Deep Learning Representations in Automatic Misogyny Identification: What Do We Gain and What Do We Miss?

Elisabetta Fersini, Luca Rosato, Antonio Candelieri, Francesco Archetti, Enza Messina
University of Milano-Bicocca, Milan, Italy
{elisabetta.fersini, antonio.candelieri}@unimib.it
{francesco.archetti, enza.messina}@unimib.it,
l.rosato1@campus.unimib.it

Abstract

In this paper, we address the problem of automatic misogyny identification focusing on understanding the representation capabilities of widely adopted embeddings and addressing the problem of unintended bias. The proposed framework, grounded on Sentence Embeddings and Multi-Objective Bayesian Optimization, has been validated on an Italian dataset. We highlight capabilities and weaknesses related to the use of pre-trained language, as well as the contribution of Bayesian Optimization for mitigating the problem of biased predictions.

1 Introduction

Nowadays, although women, girls and teenagers have a strong presence in online social environments, they are strongly exposed to hateful comments. In 2021, a survey provided by the Pew Research Center has shown that females are targeted for severe types of online gender-based attacks[1]: women are more likely than men to report having been sexually harassed online (16% vs. 5%) or stalked (13% vs. 9%). These phenomena can be found under the umbrella of online misogyny, which can be generally defined as hate, violence or prejudice against women (Ging and Siapera, 2018).

2 State of the Art

In order to counter online misogyny, several computational approaches have been presented in the literature ranging from natural language processing models to machine learning classifiers, denoting quite promising recognition performance. The earliest investigation about computational models for automatic misogyny identification has been presented in Anzovino et al. (2018), where the authors proposed the adoption of several linguistic cues and baseline classifiers for addressing three main problems, i.e., misogyny identification, misogynistic behaviour recognition and target classification. After this seminal paper, several approaches have been presented in the literature distinguishing them according to the feature representations that have considered for representing the textual contents and the machine learning models adopted as classifiers. Most of the approaches experimented a high-level representation of the word and/or sentence (García-Díaz et al., 2021; Pamungkas et al., 2020; Farrell et al., 2020; Lees et al., 2020), coupled with fine-tuning, while few of them adopted shallow models or trained deep architectures from scratch (Fabrizi, 2020; Ou and Li, 2020; da Silva and Roman, 2020; El Abassi and Nisioi, 2020; Koufakou et al., 2020).

Recently, an increasing interest has been focused on the problem of unintended bias (Dixon et al., 2018). In particular, it is important to focus on a given error induced by the training data, i.e., the bias injected in the model by a set of *identity terms* that are frequently associated to the misogynous class. For example, the term *women*, if frequently used in misogynous messages, would lead most of the supervised classification models to over-generalization and to disproportionately associate this identity term to the misogynous label. To this purpose, only few approaches have been dedicated to the unintended bias problem for misogyny identification (Nozza et al., 2019; Lees et al., 2020; Gencoglu, 2020; Zueva et al., 2020), denoting a research panorama that is in its infancy. Although

[1] https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

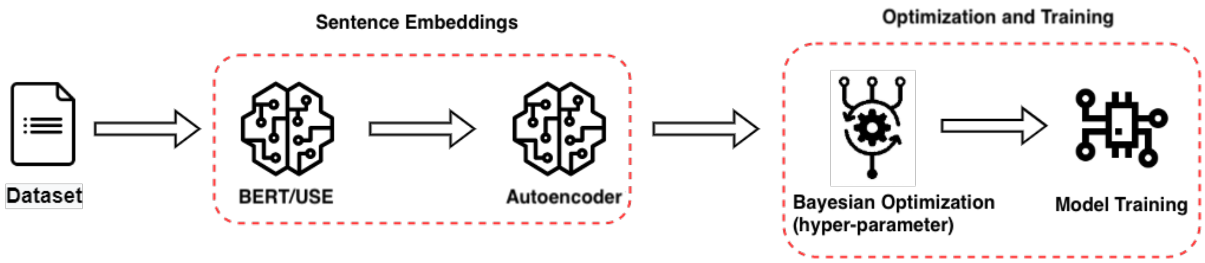

Sentence Embeddings
Optimization and Training
Dataset
BERT/USE
Autoencoder
Bayesian Optimization (hyper-parameter)
Model Training

Figure 1: Workflow of the proposed investigation.

the above mentioned approaches represent a fundamental contribution to the problem of automatic misogyny identification in online social environments, they do not focus on two main research questions:
(RQ1) Do embeddings always success when representing different misogyny related problem such as the type of misogyny and the target?
(RQ2) Could classification models be constrained to be less biased by the optimization of their hyper-parameters, therefore having good generalization capabilities also on uncommon expressions?

