Vector Space Models for Automatic Misogyny Identification (Short Paper)

Amir Bakarov
National Research University Higher School of Economics, Moscow, Russia
amirbakarov at gmail.com

Abstract

English. The problem of hate speech and, especially, of misogynous language is one of the most crucial problems of contemporary Internet communities. Therefore, automatic detection of such language becomes one of the most actual natural language processing tasks. The most ubiquitous tools for resolving this task are based on vector space models of texts. In this paper we describe our system that exploits such tools and have shown the best performance on the Italian AMI task of EVALITA 2018.

Italiano. *Il problema dell'uso di discorsi che incitano l'odio, e specialmente dell'uso di linguaggio misogino, è uno dei problemi più cruciali delle comunità di internet al giorno d'oggi. Pertanto, il rilevamento automatico di tale linguaggio diventa uno degli obiettivi più attuali per l'elaborazione del linguaggio naturale. I sistemi più diffusi atti ad affrontare questo obiettivo sfruttano l'ipotesi distributiva. In questo articolo, descriviamo il sistema proposto basato su quest'ipotesi che hanno dimostrato le migliori performance nel task AMI di EVALITA 2018 nella lingua italiana.*

1 Introduction

As the Internet community and several online discussions grow, the number of manifestations of *hate speech* on open web resources also increases. Such type of speech (also called *abusive language* or *textual harassment*) could get different forms depending on its focus on the person's ethnicity, gender identity, religion, or sexual orientation. Probably, one of the most destructive forms of hate speech is the one that abuses a person's gender identity. Such form of hate speech is called *misogynous language* since misogyny is a specific case of hate whose targets are women. Misogyny on the Internet (*cybermisogyny*, or *online sexual harassment*) is one of the crucial problems of contemporary Internet communities, especially from the perspective of the societal impact of this phenomenon.

Thus, the problem of automatic misogyny identification could be considered as one of the most important branches of a hate speech detection task. The successful solution of this problem could lead to the significant limitation of the diffusion for the hate speech against women. The problem of automatic misogynous language detection got attention from the research community fairly recently, and the shared task on *automatic misogyny identification* held as a part of the EVALITA-2018 campaign is one of the first works trying to deal with this problem (Fersini et al., 2018b). The aim of this task is to automatically identify misogynous content in tweets for the Italian and English languages.

This paper describes our system that has outperformed all other systems for the Italian language and also has shown fairly good results for the English language. This system is based on using semantic features of tweets as an input of a supervised classifier. The semantic features are considered as latent vectors produced by a vector space model.

Our work is organized as follows. Section 2 briefly describes related work on the proposed task. Section 3 describes the setup of our system, while Section 4 discusses the results and proposes an analysis of them. Section 5 concludes the paper.

	Task A (Italian)	Task B (Italian)	Task A (English)	Task B (English)
Baseline	0.830	0.487	0.605	0.370
TFIDF+LR	0.842	0.443	**0.649**	0.241
TFIDF+XGB	0.836	**0.493**	0.604	**0.309**
TFIDF+SVD+LR	**0.844**	0.478	0.628	0.275
TFIDF+SVD+XGB	0.833	0.463	0.605	0.254

Table 1: Performance of each of the compared vectorizers and supervised classifiers on each of the tasks. Task A reports accuracy, Task B reports macro F1-measure.

2 Related Work

The first notorious works of the task of automatic misogyny identification were described as shared task proposed at IverEval 2018 workshop (Fersini et al., 2018a) (a shared task organized jointly with SEPLN-2018 Conference for Iberian languages), and SemEval-2019[1]. These tasks proposed certain baselines based on ubiquitous text classification techniques (for example, SVM). The automatic misogyny identification task considered in our research is the third shared task on this topic (Anzovino et al., 2018). We are also aware of certain other attempts to computationally resolve the task of automatic misogyny identification, but most of them were published only as some exploratory analysis (Hewitt et al., 2016). Most of the state-of-the art approaches to this problem were described as system reports for the aforementioned IberEval-2018 shared task. As far as we know, there were no other scholarly works trying to resolve or to formalize this task.

In the natural language processing community very similar tasks were also considered in other hate speech online challenges and scholarly works (Davidson et al., 2017). An extensive overview of all the research related to hate speech detection goes beyond the scope of this work, and an interested reader could be referred to a survey paper specialized on this topic (Schmidt and Wiegand, 2017).

