Overview of the Task on Automatic Misogyny Identification at IberEval 2018

E. Fersini[1], P. Rosso[2], and M. Anzovino[1]

[1] DISCo, University of Milano-Bicocca
[2] PRHLT Research Center, Universitat Politècnica de València

Abstract. Automatic Misogyny Identification (AMI) is a new shared task proposed for the first time at the IberEval 2018 evaluation campaign. The AMI task proposes misogyny identification, misogynistic behaviour categorization and target classification both from Spanish and English tweets. We have received a total of 32 runs for English and 24 for Spanish, submitted by 11 different teams from 5 countries. We present here the datasets, the evaluation methodology, an overview of the proposed systems and the obtained results. Finally, we draw some conclusions and discuss future work.

Keywords: Automatic Misogyny Identification · Twitter · Spanish · English.

1 Introduction

During the last years, the role of the women within the society has been given more attention, unfortunately even because of several cases of real hatred against them. According to the Pew Research Center Online Harassment report (2017) [1], we can highlight that 41% of people were personally targeted, whose 18% were subjected to serious kinds of harassment because of the gender (8%) and that women are more likely to be targeted than men (11% vs 5%). With the exponential growth of social media and microblogging platforms, hate against women is taking place even more frequently, highlighting how misogyny can be spread with almost no barrier [3, 4]. Misogyny, defined as the hate or prejudice against women, can be linguistically manifested in numerous ways, including social exclusion, discrimination, hostility, threats of violence and sexual objectification [2]. Given this important social problem, the Automatic Misogyny Identification (AMI) task has been proposed at IberEval 2018. The main goal of AMI is to distinguish misogynous contents from non-misogynous ones, to categorize misogynous behaviors and finally to classify the target of a tweet written in Spanish or English. In particular, the AMI shared task is organized according to two main subtasks:

- **Subtask A - Misogyny Identification**: discrimination of misogynistic contents from the non-misogynistic ones. Examples of misogynistic and non-misogynistic tweets are reported in the following:

Table 1. Examples of misogynist and non-misogynist tweets

Target	**Text**
Misogynistic	Ive yet to come across a nice girl. They all end up being bit**es in the end.
Non-misogynistic	@RealMadrid you are a bi*ch!

– **Subtask B - Misogynistic Behavior and Target Classification**: recognition of the targets that can be either specific users or groups of women together with the identification of the type of misogyny against women.

Regarding the misogynistic behavior, a tweet must be classified as belonging to one of the following categories:

– *Stereotype & Objectification*: a widely held but fixed and oversimplified image or idea of a woman; description of women's physical appeal and/or comparisons to narrow standards.
– *Dominance*: to assert the superiority of men over women to highlight gender inequality.
– *Derailing*: to justify woman abuse, rejecting male responsibility; an attempt to disrupt the conversation in order to redirect women's conversations on something more comfortable for men.
– *Sexual Harassment & Threats of Violence*: to describe actions as sexual advances, requests for sexual favours, harassment of a sexual nature; intent to physically assert power over women through threats of violence.
– *Discredit*: slurring over women with no other larger intention.

Examples of Misogynistic Behaviors are reported in Table 2.

Table 2. Examples of tweets for each misogyny category

Misogyny Category	**Text**
Discredit	@melaniatrump stupid fuc**ing bi**ch
Stereotype	Women are good only into the kitchen..#makemeasandwich
Objectification	Big te**ts and no brain..this is what I want as wife!
Sexual Harassment	I wanna touch your te**ts..you can't immagine what I can do on your body.
Threats of Violence	Stupid bi**ch I'll put you down on the floor and I'll rape you! You should be scared!
Dominance	Women are inferior to men..so shut up please!
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

Concerning the target classification, the main goal is to classify each misogynous tweet as belonging to one of the following two target categories:

– *Active (individual)*: the text includes offensive messages purposely sent to a specific target;
– *Passive (generic)*: it refers to messages posted to many potential receivers (e.g groups of women).

Examples of targets of misogynistic tweets are reported in Table 3.

Table 3. Examples of targets.

Target	Text
Active	@JulieB stupid crazy psychopathic woman..you should die...
Passive	Women: just an inferior breed!!!