In this paper, we address the above mentioned open issues by the following main contributions:

- we perform an analysis of capabilities and weaknesses of the widely used state-of-the-art sentence encoders USE and BERT when adopted for misogyny detection;

- we investigate how to reduce the bias of the models by optimizing their hyper-parameters through a multi-objective bayesian optimization strategy.

3 Proposed Framework

In order to address the above mentioned research questions, related to the understanding of weaknesses and capabilities of pre-trained language models for misogyny identification and the reduction of unintended bias, we introduce the framework reported in Figure 1.

3.1 Sentence Embeddings

The proposed approach uses two pre-trained language models to generate a contextual representation of the data. The considered models are based on the *transformer* architecture initially presented in Vaswani et al. (2017). More specifically, the first model is the "small" version of **BERT**, uncased, consisting of 12 stacked *encoders*, 12 parallel *self-attention* and 768 units to represents text. The model is pre-trained on 102 languages, has a dictionary of 110.000 terms and provides a 768-dimensional representation of the text as output. The second model is the multi-language version of **USE** trained on 16 languages, which consists of 6 *stacked encoders*, 8 parallel *self-attention* and 512 units for the text representation. USE provides a 512-dimensional representation of the text, computed as the average over the last encoder's embeddings of each token. The pre-trained BERT and USE models have been *fine-tuned* according to the available misogyny related labels. In order to reduce the dimension of the vector representation given by the fine-tuned pre-trained models and to introduce sparsity to improve the separability of the data, an Autoencoder is used as suggested in (Glorot et al., 2011). The architecture of the Autoencoder is reported in Figura 2.

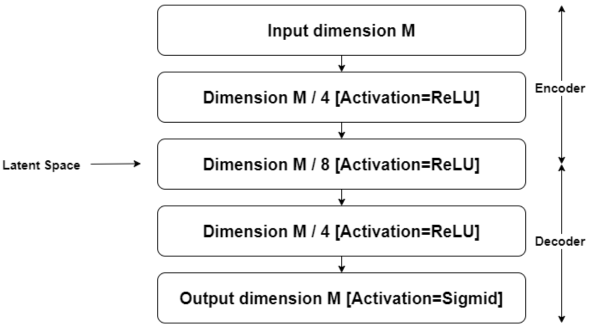

Input dimension M
Dimension M / 4 [Activation=ReLU]
Dimension M / 8 [Activation=ReLU]
Dimension M / 4 [Activation=ReLU]
Output dimension M [Activation=Sigmid]
Latent Space
Encoder
Decoder

Figure 2: Autoencoder adopted to map the original data in a more compact and sparse representation.

3.2 Training and Optimization

Once the latent representation of a sentence is obtained by means of the Autoencoder, any machine learning model could be adopted to recognize misogynous contents. To this purpose, Support Vector Machines (SVM) have been used, searching for their optimal hyper-parameter settings that are able to ensure the highest recognition performance. However, searching for hyper-parameters that maximize a specific performance metric is a computational expensive black-box optimization process. Due its sample efficiency, Bayesian Optimization (BO), has been adopted. BO works sequentially: each classifier's hyper-parameters to evaluate is chosen by dealing with the exploitation-exploration dilemma. To do this, BO relies on two key components: a *probabilistic surrogate model* approximating the performance metric to optimize - depending on SVM classifiers evaluated so far - and an *acquisition function* (*utility function* suggesting the choice of the next SVM's hyper-parameters to evaluate. The adoption of a probabilistic surrogate model, specifically a Gaussian Process (GP) in this study, allows to estimate the expected value of the performance metric (i.e., GP's predictive mean) and the associated uncertainty (i.e., GP's predictive standard deviation), for any given SVM's hyperparaters configuration. These two estimates are combined into the acquisition function, which implements the exploitation-exploration trade-off mechanism, where exploitation and exploration are associated to the surrogate's predictive mean and standard deviation, respectively. More formally, let $\mathcal{D}_{1:n} = \left\{\left(h^{(i)}, v^{(i)}\right)\right\}_{i=1,...,n}$ be the set of n possible configuration, where $h^{(i)}$ is a d-dimensional vector whose component $h_j^{(i)} \in H_j$ is the value of the j-th hyperparameter of the i-th SVM classifier, and $v^{(i)}$ is the associated value of the target performance measure. The overall search space H is usually a subspace of the cartesian product of the hyper-parameters's ranges: $H \subseteq H_1 \times ... \times H_j \times ... \times H_d$. In this study the search space H is spanned by $d = 2$ hyper-parameters whose values can vary into the following ranges:

