Profiling Italian Misogynist: An Empirical Study

Elisabetta Fersini, Debora Nozza, Giulia Boifava
University of Milano-Bicocca, Bocconi University, University of Milano-Bicocca
elisabetta.fersini@unimib.it, debora.nozza@unibocconi.it, g.boifava1@campus.unimib.it

Abstract

Hate speech may take different forms in online social environments. In this paper, we address the problem of automatic detection of misogynous language on Italian tweets by focusing both on raw text and stylometric profiles. The proposed exploratory investigation about the adoption of stylometry for enhancing the recognition capabilities of machine learning models has demonstrated that profiling users can lead to good discrimination of misogynous and not misogynous contents.

Keywords: Automatic Misogyny Identification, Stylometry

1. Introduction

The problem of identifying misogynist language in online social contexts has recently attracted significant attention. Social networks need to update their policy to address this issue and due to the high volume of texts shared daily, the automatic detection of misogynist and sexist text content is required. However, the problem of automatic misogyny identification from a linguistic point of view is still in its early stage. In particular, trivial statistics about the usage of misogynous language in Twitter have been provided in (Hewitt et al., 2016), while in (Anzovino et al., 2018) a first tentative of defining linguistic features and machine learning models for automatically recognizing this phenomenon has been presented. Given this relevant social problem, several shared tasks have been recently proposed for different languages (i.e. Italian, Spanish and English) to discriminate misogynous and not misogynous contents, demonstrating the interest of the Natural Language Processing community on investigating the linguistic and communication behaviour of this phenomenon. The Automatic Misogyny Identification (AMI) challenge (Fersini et al., 2018a; Fersini et al., 2018b) has been proposed at Ibereval 2018[1] for Spanish and English, and in Evalita 2018 (Caselli et al., 2018) for Italian and English. The main goal of AMI is to distinguish misogynous contents from non-misogynous ones, to categorize misogynistic behaviors and finally to classify the target of a tweet. Afterwards, (Basile et al., 2019) proposed HatEval, the shared task at SemEval 2019 on multilingual detection of hate speech against immigrants and women in Twitter for Spanish and English. The aim of HatEval is to detect the presence of hate speech against immigrants and women, and to identify further features in hateful contents such as the aggressive attitude and the target harassed, to distinguish if the incitement is against an individual rather than a group. This challenges offered the unique opportunity to firstly address the problem of hate speech against women in online social networks.

2. State of the art

During the above mentioned challenges, several systems have been presented to obtain the best performing solution in terms of recognition performance. Most of the participants to the AMI challenge considered a single type of text representation, i.e. traditional TF-IDF representation, while (Bakarov, 2018) and (Buscaldi, 2018) considered only weighted n-grams at character level for better dealing with misspellings and capturing few stylistic aspects. Additionally to the traditional textual feature representation techniques, i.e. bag of words/characters, n-grams of words/characters eventually weighted with TF-IDF, several approaches used specific lexical features for improving the input space and consequently the classification performances. In (Basile and Rubagotti, 2018) the authors experimented feature abstraction following the bleaching approach proposed by Goot et al. (Goot et al., 2018) for modelling gender through the language. Finally, specific lexicons for dealing with hate speech language have been included as features in several approaches (Frenda et al., 2018), (Ahluwalia et al., 2018) and (Pamungkas et al., 2018). Few participants to the AMI challenge, (Fortuna et al., 2018) and (Saha et al., 2018) considered the popular Embeddings techniques both at word and sentence level. More recently, (Nozza et al., 2019) investigated the use of a novel Deep Learning Representation model, the *Universal Sentence Encoder* introduced in (Cer et al., 2018) built using a transformer architecture (Vaswani et al., 2017) for tweet representation. The use of this more sophisticated model for textual representation coupled with a simple single-layer neural network architecture allowed the authors to outperform the first-ranked approach (Saha et al., 2018) at Evalita 2018. Thus, in the HatEval challenge, more than half of the participants exploited Word Embeddings or Deep Learning models (Sabour et al., 2017; Cer et al., 2018) for textual representation.

Concerning the machine learning models, the majority of the available investigations in the state of the art are usually based on traditional Support Vector Machines and Deep Learning methods, mainly Recurrent Neural Networks.

Several works have been done for adopting or even enlarging some lexical resources for misogyny detection purposes. The lexicons for addressing misogyny detection for the Italian language have been mostly obtained from lists available online, i.e. "Le parole per ferire" given by Tullio De Mauro[2], and the HurtLex multilingual lexicon (Bassignana et al., 2018).