2 Training and Testing Data

In order to provide training and testing data both for Spanish and English, three approaches were employed to collect misogynistic text on Twitter:

– Streaming download using a set of representative keywords, e.g. *bi**h, w**re, c*nt*
– 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 collection phase started on 20th of July 2017 and ended on 30th of November 2017, leading to a final corpus of 83 million tweets for English and 72 millions for Spanish. Next, among all the collected texts we selected a subset of tweets querying the database with the co-presence of keywords. The labeling phase involved two steps: firstly, a gold standard was composed and labeled by two annotators, whose cases of disagreement were solved by a third experienced contributor. Secondly, the remaining tweets were labeled through a majority voting approach by external contributors on the CrowdFlower[3] platform. The gold standard has been used for the quality control of the judgements throughout the second step.

For the AMI task, at the end of the labelling phase, we provided one corpus for Spanish and one corpus for English to all the participants. Each corpus is distinguished in Training Set and Test datasets. Regarding the training data, the Spanish corpus is composed of 3307 tweets, while the English one is composed of 3251 tweets. Concerning the test data, we provided 831 tweets for Spanish and 726 for English. The training data provided are tab-separated, reporting the following fields:

[3] Now Figure Eight: https://figure-eight.com/

"id" "text" "misogynous" "misogyny_category" "target"

where:

- **id** denotes a unique identifier of the tweet.
- **text** represents the tweet text.
- **misogynous** defines if the tweet is misogynous or not misogynous; it takes values as 1 if the tweet is misogynous, 0 if the tweet is not misogynous.
- **misogyny_category** denotes the type of misogynistic behaviour; it takes value as:
  - *stereotype*: denotes the category Stereotype & Objectification;
  - *dominance*: denotes the category Dominance;
  - *derailing*: denotes the category Derailing;
  - *sexual_harassment*: denotes the category Sexual Harassment & Threats of Violence;
  - *discredit*: denotes the category Discredit;
  - *0* if the tweet is not misogynous.
- **target** denotes the subject of the misogynistic tweet; it takes value as:
  - *active*: denotes a specific target (individual);
  - *passive*: denotes potential receivers (generic);
  - *0* if the tweet is not misogynous.

Concerning the test data, only "id" and "text" have been provided to the participants. Examples of all possible allowed combinations are reported in the following. Additionally to the field "id", we report all the combinations of labels to be predicted, i.e. "misogynous", "misogyny_category" and "target":

0 0 0
1 stereotype active
1 stereotype passive
1 dominance active
1 dominance passive
1 derailing active
1 derailing passive
1 sexual_harassment active
1 sexual_harassment passive
1 discredit active
1 discredit passive

The label distribution related to the Training and Test datasets are reported in Table 4. While the distribution of labels related to the field "misogynous" is balanced (for both languages), the classes for related to the other fields are quite unbalanced. Regarding the "misogyny_category", the most frequent label is related to the category *discredit* both for Spanish and English. Concerning the "target", the most predominant victims are specific users (*active*) for Spanish with a strong imbalanced distribution.

Table 4. Distribution of labels for "misogynous", "misogyny_category" and "target" on the Training and Test datasets

	Training		**Testing**	
	Spanish	**English**	**Spanish**	**English**
Misogynistic	1649	1568	415	283
Non-misogynistic	1658	1683	416	443
Discredit	978	943	287	123
Sexual Harassment & Threats of Violence	198	410	51	32
Derailing	20	29	6	28
Stereotype & Objectification	151	137	17	72
Dominance	302	49	54	28
Active	1455	942	370	104
Passive	194	626	45	179

3 Evaluation Measures and Baseline

Considering the distribution of labels of the dataset, we have chosen different evaluation metrics. In particular, we distinguished as follows:

Subtask A. Systems have been evaluated on the field "misogynous" using the standard accuracy measure, and ranked accordingly. Accuracy has been computed as follows:

$$Accuracy = \frac{\text{number of correctly predicted instances}}{\text{total number of instances}} \tag{1}$$

Subtask B. Each field to be predicted has been evaluated independently on the other using a Macro F1-score. In particular, the Macro F1-score for the "misogyny_category" field has been computed as average of F1-scores obtained for each category (stereotype, dominance, derailing, sexual_harassment, discredit), estimating $F_1(misogyny\_category)$. Analogously, the Macro F1-score for the "target" field has been computed as average of F1-scores obtained for each category (active, passive), $F_1(target)$. The final ranking of the systems participating to subtask B was based on the Average Macro F1-score (F_1), computed as follows:

$$F_1 = \frac{F_1(misogyny\_category) + F_1(target)}{2} \tag{2}$$

In order to compare the submitted runs with a baseline model, we provided a benchmark (AMI-BASELINE) based on Support Vector Machine trained on a unigram representation of tweets. In particular, we created one training set for each field to be predicted, i.e. "misogynous", "misogyny_category" and "target", where each tweet has been represented as a bag-of-words (composed of 1000 terms) coupled with the corresponding label. Once the representations have been obtained, Support Vector Machines with linear kernel have been trained, and provided as AMI-BASELINE.