- $h_1 \in H_1 := [10^{-1}, 10^5]$, that is the *regularization* hyperparameter C of the SVM classifier (i.e., soft margin SVM)
- $h_2 \in H_2 := [10^{-5}, 10^1]$, that is the hyperparameter γ of the Radial Basis Function kernel of the SVM classifier (i.e., $k(x, x') = e^{-\gamma||x-x'||^2}$)

In this study we consider two different cases (on stratified 10-fold cross-validation):

- tuning the SVM classifier's hyper-parameters to maximize the accuracy, irrespectively to any measure of bias;
- tuning the SVM classifier's hyper-parameters to optimize an objective function aimed at maximizing accuracy and minimizing a *bias-related metric*.

Measuring the Bias In this paper, we measure the model bias by referring to the specific definition of *unintended bias* presented in (Dixon et al., 2018):

> A model contains unintended bias if it performs better for comments containing some particular identity terms than for comments containing others.

In order to measure the level of unintended bias of a given model, *identity terms* (terms related to the *woman* concept) and templates (pre-defined skeleton used to create synthetic samples) are used to generate sentences referred to women, which however can be unreasonably classified as misogynous with high scores. To this purpose, identity terms and templates available for the AMI at Evalita 2020 challenge (Fersini et al., 2020) have been used. Identity terms have been listed using a set of 37 concepts related to "woman", considering both their singular and plural form for the Italian language. Since unintended bias of identity terms cannot be measured on the original dataset set due to class imbalance and highly different identity term contexts, a *synthetic dataset* is generated by means of templates. Following (Nozza et al., 2019), we defined several templates that are filled out with identity terms and with verbs and adjectives that are divided into negative (e.g. hate, inferior) or positive (e.g. love, awesome) forms to convey misogyny or not. Table 1 reports examples of templates. The generated synthetic dataset comprises 3,923 instances, of which 50% misogynous and 50% non-misogynous, where each identity term appears in the same contexts.

Template Examples	Label
<identity_term>devono essere protette	Non-Misogynous
<identity_term>devono essere torturate	Misogynous
adorare <identity_term>	Non-Misogynous
umiliare <identity_term>	Misogynous
<identity_term>stimabile	Non-Misogynous
<identity_term>rivoltante	Misogynous

Table 1: Template examples.

Identity terms, templates and synthetic dataset are available at https://github.com/MIND-Lab/ItalianBias.

In order to evaluate the performance of the classification in terms of bias, an AUC-related measure has been used (AUC_{final}). In what follow, the higher is the AUC_{final}, the lower is the bias of the model. In particular, a weighted combination of AUC estimated on the raw dataset AUC_{raw} (original tweets) and three per-term AUC-based scores computed on the synthetic dataset ($AUC_{Subgroup}$, AUC_{BPSN}, AUC_{BNSP}) is adopted (Borkan et al., 2019). Let s be an identity-term (e.g. "donna" and "moglie") and N be the total number of identity-terms, the AUC_{final} is defined as:

$$\begin{aligned} AUC_{final} \; &= \tfrac{1}{2}AUC_{raw} + \\ &+\tfrac{1}{2N}\Big[\textstyle\sum_s AUC_{subgroup}(s) \\ &+\textstyle\sum_s AUC_{BPSN}(s) \\ &+\textstyle\sum_s AUC_{BNSP}(s)\Big] \end{aligned} \tag{1}$$

where:

- $AUC_{Subgroup}(s)$: computes AUC only on the data within the subgroup containing a given identity term s. This represents model understanding and separability within the subgroup itself. A low value means that the model does not distinguish properly misogynous and non-misogynous comments containing a give identity term s.

