Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI)

Elisabetta Fersini[1], **Debora Nozza**[1], **Paolo Rosso**[2]
[1]DISCo, Universitá degli Studi di Milano-Bicocca
[2]PRHLT Research Center, Universitat Politècnica de València
{fersini, debora.nozza}@disco.unimib.it
prosso@dsic.upv.es

Abstract

English. Automatic Misogyny Identification (AMI) is a new shared task proposed for the first time at the Evalita 2018 evaluation campaign. The AMI challenge, based on both Italian and English tweets, is distinguished into two subtasks, i.e. Subtask A on misogyny identification and Subtask B about misogynistic behaviour categorization and target classification. Regarding the Italian language, we have received a total of 13 runs for Subtask A and 11 runs for Subtask B. Concerning the English language, we received 26 submissions for Subtask A and 23 runs for Subtask B. The participating systems have been distinguished according to the language, counting 6 teams for Italian and 10 teams for English. We present here an overview of the AMI shared task, the datasets, the evaluation methodology, the results obtained by the participants and a discussion of the methodology adopted by the teams. Finally, we draw some conclusions and discuss future work.

Italiano. *Automatic Misogyny Identification (AMI) è un nuovo shared task proposto per la prima volta nella campagna di valutazione Evalita 2018. La sfida AMI, basata su tweet italiani e inglesi, si distingue in due sottotask ossia Subtask A relativo al riconoscimento della misoginia e Subtask B relativo alla categorizzazione di espressioni misogine e alla classificazione del soggetto target. Per quanto riguarda la lingua italiana, sono stati ricevuti un totale di 13 run per il Subtask A e 11 run per il Subtask B. Per quanto riguarda la lingua inglese, sono stati ricevuti 26 run per il Subtask A e 23 per Subtask B. I sistemi partecipanti sono stati distinti in base alla lingua, raccogliendo un totale di 6 team partecipanti per l'italiano e 10 team per l'inglese. Presentiamo di seguito una sintesi dello shared task AMI, i dataset, la metodologia di valutazione, i risultati ottenuti dai partecipanti e una discussione sulle metodologie adottate dai diversi team. Infine, vengono discusse conclusioni e delineati gli sviluppi futuri.*

1 Introduction

During the last years, the phenomenon of hate against women increased exponentially especially in online environment such as microblogs (Hewitt et al., 2016; Poland, 2016). According to the Pew Research Center Online Harassment report (2017) (Duggan, 2017), 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%). Misogyny, defined as the hate or prejudice against women, can be linguistically manifested in numerous ways, ranging from less aggressive behaviours like social exclusion and discrimination to more dangerous expressions related to threats of violence and sexual objectification (Anzovino et al., 2018). Given this relevant social problem, the Automatic Misogyny Identification (AMI) task has been proposed first at IberEval 2018 (Spanish and English) (Fersini et al., 2018) and later at Evalita 2018 (Italian and English) (Caselli et al., 2018). The main goal of AMI is to distinguish misogynous contents from non-misogynous ones, to categorize misogynistic behaviours and finally to classify the target of a tweet.

Table 1: Examples of misogynous and non-misogynous tweets

Misogynous	**Text**
Misogynous	I've yet to come across a nice girl. They all end up being bit**es in the end.
Non-misogynous	@chiellini you are a bi*ch!

2 Task Description

The AMI shared task is organized according to two main subtasks:

- **Subtask A - Misogyny Identification**: a system must discriminate misogynistic contents from the non-misogynistic ones. Examples of misogynous and non-misogynous tweets are reported in Table 1.
- **Subtask B - Misogynistic Behaviour and Target Classification**: a system must recognize 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 behaviour, 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 Behaviours are reported in Table 2.

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 misogynous tweets are reported in Table 3.

3 Training and Testing Data

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

- Streaming download using a set of manually defined representative keywords, e.g. *bi**h, w**re, c*nt* for English and *pu****a, tr**a, f**a di legno* for Italian;
- 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*.

Among all the collected tweets we selected a subset of text querying the database with the copresence of keywords, originating two corpora initially composed of 10000 tweets for each language. In order to label both the Italian and English datasets, we involved a group of 6 experts exploiting the CrowdFlower[1] platform for internal use. At the end of the labelling phase, we provided one corpus for Italian and one corpus for English to all the participants. The inter-rater annotator agreement on the English dataset for the fields of "misogynous", "misogyny_category" and "target"

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

Table 2: Examples of tweets for each misogyny category

Misogyny Category	**Text**
Discredit	@AsiaArgento brainless fuc**ing bi**ch
Stereotype	Women play your role..#gobacktothekitchen
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

Table 3: Examples of targets.