4 Overview of the Submitted Approaches

As far is concerned with the participants, we have received a total of 32 runs for English and 24 for Spanish, submitted by 11 different teams from 5 countries (Spain, Italy, United States, Ireland and United Kingdom). Table 5 provides an overview of the teams, the number of submitted runs for Spanish and English, and finally the subtasks addressed.

Table 5. Team overview

Team Name	English Runs	Spanish Runs	SubTask A	SubTask B
14-exlab [6]	5	5	✓	✓
IxaTeam [8]	1	1	✓	✗
GrCML2016 [11]	3	-	✓	✓
JoseSebastian [5]	1	1	✓	✓
vic [12]	5	3	✓	✓
ITT [9]	2	-	✓	✓
SB [10]	5	5	✓	✓
meybelraul [14]	5	5	✓	✓
AnotherTeam [15]	1	1	✓	✓
resham [7]	1	-	✓	✓
Amrita_CEN [13]	3	3	✓	✗

Each team had the chance to submit up to five runs for English and five runs for Spanish. Runs could be constrained, where only the provided training data and lexicons were admitted, and unconstrained, where additional data for training were allowed.

Concerning the English language, all the teams participated in Subtask-A and nine of them in Subtask-B. Regarding the Spanish language, eight teams submitted at least one run for Subtask-A and seven of them in Subtask-B. All the teams submitted constrained runs (both for Spanish and English), while only one team has provided unconstrained runs. Following, we provide an outline of the systems participating at the AMI task, focusing on the proposed classification approaches and features used for training the models.

Regarding the classification approaches, the majority of participants exploited Support Vector Machines (SVM) and Ensemble of Classifiers (EoC) both for Subtask-A and Subtask-B. SVMs have been experimented by meyberaul, AnotherTeam, _vic_, SB, 14-exlab and JoseSebastian, while EoC have been investigated by SB, ITT, resham and GrCML2016. Deep learning classification approaches have been adopted by a subset of participants, i.e. resham, IxaTeam and Amrita_CEN.

Concerning the feature set, n-grams and embeddings are the most used ones. Teams using SVM represented the tweets with n-gram based approaches, whereas teams using different kinds of deep learning methods basically used word embeddings. N-grams representations have been experimented by AnotherTeam, ITT, resham, JoseSebastian, _vic_, SB, meyberaul and 14-exlab. Embeddings have been investigated by Amrita_CEN, GrCML2016, resham, IxaTeam and AnotherTeam.

Systems using n-gram representations have frequently adopted several additional linguistic characteristics such as stylistic, structural, lexical and affective features.

5 System Results

We evaluated both Subtask-A and Subtask-B independently. In the following subsections, we will show results separately for the evaluation of each subtask and for each language. Results are given in terms of accuracy for Subtask-A and Maccro Average F-Measure for Subtask-B. Concerning Subtask-B, also detailed results for each considered label are provided.

5.1 Subtask A

Eleven teams participated in Subtask-A for English, presenting 32 runs, and 9 teams participated for Spanish, presenting 24 submissions. In Table 6, the Accuracy achieved by all runs is shown, as well as the AMI-BASELINE. At the bottom of the table some basic statistics are provided: minimum (min), maximum (max), mean, median, standard deviation (stdev), first quartile (q1) and third quartile (q3).

Among the 32 runs for English, 14 teams achieved an Accuracy above the AMI-BASELINE, while 18 teams are below the benchmark model. The best performing team for English is 14-exlab, which achieved an overall accuracy of 0.913223 by their constrained run1. In 14-exlab.c.run1 the participants exploited SVM trained with a combination of stylistic, structural and lexical features, i.e. Hashtag Presence, Link Presence, Swear Word Count, Swear Word Presence, Sexist Slurs Presence and Woman-related Words Presence. The worst results have been obtained by GrCML2016 and Amrita_CEN, both exploiting embedding representation of tweets. As classification models they used EoC (GrCML2016) and Deep Learning approaches (Amrita_CEN).

