Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study

Information Processing and Management xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study

Endang Wahyu Pamungkas*, Valerio Basile, Viviana Patti

Department of Computer Science, University of Turin, Italy

ARTICLE INFO

Keywords:
Automatic misogyny identification
Abusive language online
Cross-domain classification
Cross-lingual classification
Social media

ABSTRACT

The freedom of expression given by social media has a dark side: the growing proliferation of abusive contents on these platforms. Misogynistic speech is a kind of abusive language, which can be simplified as hate speech targeting women, and it is becoming a more and more relevant issue in recent years. AMI IberEval 2018 and AMI EVALITA 2018 were two shared tasks which mainly focused on tackling the problem of misogyny in Twitter, in three different languages, namely English, Italian, and Spanish. In this paper, we present an in-depth study on the phenomena of misogyny in those three languages, by focusing on three main objectives. Firstly, we investigate the most important features to detect misogyny and the issues which contribute to the difficulty of misogyny detection, by proposing a novel system and conducting a broad evaluation on this task. Secondly, we study the relationship between misogyny and other abusive language phenomena, by conducting a series of cross-domain classification experiments. Finally, we explore the feasibility of detecting misogyny in a multilingual environment, by carrying out cross-lingual classification experiments. Our system succeeded to outperform all state of the art systems in all benchmark AMI datasets both subtask A and subtask B. Moreover, intriguing insights emerged from error analysis, in particular about the interaction between different but related abusive phenomena. Based on our cross-domain experiment, we conclude that misogyny is quite a specific kind of abusive language, while we experimentally found that it is different from sexism. Lastly, our cross-lingual experiments show promising results. Our proposed joint-learning architecture obtained a robust performance across languages, worth to be explored in further investigation.

1. Introduction

In the digital era, social media has an integral role in online communication, facilitating its users to publish and share contents providing accessible ways to express their feelings and opinions about anything anytime. Within the fields of artificial intelligence and natural language processing, this abundance of data allowed the research community to tackle more in depth long standing questions such as understanding, measuring and monitoring the sentiment of the users towards certain topics or events Cambria, Poria, Gelbukh, and Thelwall (2017), expressed in mere texts or also by relying on other visual and vocal modalities Poria et al. (2018). Robust and effective approaches are made possible by the rapid progress in supervised learning technologies and by the huge amount of user-generated content available online, especially on social media. Such techniques are typically motivated by purposes such as extracting user opinions on a given product or polling political stance. There is ever increasing awareness of the

* Corresponding author.
E-mail addresses: pamungka@di.unito.it (E.W. Pamungkas), valerio.basile@unito.it (V. Basile), viviana.patti@unito.it (V. Patti).

https://doi.org/10.1016/j.ipm.2020.102360
Received 7 February 2020; Received in revised form 29 May 2020; Accepted 15 July 2020

need to take a holistic approach to sentiment analysis by handling the many finer-grained tasks involved in extracting meaning, polarity and specific emotions from text, like the detection of sarcasm Majumder et al. (2019); Sulis, Farías, Rosso, Patti, and Ruffo (2016).

However, there is a downside to the freedom of expression given by social media, as more and more episodes of hate speech and online harassment happen in social media. This has determined a growing interest in artificial intelligence and natural language processing tasks related to social and ethical issues, also encouraged by the global commitment to fighting extremism, violence, fake news and other plagues affecting the online environment. In this perspective, let us mention the latest trends of "AI for social good", with emphasis on developing applications for maximizing the "good" social impacts, while minimizing the likelihood of harm, e.g., suicidal ideation detection for early intervention Gaur et al. (2019) and recent works on the prevention of sexual harassment Khatua, E., and Khatua (2018), sexual discrimination Khatua, Cambria, Ghosh, Chaki, and Khatua (2019), and cyberbullying and trolling Cambria, Chandra, Sharma, and Hussain (2010); Menini et al. (2019), or on hate speech counter-narratives Chung, Kuzmenko, Tekiroglu, and Guerini (2019), with focus on generating positive responses, after tackling with detection of abusive content published online, encouraging the community to adopt a proactive approach to transform the toxic environments into positive ones Jurgens, Chandrasekharan, and Hemphill (2019).

In recent years, hateful language and in particular the phenomenon of hate against women are exponentially increasing in social media platforms such as Twitter and Facebook Hewitt, Tiropanis, and Bokhove (2016); Poland (2016), becoming a relevant social problem that needs to be monitored. Misogyny, defined as the hate or prejudice against women, can be linguistically manifested in different and various ways, including social exclusion, sex discrimination, hostility, androcentrism, patriarchy, male privilege, belittling of women, disenfranchisement of women, violence against women, and sexual objectification Code (2002); Kramerae and Spender (2000). Based on the recent Online Harassment report from Pew Research Center[1], women are more likely to be targeted as subject of online harassment because of gender than men (11% vs. 5%). This is a concerning issue, since the study from Fulper et al. (2014) found that there is a strong association between the number of misogynistic tweets and the rape crime statistics in the United States.

Given the enormous amount of social media data produced in different regions of the world and also different languages[2], we also face the significant opportunity to develop tools that are able to detect and identify hateful language toward women across different languages.

The work of Hewitt et al. (2016) is a first study that attempts to detect misogyny in Twitter manually. The authors used several terms related to slurs against women to gather data from Twitter. The Automatic Misogyny Identification (AMI) campaign started by Anzovino, Fersini, and Rosso (2018) proposed a first benchmark dataset, capturing misogyny phenomena in Twitter. This dataset is a starting point for automatic misogyny identification, leading to two shared tasks focused on the detection of misogyny online, namely AMI IberEval 2018 Fersini, Rosso, and Anzovino (2018b) and AMI EVALITA 2018 Fersini, Nozza, and Rosso (2018a). AMI IberEval 2018 proposed an automatic misogyny identification task in two languages, Spanish (ES) and English (EN), while AMI EVALITA 2018 included Italian (IT) and English (EN). The task comprises two sub-tasks: i) classification of tweets as either misogynistic or not-misogynistic; ii) classification of misogynistic behaviour into 5 categories (derailing, dominance, discredit, sexual harassment and stereotype), and classification of the target of misogyny as active (individual) or passive (generic or group or women). These shared tasks succeeded in highlighting the barriers and difficulties of automatically detecting misogyny in social media.

Considering that more and more episodes of misogynistic hate speech and online harassment happen in social media, which stems from sexist stereotypes, prejudices and intolerance and which can lead to episodes of violence, discrimination and persecution also offline, our contribution is devoted to advance the understanding of online misogynistic behaviours. We propose for the first time a computational and multilingual study where the emphasis on a better conceptualization of misogyny and its relation to other abusive phenomena such as sexism, which are more subtle but contribute to a negative environment, is combined with the development of models for detecting misogynous contents in different languages and domains. In this way, we address the open challenge to enhance the robustness and accuracy of tools to contrast the harmful effects of misogynistic behaviors, e.g., tools for automatic support to moderation or for monitoring and mapping the dynamics and the diffusion of hate speech dynamics over a territory, which is only possible at a large scale by employing computational methods.

More specifically, we provide a deep analysis of the automatic misogyny identification task. In particular, we investigate the most predictive features for capturing misogynistic content in social media. In this direction, we explore the state of the art approaches on several available benchmark datasets provided by shared tasks. We experiment with three families of supervised classification models: i) Support Vector Machines using word ngrams as features; ii) Recurrent Neural Networks initialized with pre-trained word embeddings; iii) Transformer-based Neural Models, with pre-trained multilingual language models and fine-tuned for each classification task. We further include our own novel method, augmenting both the SVM and the deep learning models with knowledge from a multilingual abusive lexicon. We aim at studying the relation between misogyny and other kinds of hateful language online such as sexist and hate speech in the datasets we collected. To this aim, we experiment in a cross-domain classification setting, to explore the interaction between misogyny and other kinds of hateful language phenomena in terms of what information can be retained across tasks (transfer learning). Finally, as corpora on misogyny are only available in a limited number of languages, developing tools which work cross-lingually is particularly important. To this aim, we conduct experiments on automatic misogyny identification in a multilingual setting.

[1] https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/

[2] Almost 6,000 tweets per second: https://www.internetlivestats.com/twitter-statistics/

1.1. Research Questions

In this paper, we address the following research questions:

RQ1 *What are the most predictive features to distinguish between misogynistic and non-misogynistic content in social media?*
We investigated several state-of-the-art systems on available benchmark datasets from the AMI shared tasks. We found that most of the submitted systems used traditional machine-learning approaches. Therefore, we are interested to explore more deeply what are the most predictive features for the classifiers to detect misogyny content in social media.

RQ2 *How is misogyny related to other abusive phenomena, and how do they inform each other towards detection of abusive language at large?*
To answer this question, we collected several datasets of hateful language in social media, covering phenomena such as hate speech, sexism and offensive language. To further probe our hypothesis, we select datasets that are somewhat related to each other in terms of topic and target (i.e. women are the main target of the abusive attitude), as well as datasets very different in nature.

RQ3 *Is the knowledge about misogyny learned from one language informative to predict misogyny in other languages?*
The problem of hateful language, and specifically misogyny, is not constrained to the English language. Therefore, we experiment with a cross-lingual environment to detect misogyny. We use available datasets in three different languages, namely English, Spanish, and Italian, and build a system to classify misogyny in cross-lingual setting.

1.2. Contribution

The contribution of this paper can be summarized as following:

1. We present an extensive review of the state of the art in misogyny detection.
2. We propose a state-of-the-art model to detect misogyny in social media, and test it on several benchmark datasets.
3. We investigate the most predictive features to distinguish misogynistic content from not-misogynistic content.
4. We investigate the relationship between misogyny and other abusive phenomena by conducting a cross-domain classification setting, leveraging the knowledge transfer from other kinds of hateful language to detect misogyny and vice versa.
5. We present the results of experiments in a cross-lingual setting, aiming at learning and generalizing knowledge about misogyny over datasets in different languages.

This study extends and summarizes several previous works on automatic misogyny identification Pamungkas, Cignarella, Basile, and Patti (2018b,c); Pamungkas and Patti (2019), by providing further analysis and additional experiments to get a deeper insight on the AMI task.

The article is organized as follows. Section 2 introduces related work on automatic misogyny identification tasks, and some other tasks having a topical focus related to misogyny, such as sexism. We also review recent previous works which focus on both cross-domain and cross-lingual experiments in abusive language detection. Section 3 describes the AMI task and dataset. In addition, we describe other datasets which will be used in this work in Section 3. The proposed experimental settings for the AMI task as well as the results are presented in Section 4. In Section 5, we investigate the relationship between misogyny and other related phenomena by conducting a cross-domain classification experiment. Meanwhile, Section 6 presents the experimental settings and results of our cross-lingual experiments on the AMI task. Section 7 discusses, analyses, and highlights the results and the main findings of the experiments presented in previous sections. Finally, Section 8 includes conclusive remarks and ideas for future work.

2. Related Work

In this section, we give a review of the recent literature on automatic misogyny identification. Several studies are connected to descriptive reports of the systems participating in shared tasks, in particular, we identified two closely related shared tasks, reviewed and discussed below. We also provide an overview of the recent literature on abusive language detection, specifically in a cross-domain and cross-lingual perspective.

2.1. Detecting Sexism, Misogyny and Related Phenomena in Social Media

Misogynistic speech is a well-studied phenomenon in social media, and its detection is often cast as a text classification problem. Misogyny, defined as the hate or prejudice against women, can be linguistically manifested in various ways, including social exclusion, discrimination, hostility, threats of violence and sexual objectification. Misogynistic language is a multifaceted phenomenon with its own specificity and it is often imbued with expressions of sexism and offensive language. Moreover, since misogyny is a form of hate, the current studies on the automatic identification of this phenomenon are related to the field of automatic hate speech detection, and studied also in conjunction with other expressions of hate, as in the HatEval shared task proposed in 2019 at SemEval Basile et al. (2019). HatEval provided a Twitter data set annotated for hate speech against women and immigrants, where a contrastive comparison of misogynist and xenophobic messages in English and Spanish is possible. While the woman-targeted section of the HatEval dataset could be considered compatible to a misogyny detection benchmark, the distinction was not made explicit and the “target” label was not published. During the competition, the participant systems were evaluated on their capacity to predict hate speech on the whole set of tweets, regardless of the target.

