Multi-class sexist language detection

Gogu Razvan-Costinel University of Bucharest
Sasu Alexandru-Cristian University of Bucharest
Dina Andrei-Constantin University of Bucharest
Rosca Alexandru University of Bucharest

Abstract

In this paper we tackle the problem of classifying text from social media networks, web articles, and books, into one of the following sexist or non-sexist subcategories: sexist direct, sexist descriptive, sexist reporting, non-sexist offensive, non-sexist non-offensive. We trained, tested, and compared a multitude of models, both from the sphere of Machine Learning, as well as Deep Learning. The best results were obtained using a fine-tuned version of a Romanian BERT (RoBERT) which features balanced weights.

1 Introduction

Sexism has a profound and detrimental impact on individuals and society as a whole. By perpetuating harmful stereotypes, devaluing and objectifying women, and reinforcing gender inequality, offensive texts in social media contribute to a culture of sexism and discrimination. These behaviours have serious consequences on various levels, influencing people's mental health and well-being.

In this paper we will be talking about how we tackled the problem of detecting sexism in a text message. We trained several models using an annotated corpus of Romanian sexist and offensive tweets (Hoefels and Mădroane, 2022). Each model was tested on a dataset made of texts from social media networks, web articles, and books, and the results were determined using the weighted accuracy metric. Our results showed that fine-tuning a BERT pretrained on romanian language with weight-loss outperformed any other classic NLP method.

2 Related work

The field of sexist language detection within the domain of Natural Language Processing (NLP) is increasingly garnering attention due to its potential for mitigating online harassment and promoting healthier digital communication environments. Multi-class sexist language detection represents a captivating and relatively uncharted niche within this broader field.

Previous studies in this area have primarily focused on both binary and multiclass classifications, such as distinguishing between sexist and non-sexist language or classifying sexist language into different categories. However, there is a gap in our understanding of the complex manifestations of sexist language within the context of studies conducted in the Romanian language.

In this context, several works have been pivotal in shaping our methodology and approach. Aristotle's "Rhetoric" (Aristotle, 1991) lays the groundwork for understanding the dynamics and patterns of debates. Moreover, research in the NLP field, particularly argumentation mining (Stede and Schneider, 2018), aids in identifying and understanding the subtleties inherent in sexist language.

Walton's work (Walton, 1998) provides an exhaustive view of ad hominem attacks, a category that often includes sexist language. His research has been instrumental in understanding the different subtypes of ad hominem attacks and has influenced our methodological approach.

Habernal et al.'s work (Habernal et al., 2018) on detecting ad hominem attacks, including sexist language within online platforms and social media, has been particularly informative. Their exploration of methodologies involving a two-stacked bi-directional LSTM network and a convolutional neural network establishes a valuable baseline for our work.

Additionally, online platforms such as Reddit have served as source platforms for data in related works. For instance, (Saleem et al., 2016) detected hateful speech on Reddit by exploiting specific sub-communities to automatically obtain training data. In a different study, (Zhang et al., 2017) proposed a set of nine comment-level dialogue act categories and annotated threads, constructing a CRF classifier for dialogue act labeling. Moreover, (Tan et al., 2016) examined persuasion strategies on Change My View using word overlap features, unlike our work, they focused solely on successful strategies with delta-awarded posts. (Musi, 2017) utilized the same dataset to study concession in argumentation. This highlights the versatility and potential of online platforms as data sources for sexist language detection and related fields.

While these foundational studies provide valuable

insights into the detection and moderation of sexist language, our work aims to contribute to this field by offering new perspectives within a multi-class framework. Leveraging the advancements in computational linguistics, our study seeks to enhance the detection and moderation of sexist language within the digital space.

3 Methods

3.1 Dataset

The dataset that we are working with is composed of a training dataset and a testing dataset. The training dataset contains 39,007 text corpora from CoRoSeOf: An annotated Corpus of Romanian Sexist and Offensive Language (Hoefels and Mădroane, 2022), while the test dataset contains 3,130 texts made by the organizing team of the NitroNLP competition (Dana Dăscălescu, 2023).

The data was divided using a two-level labelling scheme. The first level consisted of dividing the content into the 'sexist' category and 'non sexist' category. Furthermore, the sexist content was divided in the 'sexist direct', 'sexist descriptive', and 'sexist reporting' subcategories, while the items in the non-sexist category were divided in 'non-offensive' and 'offensive' posts.

	Data
Sexist Direct	2,156
Sexist Descriptive	1,494
Sexist Reporting	219
Non-Sexist Offensive	4301
Non-Sexist Non-Offensive	30,837
Total Data	39,007

Table 1: Statistics of training dataset

The given classifications can be defined as bellow:

Sexist direct: The text includes sexist elements and it is directly addressed to a particular gender, usually women or groups of women.
Sexist descriptive: The text describes one or more people, usually a woman or women, in a sexist way, but not being addressed to them directly.
Sexist reporting: The text is reporting an act of sexism witnessed or heard from other sources.
Non-sexist offensive: The text has no sexist elements, but it contains offensive language.
Non-sexist non-offensive: there are no sexist or offensive elements in the text, nor any sexist or offensive connotations. Texts may contain sexist or offensive elements or hashtags, (providing little context or no context at all) but the overall message of the text is neither offensive nor sexist.

3.2 Data preparation

We started preparing the training data by excluding noisy parts from the dataset. These noisy parts consisted of samples that had in their composition a multitude of other samples, thus skewing with the training and testing processes of the models. Next, we removed usernames and links, as we thought they didn't contribute much to the process of classification.