Concerning the 24 runs for Spanish, 17 of them are above the AMI-BASELINE, while the remaining 7 are below. The best performing teams for Spanish are 14-exlab and JoseSebastian achieving an accuracy of 0.814681, with their constrained run3 and run1 respectively. 14-exlab.c.run3 is based on on Bag of Word, Bag of Hashtags, Bag of Emojis, Sexist Slurs Presence, Woman Word Presence, count of negative stereotypes words and count of hate words and slurs beyond stereotypes. JoseSebastian.c.run1 used a TF-IDF representation of words obtained after a pre-processing step mostly focused on maintaining specific hashtags. The

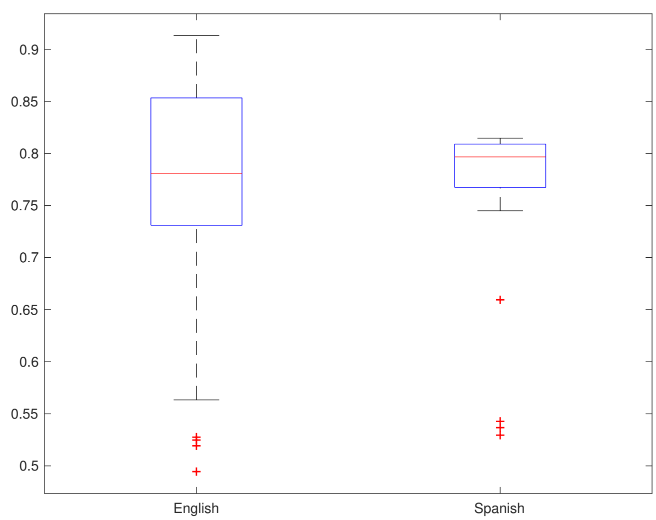
Table 6. Subtask A - Rankings

ENGLISH			**SPANISH**		
Rank	**Team**	**Accuracy**	**Rank**	**Team**	**Accuracy**
1	14-exlab.c.run1	0.913223	1	14-exlab.c.run3	0.814681
2	14-exlab.c.run2	0.902204	2	JoseSebastian.c.run1	0.814681
3	14-exlab.c.run4	0.898072	3	SB.c.run4	0.813478
4	14-exlab.c.run3	0.878788	4	14-exlab.c.run1	0.812274
5	SB.c.run4	0.870523	5	14-exlab.c.run2	0.812274
6	SB.u.run1	0.866391	6	14-exlab.c.run4	0.809868
7	SB.u.run3	0.862259	7	SB.c.run2	0.808664
8	SB.u.run2	0.859504	8	SB.c.run5	0.806258
9	SB.c.run5	0.851240	9	_vic_.c.run1	0.805054
10	14-exlab.c.run5	0.823691	10	SB.c.run3	0.805054
11	AnotherTeam.c.run1	0.793388	11	SB.c.run1	0.803851
12	meybelraul.c.run2	0.793388	12	AnotherTeam.c.run1	0.802647
13	ixaTeam.c.run1.txt	0.789256	13	meybelraul.c.run5	0.796631
14	resham.c.run1.txt	0.785124	14	meybelraul.c.run2	0.788207
15	*AMI-BASELINE*	*0.783747*	15	meybelraul.c.run3	0.787004
16	_vic_.c.run2	0.780992	16	meybelraul.c.run4	0.782190
17	_vic_.c.run3	0.780992	17	ixaTeam.c.run1	0.768953
18	_vic_.c.run4	0.780992	*18*	*AMI-BASELINE*	*0.767750*
19	meybelraul.c.run3	0.779614	19	meybelraul.c.run1	0.767750
20	meybelraul.c.run1	0.771350	20	_vic_.c.run2	0.766546
21	meybelraul.c.run4	0.769972	21	Amrita_CEN.c.run3	0.744886
22	meybelraul.c.run5	0.760331	22	_vic_.c.run3	0.659446
23	ITT.c.run2	0.758953	23	Amrita_CEN.c.run1	0.542720
24	JoseSebastian.c.run1	0.749311	24	14-exlab.c.run5	0.536703
25	Amrita_CEN.c.run3	0.738292	25	Amrita_CEN.c.run2	0.529483
26	_vic_.c.run1	0.709366			
27	ITT.c.run1	0.706612			
28	_vic_.c.run5	0.646006			
29	Amrita_CEN.c.run2	0.563361			
30	GrCML2016.c.run3.txt	0.527548			
31	GrCML2016.c.run2.txt	0.524793			
32	Amrita_CEN.c.run1	0.519284			
33	GrCML2016.c.run1.txt	0.494490			
	min	0.494490		min	0.529483
	q1	0.738292		q1	0.767750
	median	0.780992		median	0.796631
	mean	0.758578		mean	0.757882
	stdev	0.114780		stdev	0.087896
	q3	0.851240		q3	0.808664
	max	0.913223		max	0.814681

