# Exploration of Misogyny in Spanish and English tweets

Simona Frenda[1,2], Bilal Ghanem[2], and Manuel Montes-y-Gómez[3]

[1] University of Turin, Italy
sfrenda@unito.it
[2] Universitat Politècnica de València, Spain
bigha@doctor.upv.es
[3] National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico
mmontesg@inaoep.mx

**Abstract.** Nowadays, misogynistic abuse online has become a serious issue due, especially, to anonymity and interactivity of the web that facilitate the increase and the permanence of the offensive comments on the web. In this paper, we present an approach based on stylistic and specific topic information for the detection of misogyny, exploring the several aspects of misogynistic Spanish and English user generated texts on Twitter. Our method has been evaluated in the framework of our participation in the AMI shared task at IberEval 2018 obtaining promising results.

**Keywords:** Misogyny Detection · NLP · Linguistic Analysis



## 1 Introduction

On the web, and especially on social platforms, the freedom to express opinions and sentiments can turn into an uncontrolled flow of thoughts that gives rise to negative online behaviors, such as cases of hate speech towards specific targets that can morally harm or incite physical violence. Indeed, it is not uncommon to find user generated online contents impregnated with hate, especially, towards women which sometimes can also be translated into actions of violence. For instance, some social studies, like [14], demonstrate the existence of a correlation between the number of rape and the number of misogynistic tweets per state on the USA, suggesting the fact that social media can be used as a social sensor of sexual violence. However, any misogynistic content hurts and distresses the victim, also without the physical attacks. Actually, the amount, persistence and diffusion of the messages, especially on social networks, support the increase of case of victims online. For this reason, many Internet companies use blacklists to block this kind of contents, but, unfortunately, the problem is far from being solved because of the complexity of language. Especially online comments, like tweets, are difficult to manage considering the variety of informal language devices used by users. In this paper, we present our approach based on stylistic

features and especially on topic information modeled by the use of specific lexicons, built exploring the traits of misogynistic speech into the dataset provided by the organizers of the AMI (Automatic Misogyny Identification) shared task [10] at IberEval 2018 evaluation campaign. In the framework of our participation in this task, we evaluated our approach using machine learning algorithms in a multilingual context.
The rest of the paper is structured as follow. In Section 2 we introduce the literature, describing synthetically the multidisciplinary contributions until today. In Section 3 we explain the methodology used and the experiments carried out, describing the results obtained in the competition. Finally, in Section 4 we draw some conclusions about our analysis.

## 2 Related Work

In the natural language processing (NLP) field the approaches to misogyny detection are very recent [1] and to our knowledge, the majority of them in this field have concentrated especially on abusive and aggressive language detection. For instance, the actual researches are directed at proposing solutions able to investigate deeper the abusive language especially into social media context. Therefore, they exploit principally: simple surface characteristics ([7], [19]), linguistic features that take into account POS-tags or dependencies relationships ([28], [4], [29]), semantic knowledge using word embedding techniques ([20], [22]), conceptual and polarity information [17], or consider also the profile information of the authors in the perspective to catch the sexual predators into chats online [9]. Moreover, [8] use an ontology-based approach in order to predict the anti-LGBT hate speech, and other scholars combine sentiment lexicons [27] and subjectivity detection [15] with other features. Taking into account the previous works, we propose an approach that exploit stylistic, semantic and topic information about the misogynistic speech. Actually, misogyny is a kind of aggressive and abusive language that aims to offend women, involving all prejudices and myths about woman. Commonly, typical expressions of misogynistic comments online are the sexist utterances, and various studies have focused on this issue. For instance, [13] and [18] investigated the aspects like interactivity and anonymity which minimize the authority and the inhibition of the user[4], facilitating sexual attacks, especially in video games context ([12], [24]). Moreover, the users usually disguise misogynistic comments as humorous, involving sarcasm or irony in their utterances. However, these are experienced by women like sexual harassment [3] and, also, the continue exposition to sexist jokes can also modify the perception of sexism like norm and not like negative behavior [11]. Finally, considering also these last observations we modeled lexicons focused on several aspects of misogyny, such as sexual and stereotype language.

