form of discourse is critical because it can perpetuate harmful stereotypes, contribute to gender discrimination, and undermine the well-being of individuals and communities. Identifying and mitigating misogynistic speech can improve the safety of online and offline environments, promoting more respectful and equitable discourse. In addition, the study of vocabulary related to misogynistic discourse in different languages, such as English and Spanish, can provide insights into cultural nuances and differences in gender-related discourses, creating also globally applicable solutions. The aim of this project is to explore the possibility of using a neural language model based on Transformers [1] to detect misogynistic speech and to provide global explanation of the decisions of the models, enhancing interpretability [2].

2 Research Question and Methodology

2.1 Binary classification

The adopted approach for binary text classification is a neural model made by a pretrained encoder, to build an informative representation of the input text with a classification head (e.g. a linear layer), mapping the output layer features to a bi-dimensional vector, where 1 indicates the presence of misogynistic speeech and 0 the absence.

2.2 SHAP explanation

SHAP (SHapley Additive exPlanation), introduced in (S. Lundberg and S.-I. Lee, 2017) [2], is a popular post-hoc interpretability technique derived from coopera-

tive game theory that provides importance scores of features that explain model predictions. The Python shap package facilitates these explanations by helping to discover which elements of the text sequence significantly influence a model's classification decisions, in this context of misogynistic discourse detection. In addition, the shap package, offers ways to aggregate local explanations (i.e. concerning individual examples) into a global explanation.; in this project it is used to investigate what vocabulary is related to misogynistic speech.

3 Experimental Results

3.1 Data

3.2 English dataset

The dataset consists of 6567 Reddit posts and comments, annotated by expert annotators as misogynistic or not [3].

After removing null rows and splitting into train, validation and test, the following splits were obtained.

Split	Cardinality
Train	5254
Validation	651
Test	650

As is shown in Figure 1, the label distribution is quite imbalanced for all the three splits.

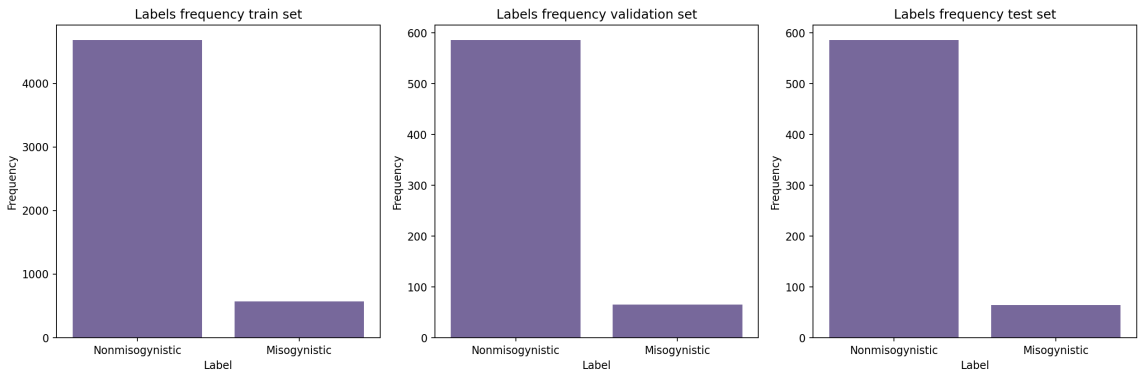

Figure 1: Labels distribution English dataset

3.3 Spanish dataset

The dataset consists of 10029 tweets in Spanish language. It was downloaded from github.comvaleriavlaworkshop_misogyny. For the experiment with Spanish language, the dataset was only split in train and test set, obtaining the following splits, for which the label distribution is shown in Figure 2.

