MDD @ AMI: Vanilla Classifiers for Misogyny Identification

Samer El Abassi
Faculty of Mathematics and Computer Science
University of Bucharest
samer.el-abassi@s.unibuc.ro

Sergiu Nisioi
Human Language Technologies Research Center,
University of Bucharest
sergiu.nisioi@unibuc.ro

Abstract

In this report[1], we present a set of vanilla classifiers that we used to identify misogynous and aggressive texts in Italian social media. Our analysis shows that simple classifiers with little feature engineering have a strong tendency to overfit and yield a strong bias on the test set. Additionally, we investigate the usefulness of function words, pronouns, and shallow-syntactical features to observe whether misogynous or aggressive texts have specific stylistic elements.

1 Introduction

This paper discusses our submission (team MDD) to the Evalita 2020 Automatic Misogyny Identification Shared Task (Elisabetta Fersini, 2020; Basile et al., 2020) (Task A). Our methods consist of a set of simple vanilla classifiers that we employ to assess their effectiveness on the datasets provided by the organizers. The systems we submitted for evaluation use a logistic regression classifier with little hyperparameter tuning or feature engineering, being trained on tf-idf and average word embeddings pooling. Previous reports on misogyny (Fersini et al., 2018b,a) and aggressiveness (Basile et al., 2019) detection indicate that support vector machines and logistic regression classifiers effectively identify these patterns in social media posts. Furthermore, vanilla classifiers with little feature engineering were successfully used for other shared tasks, such as identifying dialectal varieties (Ciobanu et al., 2016; Zampieri et al., 2017) or native language identification (Malmasi et al., 2017), where high scores were obtained by simple approaches using SVMs or logistic regression classifiers.

The classifiers we built achieved a relatively good accuracy on our cross-validation tests; however, for this competition, the results obtained by our systems are not among the top-scoring ones and show to be misfit, with a significant tendency towards biased results.

In addition to the description of our submissions, in this report, we analyze the errors of our systems, and we bring into discussion several and topic-independent features to: 1) test the effectiveness of part-of-speech n-grams, function words, and pronouns on the task of identifying misogynous and aggressive texts on social media and 2) observe whether texts labeled as misogynous or aggressive have a particular bias towards certain grammatical structures.

2 System Description

At the basis of submissions is the logistic regression classifier with liblinear (Fan et al., 2008) optimizer, l2 penalty, and regularization constant $C = 3$ that we chose based on different cross-validation iterations. In addition, we introduced a heuristic at the prediction time in which we predict a text not to be aggressive if it was not categorized as misogynous.

The difference between our three submissions for Task A consist in the feature extraction process, where:

MDD.A.r.c.run1 is the logreg model trained on td-idf of word n-grams, n ranging from 1 to 5

MDD.A.r.u.run2 is the logreg model trained on pre-trained glove twitter embeddings of size 200 on 27 billion words[2]

MDD.A.r.u.run3 is the logreg model using spaCy (Honnibal and Montani, 2017) FastText

[1]Copyright c 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[2]English model GloVe.twitter.27B.200d https://nlp.stanford.edu/projects/GloVe/

CEUR-WS.org/Vol-2765/paper149.pdf

CBOW embeddings pre-trained on Wikipedia and OSCAR (Common Crawl)[3].

The second run is trained on English glove embeddings that surprisingly contain the representation of more than half of our Italian vocabulary, i.e., approximately 9500 words out of the total 15,000 size of the vocabulary of our data. The English glove embeddings cover code-switching, emojis, and basic Italian words. Despite having the lowest evaluation score of our submissions (0.666 macro f1), we believe it provides a decent estimation for identifying non-misogynous texts.

2.1 Feature Extraction

Our feature extraction processes for the submissions are simple, the first one uses the tf-idf vectorizer (Buitinck et al., 2013) on word n-grams, with n ranging from 1 to 5, to cover more of word context. Tf-idf features were used for their ability to categorize the importance of an n-gram with respect to the entire corpus. The second feature set is based on pre-trained word representations by calculating every word's embeddings in the text to eventually get an average representation. For words not present in the embeddings, an array of zeroes with the same dimensions was used.

Preprocessing

Our submissions use raw, un-processed texts, including tags and URLs. We have also experimented with different preprocessing and feature extraction steps for which we did not make any submission. We consider multiple approaches in this direction:

1. **clean** - changing the entire text to lowercase, removing hashtags, and links
2. **nps** - replacing the text with the noun phrases; these features contain the nouns and surrounding attributes that can highlight misogynous remarks
3. **fct words** - classification based on function word occurrence; these words cover stylistic and information of texts. We have collected a list of conjunctions, prepositions, connectors, etc. for Italian for this purpose.
4. **POS n-grams** - n-grams with n ranging from 1 to 5 over part-of-speech tags; these features would indicate a certain syntactic and stylistic pattern in misogynous or aggressive texts
5. **pronouns** - n-grams with n ranging from 1 to 5 over the pronouns and pronoun properties from the texts; we observed an increased usage in aggressive expressions of second-person pronouns
6. **filter POS** - n-grams over a filtered set of words and POS tags.