- $AUC_{BPSN}(s)$: Background Positive Subgroup Negative (BPSN) estimates AUC on the misogynous examples using the background and the non-misogynous examples belonging the subgroup. A low value means that the model mislead non-misogynous examples that mention the identity-term with misogynous examples that do not, likely meaning that the model predicts higher misogynous scores than it should for non-misogynous examples mentioning the identity-term.

- $AUC_{BNSP}(s)$: Background Negative Subgroup Positive (BNSP) calculates AUC on the non-misogynous examples from the background and the misogynous examples from the subgroup. A low value means that the model confuses misogynous examples that mention the identity with non-misogynous examples that do not, likely meaning that the model predicts lower misogynous scores than it should for misogynous examples mentioning the identity.

4 Experimental Investigation

In this section we report the experimental investigation performed on the Italian version of the Automatic Misogyny Detection (AMI) dataset (Fersini et al., 2020), comparing the results obtained with the proposed framework with the ones obtained by the baseline model (i.e. SVM trained on a TF-IDF representation). The AMI dataset is composed of 5,000 tweets, labelled according to "misogyny" (i.e., indicating if a Tweet is misogynous or not), "misogyny category" (i.e., Stereotype&Objectification, Dominance, Derailing, Sexual Harassment&Threats of Violence, Discredit) and "target" (i.e., individual or generic).

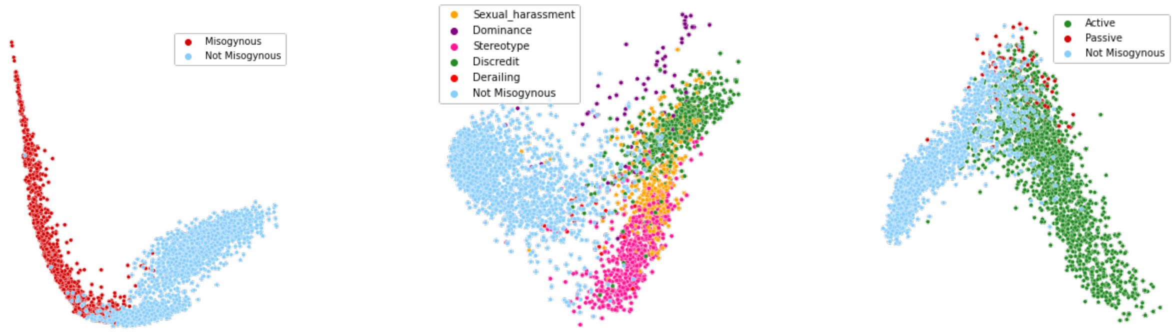
Regarding the first research question **(RQ1)**, we tuned the SVM classifier's hyper-parameters to maximize only the performance measure related to each label (i.e. Accuracy for misogyny labels, F-Measure for category and target labels). First of all, we reported in Table 2 the results comparing different models. It can be easily noted that, although BERT and USE allow SVM to achieve better performance than TFIDF, there is no difference between them achieving similar results. Moreover, while the recognition performance on the misogyny labels are satisfactory, the capabilities on discriminating the misogyny category and the target are still far from being acceptable.

In order to understand if the low performance can be due to the embedding, we investigated the class overlapping originated by USE and BERT. We report in Figure 3, a 2D representation of the embeddings obtained by USE (similar results have been obtained for BERT). We can immediately highlight that while the embeddings tuned for recognizing misogyny are quite distinguishable between misogynous and not misogynous tweets, for the category and target embeddings there is an overlapping among the classes. This makes the learned representations not ready for being used to recognize the specific form of misogyny and the subject of misogynous comments.

Regarding the second research question **(RQ2)**, we determined the SVM optimal hyper-parameters to maximize both $Accuracy$ and AUC_{final} (i.e. the bias-related metric). In order

	Baseline (TFIDF + Opt. SVM)	OUR (BERT + Opt. SVM)	OUR (USE + Opt. SVM)	Absolute Improvement
Misogyny [Accuracy]	0.8390	0.8670	0.8640	+2.8%
Misogyny Category [F-measure]	0.5427	0.5988	0.5991	+5.64%
Target [F-measure]	0.4217	0.4599	0.4537	+3.82%

Table 2: Performance comparison of different approaches. Underlined numbers denote the best result.



