AMI @ EVALITA2020: Automatic Misogyny Identification

Elisabetta Fersini[1], **Debora Nozza**[2], **Paolo Rosso**[3]
[1]DISCo, University of Milano-Bicocca
[2] Bocconi University
[3]PRHLT Research Center, Universitat Politècnica de València
elisabetta.fersini@unimib.it
debora.nozza@unibocconi.it
prosso@dsic.upv.es

Abstract

English. Automatic Misogyny Identification (AMI) is a shared task proposed at the Evalita 2020 evaluation campaign. The AMI challenge, based on Italian tweets, is organized into two subtasks: (1) Subtask A about misogyny and aggressiveness identification and (2) Subtask B about the fairness of the model. At the end of the evaluation phase, we received a total of 20 runs for Subtask A and 11 runs for Subtask B, submitted by 8 teams. In this paper, we present an overview of the AMI shared task, the datasets, the evaluation methodology, the results obtained by the participants and a discussion about the methodology adopted by the teams. Finally, we draw some conclusions and discuss future work.

Italiano. *Automatic Misogyny Identification (AMI) é uno shared task proposto nella campagna di valutazione Evalita 2020. La challenge AMI, basata su tweet italiani, si distingue in due subtasks: (1) subtask A che ha come obiettivo l'identificazione di testi misogini e aggressivi (2) subtask B relativo alla fairness del modello. Al termine della fase di valutazione, sono state ricevute un totale di 20 submissions per il subtask A e 11 per il subtask B, inviate da un totale di 8 team. 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 le conclusioni e delineati gli sviluppi futuri.*

1 Introduction

The expressions of people about thoughts, emotions, and feelings by means of posts in social media have been widely spread. Women have a strong presence in these online environments: 75% of females use social media multiple times per day compared to 64% of males. While new opportunities emerged for women to express themselves, systematic inequality and discrimination take place in the form of offensive content against the female gender. These manifestations of misogyny, usually provided by a man to a woman for dominating or using a sort of power against the female gender, is a relevant social problem that has been addressed in the scientific literature during the last few years. Recent investigations studied how the misogyny phenomenon takes place, for example as unjustified slurring or as stereotyping of the role/body of a woman (i.e., the hashtag #getbacktokitchen), as described in the book by Poland (Poland, 2016). Preliminary research work was conducted in (Hewitt et al., 2016) as the first attempt of manual classification of misogynous tweets, while automatic misogyny identification in social media has been firstly investigated in (Anzovino et al., 2018). Since 2018, several initiatives have been dedicated as a call-to-action to stop hate against women both from a machine learning and computational linguistics points of view, such as AMI@Evalita 2018 (Fersini et al., 2018a), AMI@IberEval2018 (Fersini et al., 2018b) and HatEval@SemEval2019 (Basile et al., 2019). Several relevant research directions have been investigated for addressing the misogyny identification challenge, among which approaches focused on effective text representation (Bakarov, 2018; Basile and Rubagotti, 2018), machine learning models (Buscaldi, 2018; Ahluwalia et al., 2018) and domain-specific lexical resources (Pamungkas et al., 2018; Frenda et al., 2018).

During the AMI shared task organized at the Evalita 2020 evaluation campaign (Basile et al., 2020), the focus is not only on misogyny identification but also on aggressiveness recognition, as well as to the definition of models able to guarantee fair predictions.