worst results for Spanish in Subtask-A have been obtained by 14-exlab.c.run5 and Amrita_CEN.c.run2.

As can be seen in Figure 1, results are similar for mean and median for both languages, although the standard deviation for English is higher than for Spanish. Moreover, for the English language, we can highlight some outliers that denote those approaches achieving an Accuracy below 56%. Results for English are between 0.494490 and 0.913223, with an average value of 0.758578. Results for Spanish are between 0.529483 and 0.814681, with an average value of 0.757882.

Fig. 1. Distribution of results (Accuracy) for Subtask-A.



5.2 Subtask B

Nine teams participated in Subtask-B for English, presenting 28 runs, and 6 teams participated for Spanish, presenting 20 submissions. In Table 7, the F-scores achieved by all runs on English are shown, as well as the AMI-BASELINE. In particular, we reported the Macro Average F-Measure used for the final ranking, together with the F-Measures computed on "misogyny_category" and "target". Among the 28 runs for English, 15 teams achieved an accuracy above the AMI-BASELINE, while 13 teams are below the benchmark model.

It is interesting to highlight the strong difference between the best and the worst systems, underlying Macro Average F-Measure ranging from 0.442483 to 0.083040. The best performing team for English is SB, which achieved an overall

Macro Average F-Measure of 0.442483 by their unconstrained run3. In SB.u.run3 the participants exploited SVM trained with a combination of lexicons concerning sexuality, profanity, femininity and human body.

Table 7. Subtask B - English Ranking

English				
Rank	**Team**	**Macro Average F-Measure**	**Macro F-Measure (misogyny_category)**	**Macro F-Measure (target)**
1	SB.u.run3	0.442483	0.292499	0.592467
2	SB.u.run1	0.437201	0.274798	0.599603
3	SB.u.run2	0.431865	0.265948	0.597781
4	SB.c.run5	0.408758	0.222102	0.595414
5	SB.c.run4	0.401897	0.215547	0.588247
6	14-exlab.c.run5	0.369819	0.158329	0.581310
7	resham.c.run1.txt	0.351468	0.148219	0.554718
8	14-exlab.c.run3	0.351380	0.177154	0.525606
9	meybelraul.c.run3	0.349342	0.153617	0.545066
10	14-exlab.c.run4	0.343282	0.180558	0.506006
11	meybelraul.c.run2	0.342323	0.146600	0.538045
12	14-exlab.c.run2	0.341632	0.182421	0.500842
13	_vic_.c.run4	0.339590	0.138319	0.540861
14	_vic_.c.run3	0.339141	0.137421	0.540861
15	14-exlab.c.run1	0.337913	0.175096	0.500730
16	*AMI-BASELINE*	0.337382	0.156794	0.517971
17	_vic_.c.run2	0.336434	0.132007	0.540861
18	meybelraul.c.run1	0.336143	0.159844	0.512442
19	meybelraul.c.run4	0.333332	0.121221	0.545442
20	meybelraul.c.run5	0.328451	0.130986	0.525915
21	JoseSebastian.c.run1	0.326309	0.147691	0.504927
22	ITT.c.run2	0.318026	0.179529	0.456523
23	_vic_.c.run1	0.316368	0.128582	0.504155
24	AnotherTeam.c.run1	0.305317	0.111295	0.499339
25	ITT.c.run1	0.279130	0.155886	0.402374
26	_vic_.c.run5	0.236876	0.160454	0.313297
27	GrCML2016.c.run1.txt	0.178087	0.085939	0.270234
28	GrCML2016.c.run3.txt	0.091724	0.064585	0.118864
29	GrCML2016.c.run2.txt	0.083040	0.052761	0.113318

It can be easily noted by looking at the Macro F-Measure of all the approaches, that the problem of recognizing the *misogyny_category* and the *target* is more difficult than the misogyny identification task. The best results for misogyny_category is 0.292499, while for target the highest performance is 0.599603. The main reason of these poor results can be grasped by analysing the detailed results in Table 8, where the F-Measure for each label is reported. We can easily note that the less frequent misogyny_category labels have not been recognized by almost all the participants, i.e. derailing and dominance.