[1] https://sites.google.com/view/ibereval-2018

[2] https://www.internazionale.it/opinione/tullio-de-mauro/2016/09/27/razzismo-parole-ferire

Resources and Techniques for User and Author Profiling in Abusive Language (ResT-UP 2020) , pages 9–13
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Although the above mentioned approaches represent a fundamental step towards the definition of mechanisms able to distinguish between misogynous and not misogynous contents, it is still pending the verification of the hypothesis that the writing style of authors could be a strong indication of misogynous profiles that therefore are likely inclined to produce misogynous contents.
To this purpose, in this paper, we propose to investigate the ability of some stylometric features to characterize misogynous and not misogynous profiles.

3. The Proposed Approach

The traditional feature vector representing a message m (used to train a given classifier) usually includes only terms that belong to a common vocabulary V of terms derived from a message collection:

$$\vec{m} = (w_1, w_2, ..., w_{|V|}, l) \quad (1)$$

where w_t denotes the weight of term t belonging to m with label l. However, some stylometric signals can be used to enhance the traditional feature vector and therefore learning models to distinguish between misogynous and not misogynous contents. The expanded feature vector of a message is defined as:

$$\vec{m}_s = (w_1, w_2, ..., w_{|V|}, s_1, s_2, \ldots, s_n, l) \quad (2)$$

where $s_1, s_2, \ldots, s_n$ represent the n additional stylometric features. The stylometric features investigate in this paper can be broadly distinguished as follow:

- *Pragmatic particles*: to better capture non-literal signals that could convey misogynous expressions, several valuable pragmatic forms could be taken into account. Pragmatic particles, such as emoticons, mentions and hashtags expressions, represent those linguistic elements typically used on social ratio to elicit, remark and make direct a given message.
- *Punctuation*: as stated in (Watanabe et al., 2018), how an internet user uses exclamation, interjections, and other punctuation marks is not necessarily an explicit cue indicating misogyny, they can be used to implicitly elicit a misogynous message (e.g. "Women rights? come on...go back to the kitchen!!!").
- *Part-Of-Speech (POS) lexical components*: the way of using some specific part of speech could be a relevant indicator of misogyny. For this reason, a POS tagger could be applied in order to assign lexical functions and derive some stylometric features related to them.

The above mentioned stylometric categories have led us to investigate the following features as candidates to capture misogynous profile and therefore to be included as additional features s_i reported in Eq. (2):

- average number of sentences
- average number of words
- frequency of the number of unique words
- frequency of complex words (more than 5 characters)
- average of the number of characters in a word
- frequency of the number of verbs
- frequency of the number of auxiliary verbs
- frequency of the number of adjectives
- frequency of the number of superlative adjectives
- frequency of the number of superlative relative adjectives
- frequency the number of comparative adjectives
- frequency of the number of nouns
- frequency of the number of conjunctions
- frequency of the number of adverbs
- frequency of articles
- frequency of indefinite articles
- frequency of definite articles
- frequency of indefinite articles prepositions
- frequency of pronouns
- frequency of numbers
- frequency of special characters
- frequency of emoji
- frequency of unigrams
- frequency of bigrams
- frequency of trigrams
- frequency of offensive words
- frequency of punctuation
- frequency of commas
- frequency of colon
- frequency of semi-comma
- frequency of exclamation mark
- frequency of question mark
- frequency of quotes
- frequency of upper-case words
- frequency of words starting with upper case
- frequency of stretched words
- frequency of the first singular person pronouns
- frequency of the first plural person pronouns
- frequency of the second singular person pronouns

- frequency of the second plural person pronouns
- frequency of the third singular person pronouns related to male
- frequency of the third singular person pronouns related to female
- frequency of the third plural person pronouns related to male
- frequency of the third plural person pronouns related to female
- frequency of the # symbol
- frequency of the @ symbol
- frequency of proper nouns

To validate the hypothesis that a stylistic profile can help to detect misogynous contents from the not misogynous ones, we trained several machine learning models both on the traditional feature vector (Eq. 1) and on the expanded feature vector (Eq. 2).

4. Experimental Investigation

4.1. Dataset

In order to validate our hypothesis that a stylistic profile of Italian misogynist can improve the generalization capabilities of machine learning models trained for misogyny detection purposes, we adopted the Italian benchmark dataset provided for the AMI@Evalita Challenge. The dataset has been collected by following the subsequent policies:

- Streaming download using a set of representative keywords, e.g. *pu****a, tr**a, f**a di legno*
- Monitoring of potential victims' accounts, e.g. *gamergate victims and public feminist women*
- Downloading the history of identified misogynist, i.e. *explicitly declared hate against women on their Twitter profiles*

The annotated Italian corpus is finally composed of 5000 tweets, almost balanced between misogynous and not misogynous labels.

