# Misogyny identification through SVM at IberEval 2018

Jose Sebastián Canós

Universitat Politècnica de València
josecan@inf.upv.es

**Abstract.** This paper explains the author's approach to the task Automatic Misogyny Identification at IberEval 2018 whose objective is to identify cases of aggressiveness and hate speech towards women. It describes the system submitted to the task, which is based on a support vector machine model, its development and the results obtained.

**Keywords:** Misogyny identification, Twitter, Support vector machine.

## 1 Introduction

The objective of the task on Automatic Misogyny Identification at IberEval 2018 is to identify cases of aggressiveness and hate speech towards women. Nowadays in social media people can develop and express their hatred easily without endangering themselves, leaving many cases of harassment. One regular target of this hatred are women, and there are investigations that study how and in which forms this misogyny is expressed [3], but there have not been many published efforts to automatically identify misogyny in social media as the one by [5].

The task of misogyny identification shares some similarities with sentiment analysis or stance detection [1], in the sense that it must detect whether a text is positive or negative to a target, in this case to identify misogyny. The complexity of the task can strive in the different meanings of some keywords where they are used in different contexts.

In the following sections this paper will describe the task (Section 2), the submitted system's approach (Section 3), and the results obtained on the task (Section 4). At the end (Section 5) some conclusions will be presented.

## 2 Task Description

The task Automatic Misogyny Identification is performed over a collection of tweets, two corpora are available in different languages, Spanish and English, whose results are evaluated separately. The task includes two subtasks. Subtask A is to classify the tweets as either misogynous or not misogynous, it is evaluated using the standard accuracy measure. Subtask B is to classify according to the type of misogyny, and the target of the message. The categories of misogynistic behavior are five: *Stereotype & Objectification*, *Dominance*, *Derailing*, *Sexual Harassment & Threats of Violence*, *Discredit*. The target classification is binary, *Active* if the target is specific and *Passive* if the target is generic. The evaluation of the second subtask is the macro average $F_1$ score as follows:

$$F_1 = \frac{F_1(behavior\ category) + F_1(target)}{2} \tag{1}$$

The distribution of misogynistic tweets, behavior categories and target for both Spanish and English tweets in the training corpora is shown in Table 1. Some behavior categories are not very represented, <2% of misogynistic tweets pertain to the category *Derailing*. There are some differences between the two languages in the presence of misogynistic tweets of the category *Dominance*, and >85% of the Spanish misogynistic tweets are *Active,* i.e. refer to a specific target.

**Table 1.** Training corpora statistics.

| | **Spanish** | **English** |
|---|---|---|
| Total tweets | 3307 | 3251 |
| Non-misogynistic | 1658 (50.1%) | 1683 (51.7%) |
| Misogynistic | 1649 (49.1%) | 1568 (48.2%) |
| Stereotype & Objectification | 151 (9.2%) | 137 (8.7%) |
| Dominance | 302 (18.3%) | 49 (3.1%) |
| Derailing | 20 (1.2%) | 29 (1.8%) |
| Sexual Harassment & Threats of Violence | 198 (12%) | 410 (26.1%) |
| Discredit | 978 (59.3%) | 943 (60.1%) |
| Active | 1455 (88.2%) | 942 (60.1%) |
| Passive | 194 (11.8%) | 626 (39.9%) |

## 3 System Description

The system’s approach starts with a preprocessing of the tweets, then each tweet is vectorized by tf-idf features. At the next step the system classifies each tweet for subtask A, the tweets classified as misogynistic in this first subtask are then finally classified for subtask B, a classification by type of misogyny and target of the message.

### 3.1 Preprocessing

The system employs a similar preprocessing to the one in [2], as follows:

- Convert all letters to lowercase.
- Replace multiple concatenated exclamation marks by the keyword MULT_EXCLAMATION.
- Replace multiple concatenated question marks by the keyword MULT_QUESTION.

- Replace multiple concatenated exclamation and question marks by the keyword MIXED_MARKS.
- Replace URLs by the keyword URL.
- Replace user mentions by the keyword USER.
- Replace misogynistic hashtags by the keyword MISO_HASHTAG. A hashtag was considered misogynistic if it appears only in several misogynistic tweets of the training corpora. Misogynistic hashtags in English are those that contain any of the words: *bitch*, *whore*, *hoe*, *cunt*, *womenare*, *womensuck*. Misogynistic hashtags in Spanish are those that contain any of the words: *feminazi*, *perra*.
- Replace the rest of the hashtags by the keyword HASHTAG.

