Unmasking Misogyny: Deep Learning for Detecting Gender Bias Online

Albias Havolli, Théodore Maradan, Armance Nouvel

Group 18

Problem definition

- Hate speech creates a toxic debate environment, discouraging participation and undermining freedom of speech as people are less likely to share their views when faced with hostility.
- The European Union Agency for Fundamental Rights (FRA) notices that current content moderation algorithms fail to accurately identify misogynistic content.
- This work aims at developing novel algorithms based on recent advances in Natural Language Processing (NLP), capable of capturing misogyny in online posts.

Key Related Works

- In a report published in 2023, the FRA provides a comprehensive analysis of the hate speech situation on online platforms [1]. The authors highlight the prevalence of misogyny on different online platforms.
- Recent advances in NLP such as Transformer models show a great potential in capturing the meaning of a message. Therefore, BERT [2] and RoBERTa [3] are chosen in this work for their ability to capture the context of a text.
- Other studies achieved a great performance using these two models to detect hate speech. However, the researchers did not display the performance achieved on misogyny.

Method

- The models described above are trained under various conditions described below after a pre-processing of the input texts. These set-ups aim at increasing the performance of the models and address some of the issues inherent to the datasets.
- Because of the nature of the study, the datasets used here are inevitably imbalanced. Only a minority of online posts includes misogynistic content. A weighted random sampler is used to correct for the initial imbalance of the dataset.
- The dataset is augmented through paraphrasing to increase the number of available misogynistic posts. This procedure also allows a correction for the small size of the dataset.
- The models are pre-trained on a dataset having a structure similar to the main dataset (see below). Pre-training generally leads to a higher generalization ability of the model.

Datasets

- The models presented above are trained on a dataset developed by Guest et al. (2021) and annotated by trained experts. The latter collects threads from several subreddits, hence guaranteeing variety in the language used and the availability of the context.
- Some of the experiments involved a pre-training on a dataset created by Kirk et al. (2023) regrouping posts from Reddit and Gab. The texts are classified as either sexist or non-sexist.

Validation

- Fig. 1 shows the results obtained under the different experimental set-ups. The results focus on precision and recall as they are the principal metrics of interest for online content moderation. An efficient content moderation algorithm should reach high precision and high recall. This situation would translate into the non-misogynistic posts to be allowed through and the misogynistic posts to be stopped.
- The Pareto frontier encapsulates the optimum (non-dominated) set-ups. The distance between this line and the baselines highlights the enhanced performance brought by the different techniques discussed earlier.

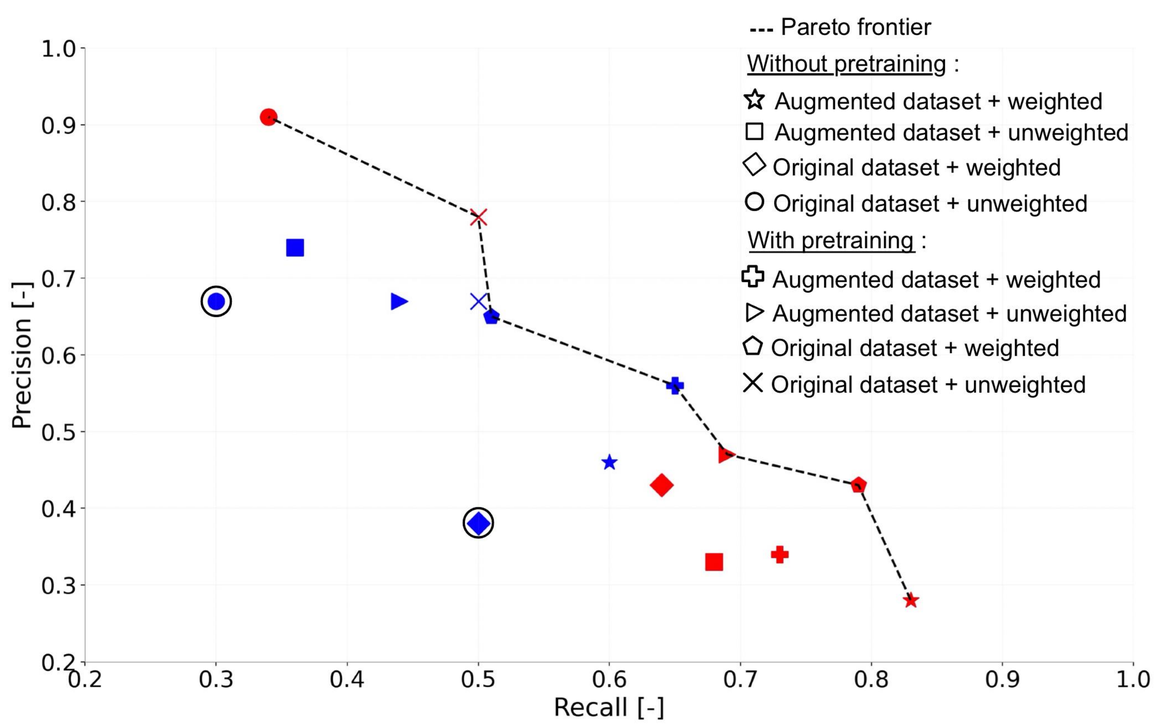

Figure 1: Precision vs. Recall under various experimental set-ups. Results obtained with BERT or RoBERTa are presented in blue or red respectively. Baselines are encircled.

Limitations

- While recall and precision metrics offer insights into model performance in identifying misogynistic posts, they may not fully capture the complexities of detecting hate speech and misogynistic content on social media platforms.
- Because of time constraints, the hyperparameters of both models were not optimized. A higher performance can be expected after such a procedure.

Conclusion

- Two different models (BERT and RoBERTa) were investigated to address the issue of the prevalence of misogyny on online platforms.
- The experimental set-ups (weighted random sampler, paraphrasing, pre-training) studied in this work bring a significant increase in performance compared to our baseline.
- Some limitations to this study were identified and could be addressed in further work.

References

[1] European Union Agency for Fundamental Rights, “Online Content Moderation – Current challenges in detecting hate speech,” Publications Office of the European Union, Luxembourg, 2023.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” CoRR, vol. abs/1810.04805, 2018.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A robustly optimized BERT pretraining approach,” CoRR, vol. abs/1907.11692, 2019.