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

is 0.81, 0.45 and 0.49 respectively, while for the Italian dataset is 0.96, 0.68 and 0.76. Each corpus is distinguished in Training and Test datasets. Regarding the training data, both the Italian and English corpora are composed of 4000 tweets. Concerning the test data, we provided 1000 tweets for each language. The training data has been provided as tab-separated, according to the following fields:

- **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 misogynous 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 below. 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 is reported in Table 4. While the distribution of labels related to the field "misogynous" is almost balanced (for both languages), the classes related to the other fields are quite unbalanced. Regarding the "misogyny_category", we

can distinguish between the two considered languages. In particular, for the Italian language, the most frequent label is related to the category *Stereotype & Objectification*, while for English the most predominant one is *Discredit*. Concerning the "target", the most predominant victims are specific users (*active*) with a strong imbalanced distribution on the Italian corpus, while it is almost balanced for the English training dataset and strongly imbalanced on the (*active*) targets for the corresponding test dataset.

4 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.

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:

$$F_1 = \frac{F_1(misogyny\_category)+F_1(target)}{2} \quad (1)$$

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.

5 Participants and Results

A total of 6 teams for Italian and 10 teams for English from 10 different countries participated in at least one of the two subtasks of AMI. Each team had the chance to submit up to three runs for English and three runs for Italian. Runs could be constrained, where only the provided training data and lexicons were admitted, and unconstrained, where additional data for training were allowed. Table 5 shows an overview of the teams[2] reporting their affiliation, their country, the number of submissions for each language and the subtasks they addressed.

5.1 Subtask A: Misogyny Identification

Table 6 reports the results for the Misogyny Identification task, which received 13 submissions for Italian and 26 runs for English submitted respectively from 6 and 10 teams. The highest Accuracy has been achieved by *bakarov* at 0.844 for Italian and by *hateminers* at 0.704 for English, both in a constrained setting. Most of the systems have shown an improvement with respect to the *AMI-BASELINE*. While the *bakarov* team submitted only one run based on TF-IDF coupled with Singular Value Decomposition and Boosting classifier, *hateminers* achieved the highest performance with a run based on vector representation that concatenates sentence embedding, TF-IDF and average word embeddings coupled with a Logistic Regression model.

5.2 Subtask B: Misogynistic Behaviour and Target Classification

Table 7 reports the results for the Misogynistic Behaviour and Target Classification task, which received 11 submissions by 5 teams for Italian and 23 submissions by 9 teams for English. The highest Average Macro F1-score has been achieved by *bakarov* at 0.501 for Italian (even if the amended run of CrotoneMilano achieved the highest effective performance) and by *himani* at 0.406 for English, both in a constrained setting. On the contrary of the previous task, most of the systems have shown lower performance compared to the *AMI-BASELINE*. It can be easily noted by looking at the Average Macro F1-score of all the approaches, that the problem of recognizing the misogyny category and the target is more difficult than the

[2]The teams *himani* and *resham* described their systems in the same report (Ahluwalia et al., 2018).

Table 4: Distribution of labels for "misogynous", "misogyny_category" and "target" on the Training and Test datasets. Percentages for "misogyny_category" and "target" are computed with respect to the number of misogynous tweets.

	Training		Testing	
	Italian	**English**	**Italian**	**English**
Misogynous	1828 (46%)	1785 (45%)	512 (51%)	460 (46%)
Non-misogynous	2172 (54%)	2215 (55%)	488 (49%)	540 (54%)
Discredit	634 (35%)	1014 (57%)	104 (20%)	141 (31%)
Sexual Harassment & Threats of Violence	431 (24%)	352 (20%)	170 (33%)	44 (10%)
Derailing	24 (1%)	92 (5%)	2 (1%)	11 (2%)
Stereotype & Objectification	668 (37%)	179 (10%)	175 (34%)	140 (30%)
Dominance	71 (3%)	148 (8%)	61 (12%)	124 (27%)
Active	1721 (94%)	1058 (59%)	446 (87%)	401 (87%)
Passive	107 (6%)	727 (41%)	66 (13%)	59 (13%)

