Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 65-76 recibido 19-11-2020 revisado 22-01-2021 aceptado 01-02-2021

Categorizing Misogynistic Behaviours in Italian, English and Spanish Tweets

Categorización de comportamientos misóginos en tweets en italiano, inglés y español

Silvia Lazzardi,[1] **Viviana Patti,**[2] **Paolo Rosso**[3]

[1]Dipartimento di Fisica, University of Turin, Italy
[2]Dipartimento di Informatica, University of Turin, Italy
[3]PRHLT research center, Universitat Politècnica de València, Spain
silvia.lazzardi@unito.it, patti@di.unito.it, prosso@dsic.upv.es

Abstract: Misogyny is a multifaceted phenomenon and can be linguistically manifested in numerous ways. The evaluation campaigns of EVALITA and IberEval in 2018 proposed a shared task of Automatic Misogyny Identification (AMI) based on Italian, English and Spanish tweets. Since the participating teams' results were pretty low in the misogynistic behaviour categorization, the aim of this study is to investigate the possible causes. We measured the overlap and the homogeneity of the clusters by varying the number of categories. This experiment showed that the clusters overlap. Finally, we tested several machine learning models both using the original data sets and merging together some categories according to their overlap, obtaining an increase in terms of macro F1.
Keywords: automatic misogyny identification, hate speech online.

Resumen: La misoginia es un fenómeno con múltiples facetas y puede manifestarse lingüísticamente de muchas formas. Las campañas de evaluación de EVALITA e IberEval en 2018 propusieron una tarea compartida de Identificación Automática de Misoginia (AMI) basada en tweets en italiano, inglés y español. Dado que los resultados de los equipos participantes fueron bastante bajos en la categorización del comportamiento misógino, el objetivo de este estudio es investigar las posibles causas. Medimos el solape y la homogeneidad de los clústeres variando el número de categorías. Este experimento mostró que los grupos se solapan. Finalmente probamos varios modelos de aprendizaje automático utilizando los conjuntos de datos originales y fusionando algunas categorías de acuerdo con consideraciones basadas en medidas de similitud y las matrices de confusión de los modelos, obteniendo un aumento de la F1 macro.
Palabras clave: identificación automática de misoginia, mensajes de odio online.

1 Introduction

During the last years, hateful language and in particular the phenomenon of hate speech against women, are exponentially increasing in social media platforms such as Twitter and Facebook (Poland, 2016), spreading across languages and countries. It is becoming a relevant social problem to be monitored, especially considering the results of studies on the usefulness of monitoring contents published in social media in foreseeing sexual crimes, such as the one in (Fulper et al., 2014), which confirms a correlation between the yearly per capita rate of rape and the misogynistic language used in Twitter. As women have often been targets of abusive language and hate speech, they also started to react on social media to both off and online abusive behaviours, for example through the viral use of the #mencallmethings hashtag to share experiences of sexual harassment in online environments and reflections on the perception of women on their freedom of expression (Megarry, 2014). Misogyny, defined as the hate or prejudice against women, can be linguistically manifested in different and various ways, including social exclusion, discrimination, hostility, threats of violence and

ISSN 1989-7553. DOI 10.26342/2021-66-5

sexual objectification (Anzovino, Fersini, and Rosso, 2018).

Given the huge amount of social media data in many languages, the urgency of monitoring misogynistic behaviours and contents online calls the computational linguistics community for a deeper effort on developing tools to automatically identify and categorize hateful content online against women, possibly bringing a multilingual perspective to highlight different viewpoints on how misogyny is not only perceived, but also expressed in different languages and cultures. The recent proposals of the AMI and HatEval shared tasks focusing on the detection of misogyny and hate speech at EVALITA (Fersini, Nozza, and Rosso, 2018) and IberEval (Fersini, Rosso, and Anzovino, 2018), and SemEval (Basile et al., 2019) respectively, can be read in the light of this urgency. Our starting point in this work is the Automatic Misogyny Identification (AMI) shared task proposed in 2018 first at IberEval for Spanish and English (Fersini, Rosso, and Anzovino, 2018), and then at EVALITA for Italian and English (Fersini, Nozza, and Rosso, 2018), to identify misogyny in Twitter texts at different levels of granularity. The task has been articulated in two sub-tasks: a first one devoted to distinguish between messages with misogynous content from not misogynous ones, and a second one with the goal to categorize the misogynous content at a finer grained level. The teams who participated in this shared task, despite having obtained excellent results in terms of the coarse grain binary classification of messages, didn't achieve a good performance in the finer grain sub-task on categorizing misogynistic behaviour, which is important to get a deeper understanding of the multi-faceted misogynistic attitudes. The aim of this study is to shed some light on the reasons behind this.