For POS tagging and noun phrases extraction, we use the default outputs from the Italian model for spaCy trained on the dataset provided by Bosco et al. (2014). In addition, we use the tag for each word that covers an entire set of features separated by whitespace; e.g., "Gender=Masc, Number=Sing, Person=2, PronType=Prs" becomes: "Masc Sing 2 Prs".

We expect the noun phrases to be less effective at detecting aggressive behaviour because aggressiveness often involves *verbal constructs* and actions.

3 Results and Discussion

In our work, we only describe the submissions for Task A of the competition, which is a classification task for the identification of misogynous and aggressive texts. Task B measures the bias of such classifiers with respect to certain concepts. Our submissions for task B are extracted from tf-idf representations of word n-grams and obtain the smallest scores of the competition.

Table 1 contains the submitted runs for Task A and the experiments we did to get a better understanding of the subtleties misogynistic and/or aggressive tweets contain. The columns *CV F1* contain the average F_1 scores computed for 10-fold cross-validation carried for ten iterations. Each cross-validation train-test split is stratified to preserve the proportions of misogynoys and/or aggressive texts in both splits. The *Test F1* columns are the results obtained on the gold standard test set. In the last column, we provide the macro F1 resulting from the average F1 between aggressiveness and misogyny predictions.

The submitted runs show that the tf-idf vectorizer from run1, although it scored better during the cross-validation stage, ended up being outperformed by the word embeddings extracted from spaCy (run3, 0.684 macro F_1), being unable to

[3]Model it_core_news_lg, version 2.3.0 released from spaCy https://spaCy.io/models/it

Feature	Misogyny		Aggressiveness		
	CV F1	Test F1	CV F1	Test F1	F1 Macro
tf-idf, run1	**0.883**	0.71	**0.8**	**0.652**	0.681
glove, run2	0.818	0.717	0.741	0.616	0.666
spacy, run3	0.842	**0.733**	0.767	0.635	**0.684**
clean tf-idf	**0.881**	0.706	**0.791**	**0.669**	0.688
clean, glove	0.847	0.722	0.766	0.618	0.67
clean spacy	0.846	**0.746**	0.784	0.655	**0.7**
nps, clean, tf-idf	**0.876**	0.714	**0.79**	**0.654**	0.684
nps, clean, spacy	0.837	**0.728**	0.768	0.646	**0.687**
fct words	0.672	0.628	0.614	0.564	0.596
POS n-grams	0.754	0.573	0.723	0.607	0.59
pronouns	0.594	0.596	0.656	0.636	0.616
filter POS	**0.832**	**0.731**	**0.765**	**0.657**	**0.694**

Table 1: Cross-validation and test-set results of logistic regression classifier with different feature extraction processes.

generalize to the new texts. The second run (run2, 0.666 macro F_1) uses the glove pre-trained embeddings for English. This result represents the biggest surprise of the three since it did not use Italian embeddings. We observe that the English glove representations cover more than 60

Cleaned texts aid the classifier by a significant threshold. In our experiments, we removed tags and URLs to observe a significant increase in macro scores for the same approaches over the cleaned texts. The best result we obtained so far (0.7 macro score) uses the Italian spaCy average vector representations extracted from clean texts.

Noun phrases extracted from each cleaned text do not indicate significant increases in misogynous or aggressive texts detection. Using these features yields comparable scores to the best of our methods, surpassing the classification attempts on uncleaned texts. This indicates that noun phrases alleviate the noise extracted by the tf-idf vectorizer. The model was less prone to overfitting and, therefore, more able to adapt to the unseen data.

Function words are features with grammatical roles, consisting of conjunctions, prepositions, articles, etc. encompassing stylistic aspects of the texts. We tested the accuracy of a simple logistic regression using function words, and the results were higher than 50% by a non-trivial amount. This is a potential indicator that misogynistic and/or aggressive tweets have a slightly different syntax than those that do not fit in either of the two. Moreover, using the tf-idf vectorizer on plain function words achieved 0.628 F_1 on the test set for misogyny identification, a result that is not at all negligible, given that these words do not encapsulate meaning.

POS n-grams are yet another set of features capable of capturing shallow syntactic constructs. Using this feature set, we observed a strong overfitting tendency on the cross-validation scenarios (average F_1 0.754 for misogyny and 0.723 for aggressiveness) while on the gold test set, the macro F_1 score is 0.59. This is an indicator that certain syntactic patterns are indeed occurring in the misogynistic and aggressive texts, weakly differentiating them from other types of texts. However, these features have little power to generalize on new samples.