Table 5: Team overview

Team Name	**Affiliation**	**Country**	**Runs**	**Subtask**
14-exlab (Pamungkas et al., 2018)	University of Turin Universitat Politècnica de València	IT ES	3 (EN), 3 (IT)	A, B
bakarov (Bakarov, 2018)	Huawei Technologies	RUS	3 (EN), 3 (IT)	A, B
CrotoneMilano (Basile and Rubagotti, 2018)	Symanto Research Independent Researcher	DE IT	1 (EN), 1 (IT)	A, B
hateminers (Saha et al., 2018)	Indian Institute of Technology	IND	3 (EN), 0 (IT)	A, B
himani (Ahluwalia et al., 2018)	University of Washington Tacoma	USA	3 (EN), 0 (IT)	A, B
ITT (Shushkevich and Cardiff, 2018)	Institute of Technology Tallaght Yandex	IRL RUS	3 (EN), 0 (IT)	A, B
RCLN (Buscaldi, 2018)	Université Paris 13	FR	1 (EN), 1 (IT)	A, B
resham (Ahluwalia et al., 2018)	University of Washington	USA	3 (EN), 0 (IT)	A, B
SB (Frenda et al., 2018b)	University of Turin Universitat Politècnica de València INAOE	IT ES MEX	3 (EN), 3 (IT)	A, B
StopPropagHate (Fortuna et al., 2018)	INESC TEC Eurecat Porto University	PT ES	3 (EN), 2 (IT)	A

misogyny identification task.

This is due to the fact that there can be a high overlapping between textual expressions of different misogyny categories, therefore it is highly subjective for an annotator (and consequently for a system) to select a category rather than another one. Regarding the target classification, systems can be easily misled by the presence of mentions that are not the target of the misogynous content.

While for the *bakarov* team the system for Subtask B is the same one of Subtask A, *himani* achieved the highest performance on the English language with a run based on a Bag of N-Gram representation coupled with an Ensemble of 5 models for classifying the Misogynistic Behaviour and 2 models for Target Classification.

6 Discussion

The submitted systems can be compared by taking into consideration the kind of input features that they have considered for representing tweets and the machine learning model that has been used as classification model.

Textual Feature Representation. The systems submitted by the challenge participants' consider various techniques for representing the tweet contents. Some teams have concentrated the effort on considering a single type of representation, i.e. the team *ITT* adopted the traditional TF-IDF representation, while *bakarov* and *RCLN* proposed systems considering only weighted n-grams at character level for better dealing with misspellings and capturing few stylistic aspects.

Additionally to the traditional textual feature representation techniques (i.e. bag of words/characters, n-grams of words/characters eventually weighted with TF-IDF) several teams proposed specific lexical features for improving the input space and consequently the classification performances. The team of *CrotoneMilano* experimented feature abstraction following the bleaching approach proposed by Goot et al. (Goot et al.,

Table 6: Results of Subtask A. Constrained runs are marked as *.c*, while the unconstrained ones with *.u*. After the deadline one team reported a format error. The resubmitted amended runs are marked with **.

ITALIAN			ENGLISH		
Rank	**Team**	**Accuracy**	**Rank**	**Team**	**Accuracy**
1	bakarov.c.run2	0.844	1	hateminers.c.run1	0.704
**	CrotoneMilano.c.run1	0.843	2	hateminers.c.run3	0.681
2	bakarov.c.run1	0.842	3	hateminers.c.run2	0.673
3	14-exlab.c.run3	0.839	4	resham.c.run3	0.651
4	bakarov.c.run3	0.836	5	bakarov.c.run3	0.649
5	14-exlab.c.run2	0.835	6	resham.c.run1	0.648
6	StopPropagHate.c.run1	0.835	7	resham.c.run2	0.647
7	AMI-BASELINE	0.830	8	ITT.c.run2	0.638
8	StopPropagHate.u.run2	0.829	9	ITT.c.run1	0.636
9	SB.c.run1	0.824	10	ITT.c.run3	0.636
10	RCLN.c.run1	0.824	11	himani.c.run2	0.628
11	SB.c.run3	0.823	12	bakarov.c.run2	0.628
12	SB.c.run2	0.822	13	14-exlab.c.run3	0.621
13	14-exlab.c.run1	0.765	14	himani.c.run1	0.619
			**	CrotoneMilano.c.run1	0.617
			15	himani.c.run3	0.614
			16	14-exlab.c.run1	0.614
			17	SB.c.run2	0.613
			18	AMI-BASELINE	0.605
			19	bakarov.c.run1	0.605
			20	StopPropagHate.c.run1	0.593
			21	SB.c.run1	0.592
			22	StopPropagHate.u.run3	0.591
			23	StopPropagHate.u.run2	0.590
			24	RCLN.c.run1	0.586
			25	SB.c.run3	0.584
			26	14-exlab.c.run2	0.500

