No Place For Hate Speech @ AMI: Convolutional Neural Network and Word Embedding for the Identification of Misogyny in Italian

Adriano dos S.R. da Silva
Schoool of Arts, Sciences and Humanities – University of Sao Paulo
Sao Paulo - Brazil
adriano.santos.silva@usp.br

Norton T. Roman
Schoool of Arts, Sciences and Humanities
University of Sao Paulo
Sao Paulo - Brazil
norton@usp.br

Abstract

English. In this article, we describe two classification models (a Convolutional Neural Network and a Logistic Regression classifier), arranged according to three different strategies, submitted to subtask A of Automatic Misogyny Identification at EVALITA 2020. Results were very encouraging for detecting misogyny, even though aggressiveness was less accurate. Our second strategy, consisting of a Convolutional Neural Network and logistic regression to identify misogyny and aggressiveness, respectively, won the sixth place in the competition.

Italiano. *In questo articolo, descriviamo due modelli di classificazione (i.e., Convolutional Neural Network e Regressione Logistica), organizzati secondo tre diverse strategie, per il subtask A dello shared task Automatic Misogyny Identification a EVALITA 2020. I risultati sono stati molto incoraggianti nel rilevamento della misoginia, anche se l'aggressività viene riconosciuta con una precisione più basse. La nostra seconda strategia (Convolutional Neural Network per misoginia e Regressione Logistica per aggressività) ci ha permesso di ottenere il sesto posto nella competizione.*

1 Introduction

Hate speech is a problem that has been gaining space both in the media and in academic research. Political organizations have been working to combat this type of discourse. As is the case with the code of conduct[1] created by the European Union Commission, and signed by some of the main social networks, such as Facebook, YouTube, Twitter, which aims to monitor and remove this type of content within 24 hours of its disclosure.

The subject has even become a marketing problem, to the extent that recently several companies stopped advertising on Facebook[2], only to put some pressure at the network to have it remove this type of publication from the posts within it. Advertisers point, in this case, is that they do not want their brand to be linked to this type of speech.

Defined as "language which attacks or demeans a group based on race, ethnic origin, religion, gender, age, disability, or sexual orientation/gender identity" (Nobata et al., 2016), hate speech represents a problem that cannot be allowed to grow, under the risk of having it lead to more concrete actions, by some people, with truly undesired results.

When this hate speech is targeted specifically at women, it is called misogyny (Manne, 2017). The problem with misogyny is such an issue that it has already been related to real crime cases and cybercrimes (Fulper et al., 2014). In this case, correlations were found between rape cases and the amount of misogynous tweets per state in the United States.

Some academic work and several competitions have proposed some tasks to promote studies and advances in the area. Much of this work and data sets focus on English (Fortuna and Nunes, 2018) only, even though this is a widespread phenomenon that happens in any language.

It is extremely important, therefore, to encourage the development of this kind of study

[1] https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

[2] https://www.nytimes.com/2020/08/01/business/media/facebook-boycott.html

in different languages and competitions, such as IberEval (Fersini et al., 2018b), SemEval (Basile et al., 2019) and EVALITA (Fersini et al., 2018a), which have already proposed activities to identify misogynous discourse in Spanish, English, and Italian.

In this work, we help address this problem by testing two classification models as part of EVALITA 2020's subtask A on Automatic Misogyny Identification (AMI). Tested models were a Convolutional Neural Network (CNN) and a Logistic Regression (LR) classifier. Three different strategies were designed and tested, with one of them scoring 6th in the competition.

The rest of this article is organized as follows. Section 2 presents some related work in the identification of misogyny or hate speech. Section 3, in turn, gives an overview of EVALITA's AMI. Next, in section 4, we describe our experimental set-up, giving details of the implemented methods and tested strategies. Finally, in Section 5 we discuss our results, whereas in Section 6 we present our final remarks on this task.

2 Related Work

IberEval (Fersini et al., 2018b) proposed a task to identify misogynous discourse in tweets in English and Spanish. Several teams participated in this competition and the best team reached an accuracy of 0.91 and 0.81 for Spanish and English, respectively, with the use of an SVM as a classifier and with the addition of some lexical features to characterize the tweets.