Split	Cardinality
Train	8023
Test	2006

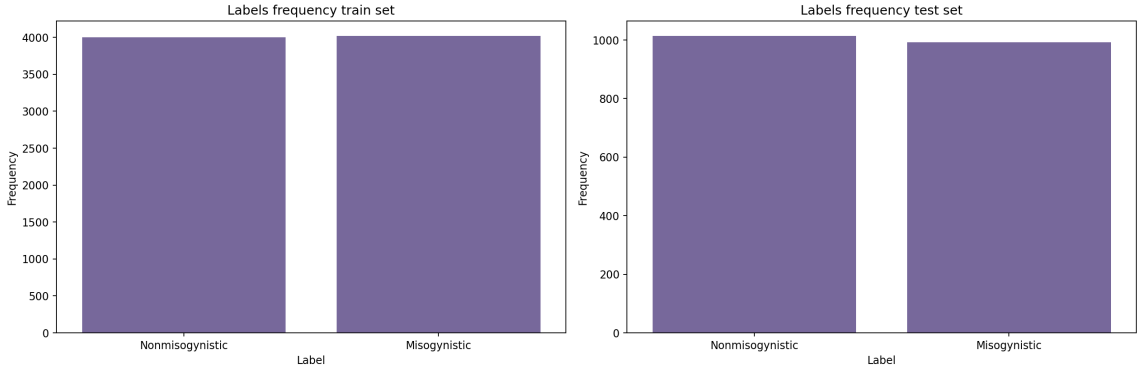

Figure 2: Labels distribution Spanish dataset

In contrast to the English-language dataset, this dataset has an extremely balanced distribution of labels.

3.4 Metrics

The performance of the model is evaluated in terms of accuracy, F1 score, precision, and recall.

Accuracy being the proportion of correct predictions among the total number of cases processed.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

F1 score being the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Precision is the fraction of correctly labeled positive examples out of all the examples that were labeled as positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall is the fraction of the positive examples that were correctly labeled by the model as positive.

$$\text{Recall} = \frac{TP}{TP + FN}$$

3.5 Detection in English posts

The pre-trained encoder employed for misogyny detection in English texts was bert-base-uncased [4], as described in Section 2 the binary classification is obtained adding to it a classification head.

The data was first prepared removing unnecessary columns. The dataset is then split into training, validation, and test sets. Subsequently, the text data is tokenized using the *DistilBERT tokenizer*, after being batched and padded to ensure a uniform input size.

A callback implementing **early stopping** was defined to halt training early if performance does not improve. The metric to be checked is *recall*, this is done to minimize false negatives, which is essential for detecting sensitive content.

Additionally, Optuna is used for hyperparameter tuning, specifically optimizing the learning rate to maximize recall. The number of training epochs is set to 12, the weight decay to 0.01, and the batch size is automatically selected thanks to the *auto_find_batch_size* from *Hugging Face Trainer*.

Since was not possible to directly save the best model during the hyperparameter optimization phase, the best model is trained with the selected learning rate, $1.5e-5$, keeping all other hyperparameters the same as before. Early stopping is called at epoch 5, with the following evaluation metrics:

- **Accuracy**: 0.93
- **F1**: 0.59
- **Precision**: 0.83
- **Recall**: 0.46

The model is now used to get prediction on the unseen data in the test set. In Figure 3, the resulting confusion matrix is reported; in Figure 4 the normalized version.

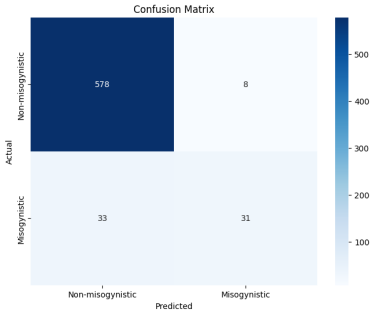

Figure 3: Confusion matrix English test set

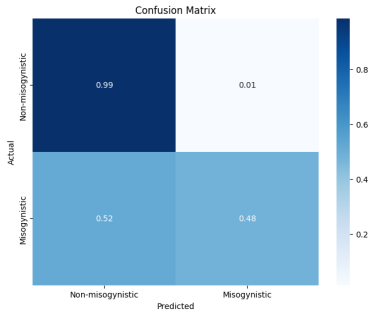

Figure 4: Confusion matrix English test set normalized

Given the high imbalance in the data, these results are encouraging, suggesting that by increasing the sensitive texts to be identified, the performance of the model may improve.