Regarding the BERT model, we built a custom dataset and dataloader. For the model, we also used a standard tokenizer, which we customized so that it replaces emojis with special tokens. Through the dataloader, we grouped the training and validation data into batches of 16 samples and fed these batches to the model. The data is stored within the custom dataset through the use of tensors.

3.3 Decision Tree, KNN, MLP

At first, we began with a Decision Tree model with a max depth of 5, accompanied by a Bag-of-Words model. The BoW model used as preprocessing and tokenization techniques, the following: converting text to lower case, removing punctuation, removing links and references, removing emojis, splitting texts into words by white spaces, removing stop words, and applying stemming. Moreover, so as not to confuse and make it difficult for the ML model to train on and classify data, we limited the max number of features retained by the BoW model to 1000.

Afterwards, we used a KNN model with 3 neighbors, followed by an MLP model using the ReLU activation function and the Adam optimizer, running for a maximum of 300 iterations. These models, like the Decision Tree model, were used alongsied the previously defined BoW model.

3.4 Fine-tuning a BERT model

Next we tried using a BERT transformer. The standard BERT wouldn't give a good result, being made for the english language, so we had to use a more customized model. We used the bert-base-romanian-cased model (Dumitrescu et al., 2020). This model is a multilingual BERT which has been pre-trained on several text corpora. The text was extracted from the romanian text found in the OPUS and OSCAR corpora (Abadji et al., 2021; Tiedemann and Thottingal, 2020) and in the romanian Wikipedia.
We fine-tuned the BERT model by adding a classification

layer with a size of 5. We used an Adam optimizer with a learning rate of 3×10^{-6}, and for calculating the loss we used the CrossEntropyLoss function. For optimal results the model was trained for 3 epochs.
After that we tried optimizing the model for the imbalanced dataset, so we added weights to the loss function. The values for the weights were selected inverse to how much the respective class is represented in the dataset.

Class	Weight
Sexist Direct	14
Sexist Descriptive	20
Sexist Reporting	80
Non-Sexist Offensive	10
Non-Sexist Non-Offensive	1.5

Table 2: Final weights for each class

4 Results

In order to determine the performance of our proposed models on this multinomial classification problem, we considered the weighted accuracy metric.

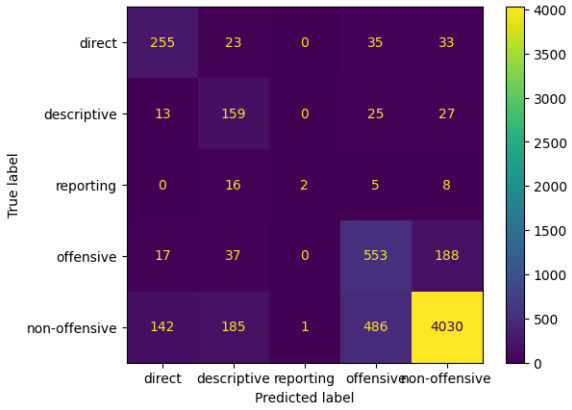
Model	Accuracy
Decision Tree	0.31
KNN	0.29
MLP	0.34
BERT	0.36
BERT with weight-loss	**0.43**

Table 3: Weighted accuracy of proposed models

In Table 3 we can draw several conclusions.

First is that the fine-tuned RoBERT model clearly outperformed the other models. Using a transformer pretrained on the respective language is significantly more optimal than using a Bag-of-Words model.

The other observation is how balancing the dataset offers a great improvement, reducing the bias of the more represented classes over the sparcer ones.



Observing the confusion matrix it can be noted how the model is not very good at detecting sexism reporting, which can be due to the class not being very well represented or the class being hard to detect.

5 Conclusions

We trained and tested a diversity of models in order to solve the problem of classifying texts into sexist or non-sexist texts. Out of all the models used, the BERT model with the weight-loss property specialized on romanian texts gave the best weighted accuracy score. We consider that for future works, the same train dataset could be used, but there could be removed the oversampling of non-sexist non-offensive texts.

References

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event), pages 1 – 9, Mannheim. Leibniz-Institut für Deutsche Sprache.

Aristotle. 1991. Aristotle and george kennedy (translator). In *On Rhetoric: A Theory of Civil Discourse.* Oxford University Press, USA.

Lucian Istrati@FMI Zavelca Miruna-Andreea Dana Dăscălescu, Livia Magureanu. 2023. Nitro language processing - 2nd edition - sexism.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Pa-*

pers), page 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

Çöltekin Çağrı Hoefels, Diana Constantina and Irina Diana Mădroane. 2022. Coroseof - an annotated corpus of romanian sexist and offensive tweets. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2269–2281, Marseille, France. European Language Resources Association.

Elena Musi. 2017. How did you change my view? a corpus-based study of concessions' argumentative role. In *Discourse Studies*, page 357–366.

Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. 2016. A web of hate: Tackling hateful speech in online social spaces. in guy de pauw, ben verhoeven, bart desmet, and els lefever, editors. In *Proceedings of the First Workshop on Text Analytics for Cybersecurity and Online Safety*, pages 1–9, Portoroz, Slovenia. European Language Resources Association(ELRA).

Manfeld Stede and Jodi Schneider. 2018. In *Argumentation Mining*, San Rafael, United States.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, Montreal, CA.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Douglas Walton. 1998. Douglas walton. 1998. In *Ad Hominem Arguments (Studies in Rhetoric & Communication)*. The University of Alabama Press, Alabama.

Amy Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017). AAAI Press*, page 357–366, Montreal, Canada.