Despite the philosophical debate on whether *sexism* and *misogyny* are distinct concepts Manne (2017), or whether misogynistic speech is *hate speech* Richardson-Self (2018), there is a strong relation between those phenomena. One of the first studies on sexism detection was proposed by Waseem and Hovy (2016), in conjunction with another abusive phenomenon, namely racism. The dataset has been widely adopted in the broader context of abusive language detection. Jha and Mamidi (2017) proposed another benchmark dataset, providing a distinction of sexism utterances into two forms: hostile (when sexism is characterized by an explicitly negative attitude) and benevolent (when sexism is more subtle, often expressed as a compliment). A more recent study by Sharifirad and Jacovi (2019) presented a new categorization of sexism including indirect, sexual, and physical sexism, building a CNN model to automatic classify tweets into these three categories. Several studies on abusive language detection used the datasets mentioned above, with different focuses such as hate speech detection Badjatiya, Gupta, Gupta, and Varma (2017); Fehn Unsvåg and Gambäck (2018); Kshirsagar, Cukuvac, McKeown, and McGregor (2018); Qian, ElSherief, Belding, and Wang (2018), author profiling in abuse detection Mishra, Tredici, Yannakoudakis, and Shutova (2018), bias in abusive language detection Davidson, Bhattacharya, and Weber (2019); Park, Shin, and Fung (2018); Wiegand, Ruppenhofer, and Kleinbauer (2019), and cross-domain abusive language detection Karan and Snajder (2018); Pamungkas and Patti (2019); Swamy, Jamatia, and Gambäck (2019); Waseem, Thorne, and Bingel (2018).

The earliest work we found specifically on *misogyny* in social media was proposed by Hewitt et al. (2016), where misogynistic tweets are collected by using several terms used to attack women, and coded manually by a single annotator. Research on automatic misogyny identification was boosted by Anzovino et al. (2018), introducing a new benchmark dataset annotated on two levels: i) misogyny identification, and ii) misogynistic behavior and target classification. They also built systems to detect misogynistic tweets automatically, employing several classifiers including random forest, naive bayes, support vector machine, and multi-layer perceptron. Two shared tasks investigate the misogyny phenomenon in social media on multiple languages, namely AMI IberEval 2018 Fersini et al. (2018b) (Spanish and English) and AMI EVALITA 2018 Fersini et al. (2018a) (Italian and English). Table 1 summarizes the participating systems in these shared tasks. Several approaches were proposed, from traditional supervised classifiers such as naive bayes, SVM, and random forest, to deep learning techniques such as Bi-LSTM. Some participants to the shared tasks proposed ensembles of classifiers, by aggregating the output from several classifiers to make the final prediction. However, the best systems in both campaigns are simple classifiers (SVM for AMI IberEval and logistic regression for AMI EVALITA) with manually engineered features.

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

2.2. *Cross-Domain Classification of Abusive Language*

We conducted experiments in a cross-domain setting to investigate the relationships between misogyny and related phenomena, including sexism and offensive language. Therefore, in this section we present relevant works which deal with cross-domain

Table 1
Summary of the AMI shared task systems.

Authors	Shared Task	Approach
Pamungkas et. al. Pamungkas et al. (2018c)	AMI IberEval	SVM with a combination of handcrafted stylistic, structural, and lexical features.
Goenaga et. al. Goenaga et al. (2018)	AMI IberEval	Bi-LSTM with pretrained word embeddings.
Liu et. al. Liu, Chiroma, and Cocea (2018)	AMI IberEval	Average probability of two traditional classifiers trained on doc2vec.
J. S. Canos Canós (2018)	AMI IberEval	SVM with tf-idf unigrams.
V. Nina-Alcocer Nina-Alcocer (2018)	AMI IberEval	SVM, Multi-layer Perceptron (MLP) and Multinomial Naive Bayes, with structural, lexical, and syntactical features.
E. Shushkevich Shushkevich and Cardiff (2018a)	AMI IberEval	Logistic regression, naive bayes, SVM, and ensemble classifier, with tf-idf.
Frenda et. al. Frenda, Ghanem, and Montes-y-Gómez (2018b)	AMI IberEval	Ensemble of SVM classifiers with character n-grams, sentiment, and lexicons.
Pamungkas et. al. Pamungkas et al. (2018b)	AMI EVALITA	Linear and RBF kernel SVM with structural and lexical features, including a multilingual hate lexicon.
A. Bakarov Bakarov (2018)	AMI EVALITA	Single Value Decomposition and boosting classifier with tf-idf.
Basile and Rubagotti Basile and Rubagotti (2018)	AMI EVALITA	SVM with n-grams and cross-lingual classification with bleaching.
Saha et. al. Saha, Mathew, Goyal, and Mukherjee (2018)	AMI EVALITA	Logistic regression trained on concatenated sentence embeddings, tf-idf, and average word embeddings.
Ahluwalia et. al. Ahluwalia, Soni, Callow, Nascimento, and Cock (2018)	AMI EVALITA	Voting ensemble with handcrafted features.
E. Shushkevich Shushkevich and Cardiff (2018b)	AMI EVALITA	Ensemble of logistic regression, SVM, and naive bayes, with tf-idf.
D. Buscaldi Buscaldi (2018)	AMI EVALITA	Bi-LSTM with character embedding and random forest with weighted n-grams.
Frenda et. al. Frenda, Ghanem, Guzmán-Falcón, Montes-y-Gómez, and Pineda (2018a)	AMI EVALITA	SVM and random forest with stylistic and lexical features and lexicons.

classification of abusive language. In line with previous works, we consider different abusive language phenomena as different domains — including sexism, misogyny, racism, hate speech, offensive language, where each abusive language phenomenon has different topical focus. Several studies have been carried out on cross-domain classification of online abusive language. In Waseem et al. (2018) the first attempt to deal with cross-domain classification in an abusive language detection task is reported, by proposing a multi-tasks learning (MTL) approach. They argue that MTL has the ability to share knowledge between two or more objective functions, so that it can leverage information encoded in one abusive language dataset to better fit others. They found that the difference of approaches in collecting and annotating datasets is the main factor which influences the performance of such model. Karan and Snajder (2018) proposed to use a traditional machine learning approach for classifying abusive language in a cross-domain setting, in order to get better interpretability of the system. This work also explored the use of the *frustratingly simple domain adaptation* (FEDA) framework Daumé III (2007) to facilitate domain sharing between different datasets. The main finding of this work is that the model did not generalize well when applied to different domains, even when trained on a much bigger out-domain dataset. In addition, the use of FEDA is able to improve the classifiers performance in most of the cases, indicating that a more sophisticated domain-informed approach might be useful in this scenario. Similarly, Pamungkas and Patti (2019) proposed a cross-domain classification of abusive language, employing a Long Short Term Memory (LSTM) netword and a list of abusive keywords from the lexicon HurtLex Bassignana, Basile, and Patti (2018), as a proxy to transfer knowledge across different datasets. Their main findings are that i) the model trained on more general abusive language dataset will produce more robust predictions, and ii) HurtLex is able to boost the system performance in cross-domain setting. The experiments proposed in this study can be considered a continuation on this line of research, but with a specific focus on misogyny detection. Bidirectional Encoder Representations from Transformers (BERT) Devlin, Chang, Lee, and Toutanova (2019) was also applied to the cross-domain setting in abusive language detection, as proposed by Mozafari, Farahbakhsh, and Crespi (2019); Swamy et al. (2019). Both studies found that BERT is capable to share knowledge between one domain dataset to other domains, in the context of transfer learning. They argue that the main difficulty in cross-domain classification of abusive language is caused by dataset issues and their biases, with the consequent incapability of the datasets to capture the complete phenomenon of abusive language. In this work, we experiment with BERT too, generally with positive results, possibly due to the narrower topical scope of misogynistic language.

Notice that, in most previous studies, a cross-domain experimental setting was used to test the generalization capability of models in detecting abusive language, or to test whether the knowledge can be transferred between different domains of abusive language. However, to the best of our knowledge, the use of a cross-domain experimental setting for studying the relationships among the specific phenomenon of misogyny and other abusive language domains is novel.

2.3. Cross-Lingual Classification of Abusive Language

Few works focus on cross-lingual experiments in misogyny detection. Basile and Rubagotti (2018) used the bleaching approach van der Goot, Ljubešić, Matroos, Nissim, and Plank (2018) to run cross-lingual experiments between Italian and English in the context of their participation to the AMI EVALITA 2018 Fersini et al. (2018a) evaluation campaign. Recent work by Pamungkas and Patti (2019) employs Multilingual Unsupervised or Supervised Word Embeddings (MUSE)[3] to build a joint-learning model for cross-lingual classification on the AMI task in three languages, namely Italian, English, and Spanish. In addition, there are studies on cross-lingual classification of abusive language, with a general topical focus. Most of the relevant work originates from the participation to recent shared tasks on German Offensive Language Identification Schneider, Roller, Bourgonje, Hegele, and Rehm (2018), Automatic Misogyny Identification task Basile and Rubagotti (2018), and Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC) Saha, Mathew, Goyal, and Mukherjee (2019). Schneider et al. (2018) used multilingual embeddings in a cross-lingual experiment when participating in GermEval 2018, to be able to use English dataset to train the system and test it on German data Wiegand, Siegel, and Ruppenhofer (2018). Saha et al. (2019) employed Multilingual-BERT[4] and language agnostic sentence embedding representations (LASER Artetxe & Schwenk (2019)) as feature representation, trained on Light Gradient Boosting Machine (LGBM) Ke et al. (2017) to build a language-agnostic system to detect hate speech, in their participation to the HASOC@FIRE 2019 shared task which covered three languages including English, German, and Hindi Mandl et al. (2019). Finally, Ousidhoum, Lin, Zhang, Song, and Yeung (2019) conducted a multilingual experiment on hate speech detection in three languages (i.e., English, France, and Arabic) by using Sluice Network Ruder12, Bingel, Augenstein, and Søgaard (2017) and Babylon multilingual word embeddings Smith, Turban, Hamblin, and Hammerla (2017).

Summarizing, to our knowledge, this paper is the first study where a conceptualization of misogyny is combined with the development of models for detecting misogynous contents in different languages and domains. Previous works were focusing on abusive language in general or on one dimension only, e.g., cross-domain abusive language detection, or difference between sexism and misogyny, in one domain and one language. Multilingual abusive language detection was explored in our own previous work Pamungkas and Patti (2019), however not focusing on misogyny.

3. Automatic Misogyny Detection: Task and Datasets

In this section, we present a detailed description of the Automatic Misogyny Identification shared tasks (AMI), including their

[3] https://github.com/facebookresearch/MUSE

[4] https://github.com/google-research/bert/blob/master/multilingual.md

definition, evaluation procedure, and the datasets provided to the participants. The data in particular form the basis of the experimental work of the present paper. We also include two additional datasets to further validate out hypotheses, namely one widely used benchmark for abusive language detection and the corpus from the hate speech detection evaluation campaign HatEval, both comprising subsets of messages with misogynistic content.

3.1. Task Definition

AMI is organized as a text classification task across different dimensions. The shared task comprises two subtasks, A and B. The main objective of the AMI task is to discriminate between misogynistic and not-misogynistic content in a binary classification fashion (subtask A). As a secondary goal, systems are asked to categorize misogynistic content into five different misogynistic behaviours and to classify the target of the misogynistic instances (subtask B). The five categories of misogynistic behaviours can be defined as follows:

1. **Stereotype and Objectification**: over-generalization of the women' image, including personality, preferences, and abilities to a very narrow standard.
2. **Dominance**: the intention to show that men are superior to women in a context of gender inequality.
3. **Derailing**: confirming abuse towards women by rejecting male responsibility, or an effort to disrupt conversation in order to redirect womens conversations on something more comfortable for men.
4. **Sexual harassment and threat of violence**: an action to harass women that relates to sexual and inappropriate promise of rewards in exchange for sexual favors. Also includes intent to physically assert power over women through violent threats.
5. **Discredit**: lack of respect toward women, which could also contain slurs.

Target classification is a binary classification task where the categories are defined as follows:

1. **Active**: when the misogyny is specifically target an individual.
2. **Passive**: when the misogyny targets more than one individual or a group of woman.

Subtask A is evaluated in terms of accuracy, while subtask B is evaluated by using the macro-average F-score for misogynistic behaviour and target, and their arithmetic mean. The reason for using a different metric for subtask B is the unbalance within both the Misogynistic Category Classification and the Target Classification, whereas accuracy would not measure performance as fairly in subtask B with respect to subtask A.

3.2. Dataset

The datasets for AMI IberEval and AMI EVALITA were collected from Twitter following the same procedure, consisting of three different approaches:

1. From the Streaming API, downloading tweets containing representative keywords frequently used to harass women, as introduced in Hewitt et al. (2016), such as “whore”, “cunt”, and “bitch”.
2. Monitoring a selection of Twitter account of potential victim of harassment and known feminist activists, such as personalities involved in the Gamergate scandal[5].
3. Downloading tweets from the history of misogynist accounts. These users declare that they are misogynists based on the information shown on the account profile or screen name.