Since the goal is to identify the vocabulary that leads the model to classify a text as misogynistic or not, before extracting the SHAP global explanation, the texts classified by the model as such were filtered, resulting in 39 examples. After further filtering for texts greater than 512 tokens, in order to facilitate the use of the *Hugging Face pipeline*, 30 examples are obtained, from which the SHAP explanation shown in Figure 5 is extracted.

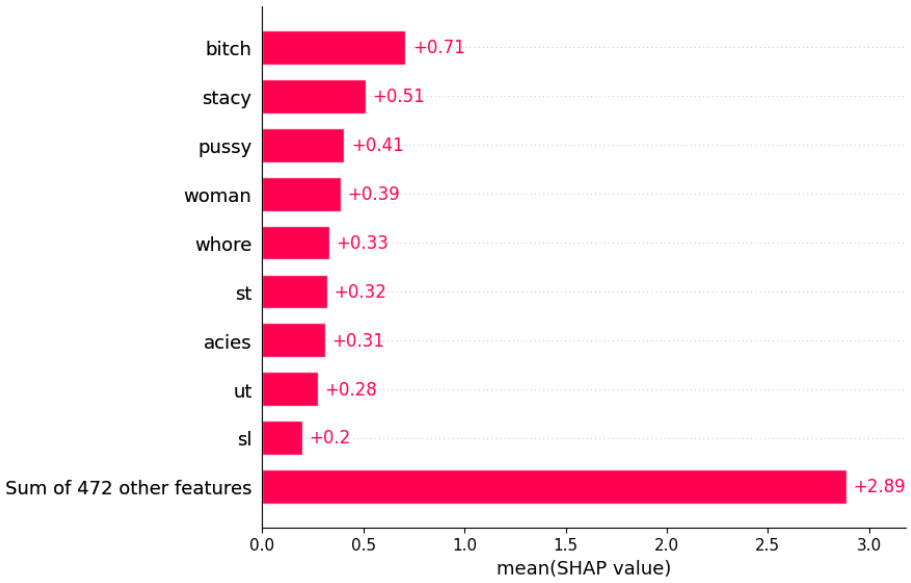

Figure 5: SHAP explanation

Despite the lack of data and the resulting poor performance of the model in terms of metrics, it can be seen that the SHAP explanation with respect to model decisions is effective in identifying misogynistic vocabulary.

3.6 Spanish texts

The pre-trained encoder employed for misogyny detection in Spanish texts was partypress/partypress-monolingual-spain, this model was preferred since it is already fine tuned on Spanish text corpora. This classification model, pretrained encoder plus a classification head, is built on dccuchile/bert-base-spanish-wwm-uncased, it was trained to identify 23 classes. Since the misogyny detection problem is framed as a binary classification problem only 2 are needed. To achieve this is sufficient to randomly initialize the weights of the classification head, setting number of classes equal to 2.

Although it is not correct to use the same hyperparameters as in the previous experiment, changing both the data and the model, this was done to speed up the process and see if anything lexically interesting is detected. Also, only the learning rate was optimized, the other hyperparameters were fixed.

Also for this experiment, early stopping is called at epoch 5, with the following evaluation metrics:

- **Accuracy**: 0.88
- **F1**: 0.88
- **Precision**: 0.8
- **Recall**: 0.90

The model is now used to get prediction on the unseen data in the test set. In Figure 6, the resulting confusion matrix is reported; in Figure 7 the normalized version.

From these confusion matrices, can be seen that with more balanced training data, the model is performing well.

As in the previous experiment, the SHAP explanation is provided for examples classified as misogynistic, in the case of the Spanish dataset 1051. Given the large amount of time required to calculate the SHAP exaplanation for each example and then aggregate it into the overall explanation, only 40 examples are considered in this experiment. The SHAP explanation is shown in Figure 8 is extracted.