2 Task Description

The AMI shared task, which is a re-run of a previous challenge at Evalita 2018, proposes the automatic identification of misogynous content in the Italian language on Twitter. More specifically, it is organized according to two main subtasks:

- **Subtask A - Misogyny & Aggressive Behaviour Identification**: a system must recognize if a text is misogynous or not, and in case of misogyny, if it expresses an aggressive attitude. In order to provide an annotated corpus for Subtask A, the following definitions have been adopted to label the collected dataset:
  - *Misogynous*: a text that expresses hating towards women in particular (in the form of insulting, sexual harassment, threats of violence, stereotype, objectification, and negation of male responsibility).
  - *Not Misogynous*: a text that does not express any form of hate towards women.
  - *Aggressive*: a message is considered aggressive if it (implicitly or explicitly) presents, incites, threatens, implies, suggests, or alludes to:
    - attitudes, violent actions, hostility or commission of offenses against women;
    - social isolation towards women for physical or psychological characteristics;
    - justify or legitimize an aggressive action against women.
  - *Not Aggressive*: If none of the previous conditions hold.
- **Subtask B - Unbiased Misogyny Identification**: a system must discriminate misogynistic contents from the non-misogynistic ones, while guaranteeing the fairness of the model (in terms of unintended bias) on a synthetic dataset (Nozza et al., 2019). To this purpose Subtask B has the goal of measuring the attitude of a model to be fair when processing sentences containing specific identity terms that likely conveyed misogyny in the training data, e.g. "girlfriend" and "wife".

3 Training and Testing Data

The data provided to the participants for the AMI shared task comprises a raw dataset and a synthetic dataset for measuring bias. Each dataset is distinguished in Training Set and Test Set.

3.1 Raw dataset

The raw dataset is a balanced dataset of misogynous and non-misogynous tweets. The *raw training set* (6,000 tweets) is derived from the data collected for the 2018 edition of the AMI shared task, where the misogynistic posts have been enriched by labelling aggressive expressions according to the definition provided in Section 2. The *raw test dataset* (approximately 1,000 tweets) has been collected from Twitter using a similar approach to the 2018 edition of the shared task. This is intentionally done to evaluate the generalization abilities of the systems on test data collected in a different time period and therefore characterized by higher language variability with respect to the training data. Examples of tweets belonging to the raw dataset are shown in Table 1.

Table 1: Examples of misogynous and non-misogynous tweets

Misogynous	Aggressiveness	Text
1	1	Stupida tr**a! Meriti di morire
1	0	Le ragazze sono tutte pu***ne
0	0	@chiellini sei una cagna

The training raw data for this dataset are provided as TSV files (tab-separated files) and report the following fields, where:

- id denotes a unique identifier of the tweet.
- text represents the tweet text.
- misogynous defines whether a tweet is misogynous or not misogynous; it takes values:
  - 0 if the tweet is not misogynous;
  - 1 if the tweet is misogynous.

Table 2: Template examples.

Template Examples	Misogynous
<identity_term>dovrebbero essere protette	0
<identity_term>dovrebbero essere uccise	1
apprezziamo <identity_term>	0
picchiamo <identity_term>	1
splendida <identity_term>	0
sporca <identity_term>	1

- aggressiveness denotes whether a misogynous tweet is aggressive or not; it takes value as:
  - 0 denotes a non-aggressive tweet (not misogynous tweets are labelled as 0 by default);
  - 1 if the tweet is aggressive.

The raw testing data are provided as TSV files reporting only id and text.

3.2 Synthetic dataset

The synthetic test dataset for measuring the presence of unintended bias has been created following the procedure adopted in (Dixon et al., 2018; Nozza et al., 2019): a list of identity terms has been constructed by taking into consideration some concepts related to the term "donna" (e.g. "moglie", "fidanzata"). Given the identity terms, several templates have been created including positive/negative verbs and adjectives (e.g. negative: hate, inferior; positive: love, awesome) both for conveying a misogynistic message or a non-misogynistic one. Some examples of such templates, used to create the synthetic dataset, are reported in Table 2.

The synthetic dataset, created for measuring the presence of unintended bias, contains template-generated text labelled according to:

- Misogyny: Misogyny (1) vs. Not Misogyny (0)

The training data for the raw dataset are provided as TSV files (tab-separated files) and report the following fields:

- id denotes a unique identifier of the template-generated text.
- text represents the template-generated text.
- misogynous defines if the template-generated text is misogynous or non-misogynous; it takes values as 1 if the tweet is misogynous, 0 if the tweet is non-misogynous.