We focus on investigating how the misogyny categories distribution in the available data sets influences models performances and the relationships among the categories themselves. In particular, we used both unsupervised and supervised machine learning algorithms to answer to the following research questions:

RQ1 Is an unsupervised clustering algorithm able to identify specific patterns and separate data into homogeneous clusters according to the labeled categories of misogynistic behaviour? How do homogeneity and distance change as the number of clusters change?

RQ2 Is it possible to extract empirical information on the similarity relationship among the categories of misogynistic behaviour, by studying a metric among such categories?

RQ3 Using the information obtained from these studies, is it possible to improve the performance of the models on the finer grain misogynistic behaviour classification task?

The paper is organized as follows. Section 2 reports on related literature. Section 3 describes the AMI shared task. Sections 4 and 5 present methods, experiments and results obtained respectively by using unsupervised clustering techniques and supervised machine learning algorithms. Section 6 provides a discussion of the results. Section **??** draws some conclusions.

2 Related Work

Misogynistic language expressed in social media is a multifaceted phenomenon with its own specificity. It has been often treated as an expression of sexist offenses. One of the first works in this area proposed a mixed data set of racism and sexism (Waseem and Hovy, 2016).

A philosophical account of misogyny and sexism has been provided by (Manne, 2017), which arguments that they are distinct. On this line, (Frenda et al., 2019) presented an approach to detect separately both misogyny and sexism analyzing collections of English tweets.

Another important contribution is due to (Farrell et al., 2019), that investigated how misogynistic ideas spread within and across Reddit communities by building lexicons of hate which describe such misogynistic behaviours. Two very recent works addressed the problem of misogyny identification in Spanish tweets comparing the performance of several models on the HatEval data set (Plaza-Del-Arco et al., 2020), and applying sentiment analysis and social computing technologies (García-Díaz et al., 2021). In the latter work, the authors have

compiled the balanced MisoCorpus-2020 corpus. Misogyny identification has been also addressed from a multimodal perspective in (Gasparini et al., 2018), where both visual and texture features are analyzed to classify sexist contents, and in a multilingual perspective in (Pamungkas, Basile, and Patti, 2020).

3 Task and Data Sets Description

The Automatic Misogyny Identification (AMI) shared task, presented for the first time at the IberEval 2018 and then at EVALITA 2018, proposes misogyny identification, misogynistic behaviour categorization, and target classification both from Spanish, English and Italian tweets.

3.1 AMI Task

The AMI shared task of EVALITA and IberEval is subdivided into two sub-tasks:

- **Sub-task A - Misogyny identification**: binary classification of misogynous contents from the non misogynistic ones.
- **Sub-task B - Misogynistic behaviour and target classification**: identification of the particular type of misogynous behaviour against women and classification of the target of the insult, that can be *active* or *passive* respectively whether it is addressed to a specific user or to a group of women.
  The different misogynous behaviours are classified into the following categories:
  - **Derailing**: to justify male abuse of women.
  - **Discredit**: slurring without any other intentions.
  - **Dominance**: to emphasize the gender superiority of men over women.
  - **Sexual harassment**: to describe actions as sexual advances, requests for sexual services.
  - **Stereotype**: to describe women using oversimplified ideas as the physical appeal and/or comparison with narrow standards.

In Table 1 are reported the different categories of misogynistic behaviour, together with an example for each of them.

Category	Example
Derailing	@yesallwomen wearing a tiny skirt is asking for it. Your teasing a (hard working, taxes paying) dog with a bone. That's cruel. #YesAllMen
Discredit	@melaniatrump stupid fuc**ing bi**ch
Dominance	Women are inferior to men..so shut up please
Sexual harassment	I wanna touch your te**ts..you can't imagine what I can do on your body.
Stereotype	Women play your role.. #gobacktothekitchen Big te**ts and no brain..this is what I want as wife!

Table 1: Misogynous categories and examples of tweets from the AMI dataset.

In this study we deal with the part of Sub-task B devoted to the misogynistic behaviour classification. Regarding the evaluation measure used to rank the submissions for this sub-task, the Macro F_1-score was used. The misogynistic behaviour categorization is obtained as the average of F_1-score of each misogynous category ($F_1\_misogyny$).

3.2 Training and Test Data Sets

The data sets used in this study are collections of tweets. We used the Italian and the English data set from EVALITA and the Spanish one from IberEval. Each data set is distinguished in a training and a test set. The EVALITA data set contains respectively 4000 and 1000 tweets for the train and the test sets, while the IberEval one includes 3302 tweets in the training set and 831 in the test set (Fersini, Nozza, and Rosso, 2018) (Fersini, Rosso, and Anzovino, 2018). Each

tweet has been provided with several fields: the *id*; the Twitter *text*; *misogynous* which defines whether the tweet is misogynous or not (1 if the tweet is misogynous, 0 otherwise); *misogyny_category*, a label to denote the misogynous behaviour (0 if it's not misogynous) and finally the target (*active*, *passive* or '0' if the tweet is not misogynous).