The collection of AMI IberEval was gathered in the span of more than 4 months, starting from 20th of July 2017 until 30th of November 2017, resulting in 83 million tweets for English and 72 millions tweets for Spanish. The shared task organizers queried subsets of tweets for English and Spanish based on co-presence of some keywords. These subsets were then partially annotated fully by two annotators, and by a third annotator to solve the disagreement cases, to build a gold standard set. The rest of tweets in these subsets were annotated by crowd-sourcing with CrowdFlower (now called Figure Eight[6]), where the gold standard was used as test question set. The labels of the crowd-sourced data were decided by using a majority vote approach. The final dataset of AMI IberEval consists of 3,977 tweets (3,251 for training and 726 for testing) for English and 4,138 tweets (3,307 for training and 831 for testing) for Spanish. The detailed distribution of the dataset is shown in Table 2.

The AMI EVALITA collection was gathered in the same time period as the one collected for AMI IberEval. The organizers queried the initial collection with a set of predefined keywords, obtaining 10,000 tweets for each language, English and Italian. The annotation process involved six experts using CrowdFlower.

The collection strategy adopted to construct the AMI dataset is partially keyword-based. As recently highlighted in Wiegand et al. (2019), the adoption of keyword-based data collection processes can introduce biases in the data, in terms of the topics they

[5] https://www.theguardian.com/technology/2016/dec/01/gamergate-alt-right-hate-trump

[6] https://www.figure-eight.com/

Table 2
AMI IberEval Dataset label distribution.

Task A			Task B		
	English	Spanish		English	Spanish
Misogynistic	1,568/283	1,649/415	Stereotype	137/72	151/17
			Dominance	49/28	302/54
			Derailing	29/28	20/6
			Sexual Harassment	410/32	198/51
			Discredit	943/123	978/287
			Active	942/104	1455/370
			Passive	626/179	194/45
Not misogynistic	1,683/443	1,658/416	No class	1,683/443	1,658/416
Total				3,251/726	3,307/831

Table 3
AMI EVALITA Dataset label distribution (training/test).

Task A			Task B		
	English	Italian		English	Italian
Misogynistic	1,785/460	1,828/512	Stereotype	179/140	668/175
			Dominance	148/124	71/61
			Derailing	92/11	24/2
			S. Harassment	352/44	431/170
			Discredit	1,014/141	634/104
			Active	1,058/401	1,721/446
			Passive	727/59	96/66
Not misogynistic	2,215/540	2,172/488	No class	2,215/540	2,172/488
Total				4,000/1,000	4,000/1,000

cover, and therefore it impacts the representativeness of the corpora. Concerning the AMI dataset, the problem is partially mitigated by embracing a combined approach where the keyword-based filtering of Twitter streams is combined with the retrieval of tweets obtained by monitoring potential victims of hate accounts and downloading the history of identified haters and filtering Twitter. However, also this combined strategy presents some limitation in terms of coverage of misogynistic behavior, probably leaving out interesting samples of misogynistic behaviours like benevolent misogyny and disfranchisement. Such phenomena, with few exceptions Jha and Mamidi (2017), are often neglected in current studies for being either too subtle or quite rare.

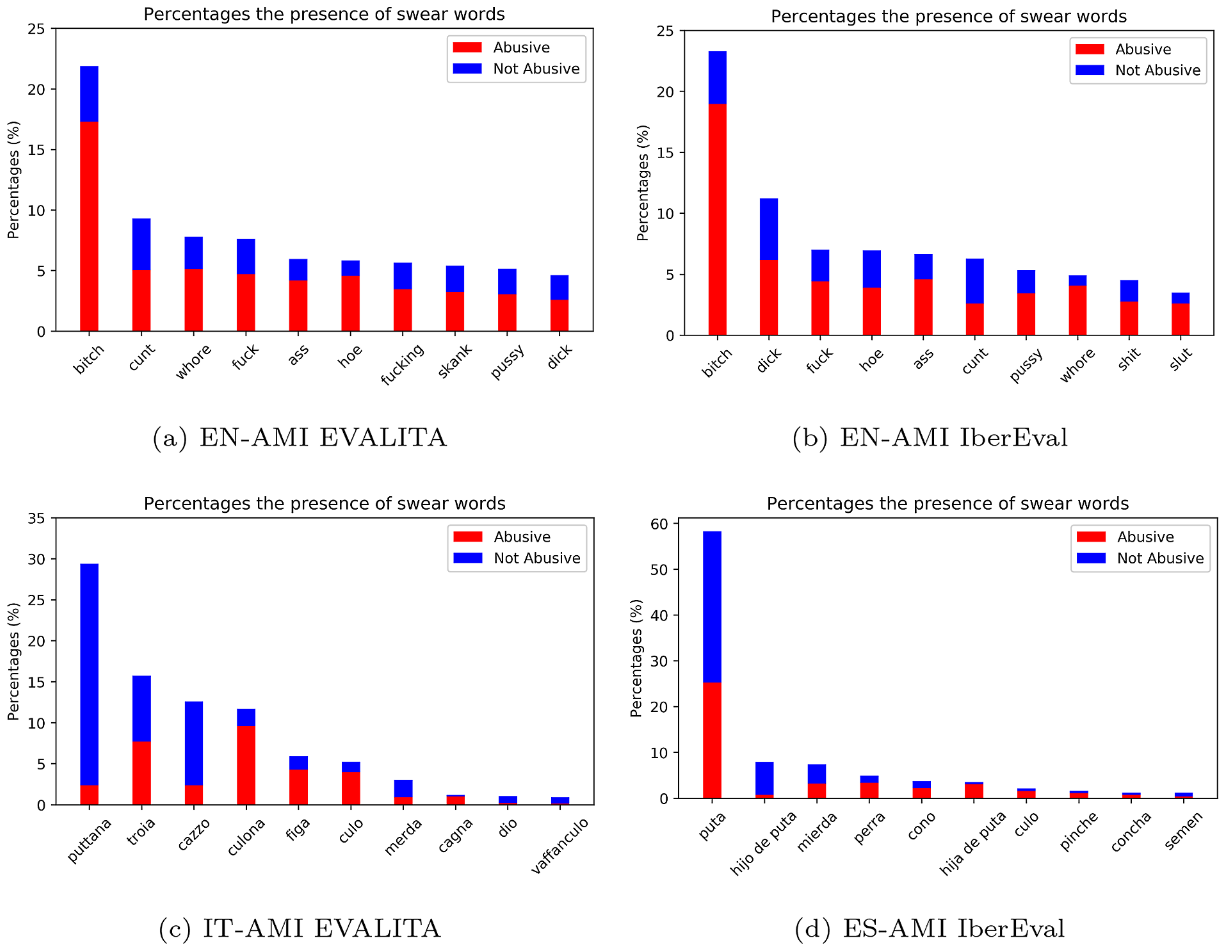
The final collection for the AMI EVALITA shared task comprises 5,000 tweets (4,000 for training and 1,000 for testing) for English and 5,000 tweets (4,000 for training and 1,000 for testing) for Italian. The overall inter-annotation agreement on the English set for “misogynistic”, “misogyny behaviour”, and “misogyny target” is 0.81, 0.45 and 0.49 respectively, while for Italian are slightly higher 0.96, 0.68 and 0.76. The detailed distribution of AMI EVALITA dataset is shown in Table 3. Interestingly, it can be observed that the label distribution are very imbalanced for the task B, where *discredit* is the most dominant category of misogynistic behaviour. Some classes were naturally under-represented in these data, such as derailing and dominance. Notice that such resulting class imbalance could have been also affected by the collection strategy applied to construct the data.

Moreover, the *active* class is definitely more represented than the *passive* one when we consider the target of misogyny. Notice that this result fits the most recent theoretical accounts of misogyny from philosophy Manne (2017): “Most misogynistic behavior is about hostility towards women who violate patriarchal norms and expectations, who aren’t serving male interests in the ways theyre expected to. So there’s this sense that women are doing something wrong: that theyre morally objectionable or have a bad attitude or they’re abrasive or shrill or too pushy”. In fact, in the AMI datasets it can be observed that it is the individual woman that, violating the patriarchal norms and expectations, triggers the misogynistic verbal attack online.

We also carried out a lexical analysis on the AMI datasets in all the languages, with the aim of gaining insight about predictive features for the detection task. Fig. 1 depicts the distribution of offensive words in all four AMI datasets of the EVALITA and IberEval 2018 evaluation campaigns. The red part of the bars shows the frequency of each swear word when it is used in misogynistic tweet, while the blue one is the frequency when used in messages labeled as not-misogynistic. We took the list of swear words from the online swear word dictionary of noswearing[7]. Based on these figures, we found that the use of specific slurs related to prostitution, female and male genitalia, physical disability and diversity (basically the same words in every languages) is very dominant in all dataset collections across languages[8]. Focusing on the English datasets, misogynistic slurs are mainly used in an abusive/misogynistic context, while in two other languages they are more evenly distributed.

[7] https://www.noswearing.com/

[8] We adopt this categorization from HurtLex Bassignana et al. (2018).



(a) EN-AMI EVALITA

(b) EN-AMI IberEval

(c) IT-AMI EVALITA

(d) ES-AMI IberEval

Fig. 1. Top 10 swear words of each dataset.

3.3. Related Datasets

Besides the AMI task datasets, in our study we considered two additional datasets with topical focus on the related notions of sexism and hate speech in social media.

The **Waseem and Hovy Hate Speech Dataset** was collected in the duration of 2 months, for a total of 136,052 tweets. This collection was bootstrapped by conducting a manual search based on several common slurs and terms related to sexual, religious, gender, and ethnic minorities, using public Twitter API search. The authors identified a final set of keywords[9] frequently used in tweets that contain hate speech and references to specific entities. The portion of collected tweets were manually annotated by the authors, and the annotation were reviewed by a student in gender studies, in order to mitigate annotator bias. The detailed annotation guideline is available in Waseem and Hovy (2016). The final annotated dataset consists of 16,914 tweets, coded in three categories: racist (1,972 tweets), sexist (3,383 tweets), and none (11,559 tweets). The overall inter-annotator agreement based on Cohen's Kappa coefficient is 0.84, where 85% of the disagreement cases occur in the annotation of sexism. Besides the tweet text, the authors also collected the demographic information of the authors, but this information was not publicly released. The tweet IDs and final annotation labels are available in the Github page[10]. Due to data decay, on our latest effort to retrieve the dataset, we were able to obtain 16,488 out of the 16,914 tweets, of which 1,957 marked as racist, 3,216 as sexist and 11,315 as none.

HatEval was introduced at SemEval 2019 Basile et al. (2019) and focuses on the detection of hate speech in Twitter on two specific targets, namely immigrants and women, in a multilingual perspective. This shared task introduced a dataset in two languages, English and Spanish. The dataset was collected by using different strategies, mainly in the time span from July to September 2018 for the subset targeting immigrants. The subset of tweets against women is mostly gathered the English part of AMI IberEval and AMI EVALITA. Similarly to AMI, this dataset was also collected using three different approaches, including : 1) monitoring accounts

[9] MKR, asian drive, feminazi, immigrant, nigger, sjw, WomenAgainstFeminism, blameonenotall, islam terrorism, notallmen, victimcard, victim card, arab terror, gamergate,jsil, racecard, race card

[10] https://github.com/zeerakw/hatespeech

of potential hate speech targets, 2) downloading the tweets from known hateful accounts, and 3) filtering the Twitter stream by using some specific keywords. The keywords include neutral words Sanguinetti, Poletto, Bosco, Patti, and Stranisci (2018), pejorative words towards the targets, and highly polarized hashtags. Based on the retrieved collection, the distribution of the keywords over the collection is skewed, with some keywords more frequently occurring than others, including “migrant”, “refugee”, “#buildthatwall”, “bitch”, “hoe”, “women” for English and “inmigraarabe”, “sudaca”, “puta”, “callate”, “perra” for Spanish. The collected tweets were annotated by non-trained contributors using Figure Eight with three binary labels: hate speech (HS), target range (TR: generic or individual), and aggressiveness (AG)[11]. The average confidence score as reported by Figure Eight for these three labels is 0.83, 0.70 and 0.73 respectively for English and 0.89, 0.47 and 0.47 for Spanish. The final dataset used for the HatEval shared task contains 13,000 (about 10,000 for training and for 3,000 testing) tweets for English and 6,600 (about 5,000 for training and 1,600 for testing) tweets for Spanish[12]. Table 4 and Table 5 show the detailed label distribution of the dataset for each target.

4. Automatic Misogyny Identification Experiment

In this section, we present our experiment at building a system with a comparable or better performance than the state of the art to detect misogyny. We use the AMI IberEval and AMI EVALITA benchmark datasets for all languages, namely English (EN), Spanish (ES), and Italian (IT), to evaluate our system. We explore several approaches, including traditional machine-learning models and more recent deep learning techniques. The system performance is evaluated along several metrics such as precision, recall, F-score, and accuracy for subtask A, and accuracy and macro-averaged F_1-score for subtask B, as explained in Section 3.