The synthetic testing data are provided as TSV files (tab-separated files) reporting only id and text.

The statistics about the raw and synthetic datasets, both for the training and testing sets, are reported in Table 3.

Table 3: Distribution of labels on the Training and Test datasets

	Training		**Testing**	
	Raw	**Synthetic**	**Raw**	**Synthetic**
Misogynous	2337	1007	500	954
Non-misogynous	2663	1007	500	954
Aggressive	1783	-	176	-
Non-aggressive	3217	-	824	-

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. Each class to be predicted (i.e. "Misogyny" and "Aggressiveness") has been evaluated independently on the other using a Macro F1-score. The final ranking of the systems participating in Subtask A was based on the Average Macro F1-score (F_1), computed as follows:

$$Score_A = \frac{F_1(Misogyny) + F_1(Aggressiveness)}{2} \quad (1)$$

Subtask B. The ranking for Subtask B is computed by the weighted combination of AUC estimated on the test raw dataset AUC_{raw} and three per-term AUC-based bias scores computed on the synthetic dataset ($AUC_{Subgroup}$, AUC_{BPSN}, AUC_{BNSP}). Let s be an identity-term (e.g. "girlfriend" and "wife") and N be the total number of identity-terms, the score of each run is estimated according to the following metric:

$$\begin{aligned} Score_B &= \tfrac{1}{2} AUC_{raw} + \\ &+ \tfrac{1}{2N} \Big[\textstyle\sum_s AUC_{subgroup}(s) \\ &+ \textstyle\sum_s AUC_{BPSN}(s) \\ &+ \textstyle\sum_s AUC_{BNSP}(s) \Big] \end{aligned} \quad (2)$$

Unintended bias can be uncovered by looking at differences in the score distributions between data mentioning a specific identity-term (subgroup distribution) and the rest (background distribution).

Table 4: Team overview

Team Name	Affiliation	Country	Runs	Subtask
jigsaw (Lees et al., 2020)	Google	US	2 (u)	A, B
fabsam (Fabrizi, 2020)	University of Pisa	IT	2 (c)	A, B
YNU_OXZ (Ou and Li, 2020)	Yunnan University	CN	2(u)	A
NoPlaceForHateSpeech (da Silva and Roman, 2020)	University of Sao Paulo	BR	3 (c)	A
AMI_the_winner (Lepri et al.,)	University of Pisa	IT	3 (c)	A
MDD (El Abassi and Nisioi, 2020)	University of Bucharest	HU	2 (u), 1 (c)	A, B
PoliTeam (Attanasio and Pastor, 2020)	Politecnico di Torino	IT	2 (c)	A, B
UniBO (Muti and Barròn-Cedeño, 2020)	University of Bologna	IT	1 (c)	A

The three per-term AUC-based bias scores are related to specific subgroups as follows:

- $AUC_{Subgroup}(s)$: calculates AUC only on the data within the subgroup related to a given identity term. This represents model understanding and separability within the subgroup itself. A low value in this metric means the model does a poor job of distinguishing between misogynous and non-misogynous comments that mention the identity.

- $AUC_{BPSN}(s)$: Background Positive Subgroup Negative (BPSN) calculates AUC on the misogynous examples from the background and the non-misogynous examples from the subgroup. A low value in this metric means that the model confuses non-misogynous examples that mention the identity-term with misogynous examples that do not, likely meaning that the model predicts higher misogynous scores than it should for non-misogynous examples mentioning the identity-term.

- $AUC_{BNSP}(s)$: Background Negative Subgroup Positive (BNSP) calculates AUC on the non-misogynous examples from the background and the misogynous examples from the subgroup. A low value here means that the model confuses misogynous examples that mention the identity with non-misogynous examples that do not, likely meaning that the model predicts lower misogynous scores than it should for misogynous examples mentioning the identity.