SVMs were also proposed to identify racism in Twitter messages in English, achieving an F1 score of 0.76 (Hasanuzzaman et al., 2017). In SemEval 2019, a Convolutional Neural Network (CNN) performed competitively in the task of identifying hate speech against immigrants and women in English (Basile et al., 2019). The team that presented this architecture ranked fourth with an F1 score of 0.535.

During the Automatic Misogyny Identification shared task at EVALITA 2018, it was proposed a subtask A, which consisted of identifying misogyny (Fersini et al., 2018a; Anzovino et al., 2018). For this subtask, Logistic Regression was the model to deliver the best performance with an accuracy of 0.704 (Saha et al., 2018).

3 Subtask

The second edition of misogyny identification at EVALITA 2020 consists of two subtasks: A and B. The purpose of subtask A is to identify the presence or absence of misogyny and aggressiveness in tweets (Elisabetta Fersini, 2020), whereas subtask B checks whether the model is capable of distinguishing misogynous from non-misogynous content, also ensuring fairness (unintended bias) (Nozza et al., 2019).

The "No Place For Hate Speech" team participated only in subtask A, and all discussions that will be followed are related to this subtask. Within EVALITA 2020, the subtask consisted of identifying the presence or absence of misogynous speech and aggressiveness in tweets in Italian (Basile et al., 2020; Elisabetta Fersini, 2020).

The training dataset consisted of 5,000 tweets. The class that determines the presence or absence of misogyny is nearly balanced. However, aggressiveness is not balanced at all, with approximately 35% of tweets containing aggressiveness. Table 1 shows the distribution of each class in the training set.

Table 1: Distribution of Tweets in relation to each class of misogyny and aggressiveness

	Mis.	**Non Mis.**	**Aggr.**	**Non aggr.**
Total	2337	2663	1783	3217

4 Materials and Methods

In subtask A, we tested two different classifiers within different configurations. These were a Convolutional Neural Network (CNN), using BERT (Devlin et al., 2018) as its language model; and a Logistic Regression (LR) classifier, with L2 regularisation.

The LR classifier used a 4-gram language model, with tf-idf (Rajaraman and Ullman, 2011) normalization. Both models were developed in Python, with the aid of the TensorFlow[3] and Sklearn[4] libraries.

Since the subtask A at EVALITA allows each team to submit up to three classifiers, we decided to approach the problem according to three different strategies, involving different combinations of these classifiers, along with different subsets of data on which they should be trained.

[3] https://www.tensorflow.org/

[4] https://scikit-learn.org/stable/

In all cases, the training set was divided in a 90% subset, used for training purposes, with the remaining 10% used for out-of-sample testing. All classifiers used this same proportion both to identify misogyny and aggressiveness. Tweets were used in their raw form and no preprocessing was used.

All CNNs used in the experiments had the same configuration, being trained for 15 epochs. They also have three convolution layers, relu activation functions, and dropout rate of 0.10, with adam optimisation. Finally, cross-entropy was used as their loss function. In what follows, we will describe, with more details, each of the strategies followed during our tests.

4.1 Strategy 1

The first strategy consisted of training two CNNs, one for each specific sub-problem separately, *i.e.* one for misogyny and another for aggressiveness classification. In both cases, the entire data set was used for training.

At the testing stage, the CNNs were arranged as a pipeline, in which the first CNN was responsible for identifying whether a tweet had some misogynous content, whereas the second CNN was responsible for identifying the presence or absence of aggressiveness only in those tweets marked as misogynous by the first CNN.

4.2 Strategy 2

Similar to Strategy 1, the second strategy also consisted of training a CNN to detect misogynous content in tweets. This time, however, the classification of aggressiveness was left to a Linear Regression classifier. As in the first strategy, both models were trained in the entire data set.

During testing, once again models were arranged in a pipeline, with the CNN coming first, to detect misogyny in tweets. In the sequence, all tweets classified as misogynous by the CNN were then fed to the LR classifier, so it could determine the presence or absence of aggressiveness.