Concerning the Spanish language rankings are reported in Table 9. Among the 20 runs for English, 13 teams achieved an accuracy above the AMI-BASELINE, while 7 teams are below the benchmark model. The best performing team for

Table 8. Subtask B - English Details

ENGLISH							
Submission	F-Measure (derailing)	F-Measure (discredit)	F-Measure (dominance)	F-Measure (sexual_harassment)	F-Measure (stereotype)	F-Measure (active)	F-Measure (passive)
vic.c.run1	0.000000	0.385455	0.000000	0.106195	0.151261	0.412698	0.595611
vic.c.run2	0.000000	0.390977	0.000000	0.123894	0.145161	0.460000	0.621723
vic.c.run3	0.000000	0.408759	0.000000	0.141593	0.136752	0.460000	0.621723
vic.c.run4	0.000000	0.427536	0.000000	0.126126	0.137931	0.460000	0.621723
vic.c.run5	0.105263	0.364964	0.121212	0.106667	0.104167	0.337531	0.289063
14-exlab.c.run1	0.000000	0.500000	0.000000	0.131579	0.243902	0.429091	0.572368
14-exlab.c.run2	0.000000	0.518950	0.000000	0.149254	0.243902	0.423611	0.578073
14-exlab.c.run3	0.000000	0.495775	0.000000	0.151899	0.238095	0.388489	0.662722
14-exlab.c.run4	0.000000	0.507375	0.000000	0.151515	0.243902	0.444444	0.567568
14-exlab.c.run5	0.000000	0.566154	0.058824	0.140351	0.026316	0.454545	0.708075
AMI-BASELINE	0.000000	0.467797	0.105263	0.090909	0.120000	0.422018	0.613924
Amrita_CEN.c.run1	-	-	-	-	-	-	-
Amrita_CEN.c.run2	-	-	-	-	-	-	-
Amrita_CEN.c.run3	-	-	-	-	-	-	-
AnotherTeam.c.run1	0.000000	0.497653	0.000000	0.058824	0.000000	0.394904	0.603774
GrCML2016.c.run1.txt	0.000000	0.233618	0.000000	0.000000	0.196078	0.131579	0.408889
GrCML2016.c.run2.txt	0.000000	0.131455	0.042553	0.068966	0.020833	0.090090	0.136546
GrCML2016.c.run3.txt	0.000000	0.136986	0.048780	0.081081	0.056075	0.060606	0.177122
ITT.c.run1	0.039216	0.407240	0.000000	0.122449	0.210526	0.326316	0.478431
ITT.c.run2	0.033333	0.451613	0.000000	0.190476	0.222222	0.355556	0.557491
ixaTeam.c.run1.txt	-	-	-	-	-	-	-
JoseSebastian.c.run1	0.000000	0.481928	0.000000	0.161290	0.095238	0.371257	0.638596
meybelraul.c.run1	0.064516	0.509259	0.060606	0.088889	0.075949	0.392157	0.632727
meybelraul.c.run2	0.000000	0.537879	0.000000	0.195122	0.000000	0.451977	0.624113
meybelraul.c.run3	0.000000	0.498195	0.000000	0.196721	0.073171	0.452632	0.637500
meybelraul.c.run4	0.000000	0.452261	0.000000	0.153846	0.000000	0.438710	0.652174
meybelraul.c.run5	0.000000	0.491667	0.000000	0.163265	0.000000	0.409938	0.641892
resham.c.run1.txt	0.000000	0.536585	0.000000	0.130435	0.074074	0.438202	0.671233
SB.c.run4	0.066667	0.587896	0.125000	0.222222	0.075949	0.454106	0.722388
SB.c.run5	0.062500	0.562500	0.166667	0.208955	0.109890	0.480769	0.710059
SB.u.run1	0.000000	0.592814	0.307692	0.181818	0.291667	0.440000	0.759207
SB.u.run2	0.000000	0.581040	0.307692	0.111111	0.329897	0.441176	0.754386
SB.u.run3	0.000000	0.607843	0.292683	0.193548	0.368421	0.448980	0.735955

Spanish is 14-exlab, which achieved an overall Average Macro F- Measure of 0.446109 by their constrained run2. In 14-exlab.c.run2 the participants exploited SVM trained with Bag of Word, Bag of Hashtags, Bag of Emojis, Sexist Slurs Presence and Woman Word Presence.