### 3.2 Features and Classifier

The system model uses tf-idf feature vectors, that are built from the tweets' unigrams. For subtask A, that is to classify if a tweet is misogynistic, all tweets are used to extract the vocabulary. On the other hand, for subtask B, that is to classify the misogynistic behavior and the target, only the misogynistic tweets are used to extract vocabulary. Which percentage of terms form the vocabulary according to its document frequency was empirically decided after the development of the system and its experimentation.

The system's classifier is based on support vector machine (SVM), it uses a linear kernel, and, except for the English part of misogynistic behavior classification, it uses a one-vs-one classifier instead of one-vs-rest. The system classifies if a tweet is misogynistic for the first subtask, only if the tweet is classified as misogynistic then the system classifies its behavior category and its target. The system employs a separately trained classifier for misogyny, behavior and target.

## 4 Development and Results

During its development the system was evaluated using 10-fold cross-validation. The range of terms that form the vocabulary according to its document frequency was tuned for each language to maximize accuracy and $F_1$ score. In Table 2, for each subtask, the percentages of minimum and maximum document frequency among the rest of terms are presented. If the value for minimum document frequency is 20%, the terms that formed the vocabulary must not be the 20% least document frequent, analogously with the maximum.

**Table 2.** Minimum and maximum tweet frequency limits for vocabulary.

| Subtask classification | Spanish | | English | |
|---|---|---|---|---|
| | min df | max df | min df | max df |
| Misogynistic tweet | 0% | 45% | 1,0% | 46% |
| Behavior | 0% | 65% | 0,3% | 80% |
| Target | 0% | 55% | 2,5% | 65% |

The exceptional case where a one-vs-rest classifier offered best results than the one-vs-one classifier was for the language English in the misogynistic behavior classification. This can be due to the use of a much larger vocabulary. Only one run of the system over the test corpora was submitted to the shared task [4] for each language under the team name JoseSebastian. The results and rankings are shown in Table 3 and Table 4. Some disparity can be seen between the results for both languages. This system obtained much better results for Spanish than English among other teams. One possible explanation could be the different choice of misogynistic hashtags for the two languages. Another possible explanation could be that misogynistic words are more accentuated than their counterparts specially in Spanish. Although among all the teams and runs submitted, the best accuracy for misogynistic tweets classification was obtained in English, with 91.32% correct guesses.

**Table 3.** Results of subtask A over the test corpora.

| Language | Accuracy | System Ranking | Team Ranking |
|---|---|---|---|
| Spanish | 0.8147 | 1 | 1 |
| English | 0.7493 | 23 | 10 |

**Table 4.** Results of subtask B over the test corpora.

| Language | $F_{avg}$ | $F_{category}$ | $F_{target}$ | System Ranking | Team Ranking |
|---|---|---|---|---|---|
| Spanish | 0.4328 | 0.3234 | 0.5422 | 6 | 3 |
| English | 0.3263 | 0.1477 | 0.5049 | 20 | 6 |

## 5 Conclusions

This paper has described the system submitted to the task on Automatic Misogyny Identification at IberEval 2018. It employs a classifier based on support vector machine, with a linear kernel, trained separately for each subtask. This system obtained pronounced differences in the results over the test corpora between the languages Spanish and English. Some possible explanations could be the different choice of misogynistic hashtags as features, or a pronounced accentuation of misogynistic words in one of the languages respect to the other.

## References

1. Taulé, M., Martí, M. A., Rangel, F. M., Rosso, P., Bosco, C., Patti, V.: Overview of the task on stance and gender detection in tweets on Catalan independence at IberEval 2017. In: Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017), Murcia, Spain, September 19, CEUR Workshop Proceedings. CEUR-WS.org, 2017 (2017)
2. Aineto-García, D., Larriba-Flor, A.M.: Stance detection at ibereval 2017: A biased representation for a biased problem. In: Proceedings of the Second Workshop on Evaluation of

Human Language Technologies for Iberian Languages (IberEval 2017), Murcia, Spain, September 19, CEUR Workshop Proceedings. CEUR-WS.org, 2017 (2017)

3. Hewitt, S., Tiropanis, T., Bokhove, C.: The problem of identifying misogynist language on Twitter (and other online social spaces). In: Proceedings of the 8th ACM Conference on Web Science, Hannover, Germany, May 22. ACM. pp. 333-335 (2016)
4. Fersini, E., Anzovino, M., Rosso, P.: 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
5. Anzovino M., Fersini E., Rosso P.: Automatic Identification and Classification of Misogynistic Language on Twitter. In: Silberztein M., Atigui F., Kornyshova E., Métais E., Meziane F. (eds) Natural Language Processing and Information Systems. NLDB 2018. Lecture Notes in Computer Science, vol 10859 (2018)