4.1. Traditional Models

We built two Support Vector Machine (SVM) models using different kernel functions, namely linear and radial basis function (RBF). The use of linear kernel is based on Joachims (1998), who argue that linear kernel has an advantage for text classification, based on the observation that text representation features are frequently linearly separable. The RBF kernel is preferable to a linear kernel for some text classification task due to its better performance, despite it having higher complexity Pamungkas, Basile, and Patti (2018a); Pamungkas, Cignarella, Basile, and Patti (2018c).

We employ several stylistic and lexical features, performing a straightforward pre-processing step including tokenization and stemming by using Natural Language Toolkit (NLTK)[13]. Specifically, we employ the features detailed in the following sections.

4.1.1. Lexical Features

This set of features aims at representing the semantic content of the tweets at the lexical level.

Bag of Words. This feature includes unigram, bigram, and trigram representation of the tweets, where all characters were changed to lower case.

Bag of Hashtags. We observed that hashtags[14] were frequently used in both AMI datasets. This feature is built by using the same technique as bag of words which includes unigram, bigrams, and trigrams (some tweets have more than one hashtags), focusing on the hashtag presence.

Bag of Emojis. Similarly to hashtags, emojis were also utilized in many instances in the AMI datasets. We normalize every emoji into its Unicode Common Locale Data Repository (CLDR) short name by using the *emoji* library[15].

Swear Words. This feature includes the presence of swear words which are often indicative of abusive content. The list of keywords is gathered from the *noswearing* website[16], an online dictionary which contains 349 English swear words. For the other languages, we translate the swear words automatically by using Google Translate[17], and including other sources such as the list of bad words from Wikipedia page[18] and a list of manually checked swear words by a popular linguist blog[19] for Italian. We encode the information about swear words into two individual features: swear word presence (binary feature) and swear word count (the number of swear words).

Sexist Slurs. We include the list of sexist words proposed by Fasoli, Carnaghi, and Paladino (2015), which are often used in hate speech messages against women. We manually translate and expand these words for Italian and Spanish. This feature has a binary value of 0 if there is no sexist slur in the tweet, or 1 if there is at least one sexist slur in the tweet.

Women-related words. We also manually built a list or words containing synonyms and related words to “woman” (for English), “donna” (for Italian), and “mujer” (for Spanish). This list of words represents a feature to detect the target of hateful content, in this

[11] the detailed description of annotation guidelines is available at https://github.com/msang/hateval/blob/master/annotation_guidelines.md.

[12] Upon manual investigation, organizers decided to exclude 1,000 tweets from the English training set, 29 tweets from the English test set due to duplicated instances, and 500 tweets from the Spanish training set, due to duplicated instances.

[13] https://www.nltk.org/

[14] We also experimented by splitting the hashtags into their constituent words using Ekphrasis Baziotis, Pelekis, and Doulkeridis (2017), but this did not improve the system performance.

[15] https://pypi.org/project/emoji/

[16] https://www.noswearing.com/

[17] https://translate.google.com/

[18] https://it.wikipedia.org/wiki/Turpiloquio_nella_lingua_italiana

[19] https://www.parolacce.org/2016/12/20/dati-frequenza-turpiloquio/

Table 4
HatEval Dataset label distribution. Hate speech target: Women.

Main class	Fine-grained class	Training	Development	Test
		English		
Hate Speech		1,985	237	623
	Aggressive	558	110	214
	Not-Aggressive	1,427	127	409
	Generic	752	27	122
	Individual	1,233	210	501
Not Hate Speech		2,515	263	849
Total (HS + not HS)		4,500	500	1,472
		Spanish		
Hate Speech		1,185	143	336
	Aggressive	1,036	127	311
	Not-Aggressive	149	16	25
	Generic	149	16	17
	Individual	1,036	127	319
Not Hate Speech		1,697	184	463
Total (HS + not HS)		2,882	327	799

Table 5
HatEval Dataset label distribution. Hate speech target: Immigrant.

Main class	Fine-grained class	Training	Development	Test
		English		
Hate Speech		1,798	190	629
	Aggressive	1,001	94	376
	Not-Aggressive	797	96	253
	Generic	1,690	181	608
	Individual	108	9	21
Not Hate Speech		2,702	310	870
Total (HS + not HS)		4500	500	1499
		Spanish		
Hate Speech		672	79	324
	Aggressive	466	49	163
	Not-Aggressive	206	30	161
	Generic	579	69	220
	Individual	93	10	104
Not Hate Speech		946	94	476
Total (HS + not HS)		1,618	173	800

case towards women. Similarly to the sexist slur feature, this feature is also represented as binary number, 0 (there is no woman-related word in the tweet) and 1 (there is at least one woman-related word in the tweet).

Hate Words Lexicon. This feature captures the presence of words contained in multilingual hate lexicon HurtLex Bassignana et al. (2018). This lexicon was built starting from a list of words compiled manually by the Italian linguist Tullio De Mauro De Mauro (2016) in Italian, then semi-automatically translated into 53 languages. The lexical items are divided into 17 categories. For our system configuration, we exploited the presence of the words in each category as a single feature, thus obtaining 17 single features, one for each HurtLex category. The full list of HurtLex categories can be seen in Table 6. We included this feature because our preliminary lexical analysis suggests that a specific subset of the HurtLex categories can be relevant to detect the misogynistic speech in social media, such as PR (words related to prostitution), ASF (words related to female genitalia), DDP (phsysical disability and diversity), and DDF (cognitive disability and diversity).

4.1.2. Stylistic Features

This set of features aims at capturing the structure of the tweets in terms of the type of some of its constituent elements.

Hashtag Count. The number of hashtags contained in the tweets.

Upper Case Count. The number of upper case characters in tweets.

Link Counts. The number of URLs in the tweets.

Tweet Length. The total number of characters of every tweet.

4.2. Deep Learning

We adopt two kind of deep learning architectures including recurrent neural networks (RNN) based and transformer based, where

Table 6
HurtLex Categories.

Category	Description
PS	Ethnic Slurs
RCI	Location and Demonyms
PA	Profession and Occupation
DDP	Physical Disabilities and Diversity
DDF	Cognitive Disabilities and Diversity
DMC	Moral Behavior and Defect
IS	Words Related to Social and Economic antage
OR	Words Related to Plants
AN	Words Related to Animals
ASM	Words Related to Male Genitalia
ASF	Words Related to Female Genitalia
PR	Words Related Prostitution
OM	Words Related Homosexuality
QAS	Descriptive Words with Potential Negative Connotations
CDS	Derogatory Words
RE	Felonies and Words Related to Crime and Immoral Behavior
SVP	Words Related to the Seven Deadly Sins of the Christian Tradition

we employ BERT. RNN was recognized as an effective architecture for learning text, also in text classification tasks. In this study we will implement two variants of RNN, namely long short term memory (LSTM) and gated recurrent unit (GRU). BERT is a transformer-based architecture which gained a lot of attention in NLP because of its superiority in most standard benchmarks. Here we describe both architectures.

4.2.1. RNN-based

We use straightforward Long Short-Term Memory (LSTM) Hochreiter and Schmidhuber (1997) and Gated Recurrent Units (GRU) Cho et al. (2014) networks. Our architecture consists of several layers, starting with an embedding layer (300 dimensions), where we experiment with and without pre-trained word embeddings. We employ the readily available embeddings provided by FastText[20] in three languages, i.e., English, Spanish, and Italian. The embedding layer is input to whether LSTM or GRU network (64 units), followed by a dense layer (16 units) with ReLU activation function. The final layer consists of a dense layer with sigmoid activation producing the final prediction. We only optimize the batch size (16, 32, 64, 128) and number of epochs (1-5) to tune our architecture in order to get the best possible result.

4.2.2. BERT

We also adapt BERT Devlin et al. (2019) for this experiment. We utilize the pre-trained models available on tensorflow-hub[21], which allows us to integrate BERT in the Keras libraryl[22]. For English, we use the `bert-cased` model, while for Italian and Spanish we use the `bert-multi-cased` model. Our network starts with the BERT layer, which takes three inputs consisting of id, mask and segment. The output of this layer connects to a dense layer with RELU activation (256 units), before passing into a dense layer with sigmoid activation as the predictor layer. We train our network with the Adam optimizer with learning rate 2^{-5}. We fine-tune the model only on the number of epochs (1-5) and batch size (16, 32, and 64).

4.3. Results

Table 8 shows the results of our experiment on subtask A. Since the task organizers only provide accuracy score as the competition baseline, we also built a baseline system with the same configuration as the competition baseline, that is, a linear SVM with word unigram representations as features. Despite not obtaining the exact score provided by the organizers, the score is still relatively comparable. We optimize this model on the training set by testing several combinations of features. We selected the best-performing system based on 10-fold cross evaluation, to be evaluated on the test set. Therefore, our best system configuration is not always containing all the features mentioned in the previous subsection. The deep learning systems were optimized by fine-tuning only on the number epochs and batch sizes. Overall, we got the best results on all benchmark datasets. The features and system configurations of our best-performing systems for the respective datasets can be found in Table 7.

For the traditional model, we use the same model as our contributed system in AMI IberEval Pamungkas et al. (2018c), which obtained the top ranking in the competition on both English and Spanish. In English the best result was obtained by a support vector machine (SVM) classifier with RBF kernel and several handcrafted features including *hashtags presence, links presence, swear words count, swear words presence, sexist slurs presence*, and *woman words presence*. We used the default hyper-parameters as defined by the

[20] https://fasttext.cc/
[21] https://www.tensorflow.org/hub
[22] https://keras.io/

Table 7
List of features of best-performing systems on each dataset.

EN-AMI IberEval	ES-AMI IberEval
SVM with RBF Kernel	SVM with Linear Kernel
- Swear words count	- Bag of words
- Swear words presence	- Bag of hashtags
- Hashtags presence	- Bag of emojis
- Links count	- Sexist slurs presence
- Sexist slurs presence	- Women words presence
- Women words presence	- ASF presence
	- ASM presence
	- DDF presence
	- DDP presence
	- PR presence

EN-AMI EVALITA	IT-AMI EVALITA
BERT	BERT
- With `bert-cased` model	- With `bert-multi-cased`
- Dense layer units = 256	- Dense layer units = 256
- Batch size = 32	- Batch size = 32
- Epoch = 2	- Epoch = 2

scikit-learn library[23]. Our system achieves an accuracy of 91.32, a significant improvement compared to the baseline. Meanwhile, our system for Spanish was also developed based on SVM but with linear kernel, coupled by some classic text representation as shown in Table 7. This model obtained 81.47 in accuracy. Our BERT models also achieved the best performance on both the English and Italian sets of AMI EVALITA, outperforming the best performing systems of the respective shared task. Our BERT model obtained 71.6 and 84.8 in accuracy on English and Italian respectively.

We also experimented with the four AMI tasks on the subtask B: Misogynistic Behaviour and Target Classification. We used the same systems as in the subtask A experiment, but different evaluation metrics are applied, namely accuracy and macro-averaged F_1-score. We need to clarify that in the official AMI shared tasks, subtask A and subtask B are treated as a pipeline process, where the prediction of subtask B will be fully dependent on the subtask A results. Rather, in this experiment, we handle subtask B as an independent multi-class classification task. Table 9 shows the full results of the experiments on the subtask B on all four AMI datasets. We compare our system performance with the AMI competition baseline and the best systems. The results show that our proposed systems were able to outperform the best performing systems on all the AMI tasks, based on the average of the macro-averaged F_1-scores on the two classification tasks of subtask B (misogynistic behaviour and target classification). Overall, BERT was the most consistent model, which gave the best performance on all dataset collections. Only in Italian AMI EVALITA, SVM with linear kernel perform slightly better than BERT. Most systems based on SVM with RBF kernel were under-performing on all datasets, compared to other systems. The big picture of the results also tells us that classifying the target of misogynistic behaviour is an easier task than determining its category, maybe due to the unbalanced distribution of classes in category of misogyny. The low annotator agreement on the “misogyny behaviour”, and “misogyny target” layers in the AMI dataset could also contribute to the difficulty of subtask B, especially on English AMI EVALITA, where the inter-annotator agreement on the dataset is only 0.45 and 0.49 for target classification and category of misogyny, respectively. On the one hand, the low annotator agreement can be a signal for the difficulty of this finer-grained tasks, especially concerning the detection of misogyny behaviours: drawing a sharp separation between the different categories has been difficult also for humans. On the other hand, it can be an alert for a possible inconsistency in the data annotation, that could cause problems to the model to learn the overall phenomena.

5. Relationship between Misogyny and other Abusive Phenomena

5.1. Experimental Setup

In this section, we present the results of an experiment carried out with the goal of studying the relationship between misogyny and other abusive language phenomena including sexism, hate speech, and offensive language. In essence, we train models on additional datasets (different abusive phenomena) and test their prediction capability for misogyny detection on the AMI benchmark. Furthermore, we train models on misogyny datasets and test their classification of other abusive phenomena. Basically, we use the same system as in the misogyny detection experiment in Section 4. We employ two classifiers: a Linear Support Vector Classifier (LSVC) and a Long Short-Term Memory (LSTM) architecture with additional features extracted from HurtLex. Our motivation is that LSVC is has a higher degree of interpretability, while deep learning is capable of better generalization. Furthermore, HurtLex, being a domain-neutral lexicon, is used as an aid for transferring knowledge between datasets with different domains. In addition to these

[23] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Table 8
Results of Automatic Misogyny Identification Experiment on AMI Dataset Task A.