2018) for modelling gender through the language. Specific lexicons for dealing with hate speech language have been included as features in the systems of *SB*, *resham* and *14-exlab*. In particular, *resham* and *14-exlab* made also use of environment-specific features, such as links, hashtags and emojis, and task-specific features, such as swear word, sexist slurs and women-related words.

Differently from these approaches, *StopPropagHate* and *hateminers* teams proposed systems that consider the popular Embeddings techniques both at word and sentence level.

Machine Learning Models. Concerning the machine learning models, we can distinguish between approaches that work with traditional Support Vector Machines and Logistic Regression, Ensemble Models and finally Deep Learning methods. Following, we report the models adopted by the systems that participated in the AMI shared task, according to the type of the machine learning model that has been adopted:

- Support Vector Machines have been exploited by *14-exlab* by using both linear and RBF kernel, by *SB* investigating only a radial basis function kernel, and by *CrotoneMilano* by adopting again a simple linear kernel;
- Logistic Regression has been used by *bakarov* and *hateminers*;
- Ensemble Models have been adopted by three teams according to different settings, i.e. *ITT* and *himani* used a Simple Voting of different classifiers, *resham* induced a Simple Voting over different input features and *RCLN* used an Ensemble based on Random Forest;
- A Deep Learning classifier has been adopted by only one team, i.e *StopPropagHate* that trained a simple dense neural network.

External Resources Several participants exploited external resources for providing task-specific lexical features.

The lexicons for addressing AMI for Italian have been mostly obtained from lists available online. The team *SB* used an available specific Italian

Table 7: Results of Subtask B. Constrained runs are marked as *.c*, while the unconstrained ones with *.u*. After the deadline one team reported a format error. The resubmitted amended runs are marked with **.

ITALIAN			ENGLISH		
Rank	**Team**	**Average Macro F1-score**	**Rank**	**Team**	**Average Macro F1-score**
**	CrotoneMilano.c.run1	0.501	1	himani.c.run3	0.406
1	bakarov.c.run1	0.493	2	himani.c.run2	0.377
2	AMI-BASELINE	0.487	3	AMI-BASELINE	0.370
3	14-exlab.c.run3	0.485	**	CrotoneMilano.c.run1	0.369
4	14-exlab.c.run2	0.482	4	hateminers.c.run3	0.369
5	bakarov.c.run3	0.478	5	hateminers.c.run1	0.348
6	bakarov.c.run2	0.463	6	SB.c.run2	0.344
7	SB.c.run3	0.449	7	himani.c.run1	0.342
8	SB.c.run1	0.448	8	SB.c.run1	0.335
9	RCLN.c.run1	0.448	9	hateminers.c.run2	0.329
10	SB.c.run2	0.446	10	SB.c.run3	0.328
11	14-exlab.c.run1	0.292	11	resham.c.run2	0.322
			12	resham.c.run1	0.316
			13	bakarov.c.run1	0.309
			14	resham.c.run3	0.283
			15	RCLN.c.run1	0.280
			16	ITT.c.run2	0.276
			17	bakarov.c.run2	0.275
			18	14-exlab.c.run1	0.260
			19	bakarov.c.run3	0.254
			20	14-exlab.c.run3	0.239
			21	ITT.c.run1	0.238
			22	ITT.c.run3	0.237
			23	14-exlab.c.run2	0.232

lexicon called "Le parole per ferire" built by Tullio De Mauro[3]. Starting from this lexicon provided by De Mauro, the HurtLex multilingual lexicon has been created (Bassignana et al., 2018). Beyond HurtLex, the team *14-exlab* gathered a swear word list from several sources[4] including a translated version of the *noswearing dictionary*[5] and a list of swear words from (Capuano, 2007).

