PoliTeam @ AMI: Improving Sentence Embedding Similarity with Misogyny Lexicons for Automatic Misogyny Identification in Italian Tweets

Giuseppe Attanasio, Eliana Pastor
Department of Control and Computer Engineering
Politecnico di Torino, Italy
{giuseppe.attanasio,eliana.pastor}@polito.it

Abstract

We present a multi-agent classification solution for identifying misogynous and aggressive content in Italian tweets. A first agent uses modern Sentence Embedding techniques to encode tweets and a SVM classifier to produce initial labels. A second agent, based on TF-IDF and Misogyny Italian lexicons, is jointly adopted to improve the first agent on uncertain predictions. We evaluate our approach in the Automatic Misogyny Identification Shared Task of the EVALITA 2020 campaign. Results show that TF-IDF and lexicons effectively improve the supervised agent trained on sentence embeddings.

Italiano. *Presentiamo un classificatore multi-agente per identificare tweet italiani misogini e aggressivi. Un primo agente codifica i tweet con Sentence Embedding e una SVM per produrre le etichette iniziali. Un secondo agente, basato su TF-IDF e lessici misogini, è usato per coadiuvare il primo agente nelle predizioni incerte. Applichiamo la soluzione al task AMI della campagna EVALITA 2020. I risultati mostrano che TF-IDF e i lessici migliorano le performance del primo agente addestrato su sentence embedding.*

1 Introduction

The increasing adoption of online communication systems we experienced in the last decades brought the rise of many public forums for our own opinions, such as forums, blogs, and social networks. In these platforms, where access cannot - and must not - be restricted to anyone, the problem of misconduct and hateful content became soon compelling. The protection of the most targeted subjects, such as races, ethnicities, religious parties, genders, and sexual orientations, is of paramount importance. Violence against women manifests in social networks every time the offensive language targets women directly or indirectly (Ellsberg et al., 2005). We refer to these cases as misogynous speech. As platform owners are updating their regulatory terms at an increasing pace[1], the high number of contents due to a fast publication rate still pose a challenge to monitoring systems.

Many recent works in the NLP community show effective results in online monitoring of hate speech (Fortuna and Nunes, 2018) and misogynous contents (Pamungkas et al. (2020), Frenda et al. (2019), Anzovino et al. (2018)). Furthermore, research communities propose evaluation initiatives (Basile et al. (2019), Bosco et al. (2018)) to challenge NLP practitioners in finding novel solutions to shared tasks. Among these, the AMI shared task proposed at EVALITA 2020 (Basile et al., 2020) focuses on automatic identification of misogynous content on Twitter in Italian (Elisabetta Fersini, 2020).

The task counts two main subtasks. The goal of the first subtask, Subtask A - Misogyny & Aggressive Behaviour Identification, is the identification of misogynous speech in tweets, and in case of misogyny, the classification of an aggressive language. Subtask B - Unbiased Misogyny Identification, aims at classifying misogynous speech while guaranteeing the fairness of the model (in terms of unintended bias) on a synthetic dataset. The unintended bias is a known phenomenon in natural lan-

[1] https://www.theverge.com/2020/3/5/21166940/twitter-hate-speech-ban-age-disability-disease-dehumanize, https://www.theverge.com/2020/8/11/21363890/facebook-blackface-antisemitic-stereotypes-ban-misinformation, https://www.theguardian.com/technology/2020/jun/29/reddit-the-donald-twitch-social-media-hate-speech

guage models and recent works address its identification and mitigation (Dixon et al. (2018), Nozza et al. (2019), Kennedy et al. (2020)).

In this work, we describe our solution to address the AMI shared task. We propose a multi-agent classification. The system uses recent Sentence Embedding techniques to encode tweets and a SVM classifier to produce initial labels. A second agent, based on TF-IDF and Misogyny Italian lexicons, is jointly adopted to improve the first agent on uncertain predictions. Results show that the TF-IDF and misogyny lexicons effectively improve sentence embeddings. For both subtasks, we chose the constrained approach, effectively using only the data provided by the organizers.

2 Description of the system

Recent work has pointed out the efficiency of sentence embeddings in many downstream tasks, such as sentiment classification. Meanwhile, NLP practitioners strive to migrate the existing solutions to languages different from English. As such, classical language models are trained on large parallel corpora, and multi-lingual, pre-trained models are published for later uses.