3.2.1 Categories Distribution

Table 2 reports the number of samples for each misogynous category while Figure 1 shows the categories distribution across the corpus. We can observe that data present many inhomogeneities both in the training and in the test sets.

In all the data sets, about half of the tweets are not misogynous (with percentages ranging from 49.1% to 55.4%). The other half contains tweets belonging to the five labeled misogyny categories according to different percentages. The less represented category is *derailing*, with percentages ranging from just 0.2% for the Italian test set to 2.3% for the English training set. It follows *dominance* which represents just the 1.8% of the Italian training set. The English and the Spanish data set contain more *dominance* tweets, with a maximum of 12.4% for the Spanish test set. About the other three categories (*discredit*, *sexual harassment* and *stereotype*), we can note that in the Italian data set tweets with *stereotype* are present in an higher percentage respect to the English and Spanish data sets (being 16.7% in the Italian train set versus 4.5% and 4.6% in the English and Spanish ones). Otherwise, in the English and Spanish data sets *discredit* is significantly more represented.

4 Clustering

In this section we describe the clustering methods used to evaluate the data sets from an unsupervised perspective and the obtained results. We extracted features from the data sets using two methods: Bag of Words (BoW) and tf-idf. Then we applied three different clustering algorithms: K-means, Spectral Clustering and Agglomerative Clustering. As a second experiment, we computed the inter-category distance matrices using the cosine similarity.

4.1 Experimental Setting

In order to answer RQ1, we carried out a a first experiment where we clustered the tweets of the data sets without considering the category label.

Our aim was to investigate: (i) how the distance among the clusters changes by varying their number; and (ii) how the misogynistic behaviour categories are distributed in the clusters. In order to address the above points, we computed two scores:

- *Silhouette score*: a measure of the distance among clusters that is computed using the mean inter-cluster distance (a) and the mean nearest-cluster distance (b) for each tweet. Given the data matrix $\mathbf{X}= \{x_1, \ldots, x_n\}$ and the label vector $\mathbf{y}$, we define it as:

$$Sil(\mathbf{X}, \mathbf{y}) = \sum_{i=1}^{n} \frac{b_i - a_i}{max(a_i, b_i)} \quad (1)$$

  The Silhouette score takes values $\in \{-1, 1\}$. Values near 0 indicate that the clusters overlap each other.

- *Homogeneity score*, a measure of how many clusters contain only tweets that belong to a misogynistic behaviour category. Its value ranges from 0 to 1, where 1 corresponds to a perfectly homogeneous labeling.

Since there are mainly two categories underrepresented in the data sets (*derailing* and *dominance*), we run each clustering algorithm by varying the number of clusters from 6 (not misogynistic plus the 5 misogynistic behaviour categories) down to 4 (not misogynistic plus 3 misogynistic behaviour categories), in order to see if there is an improvement in the scores, ignoring the two least represented categories. Given a fixed number of clusters, we computed the Silhouette and the Homogeneity scores. Moreover, we represent the distribution of the misogynistic behaviour categories within each cluster.

In order to answer RQ2, we carried out the experiment from the opposite perspective. We considered perfectly homogeneous clusters and computed distances among the misogynistic behaviour categories using the cosine similarity. We didn't use the Euclidean distance due to the high dimensionality of the feature vectors.

4.2 Results

First we calculated the Silhoutte and Homogeneity scores resulting from the analy-

	Not mis.	Derailing	Discredit	Dominance	Sexual harass.	Stereotype	**Tot**
Italian							
Train	2172	24	634	71	431	668	**4000**
Test	491	2	104	61	167	175	**1000**
English							
Train	2214	92	1014	148	351	179	**4000**
Test	540	10	141	123	43	140	**1000**
Spanish							
Train	1658	19	977	301	197	150	**3302**
Test	416	5	286	54	50	17	**831**

Table 2: Number of samples for each misogynous category for Italian, English and Spanish both for training and test data sets.

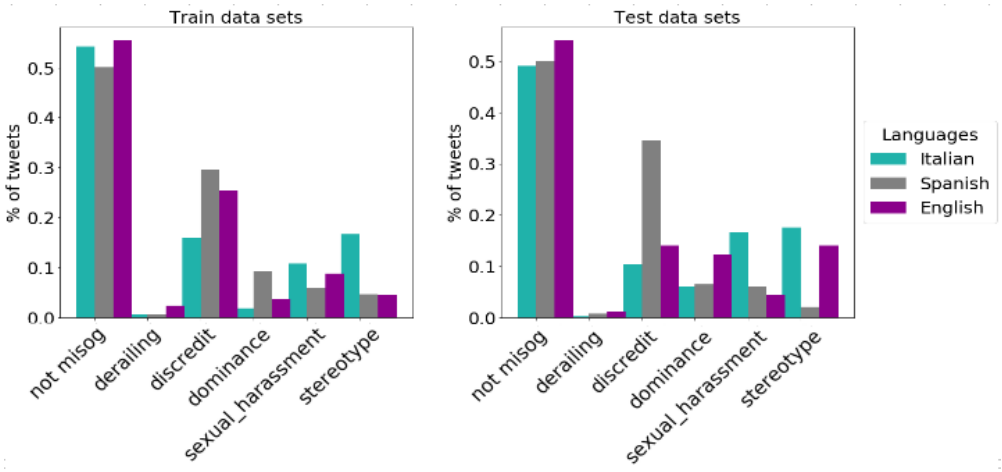

Figure 1: Categories distribution of the data sets: training (left) and test (right).