Table 9. Subtask B - Spanish Ranking

SPANISH				
Rank	**Team**	**Macro Average F-Measure**	**Macro F-Measure (category)**	**Macro F-Measure (target)**
1	14-exlab.c.run2	0.446109	0.339026	0.553192
2	14-exlab.c.run3	0.445894	0.336600	0.555187
3	14-exlab.c.run4	0.444223	0.335357	0.553090
4	SB.c.run4	0.441045	0.330355	0.551736
5	14-exlab.c.run1	0.440557	0.339512	0.541602
6	JoseSebastian.c.run1	0.432807	0.323398	0.542216
7	meybelraul.c.run3	0.431414	0.273222	0.589605
8	_vic_.c.run1	0.427225	0.320337	0.534112
9	SB.c.run3	0.426759	0.300218	0.553299
10	meybelraul.c.run5	0.423978	0.311941	0.536016
11	meybelraul.c.run1	0.415784	0.311032	0.520536
12	SB.c.run1	0.415165	0.270981	0.559349
13	_vic_.c.run2	0.411750	0.306539	0.516960
14	*AMI-BASELINE*	0.409185	0.281424	0.536946
15	meybelraul.c.run4	0.408152	0.272877	0.543426
16	meybelraul.c.run2	0.400899	0.267458	0.534341
17	AnotherTeam.c.run1	0.349718	0.256416	0.443020
18	SB.c.run5	0.337350	0.283631	0.391069
19	SB.c.run2	0.335391	0.281295	0.389488
20	14-exlab.c.run5	0.278952	0.220582	0.337322
21	_vic_.c.run3	0.272720	0.220473	0.324967

Similar results on Subtask-B have been obtained for the Spanish language about "misogyny_category" "target", as reported in Table 10. The best results for misogyny_category is 0.339026, while for target the highest performance is 0.589605.

It can be easily noted by looking at all the Average F-Measure of all the approaches reported in Table 10, that the *derailing* misogyny category is more difficult to be recognized than others because of the few examples available in the training set. An analogous consideration can be given for what concerns the prediction capabilities of the systems in distinguishing the targets between *active* and *passive*. The reduced number of examples related to the *passive* label (194 *passive* instances against 1455 *active* examples) have likely biased all the participating systems.

Table 10. Subtask B - Spanish Details

SPANISH							
Submission	F-Measure (derailing)	F-Measure (discredit)	F-Measure (dominance)	F-Measure (sexual_harassment)	F-Measure (stereotype)	F-Measure (active)	F-Measure (passive)
vic.c.run1	0.000000	0.653846	0.311475	0.545455	0.090909	0.810160	0.258065
vic.c.run2	0.000000	0.622296	0.300752	0.526316	0.083333	0.783920	0.250000
vic.c.run3	0.000000	0.470348	0.307692	0.324324	0.000000	0.595880	0.054054
14-exlab.c.run1	0.000000	0.701695	0.378788	0.530120	0.086957	0.816537	0.266667
14-exlab.c.run2	0.000000	0.702886	0.378788	0.530120	0.083333	0.816062	0.290323
14-exlab.c.run3	0.000000	0.706081	0.375940	0.517647	0.083333	0.820051	0.290323
14-exlab.c.run4	0.000000	0.705882	0.375940	0.511628	0.083333	0.815857	0.290323
14-exlab.c.run5	0.000000	0.439791	0.215054	0.361111	0.086957	0.490028	0.184615
AMI-BASELINE	0.000000	0.648464	0.285714	0.389610	0.083333	0.770861	0.303030
Amrita_CEN.c.run1	0.000000	0.006944	-	-	-	-	-
Amrita_CEN.c.run2	0.000000	0.006944	-	-	-	-	-
Amrita_CEN.c.run3	0.000000	0.006944	-	-	-	-	-
AnotherTeam.c.run1	0.000000	0.632381	0.243902	0.405797	0.000000	0.804408	0.081633
ixaTeam.c.run1	0.000000	0.006944	-	-	-	-	-
JoseSebastian.c.run1	0.000000	0.649660	0.333333	0.528736	0.105263	0.813245	0.271186
meybelraul.c.run1	0.153846	0.624506	0.326923	0.358974	0.090909	0.769886	0.271186
meybelraul.c.run2	0.000000	0.608696	0.311927	0.416667	0.000000	0.782967	0.285714
meybelraul.c.run3	0.000000	0.599628	0.318584	0.376471	0.071429	0.784844	0.394366
meybelraul.c.run4	0.000000	0.621094	0.320755	0.422535	0.000000	0.781768	0.305085
meybelraul.c.run5	0.000000	0.621881	0.342857	0.421053	0.173913	0.796170	0.275862
SB.c.run1	0.000000	0.617143	0.293103	0.359551	0.085106	0.796117	0.322581
SB.c.run2	0.000000	0.625954	0.307692	0.347826	0.125000	0.778976	0.000000
SB.c.run3	0.000000	0.630282	0.344828	0.439024	0.086957	0.798906	0.307692
SB.c.run4	0.000000	0.680702	0.396396	0.431818	0.142857	0.798387	0.305085
SB.c.run5	0.000000	0.598095	0.278689	0.359551	0.181818	0.782138	0.000000