In this work, we adopt a multi-agent classification procedure to address each proposed subtask. Firstly, we encode tweets to their sentence embeddings using a pre-trained multi-lingual sentence encoder. Next, we train a supervised classifier (the first agent) on the latent embedding space. In parallel, we extract the smoothed TF-IDF of tweets and enhance the representation with features built upon Hate Speech and Misogyny lexicons. This representation is then used to train a supervised classifier (the second agent). Finally, we propose a classification schema where uncertain predictions from the first agent are corrected with certain ones from the second agent.

The following paragraphs describe the data preprocessing step, expand on the classification system, and provide insights on its application to subtasks A and B.

2.1 Sentence embedding

Researchers devoted significant work to the empirical construction of sentence embeddings for the English language (Giorgi et al. (2020), Wang and Kuo (2020), Reimers and Gurevych (2019), Cer et al. (2018)). The most recent studies leverage high-quality language models, such as the BERT or Transformer-XL families, to build embeddings that properly transfer to several downstream tasks. Extending monolingual models, other works assess the generalization performance of language models pre-trained on multi-lingual corpora, producing sentence embeddings either aligned between languages (Reimers and Gurevych, 2020) or not (Aluru et al., 2020).

We build sentence embeddings testing two models. On the one hand, we use (Aluru et al., 2020), a monolingual BERT-based model originally fine-tuned from multilingual BERT on an Italian corpus for hate-speech detection tasks. The model is then fine-tuned on our specific subtasks. On the other hand, we choose the multilingual adaptation of Sentence-BERT (Reimers and Gurevych (2020)), which is based on the DistilBERT architecture (Sanh et al. (2019)). We use the implementation[2] built on top of the *transformers* library. Since results for the monolingual BERT were not encouraging from the beginning, in any of the subtasks, we will focus the discussion on multi-lingual Sentence-BERT.

Further, we run a fine-tuning round on multilingual Sentence-BERT to our specific subtasks. To tune the initial embeddings, we optimize a contrastive loss on pairs generated from the training set. For any pair of tweets, if the ground truth labels are the same (e.g. both misogynous or both non-aggressive) the distance between the two embeddings is decreased, while it is increased otherwise. Since computing the set of potential pairs is hard, we sample only 20% of the initial tweets, namely S, compute all the P possible pairs among those, where $|P| = (|S| \cdot |S-1|)/2$, and use them for fine-tuning. We anticipate this partial fine-tuning achieved worse results than the original model and leave other fine-tuning strategies as future work.

The final agent is then a supervised classifier trained on multi-lingual sentence embeddings (referred as the *SE* agent). We use a Support Vector Machine (SVM) with Radial Basis Function kernel, which achieves the best results on our validation set. Please refer to Section 3 for more details on parameter configuration and performance.

2.2 TF-IDF and Misogyny Lexicons

[2] https://github.com/UKPLab/sentence-transformers

Lexicons	#Words	Type of words
Sexist	138	Misogynous and sexist
Profanity	4	Vulgar and swear
Sexuality	7	Sexual references
Female body	6	Feminine body

Table 1: Description on misogynous lexicon.

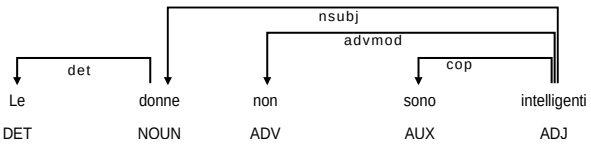

Figure 1: Example of dependency-based parse tree with sentiment polarity inversion.

Pre-processing. We firstly pre-process the data by replacing every URL found in tweets with the meta-token *LINK*. Next, we perform tokenization and lemmatization using the spaCy's[3] pre-trained Italian core model *it_core_news_lg*.

Input features. We use a smoothed TF-IDF vectorization of pre-processed tweets. We then enrich word representations using lexicons to encode misogynous speech and tweet sentiment.