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 with Tf-IDF weighing schema. In particular, we created one training set for each field to be predicted, i.e. "misogynous", "aggressiveness", 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 8 teams from 6 different countries participated in at least one of the two subtasks of AMI. Two teams participated with the same approach also in the HaSpeeDe shared task (Sanguinetti et al., 2020), addressing misogyny identification with generic models for detecting hate speech. Each team had the chance to submit up to three runs that could be constrained (c), where only the provided training data and lexicons were admitted, and unconstrained (u), where additional data for training were allowed. Table 4 reports an overview of the teams illustrating their affiliation, their country, the number and type (c for constrained, u for unconstrained) of submissions, and the subtasks they addressed.

5.1 Subtask A: Misogyny & Aggressive Behaviour Identification

Table 5 reports the results for the Misogyny & Aggressive Behaviour Identification task, which received 20 submissions submitted by 8 teams. The highest result has been achieved by *jigsaw* at 0.7406 in an unconstrained setting and by *fabsam* at 0.7342 in a constrained run. While the best results obtained as unconstrained is based on ensembles of fine-tuned custom BERT models, the one achieved by the best constrained system is grounded on a convolutional neural network that exploits pre-trained word embeddings.

By analysing the detailed results, it emerged that while the identification of misogynous text can be considered a quite simple problem, the recognition of aggressiveness needs to be properly

addressed. In fact, the score reported in Table 5 are strongly affected by the prediction capabilities mostly related to the aggressive posts. This is likely due to the subjective perception of aggressiveness captured by the variance of the data available in the ground truth.

Table 5: Results of Subtask A. Constrained runs are marked as "c", while the unconstrained ones with "u". An amended run, marked with **, has been submitted after the deadline.

Rank	Run Type	Score	Team
**	c	0.744	UniBO **
1	u	0.741	jigsaw
2	u	0.738	jigsaw
3	c	0.734	fabsam
4	u	0.731	YNU_OXZ
5	c	0.731	fabsam
6	c	0.717	NoPlaceForHateSpeech
7	u	0.701	YNU_OXZ
8	c	0.695	fabsam
9	c	0.693	NoPlaceForHateSpeech
10	c	0.687	AMI_the_winner
11	u	0.684	MDD
12	c	0.683	PoliTeam
13	c	0.682	MDD
14	c	0.681	PoliTeam
15	u	0.668	MDD
16	c	0.665	AMI_the_winner
17	c	0.665	AMI_BASELINE
18	c	0.647	PoliTeam
19	c	0.634	UniBO
20	c	0.626	AMI_the_winner
21	c	0.490	NoPlaceForHateSpeech

After the deadline the team *UniBO* submitted an amended run (**), that has not been ranked in the official results of the AMI shared task. However, we believe interesting to mention their achievement showing an Average Macro F1-score equal to 0.744.

5.2 Subtask B: Unbiased Misogyny Identification

Table 6 reports the results for the Unbiased Misogyny Identification task, which received 11 submissions by 4 teams, among which 4 unconstrained and 7 constrained. The highest Average Macro F1 score has been achieved by *jigsaw* at 0.8825 with an unconstrained run and by *PoliTeam* at 0.8180 with a constrained submission.

Similarly to the previous task, most of the systems have shown better performance compared to the *AMI-BASELINE*. By analizing the runs, we can highlight that the two best results achieved on Subtask B have been obtained by the unconstrained run submitted by *jigsaw*, where a simple debiasing technique based on data augumentation have been adopted, and by the constrained run provided by *Politeam*, where the problem of biased prediction has been partially mitigated by introducing misogynous lexicon.

Table 6: Results of Subtask B. Constrained runs are marked as "c", while the unconstrained ones with "u".