4.2. Models and Performance Measures

Concerning the machine learning models trained to distinguish between misogynous and not misogynous tweets, Naïve Bayes (NB), Support Vector Machines (SVM) and Multi-Layer Perceptron (MLP) have been adopted[3].
Regarding the *traditional feature vector*, the text of each tweet has been stemmed and its TF-IDF representation has been obtained by exploiting the *sklearn* library (Pedregosa et al., 2011). For the stylometric features, we employed the Italian models of the *spaCy* library to obtain the part-of-speech tags to collect nouns, adjectives, adverbs. We also created a manual list of prepositions and articles. The list of offensive words has been extracted from an online resource [4].
Concerning the experimental evaluation, a 10-folds cross validation has been performed. To compare the two feature spaces, traditional textual feature vector and the ones with additional stylometric features, $Precision$, $Recall$ and *F1-measure* have been estimated focusing on both labels (i.e. 0=notMisogynous, 1=misogynous).

4.3. Experimental Results

We report in Table 1 the experimental results obtained by training all the considered machine learning models on the two feature space, i.e. the first based on Tf-IDF only and the second one based on TF-IDF and stylometric features.
We can easily note that the stylometric features provide a strong contribution for discriminating between misogynous and not misogynous messages. It is interesting to note that the stylometric features are not only able to improve the performance with respect to the traditional features, but they lead to have good performance for both classes guarantying a good compromise of Precision and Recall for misogynous and not misogynous instances. In this way, we are able to provide a feature representation and a machine learning approach that is able to recognize "the easy class" related to not misogynous contents and "the difficult class" related to the misogynous text. In order to better understand the role of stylometric cues, we performed an error analysis on those messages that were wrongly classified by the best performing model, i.e. Support Vector Machines. First of all, the proposed analysis involving stylometry has led to 20% of classification error, where 43.85% of misclassified instances are not misogynous tweets that are classified as misogynous and 56.15% of misclassified instances are misogynous tweets that are classified as not misogynous.
For those instances for which the actual label was not misogynous but the classifier predicted them as misogynous, we can highlight the main types of errors:

- *Unsolved Mentions*: the model, do not solving the user mentions, is biased by adjectives. In particular, when referring to a target by using a mention (denote by the @ symbol), the stylometric features are not able to capture the gender-related to a given noun and therefore is biased by the bad words typically related to women. An example of this type of errors are represented by the following sentence:

  *@laltrodiego Mer*a schifosa lurida*

  that can be translated as:

  *@laltrodiego Bad Sh*tty Sh*t*

  The target of the tweet is an account of a male user, but the model do not have the chance to solve the uncertainty related to the mention.

- *Wrong Target*: in this case, the model is again biased by adjectives typically denoting bad words because it is not able to recognize female proper nouns. In particular, when mentioning a given entity (i.e. football

[3]The experiments have been conducted using default parameters of models implemented in sklearn: https://scikit-learn.org

[4]https://bit.ly/2HK3fYE

		Precision		Recall		F1-Measure	
		0	**1**	**0**	**1**	**0**	**1**
TD-IDF	NB	0.816	0.459	0.381	0.858	0.519	0.598
	MLP	0.840	0.745	0.844	0.735	0.841	0.738
	SVM	0.839	0.713	0.811	0.746	0.823	0.727
TF-IDF + stylometry	NB	0.816	0.524	0.559	0.793	0.662	0.631
	MLP	0.851	0.810	0.888	0.743	0.868	0.773
	SVM	0.910	0.747	0.793	0.848	0.835	0.777

Table 1: Experimental results

teams, male, locations) the stylometric features are not able to capture the gender and therefore the model is again biased by the bad words typically related to women. An example of this type of errors are represented by the following sentence:

> *Sintesi: Barcellona cul*na, De Rossi come CR7. Entrambi applauditi dai tifosi avversari. #BarcaRoma #BarcellonaRoma*

that can be translated as:

> *Summary: Barcelona big a*s, De Rossi as CR7. Both applauded by the opposing fans. #BarcaRoma #BarcelonaRoma*

The target of the tweet relates to a football team and not on a female user, but the model does not have the chance to solve the uncertainty related to the target.

- *Absence of an Explicit Target*: in this case, the model misclassify those tweets where the target is not explicitly stated. Typical examples are comments related to events, where offensive words related to female are used to complain:

  > *PORCA PUT**NA LADRA SCHIFOSA IERI HARRY STYLES ERA NELLA MIA CITTA E IO NON SAPEVO NULLA.HARRY STYLES ERA A MODENA E IO LO AVREI POTUTO INCONTRARE, ODIO TUTTI CHE VITA DI MER*A*

  that can be translated as:

  > *SHITTY BIT*H YESTERDAY HARRY STYLES WAS IN MY CITY AND I DID NOT KNOW ANYTHING.HARRY STYLES WAS IN MODENA AND I WOULD HAVE BEEN ABLE TO MEET HIM, I HATE ALL WHAT A SHIT*Y LIFE*

  In this case, the implicit target is an event and the model, observing offensive words such as *putt*na/bit*h* wrongly predict the message as misogynous.