Regarding the English language, both *resham* and *14-exlab* used the list of swear words from *noswearing dictionary* and the sexist slur list provided by (Fasoli et al., 2015). The team *resham* further investigated the sentiment polarity retrieved from SentiWordNet (Baccianella et al., 2010). Differently, the team *SB* exploited a manually modeled lexicon for the misogyny detection task proposed in (Frenda et al., 2018a). The HurtLex lexicon has been used by the team *14-exlab* also for the English task.

Finally, pre-trained Word Embeddings have been considered by *SB* and *hateminers* teams, specifically *GloVe* (Pennington et al., 2014) for the English task and Word Embeddings built on the TWITA corpus for the Italian one (Basile and Novielli, 2014).

7 Conclusions and Future Work

We presented here a new shared task about Automatic Misogyny Identification on Twitter for Italian and English. By analysing the runs submitted by the participants we can conclude that the problem of misogyny identification has been satisfactorily addressed by all the teams, while the misogynistic behaviour and target classification still remains a challenging problem. Concerning the future work, several issues should be considered to improve the quality of the collected data, especially for capturing those less frequent misogynistic behaviours such as Dominance and Derailing. The problem of hate speech against women will be further addressed in the HatEval shared task at SemEval in English and Spanish tweets[6].

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

[4] https://www.parolacce.org/2016/12/20/dati-frequenza-turpiloquio/ and https://it.wikipedia.org/wiki/Turpiloquio_nella_lingua_italiana

[5] https://www.noswearing.com/dictionary

[6] SemEval 2019 Task 5: *HatEval: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter* https://competitions.codalab.org/competitions/19935

Acknowledgements

The work of the third author was partially funded by the SomEMBED TIN2015-71147-C2-1-P research project (MINECO/FEDER). We thank Maria Anzovino for her initial help in collecting the tweets subsequently used for the labelling phase and the final creation of the Italian and English corpora used for the AMI shared task.

References

Resham Ahluwalia, Himani Soni, Edward Callow, Anderson Nascimento, and Martine De Cock. 2018. Detecting Hate Speech Against Women in English Tweets. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation*.

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

Pierpaolo Basile and Nicole Novielli. 2014. UNIBA at EVALITA 2014-SENTIPOLC Task: Predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features. In *Proceedings of Fourth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2014)*. CEUR.org.

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

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

R.G. Capuano. 2007. *Turpia: sociologia del turpiloquio e della bestemmia*. Riscontri (Milan, Italy). Costa & Nolan.

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

Maeve Duggan. 2017. Online Harassment. `http://www.pewinternet.org/2017/07/11/online-harassment-2017/`. Last accessed 2018-10-28.

Fabio Fasoli, Andrea Carnaghi, and Maria Paola Paladino. 2015. Social acceptability of sexist derogatory and sexist objectifying slurs across contexts. *Language Sciences*, 52:98–107.

Elisabetta Fersini, M Anzovino, and P Rosso. 2018. 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). CEUR Workshop Proceedings. CEUR-WS. org, Seville, Spain*.

Paula Fortuna, Ilaria Bonavita, and Sérgio Nunes. 2018. INESC TEC, Eurecat and Porto University.

Simona Frenda, Ghanem Bilal, et al. 2018a. Exploration of Misogyny in Spanish and English tweets. In *Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, volume 2150, pages 260–267. Ceur Workshop Proceedings.

Simona Frenda, Bilal Ghanem, Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez, and Luis Villaseñor-Pineda. 2018b. Automatic Lexicons Expansion for Multilingual Misogyny Detection. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

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

Sarah Hewitt, Thanassis Tiropanis, and Christian Bokhove. 2016. The Problem of identifying Misogynist Language on Twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, pages 333–335. ACM.

Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, and Viviana Patti. 2018. Automatic Identification of Misogyny in English and Italian Tweets at EVALITA 2018 with a Multilingual Hate Lexicon. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Bailey Poland. 2016. *Haters: Harassment, Abuse, and Violence Online*. Potomac Books, Incorporated.

Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. 2018. Indian Institute of Engineering Science and Technology (Shibpur), Indian Institute of Technology (Kharagpur).

Elena Shushkevich and John Cardiff. 2018. Misogyny detection and classification in English tweets. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.