	English AMI IberEval			
	P	R	F_1	Acc
Baseline of the shared task	-	-	-	78.37
Best system of the shared task Pamungkas et al. (2018c)	-	-	-	91.32
New baseline	73.63	71.02	72.30	78.79
Support vector classifier with linear kernel	82.70	69.26	75.38	82.37
Support vector classifier with RBF kernel	**87.16**	91.16	**89.12**	**91.32**
LSTM without pre-trained embedding	74.63	71.73	73.15	79.48
LSTM with FastText embedding	74.12	59.72	66.14	76.17
GRU without pre-trained embedding	65.34	81.27	72.44	75.9
GRU with FastText embedding	68.35	76.33	72.12	77.00
LSTM Attention without pre-trained embedding	65.33	69.26	67.24	73.69
LSTM Attention with FastText embedding	74.86	46.29	57.21	73.00
BERT	77.31	**91.52**	83.82	86.23

	Spanish AMI IberEval			
	P	R	F_1	Acc
Baseline of the shared task	-	-	-	76.78
Best system of the shared task Canós (2018); Pamungkas et al. (2018c)	-	-	-	81.47
New baseline	72.92	73.98	73.44	73.29
Support vector classifier with linear kernel	80.71	82.65	**81.67**	**81.47**
Support vector classifier with RBF kernel	54.26	46.02	49.80	53.67
LSTM without pre-trained embedding	76.40	75.66	76.03	76.17
LSTM with FastText embedding	76.42	78.07	77.23	77.02
GRU without pre-trained embedding	**81.95**	68.92	74.87	76.90
GRU with FastText embedding	77.03	82.41	79.62	78.94
LSTM Attention without pre-trained embedding	74.59	77.11	75.83	75.45
LSTM Attention with FastText embedding	75.22	82.65	78.76	77.74
BERT	70.84	**87.23**	78.18	75.69

	English AMI EVALITA			
	P	R	F_1	Acc
Baseline of the shared task	-	-	-	60.50
Best system of the shared task Bakarov (2018)	-	-	-	70.40
New baseline	55.70	65.87	60.36	60.20
Support vector classifier with linear kernel	44.44	34.78	39.24	50.00
Support vector classifier with RBF kernel	57.54	67.17	61.99	62.10
LSTM without pre-trained embedding	64.39	49.13	55.73	64.39
LSTM with FastText embedding	63.61	57.39	60.34	65.30
GRU without pre-trained embedding	52.12	**69.57**	59.59	56.6
GRU with FastText embedding	56.85	66.74	61.40	61.40
LSTM Attention without pre-trained embedding	56.61	63.26	59.75	60.80
LSTM Attention with FastText embedding	57.88	63.04	60.35	61.90
BERT	**70.37**	66.09	**68.16**	**71.6**

	Italian AMI EVALITA			
	P	R	F_1	Acc
Baseline of the shared task	-	-	-	83.00
Best system of the shared task Saha et al. (2018)	-	-	-	84.40
New baseline	77.92	93.75	85.11	83.20
Support vector classifier with linear kernel	77.24	**97.46**	**86.18**	83.90
Support vector classifier with RBF kernel	76.52	78.91	77.69	76.80
LSTM without pre-trained embedding	79.70	92.77	85.74	84.20
LSTM with FastText embedding	82.10	88.67	85.26	84.30
GRU without pre-trained embedding	78.05	92.38	84.62	82.80
GRU with FastText embedding	78.35	94.73	85.76	83.90
LSTM Attention without pre-trained embedding	82.28	88.87	85.45	84.50
LSTM Attention with FastText embedding	79.05	94.34	86.02	84.30
BERT	**83.93**	87.11	85.44	**84.80**

Table 9
Result of Experiment on SubTask B.

English AMI EVALITA	Category		Target		Average	
	Accuracy	Macro F_1	Accuracy	Macro F_1	Accuracy	Macro F_1
Baseline of Shared Task	-	.342	-	.399	-	.371
Best System of Shared Task	-	**.361**	-	.451	-	.406
SVM Linear Kernel	.544	.355	.579	.484	.562	.419
SVM RBF Kernel	.461	.164	.552	.446	.507	.305
LSTM without Pre-trained Emb.	.515	.299	.594	.506	.555	.403
LSTM with FastText Emb.	.489	.258	.608	.501	.549	.380
GRU without Pre-trained Emb.	.488	.295	.587	.498	.538	.396
GRU with FastText Emb.	.474	.275	.578	.475	.526	.375
LSTM Att. without Pre-trained Emb.	.496	.336	.563	.475	.530	.405
LSTM Att. with FastText Emb.	.483	.301	.559	.480	.521	.390
BERT	**.568**	.278	**.680**	**.580**	**.624**	**.429**

Italian AMI EVALITA	Category		Target		Average	
	Accuracy	Macro F_1	Accuracy	Macro F_1	Accuracy	Macro F_1
Baseline of Shared Task	-	.543	-	.440	-	.492
Best System of Shared Task	-	.501	-	.579	-	.540
SVM Linear Kernel	**.751**	**.596**	.793	.558	**.772**	**.577**
SVM RBF Kernel	.488	.109	.445	.237	.467	.173
LSTM without Pre-trained Emb.	.743	.584	.753	.564	.748	.574
LSTM with FastText Emb.	.721	.516	.770	.575	.746	.546
GRU without Pre-trained Emb.	.710	.480	.767	**.607**	.739	.543
GRU with FastText Emb.	.729	.538	**.797**	.571	.763	.554
LSTM Att. without Pre-trained Emb.	.738	.549	.783	.553	.761	.551
LSTM Att. with FastText Emb.	.721	.470	.795	.553	.758	.512
BERT	.739	.508	.777	.537	.758	.522

English AMI IberEval	Category		Target		Average	
	Accuracy	Macro F_1	Accuracy	Macro F_1	Accuracy	Macro F_1
Baseline of Shared Task	-	.157	-	.518	-	.337
Best System of Shared Task	-	**.293**	-	.593	-	.443
SVM Linear Kernel	.674	.259	.759	.642	.716	.451
SVM RBF Kernel	.674	.228	.709	.545	.692	.387
LSTM without Pre-trained Emb.	.572	.274	.674	.577	.623	.425
LSTM with FastText Emb.	.623	.256	.663	.579	.643	.417
GRU without Pre-trained Emb.	.596	.262	.606	.536	.601	.399
GRU with FastText Emb.	.606	.248	.664	.601	.635	.424
LSTM Att. without Pre-trained Emb.	.619	.229	.643	.553	.631	.391
LSTM Att. with FastText Emb.	.605	.227	.663	.527	.634	.377
BERT	**.703**	.285	**.814**	**.714**	**.758**	**.499**

Spanish AMI IberEval	Category		Target		Average	
	Accuracy	Macro F_1	Accuracy	Macro F_1	Accuracy	Macro F_1
Baseline of Shared Task	-	.281	-	.537	-	.409
Best System of Shared Task	-	.339	-	.553	-	.446
SVM Linear Kernel	**.698**	**.371**	**.770**	.566	**.734**	.469
SVM RBF Kernel	.501	.111	.460	.261	.480	.186
LSTM without Pre-trained Emb.	.633	.344	.668	.585	.650	.465
LSTM with FastText Emb.	.658	.328	.728	**.614**	.693	.471
GRU without Pre-trained Emb.	.608	.332	.706	.582	.657	.457
GRU with FastText Emb.	.661	.351	.732	.577	.696	.464
LSTM Att. without Pre-trained Emb.	.609	.311	.666	.545	.637	.428
LSTM Att. with FastText Emb.	.584	.328	.718	.568	.651	.448
BERT	.666	**.371**	.744	.577	.705	**.474**

Table 10
Dataset label distribution of OLID. OFF : Offensive; NOT : Not Offensive; TIN : Targeted Insult; UNT : Untargeted; IND : Individual; OTH : Other; GRP : Group.

Subtask A	Subtask B	Subtask C	Train	Test	Total
OFF	TIN	IND	2,407	100	2,507
OFF	TIN	OTH	395	35	430
OFF	TIN	GRP	1,074	78	1,152
OFF	UNT	-	524	27	551
NOT	-	-	8,840	620	9,460
All			13,240	860	14,100

systems, we also build a BERT-based model, which is reported as the best model in generalizing different tasks of abusive language detection Swamy et al. (2019). All these systems are trained and optimized with similar approach, as explained in Section 4.

This experiment is restricted to English datasets, namely the two collection AMI datasets from AMI IberEval and AMI EVALITA, and three other related datasets, Waseem Waseem and Hovy (2016), HatEval Basile et al. (2019), and OffensEval Zampieri et al. (2019b). Based on the description of each dataset, we assume that the Waseem and HatEval datasets are partly related to AMI topic-wise (sexism in Waseem and hate speech toward women in HatEval), while OffensEval has a very different and broader focus on offensive language.

The **OffensEval** corpus, also known as Offensive Language Identification Dataset (OLID Zampieri et al. (2019a)) is a collection of 14,200 English tweets where abuse is represented and annotated according to a hierarchical framing for the following dimensions: presence of offensiveness (binary labels OFF vs NOT, Subtask A), offensive type (binary labels TIN and UNT for targeted vs not targeted offenses, Subtask B), target type (labels IND, GRP and OTH for individual, group or other types of target). Table 10 shows the label distribution for the three layers. The data were collected by filtering Twitter with keywords for topics on which significant among of offensive language was observed (e.g., MAGA, antifa) as well as patterns correlated to direct insults (e.g.,“she is”, “you are”). The dataset was annotated by two to three annotators per instance, reporting a relatively high agreement (.83 Fleiss kappa on a trial set of 21 tweets). Notice that the class distribution for all the layers is very imbalanced, as the authors claim that did not alter the natural distribution resulting from the adopted data collection criteria.

In this work, we only use the “sexism” class of the Waseem dataset (which we will call “WaseemS” in the rest of the paper) and the “hate targeting women” subset of the HatEval dataset (which we will call “HatEvalM” in the rest of the paper), to observe the shared characteristics and relations between phenomena contained in these datasets with misogyny.

The main procedure for this experiment is to train a system in an dataset, and test it on the other datasets. In addition to the main experiment, we also experiment by combining two datasets as a training set to extend the coverage of the dataset, then test it on the test set of each dataset. Similarly to the previous experiment on the AMI task, this experiment is evaluated in terms of precision, recall, *F*-score, and accuracy. In case of the WaseemS dataset, where the partition of training and testing set is not specified, we split randomly the dataset in a 70%/30% proportion for training and testing, respectively.

5.2. Results

Table 11 shows the full result of cross-domain classification with five different classifiers on five different datasets. The systems are based on LSVC and LSTM either with and without HurtLex, and also BERT. Datasets were chosen based on their relation with misogyny phenomena, where HatEvalM contains a similar phenomena (hate speech towards women), WaseemS covers a related phenomena (sexism), and OffensEval has a quite different focus, related to offensive language in general. Based on the description of each dataset, AMI IberEval, AMI EVALITA, and HatEvalM were collected and annotated with the same approach. Based on our manual investigation on these three datasets, we found duplicate instances across the collections. We identified 489 tweets in the EN-HatEval training set identical to tweets in the EN-AMI IberEval test set and 636 tweets in the EN-AMI EVALITA test set. 656 duplicated tweets are also shared between the EN-AMI EVALITA training set and the EN-AMI IberEval test set. For cross-domain classification purposes, we excluded these duplicates from the training sets. The final HatEvalM training set contains 3,355 tweets, while the training set of EN-AMI EVALITA consists of 3,344 tweets.

The underlined numbers in the table indicate the basic classification setting, where a system is trained and tested on the same dataset. Overall, the deep learning models (LSTM and BERT) achieved better performance than traditional classifiers such as LSVC in the cross-domain classification setting (16 out of 18 runs) in terms of F_1-score. Specifically, both models almost always obtain a better score in term of recall, resulting often in a better F_1-score. LSVC obtained better results than LSTM and BERT only when the system is tested on the WaseemS dataset. Meanwhile, the comparison between BERT and LSTM shows that BERT has a better performance when tested on EN-IberEVal, EN-EVALITA, and WaseemS, while LSTM outperforms BERT when tested on HatEvalM and OffensEval. The results also indicate that our systems obtain lower results in most of out-domain settings with respect to in-domain. An exception is when LSVC is trained on WaseemS and tested on HatEvalM, where it obtained the highest performance compared to the other runs. The lowest result was obtained when our systems are tested on OffensEval, the dataset which has the most different focus from misogyny phenomena. The final important finding is that the use of HurtLex boosts the systems performance, both LSVC and LSTM. Most of the improvement was measured in the recall score.