sis of the tf-idf feature vectors by decreasing the number of clusters from 6 down to 4. The Homogeneity score was always very low, which indicates that the clustering algorithms couldn't recognize a significant difference between the different categories. Then we did the same using the BoW feature vectors. We used the statistical test of Wilcoxon, to see if the difference using tf-idf and BoW was statistically significant. With a p-value of 0.66 we can say it wasn't. In Figure 2 we show the results obtained with the BoW feature vectors (similar results were obtained with the tf-idf feature vectors).

Figure 3 shows how the tweets belonging to different misogynistic behaviour categories are distributed among the clusters. Since all clusters have at least a sample for each category, it is clear why the homogeneity score was very low.

Finally, we computed the cosine similarity among the misogynistic behaviour categories for the three data sets. The results are illustrated in Figure 4. In agreement with the clustering results, we can appreciate how difficult is to separate the misogynistic behaviour categories. Interestingly *derailing* seems to be the most well defined category, followed by *dominance*. This could be due to their smaller number of tweets (i.e., less variability in their feature vectors). An interesting insight is the overlapping between *discredit* and *dominance* and between *derailing* and *sexual harassment*.

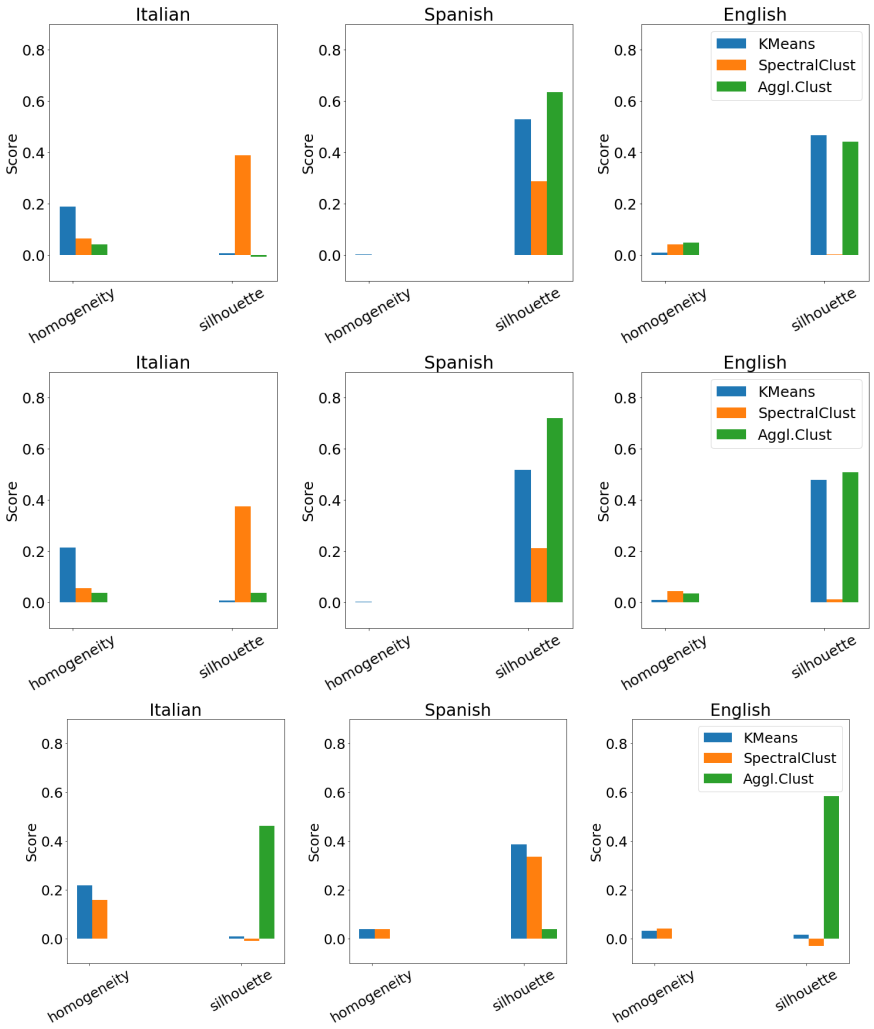

Figure 2: Silhouette and Homogeneity scores for Italian, Spanish, and English using BoW feature vectors by varying the number of clusters. From left to right: 6, 5 and 4 clusters. Train and test data sets were merged together.