4.3 Strategy 3

Our third strategy is similar to Strategy 1, in that it also consists of two CNNs trained separately over the data set. The only difference, however, lies during the training stage. In this case, whereas the first CNN (*i.e.* the one responsible for misogyny identification) was trained using the entire data set, the second CNN (the one responsible for detecting aggressiveness) was trained only on those examples labeled as misogynous.

During testing the same set-up as in Strategy 1 was followed. As such, both CNNs were arranged in a pipeline, with the first one responsible for detecting misogynous tweets, and the second one responsible for identifying aggressiveness, amongst those tweets held misogynous by the first CNN.

5 Results and Discussion

Table 2 shows the performance of each tested strategy. As expected, the results for misogyny identification were the same over all strategies, since this subtask A was left to a CNN trained over the entire data set.

Table 2: Performance of each classifier strategy in terms of F1 score in the test set.

Classifier	Misogyny	Aggressiveness
Strategy 1	0.96	0.75
Strategy 2	0.96	0.70
Strategy 3	0.96	0.85

Results for aggressiveness detection, on the other hand, varied substantially, with the Logistic Regression classifier (Strategy 2) performing worst, when compared to the CNNs used for the same task in the other strategies (7% against Strategy 1, and 18% against Strategy 3).

Interestingly, the CNN trained only on examples labeled as misogynous (Strategy 3) performed better (around 13%) than its counterpart trained over the entire data set (Strategy 1). It is important to recall that this was the only difference between both strategies.

Final results at the competition's private test set can be seen in Table 3. As it turns out, Strategy 2 was the best ranked of our models, reaching the sixth place at the competition (being only $F = 0.03$ worse than the winning model).

Table 3: Official result of the subtask A in the evaluation set is calculated by averaging the F1 measures estimated for the Misogynous and Aggressiveness classes

Classifier	Average F1
Strategy 1	0.693
Strategy 2	0.716
Strategy 3	0.490

Puzzling enough, this was the model that scored

worse in our test set. One possible explanation for this fact might be that our CNN was not capable of generalising over different data sets. Differences in the balance between misogynous and non-misogynous, and between aggressive and non-aggressive examples, in both data sets, might also explain this behaviour. Whatever the reason, we leave this investigation for future work.

6 Conclusion

In this work, we described two models submitted to EVALITA 2020's subtask A on Automatic Misogyny Identification. To this task, a CNN and an LR classifier were trained, and arranged as a pipeline following three different strategies, with one of them coming at sixth place in the competition.

Even though our classifier turned out to be competitive, we believe improvements could be made to achieve better results, such as the addition of lexical features, for example. Also, it might be that following some preprocessing strategy, such as removing stop words, for example, might result in a better performance.

As for future work, besides testing the above cited changes, it would be interesting investigating why the worst model at the test set (as distributed to all participants) turned out to be the best model at the competition's private data set. The reasons for this behaviour are something to be determined.

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.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paolo Rosso Elisabetta Fersini, Debora Nozza. 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)*, Online. CEUR.org.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the evalita 2018 task on automatic misogyny identification (ami). *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:59.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the task on automatic misogyny identification at ibereval 2018. *IberEval@ SEPLN*, 2150:214–228.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Rachael Fulper, Giovanni Luca Ciampaglia, Emilio Ferrara, Y Ahn, Alessandro Flammini, Filippo Menczer, Bryce Lewis, and Kehontas Rowe. 2014. Misogynistic language on twitter and sexual violence. In *Proc. ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM)*.

Mohammed Hasanuzzaman, Gaël Dias, and Andy Way. 2017. Demographic word embeddings for racism detection on twitter.

Kate Manne. 2017. *Down girl: The logic of misogyny*. Oxford University Press.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*.

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.

Anand Rajaraman and Jeffrey David Ullman. 2011. *Mining of massive datasets*. Cambridge.

Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. 2018. Hateminers : Detecting hate speech against women. *CoRR*, abs/1812.06700.