Table 12 depicts the results of an additional experiment where we combined the training sets of two datasets at a time, to augment

Table 11
Result of Cross-domain Automatic Misogyny Identification Experiment.

Linear Support Vector Classifier

	EN-IberEval				EN-EVALITA				WaseemS				HatEvalM				OffensEval			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
EN-Ibereval	.828	.629	.715	.804	-	-	-	-	.695	.305	.424	.815	.438	.962	.602	.461	.563	.113	.188	.728
EN-EVALITA	-	-	-	-	.584	.670	.624	.629	.621	.150	.242	.790	.442	.974	.608	.469	.553	.088	.151	.726
WaseemS	.892	.205	.333	.680	.559	.370	.445	.576	.874	.626	.730	.896	.503	.750	.602	.581	.526	.042	.072	.722
HatEvalM	.869	.537	.664	.788	.597	.665	.629	.639	.627	.150	.242	.790	.449	.973	.614	.483	.561	.096	.164	.727
OffensEval	.591	.484	.532	.668	.534	.639	.582	.577	.395	.261	.314	.746	.431	.990	.602	.442	.710	.479	.572	.800

Linear Support Vector Classifier and HurtLex

	EN-IberEval				EN-EVALITA				WaseemS				HatEvalM				OffensEval			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
EN-Ibereval	.822	.622	.708	.800	-	-	-	-	.684	.302	.419	.813	.442	.928	.605	.471	.532	.104	.174	.724
EN-EVALITA	-	-	-	-	.584	652	.616	.626	.606	.155	.247	.789	.452	.970	.617	.490	.636	.117	.197	.735
WaseemS	.848	.276	.416	.698	.565	.417	.480	.584	.869	.632	.732	.897	.464	.955	.625	.514	.615	.067	.120	.728
HatEvalM	.851	.527	.651	.780	.592	.659	.624	.634	.618	.159	.253	.790	.456	.965	620	.499	.650	.108	.186	.735
OffensEval	.569	.513	.539	.658	.521	.650	.578	.564	.391	.270	.320	.743	.429	.995	.600	.438	.707	.483	.574	.800

Long Short Term Memory

	EN-IberEval				EN-EVALITA				WaseemS				HatEvalM				OffensEval			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
EN-Ibereval	.746	.717	.732	.795	-	-	-	-	.426	.398	.412	.746	.444	.891	.593	.482	.372	.267	.311	.670
EN-EVALITA	-	-	-	-	.598	652	.624	.638	.396	.110	.172	.764	.546	.827	.657	.635	.443	.146	.219	.711
WaseemS	.750	.286	.414	.685	.612	.498	.549	.624	.855	.697	.768	.906	.458	.957	.619	.502	.511	.096	.161	.722
HatEvalM	.678	.587	.629	.730	.569	.657	.610	.613	.275	.275	.275	.676	.484	.910	.632	.552	.389	.358	.373	.664
OffensEval	.611	.555	.582	.689	.561	.678	.614	.608	.285	.215	.245	.704	.433	.986	.602	.448	.746	.513	.607	.815

Long Short Term Memory and HurtLex

	EN-IberEval				EN-EVALITA				WaseemS				HatEvalM				OffensEval			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
EN-Ibereval	.701	.739	.719	.776	-	-	-	-	.403	.430	.416	.730	.441	.920	.596	.472	.338	.288	.311	.644
EN-EVALITA	-	-	-	-	.536	.763	.630	.587	.264	.362	.306	.632	.460	.960	.622	.505	.343	.425	.380	.613
WaseemS	.695	.290	.409	.674	.606	.478	.535	.617	.855	.676	.755	.902	.454	.958	.616	.495	.542	.108	.181	.726
HatEvalM	.745	.505	.602	.740	.617	.585	.600	.642	.328	.172	.225	.736	.517	.867	.649	.601	.414	.250	.312	.692
OffensEval	.612	.804	.695	.660	.593	.786	.676	.664	.292	.271	.281	.690	.428	.995	.599	.435	.712	.567	.631	.815

BERT

	EN-IberEval				EN-EVALITA				WaseemS				HatEvalM				OffensEval			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
EN-Ibereval	.773	.915	.838	.862	-	-	-	-	.722	.229	.348	.808	.486	.913	.634	.554	.386	.142	.207	.698
EN-EVALITA	-	-	-	-	.716	.704	.661	.682	.645	.322	.430	.809	.504	.918	.651	.584	.444	.100	.163	.714
WaseemS	.864	.201	.327	.676	.589	.657	.621	.631	.846	.692	.761	.903	.532	.621	.573	.608	.406	.054	.096	.714
HatEvalM	.829	.802	.815	.858	.698	.670	.684	.715	.679	.328	.442	.815	.509	.957	.664	.590	.404	.088	.144	.709
OffensEval	.538	.748	.626	.588	.508	.811	.624	.551	.407	.574	.476	.718	.429	.995	.599	.437	.815	.550	.657	.840

the coverage. We compared the classification results of this setting with the basic setting, where only the original training set is used. The experiment results show that a performance improvement is measured on all test sets, except for OffensEval. When systems are tested on AMI EVALITA, almost all the additional training sets succeeded to enhance the classification result, whether on F_1-score or accuracy. When tested on AMI IberEval, the performance improvement is only achieved when the in-domain training sets are added. On the contrary, the addition of out-domain training sets (WaseemS and OffensEval) was be able to boost the system performance when tested on HatEvalM. When tested on WaseemS, the extra training set from OffensEval was the only one which could not improve the system performance. In the last experiment setting, testing on OffensEval, there was no additional training set able to enhance the system performance.

Table 12
Result of Experiment by Combining Two Datasets in Cross Domain Classification of Misogyny.

Test on AMI EVALITA	LSVC				LSTM				BERT			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
AMI EVALITA Only	.584	.670	.624	.629	.658	.489	.561	.648	.704	.661	.682	.716
+ AMI IberEval	.626	.639	.632	.658	.648	.504	.567	.646	.697	.639	.667	.706
+ HatEvalM	.598	.663	.629	.640	.549	.726	.626	.600	.628	.713	.668	.674
+ WaseemS	.603	.587	.595	.632	.595	.519	.555	.616	.678	.696	.687	.708
+ OffensEval	.547	.654	.596	.592	.559	.680	.614	.606	.586	.667	.624	.630

Test on AMI IberEval	LSVC				LSTM				BERT			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
AMI IberEval Only	.828	.629	.715	.804	.746	.717	.732	.795	.773	.915	.838	.862
+ AMI EVALITA	.837	.636	.723	.810	.808	.580	.675	.782	.795	.919	.853	.876
+ HatEvalM	.845	.636	.726	.813	.664	.746	.702	.754	.842	.696	.762	.831
+ WaseemS	.836	.488	.616	.763	.776	.601	.677	.777	.824	.841	.832	.868
+ OffensEval	.728	.576	.643	.751	.712	.785	.746	.792	788	.774	781	.831

Test on HatEvalM	LSVC				LSTM				BERT			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
HatEvalM Only	.449	.973	.614	.483	.484	.910	.632	.552	.479	.981	.643	.539
+ AMI EVALITA	.444	.971	.609	.472	.478	.857	.613	.543	.541	.908	.678	.635
+ AMI IberEval	.444	.954	.606	.475	.466	.925	.620	.520	.491	.974	.652	.561
+ WaseemS	.461	.979	.627	.507	.481	.968	.643	.545	.488	.971	.650	.557
+ OffensEval	.461	.958	.623	.508	.463	.934	.620	.514	.450	.990	.619	.483

Test on WaseemS	LSVC				LSTM				BERT			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
WaseemS Only	.874	.626	.730	.896	.855	.697	.768	.906	.826	.749	.785	909
+ AMI EVALITA	.874	.633	.735	.898	.854	.652	.739	.897	.847	.674	.750	.899
+ AMI IberEval	.874	.649	.745	.901	.830	.706	.763	.902	.750	.775	.762	.892
+ HatEvalM	.875	.632	.734	.897	.819	.703	.757	.899	.806	.725	763	899
+ OffensEval	.797	.636	.707	.882	.724	.701	.712	.873	.839	.690	.757	.901

Test on OffensEval	LSVC				LSTM				BERT			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
OffensEval Only	.710	.479	.572	.800	.746	.513	.607	.815	.694	.679	.686	.827
+ AMI EVALITA	.710	.388	.501	.785	.752	.492	.595	.813	.708	.617	.659	.822
+ AMI IberEval	.703	.404	.513	.786	.648	.567	.604	.793	.734	.621	.673	831
+ HatEvalM	.692	.383	.493	.780	.806	.450	.578	.816	.888	.329	.480	.801
+ WaseemS	.709	.346	.465	.778	.752	.392	.515	.794	.778	.350	.483	.791

6. Cross-Lingual Automatic Misogyny Identification Experiment

6.1. Experimental Setup

In this section, We propose an experiment in cross-lingual automatic misogyny identification. We take advantage of the AMI task datasets, which contain tweets in three different languages: English, Spanish, and Italian. In this cross-lingual classification experiment, we train models on one language and test it on datasets in a different language. Specifically, we build four systems:

1. **Linear Support Vector Classifier (LSVC).** With this classifier, we only use unigrams as features. Therefore, we need to translate the training set from the source language (the original language of the training set) to the target language (the language of the test set). We used Google Transate[24] as translation service.

[24] https://translate.google.com/

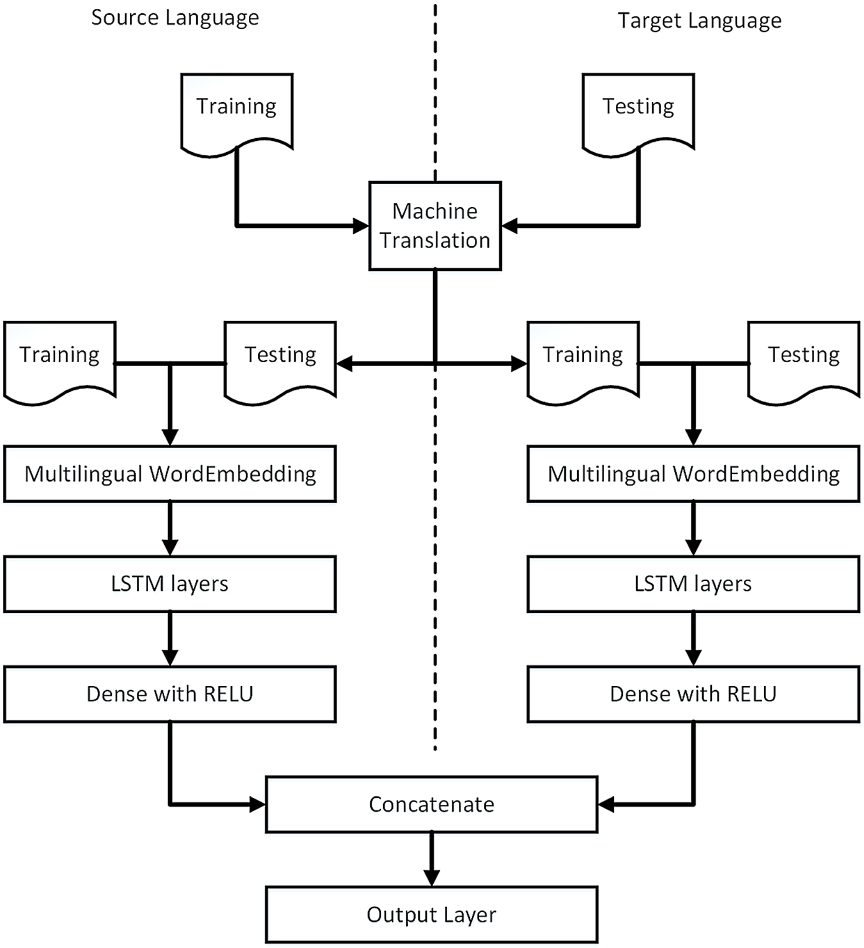


Fig. 2. Joint-Learning Model Architecture.