5 Behaviour Classification

In this section we comment on the machine learning algorithms we used on the EVALITA and IberEval data sets, and the results we obtained.

5.1 Experimental Setting

For this experiment we tried two approaches to built feature vectors, using different kinds of features. In the first case, we extracted a list of words processing each tweet of the training data set. We did not consider those words with a frequency smaller than 0.001. In the second case, we used as list of features those taken from *Hurtlex*, a multilingual lexicon of offensive, aggressive, and hateful words (Bassignana, Basile, and Patti, 2018).

Hurtlex categorizes words according to 17 categories, from which we selected the most relevant for our purpose, as reported in Table 3.

Label	Description
PR	words related to prostitution
ASM	male genitalia
ASF	female genitalia
DDP	cognitive disabilities and diversity
DDF	physical disabilities and diversity

Table 3: S *Hurtlex*: selected categories (labels and descriptions) to build the vocabulary of features.

To convert the tweets to feature vectors we used tf-idf and BoW. To classify the tweets we trained different We run these models and compared their performance, using tf-idf and BoW, as well as the features extracted from

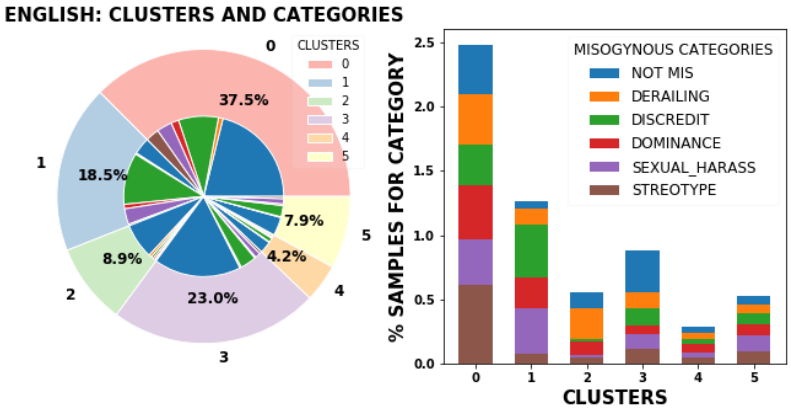

Figure 3: Example of misogynistic behaviour categories among 6 clusters using K-means on the English data set.

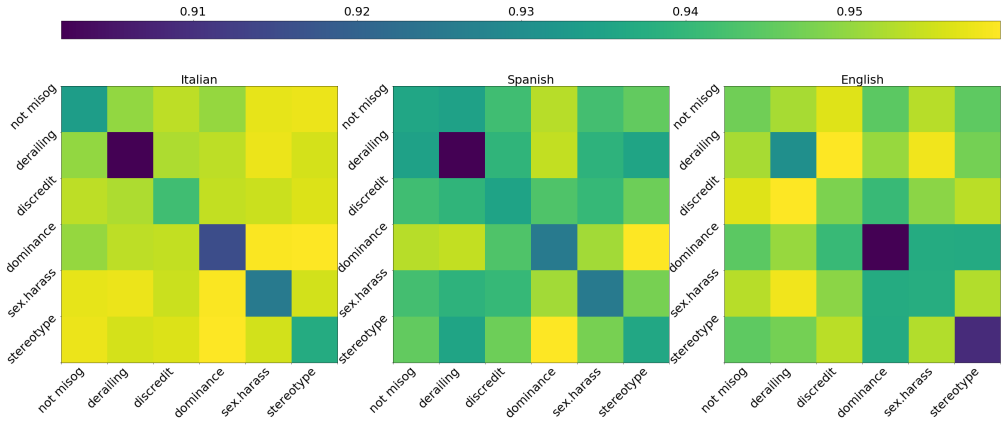

Figure 4: From left to right: cosine inter-category similarities computed respectively for Italian, Spanish and English data sets.

the tweets of the data sets considering the *Hurtex* categories of Table 3.

In order to give an answer to RQ3, we calculated the Macro F1-score considering the 5 misogynistic behaviour categories, and after merging the most similar ones: *dominance* with *discredit* (domin+discr) and *derailing* with *sexual harassment* (der+sex har). We finally show the confusion matrices of the best performing models.