[4] Also Twitter users in AMI data underline this minimization of the inhibition (*Acá en twitter si la puedo insultar ya que nunca lo verá*).

## 3 Proposed Approach

The AMI task[5] aims to detect misogyny in two datasets of English and Spanish tweets. In the case a tweet is classified as misogynistic we need to distinguish if the target is towards an individual or not, and identify the type of misogyny, according to the following classes: stereotype and objectification, dominance, derailing, sexual harassment and threats of violence, and discredit. This subdivision of misogyny allows us to explore the different aspects of this form of hate speech and compare them in two different languages. However, none of the datasets is geolocalized so there are misogynistic examples from different countries. Therefore, in order to gather the linguistic variations and consider all the aspects about misogyny discussed above, we propose an approach based on stylistic features captured by means of character n-grams, on sentiment information and on a set of lexicons built by examining the misogynistic tweets from training data provided by the organizers. For extracting the meaningful words we used information gain for identifying the relevance of words for the classification task, and frequency distributions for detecting the most informative words about misogyny. Considering the tweet context, we take into account also informal language devices used often in social media, such as slangs, abbreviations and hashtags in the two languages. This set of features is experimented employing a SVM (Support Vector Machine) algorithm and an ensemble technique, reaching promising results. In order to perform the experiments, the tweets are represented by a vector composed of: all specific topic features (set of lexicons), pondered with Information Gain, and character n-grams, weighted with TF-IDF (Term FrequencyInverse Document Frequency) measure. Finally, we preprocessed the data deleting emoticons and urls in order to treat the data for the creations of the lexicons, and for the correct extraction of the features to train our models. In addition, we used also FreeLing lemmatizer provided by [5] to face the inflectional morphology of Spanish language, and Porter stemmer provided for NLTK (Natural Language Toolkit).
In the following subsections we are going to describe the features and the systems proposed, considering the best parameters for this task.

### 3.1 Linguistic Features

Analyzing the multilingual data provided by the organizers and taking into account the traits of the misogynistic speech, we delineated a set of specific linguistic features in order to face the two classifications involved in the AMI shared task: at the first level it is required to detect misogynistic tweets, and at the second level to identify the misogynistic class and the type of target if the message is misogynistic.
For the principal task we built the lexicons concerning sexuality, profanity, femininity and human body. Along with them we used a list of hashtags, abbreviations, sentiment lexicons for both of languages and characters n-grams. For

[5] https://amiibereval2018.wordpress.com/

the subtask about misogynistic behavior, we added at these lexicons a list of words concerning the stereotypes about women, and especially for Spanish we used also emotional information. Lastly, to identify the category of the target we used characters n-grams.
Below we briefly describe the several features.

*Sexuality* One of the most frequent subjects in misogynistic tweets of provided datasets is the sexuality (*pussy*, *concha*) and in particular the desire to dominate women in sexual contexts, trend visible especially in the English language.

*Profanity* As in previous studies ([6], [16]), in this list we did not take into account of some common words like $fuck$ or *puta* which are used also without offensive purposes. Especially for Spanish this lexicon gathers several vulgarities from different countries.

*Femininity* In order to establish if the target of offense is a woman, we collected the most used terms related to women mainly in negative sense (*gallina*, *blonde*).

*Human body* A set of terms about feminine body is strongly connected with sexuality, and for this reason this kind of terms are used especially in English.

*Hashtags* As in previous works [13], we take into consideration the hashtags used as referents for shared concepts by online communities. For instance, in the datasets we can find hashtags like $\#todasputas$ or $\#womensuck$.