(a) Misogyny Embedding.

(b) Category Embedding.

(c) Target Embedding.

Figure 3: 2D embedding representation obtained by USE.

to guarantee the use different set of tokens for the hyper-parameter search and the inference phase, the synthetic samples have been split in training and testing. We compare in Table 3, the AUC_{final} values estimated on biased and unbiased SVM. We can easily note that the Unbiased SVM leads to maintain constant the Accuracy, but improve the generalization capabilities of the embeddings given by the AUC_{final} values, denoting a slightly better results for USE. This means that the SVM hyper-parameter optimization, with respect to both performance measures, leads to promising unbiased models. This ensures ensure good recognition capabilities on both common expressions (typically used on Twitter) and on uncommon comments (synthetic data). The obtained results also suggest that an SVM trained using the USE embedding is more keen to adapt the hyper-parameters to reduce its bias during training and inference.

		Biased SVM	Unbiased SVM
TFIDF	*Accuracy*	0.8390	0.8390
	AUC_{final}	0.6910	0.6950
BERT	*Accuracy*	0.8679	0.8679
	AUC_{final}	0.7197	0.7211
USE	*Accuracy*	0.8640	0.8640
	AUC_{final}	0.7181	**0.7430**

Table 3: Generalizaion capabilites on biased and unbiased models.

5 Conclusions and Future Work

In this paper we have investigated the capabilities and weaknesses of pre-trained language models, as well as the problem of the unintended bias when addressing the automatic misogyny identification for the Italian language. The proposed investigation has highlighted that, while pre-trained embeddings are able to distinguish misogynous and not misogynous comments, they still have poor discrimination capabilities related to the type of misogyny and its target. Regarding the unintended bias problem, it has been shown that an hyper-parameter search guided by Bayesian Optimization can lead to debiased models with good recognition generalization capabilities. As future work, we will investigate explainable AI techniques aimed at generating a feature score that is directly proportional to the feature's effect on inducing bias in the prediction model.

References

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum

Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 491–500.

Adriano dos S. R. da Silva and Norton T. Roman. 2020. No Place For Hate Speech @ AMI: Convolutional Neural Network and Word Embedding for the Identification of Misogyny in Italian. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Bologna, Italy. CEUR.org.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73. ACM.

Samer El Abassi and Sergiu Nisioi. 2020. MDD@AMI: Vanilla Classifiers for Misogyny Identification. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Bologna, Italy. CEUR.org.

Samuel Fabrizi. 2020. fabsam @ AMI: a Convolutional Neural Network approach. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Bologna, Italy. CEUR.org.

Tracie Farrell, Oscar Araque, Miriam Fernandez, and Harith Alani. 2020. On the use of Jargon and Word Embeddings to Explore Subculture within the Reddit's Manosphere. In *12th ACM Conference on Web Science*, pages 221–230.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. AMI@ EVALITA2020: Automatic Misogyny Identification. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*. CEUR.org.

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.

Oguzhan Gencoglu. 2020. Cyberbullying detection with fairness constraints. *IEEE Internet Computing*, 25(1):20–29.

Debbie Ging and Eugenia Siapera. 2018. Special issue on online misogyny. *Feminist Media Studies*, 18(4):515–524.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep Sparse Rectifier Neural Networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language. In *Proceedings of the fourth workshop on online abuse and harms*, pages 34–43.

Alyssa Lees, Jeffrey Sorensen, and Ian Kivlichan. 2020. Jigsaw@ AMI and HaSpeeDe2: Fine-Tuning a Pre-Trained Comment-Domain BERT Model. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Bologna, Italy. CEUR. org*.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended Bias in Misogyny Detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 149–155.

Xiaozhi Ou and Hongling Li. 2020. YNU_OXZ @ HaSpeeDe 2 and AMI : XLM-RoBERTa with Ordered Neurons LSTM for Classification Task at EVALITA 2020. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Bologna, Italy. CEUR.org.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Nadezhda Zueva, Madina Kabirova, and Pavel Kalaidin. 2020. Reducing Unintended Identity Bias in Russian Hate Speech Detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 65–69.