5.2 Results

Table 4 shows the performance of our models on the test sets in terms of Macro F1-score. Our best results overcome the official results for the sub-task B both of EVALITA and IberEval: Macro $F_1(misogyny\_category)$ from 0.534 to 0.694 for Italian, from 0.339 to 0.347 for Spanish and from 0.361 to 0.470 for English, respectively. The results in general are not high, especially for English and Spanish. This is due to the under represented misogynistic behaviour categories in all the training sets, such as in the case of *derailing* and *dominance*, that turned out to be very difficult to detect for the models. With regard to the choice of the feature vector to use (BoW vs. tf-idf), we compared the macro F1-scores obtaining significance levels above the threshold of 5%. Therefore, it is not possible to say what technique is better. Figure 5 shows the confusion matrices of the models that obtained the best results (Linear SVM for Italian and Spanish, while for English tf-idf combined with SGD_selected, which differs from the standard SGD for the parameters settings, optimized through a python

	Italian			
Model	**BoW**	**BoW[H]**	**tf-idf**	**tf-idf[H]**
Lin SVM	0.576	0.251	**0.694**	0.254
Rbf SVM	0.57	0.249	0.222	0.245
LR	0.573	0.25	0.552	0.242
SGD	0.596	0.06	0.51	0.251
SGD_ selected	0.601	0.25	0.588	0.241
DT	0.571	0.247	0.519	0.251
	Spanish			
Model	**BoW**	**BoW[H]**	**tf-idf**	**tf-idf[H]**
Lin SVM	0.29	0.206	**0.347**	0.206
Rbf SVM	0.306	0.206	0.021	0.206
LR	0.241	0.002	0.315	0.002
SGD	0.257	0.0	0.209	0.0
SGD_ selected	0.248	0.002	0.295	0.002
DT	0.227	0.002	0.248	0.002
	English			
Model	**BoW**	**BoW[H]**	**tf-idf**	**tf-idf[H]**
Lin SVM	0.383	0.2	0.468	0.2
Rbf SVM	0.322	0.197	0.009	0.197
LR	0.353	0.0	0.452	0.0
SGD	0.286	0.0	0.305	0.0
SGD_ selected	0.339	0.0	**0.47**	0.0
DT	0.259	0.0	0.271	0.0

Table 4: F_1 for each misogynistic behaviour category and Macro $F_1(misogyny\_category)$ of the different models with the BoW and ft-idf feature vectors. We compare the results using the list of words obtained from the tweets of the data sets with the one built using [H]urtlex.

library[1]. We can observe that for the Italian data set the most problematic category is *derailing*, for which half samples are misclassified as not-misogynist. Moreover, the 11% of tweets labeled as *dominance* are classified as *derailing* and the 16% of tweets from *sexual harassment* are misclassified as *stereotype*. Regarding Spanish, the most often misclassified misogynistic behaviour category is *stereotype*, mainly confused with *discredit*, which in turn is often confused with *dominance*. We found very similar misclassification patterns in the English data set. These results are particularly interesting because in line with what we found previously. Since, especially for the Spanish and the English data sets, there is a certain similarity among *discredit* and *dominance*, *derailing* and *sexual harassment*. Therefore, decided to merge these two pairs of misogynistic behaviour categories. Table 5 shows the obtained results with tf-idf feature vectors. We can see that the F_1 scores always increase, especially for the English data set. The best Macro F1 scores in this case are respectively 0.767, 0.45 and 0.663 for the three data sets, being 0.691, 0.347 and 0.47 before merging them.

6 Discussion

The aim of this study is to understand the reasons behind the low scores obtained by the participants on the misogynistic behaviour classification at the AMI shared task, proposed at the EVALITA and IberEval evaluation campaigns in 2018.

To have a better understanding of the problem, we first studied the categories distribution for the different data sets (Figure 1). First of all, we note that there are many inhomogeneities among the misogynistic behaviour categories in the data sets. *Derailing* is the most obvious example of underrepresented category, followed by *dominance*. *Discredit* and *dominance* have a greater percent-

[1]available at `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html`

	Italian				
Model	**Not Mis**	**domin+discr**	**der+sex_har**	**stereotype**	**MacroF1**
Lin SVM	0.699	0.709	0.74	0.851	**0.767**
Rbf SVM	0.055	0.988	0.041	0.069	0.366
LR	0.754	0.715	0.716	0.817	0.749
SGD	0.776	0.642	0.71	0.777	0.71
SGD_selected	0.735	0.739	0.716	0.84	0.765
DT	0.674	0.485	0.574	0.834	0.631
	Spanish				
Model	**Not Mis**	**domin+discr**	**der+sex_har**	**stereotype**	**MacroF1**
Lin SVM	0.656	0.554	0.649	0.118	0.44
Rbf SVM	0.981	0.129	0.0	0.0	0.043
LR	0.781	0.666	0.509	0.176	**0.45**
SGD	0.853	0.669	0.281	0.118	0.356
SGD_selected	0.764	0.674	0.474	0.059	0.402
DT	0.536	0.607	0.368	0.118	0.364
	English				
Model	**Not Mis**	**domin+discr**	**der+sex_har**	**stereotype**	**MacroF1**
Lin SVM	0.13	0.54	0.6	0.85	**0.663**
Rbf SVM	0.996	0.03	0.0	0.0	0.01
LR	0.254	0.672	0.564	0.721	0.652
SGD	0.537	0.706	0.4	0.179	0.428
SGD_selected	0.156	0.675	0.618	0.7	0.665
DT	0.493	0.347	0.382	0.15	0.293

Table 5: F_1 for each misogynistic behaviour category and Macro $F_1(misogyny\_category)$ for each model (using tf-idf feature vectors) on the Italian, Spanish and English data sets where *derailing* was merged with *sexual harassment* (der+sex har), and *dominance* with *discredit* (domin+discr).