2. **Long Short Term Memory (LSTM) with Monolingual Word Embedding.** We implement a LSTM architecture with monolingual word embeddings as word representation. The pre-trained word embeddings provided by FastText are used to initialize the embedding layer of the network, followed by LSTM layers consisting of 64 units. The output of the LSTM network is connected to a dense layer (16 units) with ReLu activation function. The last layer is a dense layer with sigmoid activation which provides the final prediction of the label. Similar to LSVC system, in this setting we also translate the training set to the target language by using Google Translate.
3. **Long Short Term Memory (LSTM) with Multilingual Word Embedding.** We employ the multilingual word embeddings developed by Facebook research group published as MUSE (Multilingual Unsupervised or Supervised word Embeddings), a supervised word embedding model aligned across 30 languages (Lample, Conneau, Ranzato, Denoyer, & Jégou, 2018). With this representation, we do not need to translate the training set to the target language. The rest of the configuration of this model is the same as the LSTM with monolingual word embedding.
4. **Joint-Learning Model with Multilingual Word Embeddings.** We also propose a joint-learning model with a focus on transfer knowledge between languages in a cross-lingual classification setting. Fig. 2 shows the full process of how the data is transformed and learned by architecture. We start the process by creating a bilingual dataset automatically by using Google Translate. The training and test set are translated in both directions (source to target language and target to source language), then used to train two LSTM-based models in two languages independently. We concatenate the output of the two models before the final layer (output layer), which provides the final prediction. In this architecture, we expect to reduce some of the noise from the translation while keeping the original structure of the training and test set. The configuration of each single LSTM architecture is the same to the two previous models (monolingual and multilingual LSTM).
5. **Joint-Learning Model with Multilingual Word Embedding and HurtLex.** Finally, we experiment by adding HurtLex to the joint-learning model. We concatenate a feature representation obtained with the lexicon to the input of each LSTM networks in both languages. In this architecture, HurtLex provides a 17-dimension vector, i.e., a one-hot encoding of the word presence in each of the lexicon categories.
6. **BERT with Multilingual Model.** We also propose a model similar to LSTM with multilingual embedding, by substituting LSTM

and MUSE with a pre-trained multilingual BERT model. In particular, we use the `bert-multi-cased` model. The rest of the configuration is the same as the BERT model in the previous experiment, with a dense layer with ReLU activation function, followed by a dense layer with sigmoid activation function as the final layer. This model is optimized by using Adam optimizer and trained on different combinations of batch size (16,32,64) and epochs (1-5).

7. **Joint-Learning BERT with Multilingual Model.** Adopting a similar idea to our joint-learning LSTM model, we also propose to build a model by substituting the LSTM with the BERT `bert-multi-cased` model. This model is optimized and trained with the same configuration as BERT with monolingual model.

6.2. Results

Table 13 depicts the results of cross-lingual automatic misogyny identification experiments, where we train a system on one language and test it on another language. In this analysis, we only focus on the system performance based on F_1 and accuracy score. We mark the highest F_1 in each run in bold face, and the highest accuracy by underline. We start the analysis of the result by focusing on the comparison between LSVC and LSTM with monolingual embeddings, where both systems only rely on the use of machine translation to deal with the multilingual environment. We found out that the use of traditional models does not always give a lower performance than deep learning. LSCV achieved better performance in some of the settings, including *ES-IberEval → EN-IberEval, ES-IberEval → EN-EVALITA, IT-EVALITA → EN-EVALITA, EN-IberEval → ES-IberEval*, and *EN-EVALITA → ES-IberEval*. However, the LSVC performance is much lower compared to LSTM in settings where the translation from English to Italian is needed.

The second analysis focuses on the performance comparison between LSTM with monolingual embedding and machine translation, and LSTM with multilingual embeddings where no translation is needed. Surprisingly, LSTM with monolingual embeddings are able to outperform LSTM with multilingual embeddings, which use pre-trained word embeddings that are specifically developed for cross-lingual learning. In terms of F_1-score, monolingual LSTM has a better performance in 6 out of 10 run settings, while based on accuracy it outperformed LSTM with multilingual embeddings in 9 out of 10 settings. However, a different outcome emerges when we compare LSTM with monolingual embedding against the multilingual BERT model. BERT tends to have more robust performance on two languages, namely English and Spanish, but not on Italian.

The third analysis focuses on the comparison between LSTM with multilingual embedding, the joint-learning model with multilingual embeddings, and the respective BERT-based variants, combining the machine translation ability and multilingual embeddings. In terms of accuracy, joint-learning always outperforms LSTM with multilingual embedding in all settings. Both systems achieve the best performance in half the runs, in terms of F_1-score. However, the overall results show that joint-learning has a more robust performance across the settings. We observe that in some settings, including *EN-IberEval → ES-IberEval, EN-IberEval → IT-EVALITA*, and *EN-EVALITA → IT-EVALITA*, LSTM with multilingual embeddings experienced a big drop in performance. With BERT, our joint-learning model also performs consistently better than the multilingual BERT model in term of F_1-score. Also in term of accuracy, the joint-learning models outperform the normal multilingual BERT configuration in 7 out of 10 runs.

The last analysis is a comparison between using and not using HurtLex in the joint learning model with multilingual embeddings. Based on the experimental results, the use of HurtLex succeeded to improve the model performance in term of F_1-score.

7. Discussion

In this section, we present the discussion and analysis of the results of all our proposed experiments. The discussion is organized in three subsections reflecting the different experimental settings, namely automatic misogyny identification (subtask A and subtask B of the AMI challenge), relationship between misogyny and other abusive phenomena, and cross-lingual classification.

7.1. Automatic Misogyny Identification Task

In order to get a deeper insight, we performed an ablation test on our best models on the AMI IberEval dataset, removing each feature to measure the impact on the system performance. Table 14 presents the ablation test results of our English AMI IberEval, which shows that sexist slurs and women words presence are the most predictive features on this task. These figures confirm the findings of the lexical distribution analysis in Section 3, where sexist slurs were found to be mainly used in misogynistic instances. Moreover, the importance of the women-related words feature indicates that the detection of target gender is highly informative for the detection of misogyny.

Similar to English part, our system was also top ranked in the Spanish AMI IberEval task. While the best system for English is a SVM classifier with RBF kernel, for Spanish the best system is a SVM with linear kernel including several features such as bags of words (1-gram to 3-grams), bags of hashtags, bags of emojis, sexist slurs presence, woman words presence, and the presence of some HurtLex categories, including words related to female genitalia (ASF), words related to prostitution (PR), words related to cognitive disabilities and diversity (DDF), words related to physical disabilities and diversity (DDF), and words related to male genitalia (ASM). In summary, these results show that HurtLex helps informing the model, but only some of its categories are actually related to the misogynistic action. As shown in Table 15, bags of words are the most informative feature of this model. Therefore, we decides to conduct a further analysis by extracting the SVM classifier weights when only token n-grams are used as features, to obtain a clearer picture of what is the most predictive features in Spanish AMI IberEval task. Table 16 shows the top ten features for the Spanish AMI IberEval task based on the SVM weight. The use of sexist slurs such as *zorra* (bitch), *perra* (bitch/slut), *guarra* (slut), and *coȘo* (pussy/cunt) is a clear signal of misogynistic content. This finding is consistent with the results on the English dataset, where sexist slurs is

Table 13
Result of Cross-lingual Automatic Misogyny Identification Experiment.

Linear Support Vector Classifier

	EN-IberEval				EN-EVALITA				ES-IberEval				IT-EVALITA			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
EN-Ibereval	-	-	-	-	-	-	-	-	.675	.771	.720	.704	.198	.135	.160	.277
EN-EVALITA	-	-	-	-	-	-	-	-	.640	.545	.589	.620	.205	.121	.152	.309
ES-IberEval	.409	.477	.441	.528	.566	.704	.524	.610	-	-	-	-	.621	.621	.621	.612
IT-EVALITA	.376	.686	**.486**	.434	.492	.739	.591	.529	.568	.542	.555	.566	-	-	-	-

Long-Short Term Memory with Monolingual Embedding

	EN-IberEval				EN-EVALITA				ES-IberEval				IT-EVALITA			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
EN-Ibereval	-	-	-	-	-	-	-	-	.672	.745	.706	.691	.458	.606	**.521**	.431
EN-EVALITA	-	-	-	-	-	-	-	-	.712	.246	.366	.575	.448	.295	.356	.453
ES-IberEval	.406	.237	.299	.568	.676	.404	.506	.637	-	-	-	-	.658	.846	**.740**	.696
IT-EVALITA	.339	.519	.410	.417	.536	.557	.546	.574	.589	.598	.593	.591	-	-	-	-

Long-Short Term Memory with Multilingual Embedding

	EN-IberEval				EN-EVALITA				ES-IberEval				IT-EVALITA			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
EN-Ibereval	-	-	-	-	-	-	-	-	.644	.157	.252	.536	.324	.065	.102	.454
EN-EVALITA	-	-	-	-	-	-	-	-	.554	.523	.538	.551	.257	.094	.137	.397
ES-IberEval	.376	.933	.536	.371	.481	.885	.623	.508	-	-	-	-	.571	.922	.706	.606
IT-EVALITA	.299	.558	.389	.317	.428	.315	.363	.491	.544	.774	.639	.563	-	-	-	-

Joint Learning Model with Multilingual Embedding

	EN-IberEval				EN-EVALITA				ES-IberEval				IT-EVALITA			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
EN-Ibereval	-	-	-	-	-	-	-	-	.749	.648	.695	.716	.472	.502	.487	.458
EN-EVALITA	-	-	-	-	-	-	-	-	.676	.407	.508	.607	.584	.102	.173	.503
ES-IberEval	.396	.516	.448	.504	.572	.576	.574	.607	-	-	-	-	.587	.920	.717	.628
IT-EVALITA	.423	.283	.339	.570	.566	.380	.455	.581	.409	.299	.409	.569	-	-	-	-

Joint Learning Model with Multilingual Embedding and HurtLex

	EN-IberEval				EN-EVALITA				ES-IberEval				IT-EVALITA			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
EN-Ibereval	-	-	-	-	-	-	-	-	.624	.868	.726	.673	.480	.506	.492	.466
EN-EVALITA	-	-	-	-	-	-	-	-	.724	.542	**.620**	.668	.553	.287	.377	.516
ES-IberEval	.395	.643	.490	.478	.530	.702	.604	.577	-	-	-	-	.637	.842	.725	.673
IT-EVALITA	.372	.686	.483	.427	.486	.911	**.633**	.512	.622	.448	.521	.588	-	-	-	-

Multilingual BERT

	EN-IberEval				EN-EVALITA				ES-IberEval				IT-EVALITA			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
EN-Ibereval	-	-	-	-	-	-	-	-	.648	.641	.645	.647	.226	.146	.177	.306
EN-EVALITA	-	-	-	-	-	-	-	-	.652	.451	.533	.605	.393	.324	.356	.398
ES-IberEval	.463	.647	.540	.570	.704	.491	.579	.671	-	-	-	-	.650	.443	.527	.593
IT-EVALITA	.277	.357	.312	.386	.528	.626	.573	.571	.536	.708	.610	.548	-	-	-	-

(*continued on next page*)

Table 13 (*continued*)

BERT Joint Learning																
	EN-IberEval				EN-EVALITA				ES-IberEval				IT-EVALITA			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
EN-Ibereval	-	-	-	-	-	-	-	-	.710	.754	**.731**	.723	.457	.115	.184	.477
EN-EVALITA	-	-	-	-	-	-	-	-	.754	.525	.619	.678	.574	.281	**.378**	.525
ES-IberEval	.499	.657	**.567**	.609	.589	.733	**.653**	.642	-	-	-	-	.639	.488	.554	.597
IT-EVALITA	.286	.445	.348	.350	.517	.722	.603	.562	.575	.810	**.673**	.607	-	-	-	-

Table 14
Ablation test result of the best system on English AMI IberEval.

Features	Accuracy	Δ (delta)
All features	91.32	-
- Swear words count	89.81	1.51
- Swear words presence	90.50	0.82
- Hashtags presence	90.63	0.69
- Links count	85.40	5.92
- Sexist slurs presence	75.90	15.42
- Women words presence	73.83	17.49

Table 15
Ablation test result of the best system on Spanish AMI IberEval.

Features	Accuracy	Δ (delta)
All features	81.47	-
- Bag of words	65.98	15.49
- Bag of hashtags	81.40	0.07
- Bag of emojis	80.44	1.03
- Sexist slurs presence	80.85	0.62
- Women words presence	81.13	0.34
- ASF presence	81.27	0.2
- ASM presence	81.13	0.34
- DDF presence	81.13	0.34
- DDP presence	81.27	0.2
- PR presence	81.13	0.34

Table 16
Top ten features based on SVM weight on Spanish AMI IberEval task.

Setting	Offensive	
No.	Features	Coefficient
1.	zorra	2.143
2.	perra	1.499
3.	callate	1.427
4.	hija	1.061
5.	guarra	1.028
6.	callate puta	0.935
7.	mi polla	0.905
8.	callate perra	0.904
9.	tu cono	0.806
10.	mujer	0.706

also the most important feature to detect misogyny instances.

Our BERT models performed very well on both English and Italian AMI EVALITA, outperforming the best systems on the respective tasks. In English AMI EVALITA, it achieved better performances than the best system participating in the shared task, with an accuracy of 71.60 (the best system obtained an accuracy of 70.40). Concerning the Italian part, our BERT model also able to surpass the competition best system, obtaining 84.80 in accuracy, slightly higher than best system with 84.40 in accuracy.