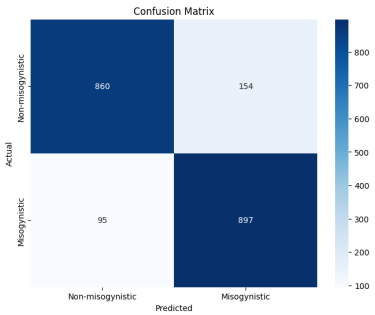

Figure 6: Confusion matrix Spanish test set

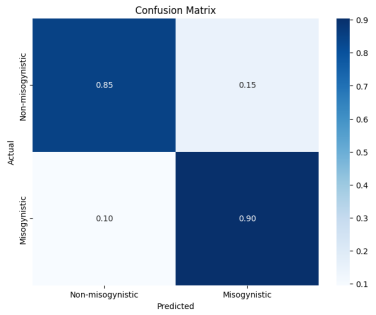

Figure 7: Confusion matrix Spanish test set normalized

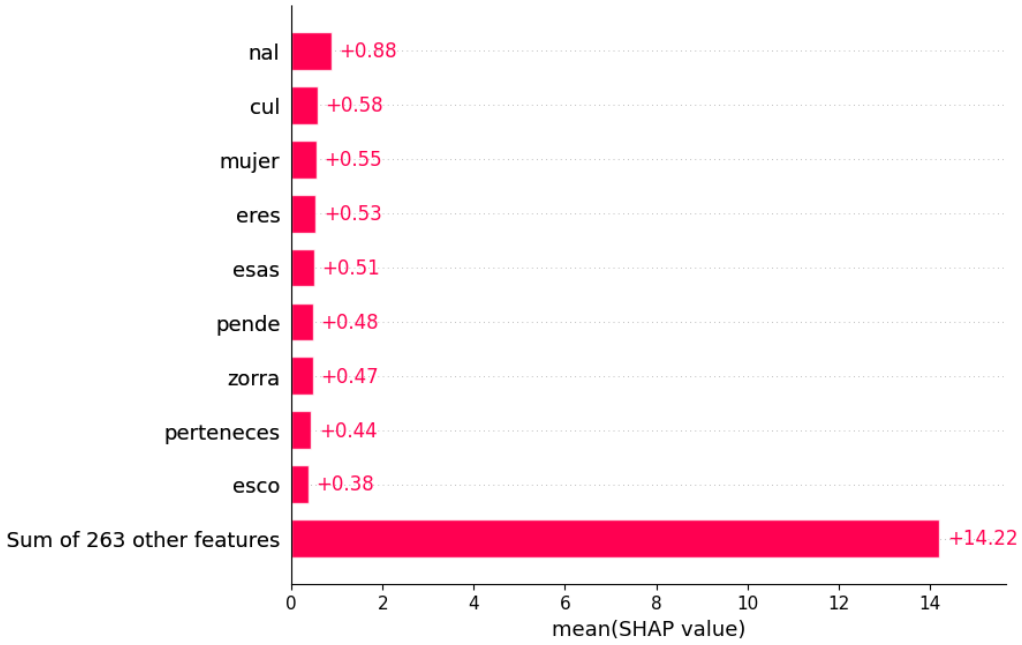

Figure 8: SHAP explanation

4 Conclusion

In the context of this project, the use of SHAP (SHapley Additive exPlanations) requires significantly more resources and time than a simpler model, such as word distribution analysis. However, the validity of this approach lies primarily in the fact that the global explanation is obtained by aggregating local explanations and that each of these local explanations assigns each token an importance score based on its role in the sequence with respect to the model decision. The great success of

the Transformer architecture in natural language processing is precisely given by giving very strong attention to the context and relationships between the tokens in the sequences, the same word therefore can have different representations, different meanings and different weights in the model decision in different sequences. SHAP (SHapley Additive exPlanations) take this into account, and because of this, the explanations provided allow investigating the lexicon related to misogynistic discourses from a deeper meaning perspective.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[2] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017.

[3] E. Guest, B. Vidgen, A. Mittos, N. Sastry, G. Tyson, and H. Margetts, "An expert annotated dataset for the detection of online misogyny," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (P. Merlo, J. Tiedemann, and R. Tsarfaty, eds.), (Online), pp. 1336–1350, Association for Computational Linguistics, Apr. 2021.

[4] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020.

1

[1]I declare that this material, which I/We now submit for assessment, is entirely my/our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my/our work. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.