(i) Misogynous lexicon. Misogynous tweets often contain sexist slurs, swear words, and sexual references. We include specific lexicons as input features for dealing with hate and misogynous speech (Frenda et al., 2018). We collect Italian lexicons from multiple online sources. We divide lexicons into the following categories: sexists, profanity, sexuality and female body as described in Table 1. The complete list of Italian lexica and sources are available at our repository[4]. As for the text of the tweet, lexicons are firstly lemmatized using spaCy. We then derive 4 features, one for each misogynous lexicon category. For a given category, we first count the occurrences of the corresponding lexicons in each tweet. We then normalize the occurrence with the tweet word count.

(ii) Sentiment Lexicon. We use a sentiment lexicon to characterize the polarity of tweets. The sentiment of words in a tweet is obtained with the OpeNER Italian Sentiment Lexicon (Russo et al., 2016). This sentiment lexicon consists of 24.293 lexical entries annotated with positive, negative and neutral polarity. In our analysis, we consider only positive and negative polarity.

Evaluating the polarity of an individual word in a tweet without considering its context, however, prevents from considering the role of negation on sentence polarity. To address this issue, we consider the following negation handling technique based on the dependency-based parse tree. We search in the parse tree extracted by spaCy for words affected by negation. For these words, we invert the polarity, if it is available. As an example, consider the phrase "le donne non sono intelligenti" (women are not intelligent). Figure 1 shows the extracted parse tree. The polarity of the word "intelligenti" (intelligent) is inverted, from positive to negative, since it is affected by negation.

Note that, as for the tweet text, we lemmatize sentiment lexicons. Finally, we extract 2 features that capture the tweet polarity. These are obtained by counting the number of words with positive and negative polarity respectively and then normalizing them by the tweet word count.

(iii) Additional features. Tweets may contain quotations of misogynous content, without being misogynous themselves. We hence consider as an additional feature the relative frequency of quotation marks. We also consider as a feature the length of the tweet (i.e. number of characters).

Finally, we train a supervised classifier (the second agent, referred as *Lex* agent) on the TF-IDF representation enriched with the additional features previously described. As for the first agent, we use a SVM with Radial Basis Function kernel model. We refer the reader again to Section 3 for details on the experimental setting.

2.3 Multi-agent prediction

We designed the multi-agent system to maximize prediction confidence by using only predictions with a high probability score. Specifically, we deem a prediction as confident if its associated probability score is above a given threshold.

We produce the final classification label by combining the outcomes of the two agents as follows. We first generate a prediction label and a score associated with it using the first agent. It entails encoding a given test point with Sentence-BERT and running the inference with SVM (*SE* agent). Afterward, we use the confidence threshold to decide whether to keep the label or not. If the *SE*'s prediction is not confident, we probe the second agent, which is built upon TF-IDF and misogyny lexicons (*Lex* agent). Finally, if *Lex*'s

[3]https://spacy.io/

[4]https://github.com/g8a9/ami20-improving-embedding

prediction is confident, we choose its label as the final one. If this is not the case, we rollback to *SE*'s class label. We kept the confidence threshold value as a hyper-parameter of the system.

By design, the proposed solution provides only confident prediction labels, either from the *SE* or the *Lex* agent. We applied the multi-agent classification procedure for both subtasks.

2.4 Approach to subtask A

In this task, participants have to assign a label indicating whether a tweet is misogynous or not. Then, limited to the misogynous ones, a second label should tell if the tweet is also aggressive.

We apply our multi-agent classification in a chained-prediction fashion. Specifically, we train a first instance of the system on the binary misogyny problem and label every tweet. In this step, we use the complete corpus. Next, we train a second instance on the binary aggressiveness problem. We feed the model with tweets predicted as misogynous on the previous step and produce a class label for those only. Finally, we label all the non-misogynous tweets as non-aggressive.

This strategy presents advantages and drawbacks since the predictions are chained. On the one hand, the two models are independent and can separately learn a simpler problem. On the other hand, this design lets errors on the misogyny prediction propagate to the aggressiveness one. We further discuss the matter in Section 4.

2.5 Approach to subtask B

For this task, we employ our multi-agent model (*SE+Lex* agents) with no modifications. Since we desire the model to encode also the structure and form of synthetic sentences, we train the model using the whole corpus.

3 Results

In this section, we firstly describe the experimental setting and the hyper-parameter tuning. We then report and comment experimental results of our multi-agent system. Further, to evaluate the effects of the two agents, we report the results of the system using only the *SE* or the *Lex* agent. The versions using only the *SE* agent or the *Lex* agent correspond to ids *run1* and *run2* respectively. The id *run3* is assigned to the multi-agent system.