Pronouns reveal the most interesting result due to two reasons: 1) the features did not overfit the data, as indicated by the cross-validation F_1 scores that are close to the actual scores on the gold test set; 2) aggressive texts can be differentiated between each other using only pronouns with an F_1 score (0.636) that is comparable with more advanced methods that use richer features such as embeddings (0.655, for the embeddings over clean texts) or tf-idf vectorizer (0.669, for tf-idf over clean texts). Therefore, in terms of aggressiveness, it is clear that certain expressions using forms of second-person pronouns are typically used to construct call-out phrases or curse-word expressions. The most common pronoun observed in aggresive

texts is *ti* - the second person singular acusative of pronoun *tu* ('you').

Filter POS account the n-grams of words and POS tags extracted from the following categories: nouns, adverbs, adpositions, determiners, adjectives, verbs, pronouns, and auxiliary verbs. The features obtain the second best result (0.694 F_1 macro score) from all our attempts. Again, in this situation, we are also facing a big difference between the cross-validation results and the released test set.

4 Discussion

The results show that the vanilla feature extraction methods suffered from a non-trivial amount of overfitting. Despite the fact that we carried a stratified 10-fold cross-validation, over ten iterations, the average F_1 scores obtained on the test set were considerably lower than the ones we obtained in our separate experiments.

The evaluation scores of said methods was over 88% in our cross-validation splits. On the cross-validation evaluation from the training set, tf-idf produced the best results. On the test set, embeddings proved to have a better power of generalization. Preprocessing the texts by removing stopwords, hashtags, links, and other types of noise proved to be beneficial for the classifier. The best results were obtained by extracting average clean text embeddings. Overall, word embeddings were more consistent when comparing cross-validation results with the test ones for misogyny detection.

At a shallow eye-check we noticed in the test set several examples labeled as misogynous with no apparent reasons: "troppo acida... non mangio yogurt", "Impiccati", "#nome?". We can only assume that the misogynistic character of these comments is given by the context in which they were posted. On the test set it also appears that the majority of misogynistic comments are remarks on different body parts, most likely as comments posted to pictures. It is, therefore, difficult to asses the misogynistic character of a short text without having at hand the full multi-modal context: to whom it was posted, what kind of relation is between the "commenter" and the "commentee", if the tweet is a reply or a single post, and so on and so forth.

It is worth noting that most text classification papers mention or use BERT (Bidirectional Encoder Representations from Transformers), as it has proven to be one of the most accurate when facing different types of data (Pamungkas et al., 2020). Other state of the art methods are LTSM (Long short-term memory) and XLNet, the latter overtaking BERT on various tasks (Yang et al., 2019). A current issue with such methods and word embeddings is that they transfer the human bias present in large corpora. This is becoming a bigger problem as AI filters are prevalent in today's society and therefore discriminatory traits of the models become discriminatory real world actions. For example, textual embeddings trained from Wikipedia data show discriminatory traits towards minorities such as associating foreigners with criminals, homosexuality with corruption, men being linked to aggression and women with the idea of the loving wife. (Papakyriakopoulos et al., 2020). Basta et al. (2019) finds that word embeddings are more likely to be discriminatory and biased than their contextualized counterparts, implying that state of the art methods are moving towards the right direction. However, as the models are getting closer to understanding language, one cannot help but wonder if this will have a negative impact on their bias if precautions aren't taken, as they might be overly impacted by the ubiquitous bias humans carry. Due to the widespread automatisation of daily tasks using machine learning models, mitigating prejudice becomes a responsibility of the developers, as it crucial for obtaining equal opportunities and treatment of minorities.

5 Conclusions

Our results indicate that simple feature engineering and vanilla classifiers cannot distinguish between misogynistic/aggressive tweets with reliable accuracy and that more research is needed to understand the important features concerning this task. However, the experiments imply a correlation between a text's syntax and its misogynistic/aggressive value. This proposes the idea that text that falls into either categories, (or maybe even hate speech in general?) does have a slightly more recognisable grammatical pattern than text that isn't. Whether it's the POS n-grams, pronouns, or just function words, the wording matters and is worth looking into for more advanced feature engineering.

References

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 *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Christine Basta, Marta Ruiz Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. *CoRR*, abs/1904.08783.

Cristina Bosco, Felice Dell'Orletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The evalita 2014 dependency parsing task. In *EVALITA 2014 Evaluation of NLP and Speech Tools for Italian*, pages 1–8. Pisa University Press.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Alina Maria Ciobanu, Sergiu Nisioi, and Liviu P Dinu. 2016. Vanilla classifiers for distinguishing between similar languages. In *Proceedings of the VarDial Workshop*.

Paolo Rosso Elisabetta Fersini, Debora Nozza. 2020. Ami @ evalita2020: Automatic misogyny identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874.

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.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel R. Tetreault, Robert A. Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *BEA@EMNLP*, pages 62–75. Association for Computational Linguistics.

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.

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 446–457, New York, NY, USA. Association for Computing Machinery.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the vardial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.