*Abbreviations* This list contains vulgar abbreviations typical of Internet slangs found in the data, perceiving a prevalence for English, such as: $idgaf$, *smh* and *hdp*.

*Stereotypes* This last list embraces various terms related to the stereotypes or myths about women, like technology, cooking or taking care of children.

*Sentiment lexicons* As previous studies about abusive language ([8], [15]), we used sentiment classification. For English we used SentiWordNet [2] and SentiStrength [26] for Python. For Spanish, instead, we used ElhPolar dictionary [23]. In both of languages, the sentiment classification helps us to increase the values confirming the intuition that hate speech expressions largely exhibit a negative polarity.

*Affective lexicons* Finally, for the second task we used Spanish Emotion Lexicon (SEL) ([25], [21]), in order to understand the impact of the emotions on specific misogynistic classes.

### 3.2 Experiments and Final Evaluation

In this section, we illustrate the experiments that we carried out. For the evaluation, the accuracy and F1 measures were used, depending on each task (Misogyny Identification and Misogynistic Behavior with Target Classification). For both of tasks, our experiments were similar.
We carry out many experiments to test different machine learning classifiers. We found that SVM has achieved the highest accuracy value for the first task, and the highest value of F1 for the second one. We employed the linear kernel of the SVM, and for each task different values of $C$ and *Gamma* parameters were chosen. Moreover, we employed an ensemble technique (majority voting) to combine the predictions from the three classifiers that obtained the highest results: the best was SVM, followed by Random Forest and Gradient Boosting classifiers. Therefore, the main difference between the tasks is in the lexicons features that we described in the previous section (Section 3.1).
In order to evaluate our classification approach, we used K-Fold Stratified method with $K$=5, and as a baseline, we used characters n-grams for each classification task with the best n-gram length (from 3 to 5 grams).

**Table 1.** The tasks baselines and the experimental results with K-Fold Stratified

| Approach | Task-1 | | Task-2 | | Task-3 | |
|---|---|---|---|---|---|---|
| | *En* | *Sp* | *En* | *Sp* | *En* | *Sp* |
| *char n-grams - SVM (Baseline)* | 76.23 | 77.76 | 25.13 | 36.03 | 71.10 | 69.01 |
| Lexicons - SVM | 79.04 | 78.83 | 27.02 | 38.11 | 52.00 | 54.48 |
| Sentiment features+SEL - SVM | 74.27 | 75.83 | 20.88 | 29.51 | 50.12 | 51.63 |
| All features - SVM | 78.19 | 79.05 | **28.44** | 40.84 | 71.32 | **69.30** |
| All features - Ensemble | **80.14** | **79.44** | 28.38 | **41.45** | **71.41** | 67.21 |

Table 1 shows the results[6] of our approach in each task. In addition, we present the results for each feature set to show separately their performance. In general, among all results, we can notice that the lexicons feature has achieved higher results compared to the other features.

In the first task, we used the same feature set for both languages, although we obtained better performance in the English language. As justification for the lower performance in Spanish, we need to take into account its high level of morphological complexity. Indeed, by employing FreeLing lemmatizer [5] we reached 79.05%, whereas without using the lemmatization process we had an accuracy value of 76.11%. Further improvements could be achieved by improving the lemmatization process. Similarly, for the English language, in order to enhance the matching process, we employed a lemmatizer for NLTK. However, the best solution for English has been to use the stemmer[7] considering the fact that the lemmatization process reduced the accuracy.

[6] Affective feature is tested only for the Spanish language.
[7] In our experiments, we used the implementation of Porter stemmer in NLTK package for Python.

For the second task, we used different types of features. Table 2 and Table 3 illustrate the results obtained in the competition, where the results for misogynistic behaviors and target classification are provided in terms of Average F1.