Table 3 shows the F1 scores for misogyny and aggressiveness classes on the test set. All our runs show lower performance in the aggressiveness identification. We analyze and discuss this aspect in Section 4.

Rank	Team	Score
1	jigsaw.u.run2	0.7406
...	...	...
12	**PoliTeam.c.run3**	**0.6835**
13	MDD.c.run1	0.6820
14	**PoliTeam.c.run1**	**0.6809**
15	MDD.u.run2	0.6679
16	AMI_the_winner.c.run1	0.6653
17	**PoliTeam.c.run2**	**0.6473**
...	...	...
20	NoPlaceForHateSpeech.c.run3	0.4902

Table 2: Official results for subtask A

Run	Misogyny	Aggressiveness
SE (run1)	0.7688	0.5931
Lex (run2)	0.7222	0.5724
SE+Lex (run3)	0.7750	0.5920

Table 3: F1 score for subtask A

3.1 Experimental setting

To perform hyper-parameter optimization and model selection, we split the input data in training and validation data using random stratified sampling on both misogyny and aggressiveness labels. We used 20% of data as validation.

We ran a grid search over multiple classifiers as Support Vector Machines (SVM), Deep Feed Forward Neural Network, Random Forest, Logistic Regression, and their input parameters. The evaluation was performed using the first agent as reference. SVM with Radial Basic Function kernel with *gamma="scale"* and *C=10* achieved highest performance on F1 score for misogynous class on the validation set. We used this configuration for the supervised classifier of the second agent.

For the TF-IDF, we tuned the n-grams from *n=1* to *n=3*, and the number of maximum tokens from 5.000 to 10.000. To estimate the best configuration, we trained the SVM classifier with tuned parameters on the vectorized data, and evaluated the classification F1 score on the binary misogyny detection problem on the validation set. We achieved the highest F1 score with unigrams and 10.000 tokens as maximum vocabulary size.

The last hyper-parameter is the confidence threshold value for the multi-agent system. We evaluated the F1 score for the misogynous class on validation data varying the confidence threshold in the range $[0.6, 0.95]$. Best performance are obtained with a confidence threshold of 0.9.

Rank	Team	Score
1	jigsaw.u.run2	0.8826
2	**PoliTeam.c.run3**	**0.8180**
3	**PoliTeam.c.run1**	**0.8137**
4	fabsam.c.run1	0.7051
5	fabsam.c.run2	0.7022
6	**PoliTeam.c.run2**	**0.6940**
...	...	...
11	MDD.u.run3	0.6013

Table 4: Official results for subtask B

The hyper-parameter settings resulting from the experimental tuning are used for both the subtasks.

3.2 Subtask A

The score for subtask A is computed by averaging the F1 measures estimated for the *misogynous* and *aggressiveness* classes. Table 2 shows the official results. Our multi-agent system (run3) achieves our highest result. It is ranked 12th out of all submissions and 7th if we consider just constrained ones. While our TF-IDF and misogyny lexicon agent (run2) reaches our worst result, its introduction improves the agent trained on sentence embedding. The average F1 score increases from 0.6809 of the *SE* agent (run1) to 0.6835.

3.3 Subtask B

The score for subtask B is the weighted combination of AUC computed on the test tweets and three per-term AUC-based bias scores computed on the synthetic dataset. We refer the reader to (Elisabetta Fersini, 2020) for the complete description of the evaluation metrics.

Table 4 shows the official results. Our multi-agent system is ranked 2nd out of all submissions and 1st if only constrained runs are considered. As for subtask A, the *Lex* agent improves the performance of the *SE* one.

4 Discussion and Conclusions

Results show that the introduction of the TF-IDF and lexicons effectively improves the solution based on sentence embedding. This finding stands as the most significant contribution of this work, and we believe that it can drive future system designs. However, results on the test set reveal that we got wrong on some choices that affected the final performance.