Rank	Run Type	Score	Team
1	u	0.882	jigsaw
2	c	0.818	PoliTeam
3	c	0.814	PoliTeam
4	c	0.705	fabsam
5	c	0.702	fabsam
6	c	0.694	PoliTeam
7	c	0.691	fabsam
8	u	0.649	jigsaw
9	c	0.613	MDD
10	c	0.602	AMI_BASELINE
11	u	0.601	MDD
12	u	0.601	MDD

6 Discussion

The submitted systems can be compared by taking into consideration the kind of input feature 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. Most of the teams experimented a high-level representation of the text based deep learning solutions. While few teams like *fabsam* and *MDD* adopted a text representation based on traditional **word embeddings** such as Word2Vec (Mikolov et al., 2013), Glove (Pennington et al., 2014) and FastText (Bojanowski et al., 2017), most of the systems. i.e *NoPlaceForHateSpeech*,*jigsaw*, *PoliTeam*, *YNU_OXZ* and *UniBO*, exploited richer **sentence embeddings** such as BERT (Devlin et al., 2019) or XLM-RoBert (Ruder et al., 2019). For enriching the space for then training the subsequent models to recognize misogyny and aggressiveness, *PoliTeam* experimented the use of additional lexical resources such as misogynous lexicon and sentiment Lexicon.

Machine Learning Models. Concerning the machine learning models, we can distinguish between approaches trained from scratch and those ones based on fine-tuning of existing pre-trained models. We report in the following the strategy adopted by the systems that participated in the AMI shared task, according to the type of machine learning model that has been adopted:

- **Shallow models** have been experimented by

MDD, where logistic regressions have been trained according to different hand-crafted features;

- **Convolutional Neural Networks** have been exploited by *NoPlaceForHateSpeech* by using two distinct models for misogyny detection and aggressiveness identification, by *fabsal* investigating the optimal hyperparameters of the model, and by *YNU_OXZ* where on top of the CNN architecture a Capsule Network (Sabour et al., 2017) has been introduced for taking advantage of spatial patterns available in short texts;
- **Fine-Tuning of pre-trained models** has been exploited by *jigsaw* by adapting BERT to the challenge domain and using a transfer multilingual strategy and ensemble learning, by *UniBO* that accommodated the BERT model using a multi-label output neuron, and by *PoliTeam* where the prediction of the fine-tuned sentence-BERT is coupled with prediction based on lexicons.

For what concerns the achieved results on the two subtasks, few considerations can be drawn considering both the errors done by the systems and the mitigation strategies adopted for reducing the bias.

Error Analysis When testing the developed systems on raw test data, the majority of the performed errors can be summarized by the following patterns:

- **Under-representation of subjective expressions**: those posts written by introducing erroneous lower case and missing spaces between adjoining words lead the models based on raw text to make errors on test predictions. An example of such common errors is the one reported in the following tweet:

  > *"Odio Sakura per il semplice motivo che qualunque cosa faccia o dica Naruto lei lo prende a schiaffi o a pugniHA CHIESTO COME STA SAI DIOSANTO BRUTTA STRONZA MA CONTRALLI MADONNA SPERO CHE TI UCCIDANOscusami Sarada"*

- **Undefined subject, but presence of aggressive terms**: for those tweets where the target is not clearly mentioned, but several aggressive terms are present, the models tend to be biased and to predict the post as misogynous and aggressive erroneously. An example of this type of misclassified posts is reported here:

  > *"Vero...ma c'e chi ti cerca, che ti vuole, più di ogni cosa al mondo.....ma non sa se viene capito..... potrebbe esser mal interpretato e di conseguenza all'abbraccio esser denunciato per molestie sessuali e/o stupro"*

- **Short hate speech sentences referred to others than women**: when the target is mentioned by using an actual account, but it is referred to men, and there are no additional indications about the gender of the target, most of the models tend to misclassify the tweet. In the following example, the target is a male football player:

  > *"@bonucci_leo19 Cagati in mano e prenditi a schiaffi. Sti post te li infili nel c*lo!"*

Concerning the errors on the synthetic test set used for estimating the bias of the models, two main errors carried out by the majority of the systems can be identified:

- **presence of unusual target**: in most of the submissions, sentences containing offensive expressions towards specific uncommon targets are misclassified. For instance, around 39% of the predictions related to the target *nonna* (i.e., grandmother) are wrong. An example of the most misclassified target is reported in the following example:

  > *"nonne belle"*

- **Presence of unusual verbs**: analogously to what has been observed for the target, sentences containing rare aggressive verbs tend to be misclassified. For instance, around 48% of the instances related to the verbs *malmenare* and *seviziare* (i.e., beat up and torture) are wrongly classified. An example of a mistaken sentence are reported here:

  > *"femmina dovrebbe essere seviziata"* (wrongly classified as non-misogynous)

Bias Mitigation strategies. Concerning the Subtask B, only one team (*jigsaw*) addressed explicitly the problem related to the unintended bias. The authors used sentences sampled from the Italian Wikipedia articles containing some of the identity terms provided with the test set. These sentences, labeled as both non-misogynous and non-aggressive, have been used to further fine-tune the model and reduce the bias given by the data. The results achieved by the *jigsaw* team highlight that a debiasing method could obtain fair predictions even using pre-trained models.

7 Conclusions and Future Work

This paper presents the AMI shared task, focused not only on identifying misogynous and aggressive expressions but also on ensuring fair predictions. By analysing the runs submitted by the participants, we can conclude that while the problem of misogyny identification has reached satisfactory results, the recognition of aggressiveness is still in its infancy. Concerning the capabilities of the systems with respect to the unintended bias problem, we can highlight that a domain-dependent mitigation strategy is a necessary step towards fair models.

Acknowledgements

The work of the last author was partially funded by the Spanish MICINN under the research project MISMISFAKEnHATE on MISinformation and MIScommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31) and by the COST Action 17124 DigForAsp supported by the European Cooperation in Science and Technology.

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 *Proceedings of 23rd International Conference on Applications of Natural Language to Information Systems (NLDB)*, pages 57–64. Springer.

Giuseppe Attanasio and Eliana Pastor. 2020. PoliTeam @ AMI: Improving Sentence Embedding Similarity with Misogyny Lexicons for Automatic Misogyny Identification in Italian Tweets. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Bologna, Italy. CEUR.org.

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.

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.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of 13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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.

Adriano dos S. R. da Silva and Norton T. Roman. 2020. No Place For Hate Speech @ AMI: Convolutional Neural Network and Word Embedding for the Identification of Misogyny in Italian. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Bologna, Italy. CEUR.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Samer El Abassi and Sergiu Nisioi. 2020. MDD@AMI: Vanilla Classifiers for Misogyny Identification. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Bologna, Italy. CEUR.org.

Samuel Fabrizi. 2020. fabsam @ AMI: a Convolutional Neural Network approach. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Bologna, Italy. CEUR.org.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the Sixth evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2018)*, Turin, Italy. CEUR.org.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *IberEval@ SEPLN*, pages 214–228.

Simona Frenda, Bilal Ghanem, Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez, and Luis Villaseñor-Pineda. 2018. 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.

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.

Alyssa Lees, Jeffrey Sorensen, and Ian Kivlichan. 2020. Jigsaw @ AMI and HaSpeeDe2: Fine-Tuning a Pre-Trained Comment-Domain BERT Model. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Bologna, Italy. CEUR.org.

Marco Lepri, Giuseppe Grieco, and Mattia Sangermano. University of Pisa, Italy.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Arianna Muti and Alberto Barròn-Cedeño. 2020. UniBO@AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AlBERTo. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Bologna, Italy. CEUR.org.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 149–155.

Xiaozhi Ou and Hongling Li. 2020. YNU_OXZ @ HaSpeeDe 2 and AMI : XLM-RoBERTa with Ordered Neurons LSTM for classification task at EVALITA 2020. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Bologna, Italy. CEUR.org.

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 D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing*, pages 1532–1543.

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

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38. Association for Computational Linguistics.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866.

Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. HaSpeeDe 2@EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.