An analogous behaviour has been observed when the actual labels of tweets are misogynous but the classifier predicted them as not misogynous. In particular, the errors are mainly related to one main lack of information:

- *Absence of Syntactic Features*: the model, which does not consider the syntactical structure of the sentence, is not able to determine the target of an offensive adjective. An example of these types of errors are represented by the following sentence:

  > *Se scrivi che Weinstein o Trump sono dei porci e dei maniaci tutti applaudono, ma se dici che Selvaggia Lucarelli è un putt*none sei sessista...*

  that can be translated as:

  > *If you write that Weinstein or Trump are pigs and maniacs everyone applauds, but if you say that Selvaggia Lucarelli is a bit*h you're sexist ...*

  The target of the offensive language is clearly a woman, but the model since it does not consider the structure of the sentence it is biased by those adjectives related to men.

The error analysis has highlighted on one side the necessity of properly dealing with the target of the message, and on the other hand, it has pointed out the needs to more additional stylometric features to obtain a better understanding on the structuring of sentences of both misogynous and not misogynous contents.

4.4. Conclusions and Future Work

In this paper, a preliminary empirical investigation about the profiling of Italian misogynous contents has been performed. A set of stylometric features have been studied for validating the hypothesis that cues about the writing style of authors can contribute to better distinguish misogynous contents from the not misogynous ones. The experimental evaluation has corroborated the hypothesis that the use of stylometric features improves the recognition capabilities of several machine learning models for misogyny detection purposes. Concerning future work, several additional syntactic features will be considered for a better understanding of the structure of the sentences. Additionally, the capabilities of the investigated features will be evaluated focusing on additional languages, i.e. Spanish and English, also investigating which set of features contributes most on the results of the classifiers. As final future work, a different paradigm for profiling misogynist will be investigated. In particular, a benchmark profile of misogynistic and not misogynistic language will be created to then enable a *learning-by-difference* approach.

5. Bibliographical References

Ahluwalia, R., Soni, H., Callow, E., Nascimento, A., and Cock, M. D. (2018). Detecting Hate Speech Against Women in English Tweets. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

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.

Bakarov, A. (2018). Vector Space Models for Automatic Misogyny Identification. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Basile, A. and Rubagotti, C. (2018). Automatic Identification of Misogyny in English and Italian Tweets at EVALITA 2018 with a Multilingual Hate Lexicon. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel, F., Rosso, P., and Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*. Association for Computational Linguistics.

Bassignana, E., Basile, V., and Patti, V. (2018). Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

Buscaldi, D. (2018). Tweetaneuse AMI EVALITA2018: Character-based Models for the Automatic Misogyny Identification Task. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Caselli, T., Novielli, N., Patti, V., and Rosso, P. (2018). EVALITA 2018: Overview of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In Tommaso Caselli, et al., editors, *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP 2018)*, pages 169–174. Association for Computational Linguistics, November.

Fersini, E., Nozza, D., and Rosso, P. (2018a). Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In Tommaso Caselli, et al., editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Fersini, E., Rosso, P., and Anzovino, M. (2018b). Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*. CEUR-WS.org.

Fortuna, P., Bonavita, I., and Nunes, S. (2018). INESC TEC, Eurecat and Porto University.

Frenda, S., Ghanem, B., Guzmán-Falcón, E., Montes-y-Gómez, M., and Villaseñor-Pineda, L. (2018). Automatic Lexicons Expansion for Multilingual Misogyny Detection. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Goot, R., Ljubešić, N., Matroos, I., Nissim, M., and Plank, B. (2018). Bleaching Text: Abstract Features for Cross-lingual Gender Prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 383–389.

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.

Nozza, D., Volpetti, C., and Fersini, E. (2019). Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 149–155.

Pamungkas, E. W., Cignarella, A. T., Basile, V., and Patti, V. (2018). Automatic Identification of Misogyny in English and Italian Tweets at EVALITA 2018 with a Multilingual Hate Lexicon. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866.

Saha, P., Mathew, B., Goyal, P., and Mukherjee, A. (2018). Indian Institute of Engineering Science and Technology (Shibpur), Indian Institute of Technology (Kharagpur).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (NIPS 2017)*, pages 6000–6010.

Watanabe, H., Bouazizi, M., and Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6:13825–13835.