Language	**BestTeam**	**OurModel**
Italian	0.555	**0.694**
Spanish	0.339	**0.347**
English	0.292	**0.470**

Table 6: Macro $F_1(misogyny\_category)$ of the best performing teams participating in the AMI shared task at EVALITA and IberEval in Italian, English and Spanish data sets vs., our best results (see Table 4).

age in the English and the Spanish data sets compared to the Italian one. The remaining categories (*derailing*, *sexual harassment* and *stereotype*) are present in a smaller percentage. We performed an unsupervised cluster analysis on the data sets, using K-means, Spectral Clustering and Agglomerative Clustering as algorithms. We measured the Silhouette and the Homogeneity scores by varying the number of clusters from 6 (not misogynistic plus the 5 misogynistic behaviour categories) down to 4 (not misogynistic, *stereotype*, *dominance + discredit*, *derailing + sexual harassment*). The obtained Silhouette scores from the cluster analysis are low, showing that the clusters are overlapping. The Homogeneity scores are near to 0, which indicate that the misogynistic behaviour categories are not easily separable in the different clusters (Figure 4), i.e., the clustering methods are not able to find hidden patterns useful to separate these categories. This is not surprising since there is a high overlap in some of them. The annotation of the tweets

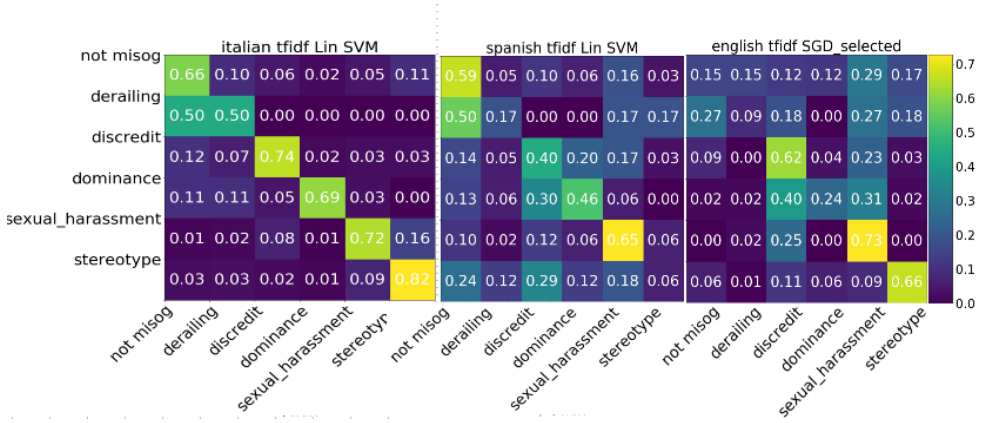

Figure 5: From left to right: confusion matrices of the models on the test sets for Italian, Spanish and English. On the x axis we have the predicted labels, while on the y axis the true ones.

may be highly subjective in some of the cases and consequently also for the models is difficult distinguish among the misogynistic behaviour categories. Doing the reverse, i.e., computing the inter-category cosine similarity among the categories, we noticed that in many cases the distance between the tweets belonging to a given category and those belonging to other misogynistic behaviour categories was pretty small. That is, in all the data sets some of the categories were closer: *derailing* to *sexual harassment*, and *discredit* to *dominance*. The best model for Italian and Spanish used tf-idf together with Lin SVM, while for English the best classification model was SGD_selected. Our highest F_1 scores are 0.694 for Italian, 0.47 for English and 0.347 for Spanish. The teams which achieved the best results respectively in the EVALITA and IberEval challenges are *bakarov* for Italian (0.555) (Bakarov, 2018), *himami* for English (0.292) (Ahluwalia et al., 2018) and *14-exlab* (0.339) (Pamungkas et al., 2018) for Spanish. Our best performance obtained better results in all the three languages (see Table 6).

Finally, what depicted in Figure 5 is in agreement with the considerations previously made on the inter-category distance matrices. Since *derailing* and *dominance* turned out to be the hardest categories to be detected, and given that the confusion matrices show that they are mostly misclassified as *sexual harassment* and *discredit*, respectively, we decided to merge them. As result, the F_1 scores improved significantly for all the data sets, increasing (for the best configurations) from 0.694 to 0.767 for Italian, from 0.47 to 0.663 for English, from 0.347 to 0.45 for Spanish.