6 Conclusion

We described a new shared task about Automatic Misogyny Identification on Twitter. By analysing the runs submitted by the participants we can conclude that the problem of misogyny identification has been easily addressed by all the teams, while the misogynous behavior and target classification still remains a very challenging problem. Concerning some potential future AMI scenarios, several issues should be considered for improving the quality of the collected data, especially for capturing those less frequent misogynistic behaviors such as Dominance and Derailing. Hate speech towards women will be also addressed at the AMI shared task[4] organized for Evalita 2018 and at the HatEval task[5] in SemEval 2019.

Acknowledgements

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

References

1. M. Duggan. Online Harassment 2017. Pew Research Center, July 2017.
2. M. Anzovino, E. Fersini, P. Rosso. Automatic Identification and Classification of Misogynistic Language on Twitter. In Proc. of the 23rd International Conference on Natural Language & Information Systems (NLDB), 2018.
3. S. Hewitt, T. Tiropanis, C. Bokhove. The problem of identifying misogynist language on Twitter (and other online social spaces). In Proc. of the 8th ACM Conference on Web Science, 2016.
4. B. Poland, 2016. Haters: Harassment, Abuse, and Violence Online. U of Nebraska
5. J.S.án Canós. Misogyny identification through SVM 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). CEUR Workshop Proceedings. CEUR-WS.org, Seville, Spain, September 18, 2018
6. E.W. Pamungkas, A. Teresa Cignarella, V. Basile, V. Patti. 14-ExLab@UniTo for AMI at IberEval2018: Exploiting Lexical Knowledge for Detecting Misogyny in English and Spanish Tweets. 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
7. R. Ahluwalia, E. Shcherbinina, E. Callow, A. Nascimento, M. De Cock. Detecting Misogynous Tweets. 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.

[4] http://amievalita2018.wordpress.com
[5] http://alt.qcri.org/semeval2019/

8. I. Goenaga, A. Atutxa, K. Gojenola, A. Casillas, A. Díaz de Ilarraza, N. Ezeiza, M. Oronoz, A. Pérez O. Perez de Viñaspre. Automatic Misogyny Identification Using Neural Networks. 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
9. E. Shushkevich, J. Cardiff. Classifying Misogynistic Tweets Using a Blended Model: The AMI Shared Task in 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). CEUR Workshop Proceedings. CEUR-WS.org, Seville, Spain, September 18, 2018.
10. S. Frenda, B. Ghanem, M. Montes-y-Gómez. Exploration of Misogyny in Spanish and English tweets. 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. H. Liu, F. Chiroma, M. Cocea. Identification and Classification of Misogynous Tweets Using Multi-classifier Fusion. 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.
12. V Nina-Alcocer. AMI at IberEval2018 - Automatic Misogyny Identification in Spanish and English Tweets. 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.
13. A. Kumar, S. Singh, B.G.H. Balakrishnan Amrita Vishwa Vidyapeetham, India.
14. M. Hernández-Bernia, R. León-Fabelo Universitat Politcnica de Valncia, Spain.
15. M. Parreño-Lara, F. Rubín-Capalbo Universitat Politcnica de Valncia, Spain.