4.1 Analysis on subtask A

Our multi-agent system missed the target on the aggressiveness detection task. As reported in Table 3, aggressiveness has a notable low F1 score. We think this is due to bad choices in training the system. (i) We used for the aggressiveness task only on the misogynous portion of the input data. This sub-set has an imbalanced class distribution with a prevalence of aggressive tweets. We did not re-balance the dataset, and our predictions produced many false positives on the test. (ii) Since we did not train the aggressiveness system on non-misogynous (and non-aggressive) tweets, whenever the misogyny system produces a false positive, the aggressiveness detector faces a completely new data point, out of its training distribution. (iii) Finally, we naively replicated the best algorithm and configuration found on the misogyny task to the aggressiveness one.

Notably, the number of misogynous false negatives which forced an aggressive tweet to be classified as non-aggressive by our chained approach (see Section 2.4) is 16 out of 365 total errors. This further enforces the conclusion that the majority of errors were due to bad training choices on the aggressiveness task and not the chained approach.

4.2 Analysis on subtask B

The multi-agent (*SE+Lex*) errors are 72 false negatives and 157 (x2.2) false positives. With a posterior error analysis on the test tweets, we identified several factors that contribute to misclassification.

Bias on parts of the body. Our system struggles with parts of the body that have sexual and misogynous reference based on the context. These words polarize the assignment to the misogynous class. As an example, 15% of false positives contain the word "gola" (throat). This behavior somewhat mimics the bias of models towards specific identity terms.

Self-mocking references. Another category hard to model is self-referencing text containing misogynous speech. While the tone of these tweets is auto-ironic or self-mocking, the model decontextualizes and produces false positives.

Targeted gender. In these tweets, the model correctly detects the hateful tone of voice but fails at identifying the gender of the target subject. As such, it predicts tweets attacking males as misogynous. This problem gets harder when the targeted gender can be only inferred by prior knowledge of tagged profiles (e.g. @bonucci_leo19, a male Italian football player).

Reported misogynous speech. Another diffi-

cult scenario to model is the reported or quoted misogynous speech. Frequently, users quote an unpleasant, misogynous passage while trying to support the exact opposite message. It can happen directly, using quotation marks, or indirectly by citing the original speaker.

We provide a list of tweets for each of the aforementioned categories as supplementary material[5].

Conclusion. In this work, we presented our solution to the AMI shared task at the EVALITA 2020 evaluation campaign. Our system is based on two models, the *SE* and *Lex* agents, which we built using sentence embedding techniques and TF-IDF enriched with misogyny lexicons respectively. We addressed both subtask A and B, limited to constrained runs. The approach fell short on the subtask A, while showed promising results on subtask B. Besides, results show the *Lex* agent effectively improves the performance of the *SE* agent.

Acknowledgments

This work was supported by the DataBase and Data Mining Group of Politecnico di Torino.

References

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep Learning Models for Multilingual Hate Speech Detection. *arXiv:2004.06465 [cs]*, April.

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 *SemEval-2019*, pages 54–63. ACL.

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.

[5]https://github.com/g8a9/ami20-improving-embedding

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the EVALITA 2018 hate speech detection task. In *EVALITA 2018*, pages 1–9. CEUR.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. *arXiv:1803.11175 [cs]*, April.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *AAAI/ACM AIES 2018*, pages 67–73, December.

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.

Mary Ellsberg, Lori Heise, World Health Organization, et al. 2005. Researching violence against women: a practical guide for researchers and activists.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):85:1–85:30, July.

Simona Frenda, Bilal Ghanem, Estefanía Guzmán-Falcón, Manuel Montes-y Gómez, Luis Villasenor-Pineda, et al. 2018. Automatic expansion of lexicons for multilingual misogyny detection. In *EVALITA 2018*, pages 1–6. CEUR-WS.

Simona Frenda, Bilal Ghanem, Manuel Montes-y Gómez, and Paolo Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752.

John M. Giorgi, Osvald Nitski, Gary D. Bader, and Bo Wang. 2020. DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations. *arXiv:2006.03659 [cs]*, June.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. pages 5435–5442, July.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *IEEE/WIC/ACM WI 2019*, pages 149–155.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv:1908.10084 [cs]*, August.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. *arXiv:2004.09813 [cs]*.

Irene Russo, Francesca Frontini, and Valeria Quochi. 2016. OpeNER sentiment lexicon italian - LMF.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Bin Wang and C.-C. Jay Kuo. 2020. SBERT-WK: A Sentence Embedding Method by Dissecting BERT-based Word Models. *arXiv:2002.06652 [cs]*, June.