Apart from computational linguistics and natural language processing, the problem of misogynous speech was also a focus of some linguistic and social science articles (Fulper et al., 2014). Most of such scholarly works were trying to understand the nature of misogynous hate speech and patterns appearing in this type of language (Poland, 2016). We think that from the perspective of natural language processing, such papers could be useful for the systems that are highly grounded to linguistic knowledge and manually crafted resources.

3 Experimental Setup

In the shared task we had two datasets (for English and for Italian) of 5000 tweets each. 4000 tweets in each dataset were considered as a training sample, and the evaluation of the system was done on 1000 tweets (their labels were hidden until the end of the competition). The classification task has included both binary and multi-label classification.

In our work we have used vectors from term-document matrix with TF-IDF values. We propose the text classification based on using semantic features obtained from vector space models of texts. We considered the terms as word n-grams, and used a factorization of the term-document matrix (we used a method of singular value decomposition, SVD) and a normalization of factorized values (in the table with the results we call it **TFIDF+SVD**). From this perspective, our approach is very close to the method of Latent Semantic Analysis (Landauer et al., 1998) (and we have also tried to resolve this task using not-factorized TF-IDF matrix, called **TFIDF** in the table). As a supervised classifier we have used a Logistic Regression classifier, therefore, our system is based on using TF-IDF n-gram word features and a Logistic Regression (LR).

For all the methods of vectorization we used a basic pipeline of text pre-processing (tokenization, lemmatization and stop-word removal based on NLTK build-in tools and resources).

We have also compared it with other classifiers (for instance, a Gradient Boosting classifier,

[1] `https://competitions.codalab.org/competitions/19935`

XGB in the table) and got worse results on the certain tasks. All in all, we have compared four different models. The exact hyperparameters of the models used in our system, and all the code for reproducing the experiments could be found at our Gitlab repository: `https://gitlab.com/bakarov/ami-evalita`.

4 Results and Discussion

The system evaluation was done on two subtasks. The first subtask had proposed a binary classification to identify whether the text is either misogynous or not misogynous (Task A). The second subtask (Task B) was to classify the misogynous tweets according to both the misogynistic behavior (multi-label classification) and the target of the message (binary-classification). The results of the system for the English and Italian subtasks for the misogyny identification task are described in Table 1. It is notable that our system has outperformed the baseline put by organizers in most of the cases, and different combinations of vectorizers and models have shown different performance in different tasks.

After an error analysis conducted on the system, we have found out that the system fails on examples where misogyny is expressed without (or with a very little use of) offensive lexis, or, vice versa, such lexis is used not in misogynous context (for example, *you pussy boy*). This could be explained by the fact that the system is too much focused on the lexicon and does not takes into account syntactic patterns or thematic roles.

5 Conclusions

The proposed work has described the system that has shown the best results for the Italian track on all the subtasks (and have also got fairly good results on English). Our system is based on a vector space model of character n-grams and a supervised gradient boosting classifier.

The system described in this paper is one of the first attempts to the problem of detecting misogynistic language for the Italian language in the natural language processing community. We think that the description of the implementation of our system could help other researchers to resolve such important and actual task. We consider this value as a main contribution of our research.

In future we plan to give more attention to some other linguistic features based on analysis of patterns that people tend to use in misogynous language. We would also like to try out more promising approaches to text classification based on deep learning (for example, convolutional neural networks).

References

Anzovino, M., Fersini, E., and Rosso, P. (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.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009.*

Fersini, E., Anzovino, M., and Rosso, P. (2018a). Overview of the task on automatic misogyny identification at ibereval. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018). CEUR Workshop Proceedings. CEUR-WS. org, Seville, Spain.*

Fersini, E., Nozza, D., and Rosso, P. (2018b). Overview of the evalita 2018 task on automatic misogyny identification (ami). In Caselli, T., Novielli, N., Patti, V., and Rosso, P., editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Fulper, R., Ciampaglia, G. L., Ferrara, E., Ahn, Y., Flammini, A., Menczer, F., Lewis, B., and Rowe, K. (2014). Misogynistic language on twitter and sexual violence. In *Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM).*

Hewitt, S., Tiropanis, T., and Bokhove, C. (2016). The problem of identifying misogynist language on twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, pages 333–335. ACM.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Poland, B. (2016). *Haters: Harassment, abuse, and violence online*. U of Nebraska Press.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.