**Table 2.** The formal results of Task 1

| Language | Approach | Accuracy | Team Rank | Run Rank |
|---|---|---|---|---|
| *English* | All features - Ensemble | **87.05** | 2 | 5 |
| *Spanish* | All features - Ensemble | **81.35** | 3 | 3 |

As we see, in both tasks our approach has obtained a good performance. Especially, in the second task, we obtained the highest results for the English language.

**Table 3.** The formal results of Task 2

| Language | Approach | Macro-F1 | Team Rank | Run Rank |
|---|---|---|---|---|
| *English* | All features - SVM | **44.25** | 1 | 1 |
| *Spanish* | All features+SEL - Ensemble | **44.10** | 2 | 4 |

## 4 Conclusions

In this work, we investigated the misogyny issue exploring its several aspects in social network context. The results obtained in the framework of the competition are promising and the error analysis done during our experiments suggests that one of the principal problem in both languages is the use of linguistic devices like irony and sarcasm. As said in Section 2, humorous utterances are common in misogynistic speech, such as: "*Cuál es la peor desgracia para una mujer? Parir un varón, porque después de tener un cerebro dentro durante 9 meses, van y se lo sacan*"; "*What's the difference between a blonde and a washing machine? A washing machine won't follow you around all day after you drop a load in it*". Moreover, by means of an Information Gain analysis, we noticed that, in both of the two tasks, sexual language is more used in misogynistic English texts, and the profanities or vulgarities are more used in misogynistic Spanish tweets. One of the points that surprises us is the fact that in the analysis of the use of affective feature in Spanish language, the Joy is the principal emotion that incites the misogynistic speech. Therefore, taking into account our final observation, as future work we want to examine deeper the linguistic phenomena like irony and sarcasm and investigate the role of the emotions in misogynistic and, more generally, in hate speech context.

## Acknowledgement

The work of Simona Frenda was partially funded by the Spanish MINECO under the research project SomEMBED (TIN2015-71147-C2-1-P).

## References


1. Anzovino, Maria, Elisabetta Fersini, and Paolo Rosso: Automatic Identification and Classification of Misogynistic Language on Twitter. In International Conference on Applications of Natural Language to Information Systems, pp. 57-64. Springer, Cham. (2018)
2. Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani: Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In LREC, vol. 10, no. 2010, pp. 2200-2204. (2010)
3. Boxer, Christie F., and Thomas E. Ford: Sexist humor in the workplace: A case of subtle harassment. Insidious workplace behaviour pp. 175-206. (2010)
4. Burnap, Peter and Williams, Matthew Leighton: Hate speech, machine classification and statistical modelling of information flows on Twitter: interpretation and communication for policy decision making. Internet, Policy Politics, Oxford, UK. (2014)
5. Carreras, Xavier, Isaac Chao, Lluis Padr, and Muntsa Padr: FreeLing: An Open-Source Suite of Language Analyzers. In LREC, pp. 239-242. (2004)
6. Clarke, Isobelle, and Jack Grieve: Dimensions of Abusive Language on Twitter. In Proceedings of the First Workshop on Abusive Language Online, pp. 1-10. (2017)
7. Chen, Ying, Yilu Zhou, Sencun Zhu, and Heng Xu: Detecting offensive language in social media to protect adolescent online safety. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), pp. 71-80. IEEE. (2012)
8. Dinakar, Karthik, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard: Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS) 2, no. **3**: 18. (2012)
9. Escalante, Hugo Jair, Esa Villatoro-Tello, Sara E. Garza, A. Pastor Lpez-Monroy, Manuel Montes-y-Gmez, and Luis Villaseor-Pineda: Early detection of deception and aggressiveness using profile-based representations. Expert Systems with Applications **89**: 99-111. (2017)
10. Fersini, Elisabetta, Anzovino MAria, Rosso Paolo: 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, September 18. (2018)
11. Ford, Thomas E., Erin R. Wentzel, and Joli Lorion: Effects of exposure to sexist humor on perceptions of normative tolerance of sexism. European Journal of Social Psychology 31, no. **6**: 677-691. (2001)
12. Fox, Jesse, and Wai Yen Tang: Sexism in online video games: The role of conformity to masculine norms and social dominance orientation. Computers in Human Behavior **33**: 314-320. (2014)
13. Fox, Jesse, Carlos Cruz, and Ji Young Lee: Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. Computers in Human Behavior **52**: 436-442. (2015)
14. Fulper, Rachael, Giovanni Luca Ciampaglia, Emilio Ferrara, Y. Ahn, Alessandro Flammini, Filippo Menczer, Bryce Lewis, and Kehontas Rowe: Misogynistic language on Twitter and sexual violence. In Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM). (2014)