7 *Conclusions*

In this paper we studied the multilingual Twitter data sets of misogynous texts released for the AMI shared task at EVALITA and IberEval 2018. Our main purpose was providing some insights on the fine-grained misogynistic behaviour, also in light of the difficulties encountered by the participants in identifying such categories. We found many inhomogeneities among the categories, which surely represented an important source of bias for the models, as it could be seen also from the confusion matrices. With respect to RQ1 and RQ2: from the clustering experiments on each data set, low Silhouette scores indicated the difficulty in separating the tweets of the different misogynistic behaviour categories (each cluster contained overlapping of categories). We also obtained very low Homogeneity scores indicating the same. We trained several machine learning algorithms improving the results of the best performing teams in both EVALITA and IbelEval. With respect to RQ3: since we found that tweets belonging to some of the categories are near in the features space for the data sets in the three languages, we trained and tested our models after merging two pairs of overlapping categories (in accordance with the insights from the inter-category cosine distance matrices and the confusion matrices). This allowed to improve even further the already good results that we obtained with the 5 misogynistic behaviour categories.

Acknowledgements

The work of S. Lazzardi was partially carried out at the Universitat Politècnica de València within the framework of the Erasmus+ program, Erasmus Traineeship 2018/19 funding. The work of P. Rosso was partially funded by the Spanish MICINN under the research project MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31). The work of V. Patti was partially funded by the research projects "STudying European Racial Hoaxes and sterEOTYPES" (STERHEOTYPES, under the call "Challenges for Europe" of VolksWagen Stiftung and Compagnia di San Paolo) and "Be Positive!" (under the 2019 "Google.org Impact Challenge on Safety" call).

References

Ahluwalia, R., H. Soni, E. Callow, A. C. Nascimento, and M. De Cock. 2018. Detecting hate speech against women in english tweets. In *EVALITA@ CLiC-it*.

Anzovino, M., E. Fersini, and P. Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, and F. Meziane, editors, *Natural Language Processing and Information Systems*, pages 57–64, Cham. Springer International Publishing.

Bakarov, A. 2018. Vector space models for automatic misogyny identification. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:211.

Basile, V., C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti. 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*, pages 54–63, Minneapolis, Minnesota, USA, June. ACL.

Bassignana, E., V. Basile, and V. Patti. 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.

Farrell, T., M. Fernandez, J. Novotny, and H. Alani. 2019. Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, pages 87–96, New York, NY, USA. ACM.

Fersini, E., D. Nozza, and P. Rosso. 2018. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Fersini, E., P. Rosso, and M. Anzovino. 2018. 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) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 214–228. CEUR-WS.org.

Frenda, S., B. Ghanem, M. Montes-y-Gómez, and P. Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on Twitter. *Journal of Intelligent and Fuzzy Systems*, 36(5):4743–4752.

Fulper, R., G. L. Ciampaglia, E. Ferrara, Y. Ahn, A. Flammini, F. Menczer, B. Lewis, and K. Rowe. 2014. Misogynistic language on twitter and sexual violence. In *Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM)*.

García-Díaz, J. A., M. Cánovas-García, R. Colomo-Palacios, and R. Valencia-García. 2021. Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114:506 – 518.

Gasparini, F., I. Erba, E. Fersini, and S. Corchs. 2018. Multimodal classifi-

cation of sexist advertisements. In *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications, ICETE 2018 - Volume 1: DCNET, ICE-B, OPTICS, SIGMAP and WINSYS, Porto, Portugal, July 26-28, 2018*, pages 565–572. SciTePress.

Manne, K. 2017. *Down Girl: The Logic of Misogyny*. Oxford University Press.

Megarry, J. 2014. Online incivility or sexual harassment? conceptualising women's experiences in the digital age. *Women's Studies International Forum*, 47:46 – 55.

Pamungkas, E. W., V. Basile, and V. Patti. 2020. Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study. *Information Processing & Management*, 57(6):102360.

Pamungkas, E. W., A. T. Cignarella, V. Basile, V. Patti, et al. 2018. 14-exlab@unito for AMI at ibereval2018: Exploiting lexical knowledge for detecting misogyny in English and Spanish tweets. In *3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2018*, volume 2150, pages 234–241. CEUR-WS.

Plaza-Del-Arco, F.-M., M. D. Molina-González, L. A. Ureña López, and M. T. Martín-Valdivia. 2020. Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies. *Rossana Damiano and Viviana Patti and Chloé Clavel and Paolo Rosso (Eds.), Special Section on Emotions in Conflictual Social Interactions, ACM Transactions of Internet Technology*, 20(2).

Poland, B. 2016. *Haters: Harassment, abuse, and violence online*. U of Nebraska Press.

Waseem, Z. and D. Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June. ACL.