15. Gitari, Njagi Dennis, Zhang Zuping, Hanyurwimfura Damien, and Jun Long: A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering 10, no. **4**: 215-230. (2015)
16. Hewitt, Sarah, Thanassis Tiropanis, and Christian Bokhove: The problem of identifying misogynist language on Twitter (and other online social spaces). In Proceedings of the 8th ACM Conference on Web Science, pp. 333-335. ACM. (2016)
17. Justo, Raquel, Thomas Corcoran, Stephanie M. Lukin, Marilyn Walker, and M. Ins Torres: Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. Knowledge-Based Systems **69**: 124-133. (2014)
18. Lapidot-Lefler, Noam, and Azy Barak: Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. Computers in human behavior 28, no. **2**: 434-443. (2012)
19. Mehdad, Yashar, and Joel Tetreault: Do Characters Abuse More Than Words?. In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 299-303. (2016)
20. Nobata, Chikashi, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang: Abusive language detection in online user content. In Proceedings of the 25th international conference on world wide web, pp. 145-153. International World Wide Web Conferences Steering Committee. (2016)
21. Daz Rangel, Ismael, Grigori Sidorov, and Sergio Surez Guerra: Creacin y evaluacin de un diccionario marcado con emociones y ponderado para el espaol. Onomazein 1, no. **29**. (2014)
22. Samghabadi, Niloofar Safi, Suraj Maharjan, Alan Sprague, Raquel Diaz-Sprague, and Thamar Solorio: Detecting Nastiness in Social Media. In Proceedings of the First Workshop on Abusive Language Online, pp. 63-72. (2017)
23. Saralegi, Xabier, and Inaki San Vicente: Elhuyar at TASS 2013. In XXIX Congreso de la Sociedad Espaola de Procesamiento de lenguaje natural, Workshop on Sentiment Analysis at SEPLN (TASS2013), pp. 143-150. (2013)
24. Shaw, Adrienne: The Internet is full of jerks, because the world is full of jerks: What feminist theory teaches us about the Internet. Communication and Critical/Cultural Studies 11, no. **3**: 273-277. (2014)
25. Sidorov, Grigori, Miranda-Jimnez Sabino, Viveros-Jimnez Francisco, Gelbukh Alexander, Castro-Snchez No, Velsquez Francisco, Daz-Rangel Ismael, Surez-Guerra Sergio, Trevio Alejandro and Gordon Juan: Empirical study of opinion mining in Spanish tweets. LNAI 7629, pp. 1-14. (2012)
26. Thelwall, Mike, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas: Sentiment strength detection in short informal text. Journal of the Association for Information Science and Technology 61, no. **12**: 2544-2558. (2010)
27. Van Hee, Cynthia, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Vronique Hoste: Detection and fine-grained classification of cyberbullying events. In International Conference Recent Advances in Natural Language Processing (RANLP), pp. 672-680. (2015)
28. Xu, Jun-Ming, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore: Learning from bullying traces in social media. In Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies, pp. 656-666. Association for Computational Linguistics. (2012)
29. Zhong, Haoti, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J. Miller, and Cornelia Caragea: Content-Driven Detection of Cyberbullying on the Instagram Social Network. In IJCAI, pp. 3952-3958. (2016)