The results on AMI subtask A show that traditional models obtain a good performance, especially on the IberEval tasks, with the advantage of being far more transparent than deep learning. However, these models fail to have a stable performance across different datasets, as highlighted by their low performance on the EVALITA tasks. Here, BERT achieves the best results, both in English and Italian. The overall result signifies that deep learning approaches have a more stable performance on both shared tasks, where they always obtain a competitive results. We also notice that SVM with RBF kernel always obtains a good performance when applied to English datasets, but a much lower performance on the other languages. Similarly, our BERT model also tend to have better performance when applied to English. Finally, SVM with linear kernel tends to achieve comparably better results when applied to languages other than English.

Error Analysis. We conducted a manual error analysis on the misclassified instances, to explore the most common pitfalls in detecting misogyny. Our investigation found that at least five issues contribute to the difficulties of this task:

1. **Presence of swear words.** We encountered a lot of “bad words” in the dataset of this shared task for both English and Italian. In case of abusive context, the presence of swear words can help to spot abusive content such as misogyny. However, they could also lead to false positives when the swear word is used in a casual, not offensive context Malmasi and Zampieri (2018); Nobata, Tetreault, Thomas, Mehdad, and Chang (2016); Van Hee et al. (2018). Consider the following two examples containing the swear word “bitch” in different contexts:

    1. Im such a fucking cunt bitch and i dont even mean to be goddammit
    2. Bitch you aint the only one who hate me, join the club, stand in the corner, and stfu.

    In Example 1, the swear word “bitch” is used just to arouse interest/show off, thus not directly insulting the other person. This is a case of *idiomatic swearing* Pinker (2007). In Example 2, the swear word “bitch” is used to insult a specific target in an abusive context, an instance of *abusive swearing* Pinker (2007). Resolving swearing context is still a challenging task for automatic system which contributing to the difficulties of this task.
2. **Reported speech.** Tweets may contain misogynistic content as an indirect quote of someone else’s words, such as in the following example:

    3. Quella volta che mia madre mi ha detto quella cosa le ho risposto “Mannaggia! Non sar mai una brava donna schiava zitta e lava! E adesso?!” Potrei morire per il dispiacere.
    → *That time when my mom told me that thing and I answered “Holy s**t! I will never be a good slave who shuts up and cleans! What now?”*

    According to task guidelines this should not be labeled as a misogynistic tweet, because it is not the user himself who is misogynistic. Therefore, instances of this type tend to confuse a classifier based on lexical features.
3. **Irony and world knowledge.** In Example 3, the sentence “Potrei morire per il dispiacere.”[25] is ironic. Humor is very hard to model for automatic systems — sometimes, the presence of figurative language even baffles human annotators. Moreover, external world knowledge is often required in order to infer whether an utterance is ironic Wallace, Kertz, Charniak et al. (2014).
4. **Preprocessing and tokenization.** In computer-mediated communication, and specifically on Twitter, users often resort to a language type that is closer to speech, rather than written language. This is reflected in less-than-clean orthography, with forms and expressions that imitate the verbal face-to-face conversation.

    4. @[redacted] @[redacted] @[redacted] @[redacted] x me glob prox2aa colpiran tutti incluso nemicinterno.. esterno colpopiduro sarlogrande che bevetropvodka e inoltre x questiondisoldi progetta farmezzofallirsudfinitestampe: ci nnvn xrchindebolis
    → *4 me glob next2aa will hit everyone included internalenemy.. external harderhit willbebigass who drinkstoomuchvodka and also 4 mattersofmoney isplanning tomakethesouthfailwithprintings: dis notgood causeweaken*

    In Example 4, preprocessing steps like tokenization and stemming are particularly hard to perform, because of the lack of spaces between one word and the other and the confused orthography. Consequently all the classification pipeline is compromised and error-prone.
5. **Gender of the target.** As defined in the Introduction, we know that misogyny is a specific type of hateful language, targeting women. However, detecting the gender of the target is a challenging task in itself, especially in Twitter datasets.

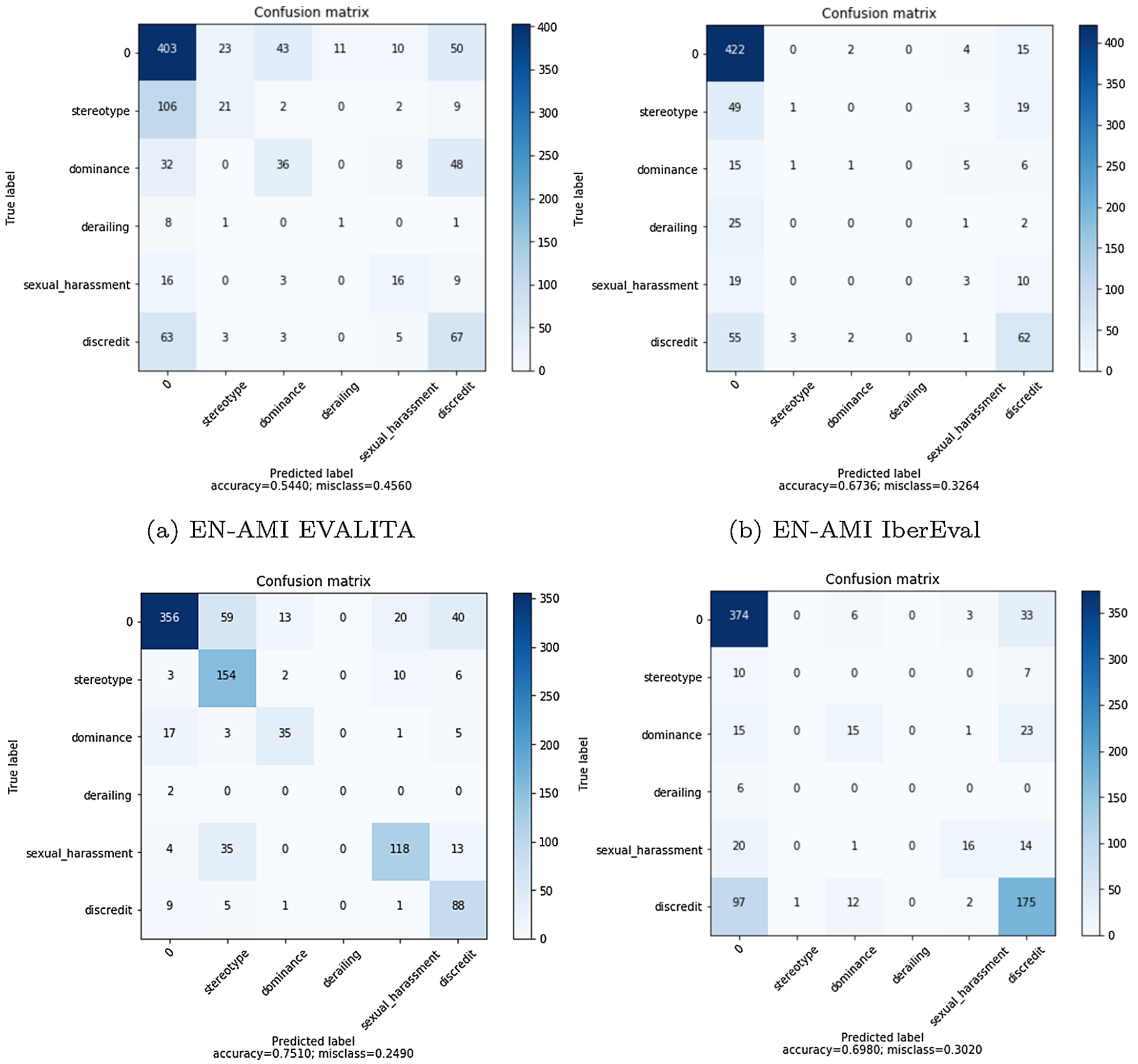
    5. @realDonaldTrump shut the FUCK up you infected pussy fungus.

    6. @TomiLahren You’re a fucking skank!

    Both examples use bad words to abuse their targets. However, the first example is labeled as not misogyny since the target is Donald Trump (man), while the second example is labeled as misogyny with the target Tomi Lahren (woman).

On subtask B, overall results indicates that treating subtask B as an independent multi-class classification is more effective than handling it as a pipeline classification task, as a sequential task to the results of subtask A, which is a mostly used approach by all AMI task participants. We argue that in a pipeline classification scenario, the results on subtask B would be highly dependant on the system performance in subtask A. In addition, we undertook a deeper analysis to get more insight regarding common issues in task B

[25] Translation: I could die for heartbreak.



(a) EN-AMI EVALITA

(b) EN-AMI IberEval

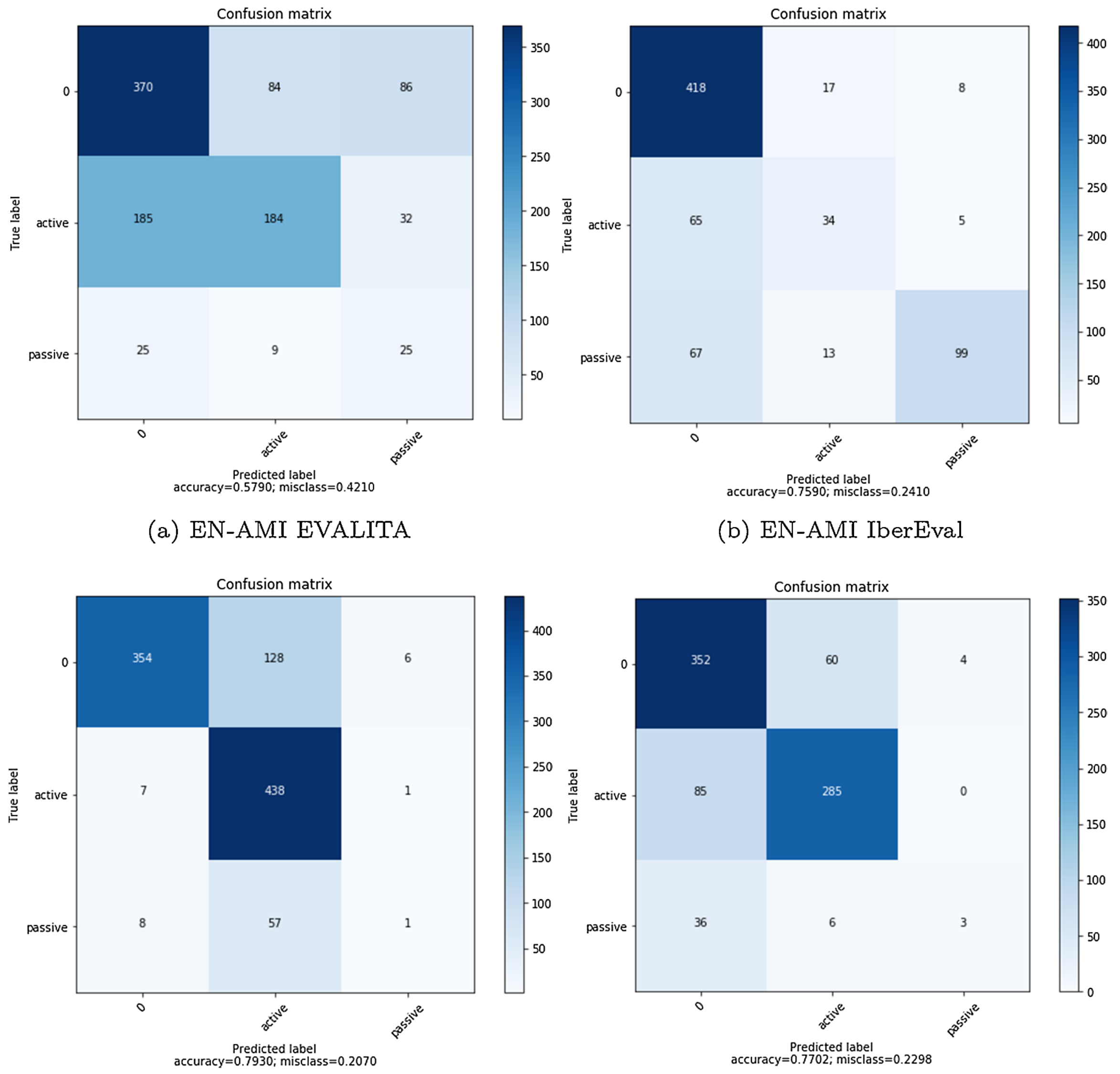
(c) IT-AMI EVALITA

(d) ES-AMI IberEval

Fig. 3. Confusion Matrix of Misogyny Behaviour Classification.

classification. We produced a confusion matrix of the classification results based on the best performing system on each dataset, consisting of two classification tasks, namely misogyny behaviour classification (in Fig. 3) and target of misogyny classification (in Figure 4). Based on Fig. 3, we can see that detecting derailing class is the most challenging aspect of this task. On English, our systems were able to classify discredit quite well. In the EN-AMI EVALITA dataset, the dominance class was also classified quite well, as well as the sexual harassment class. On the Italian dataset, our system got a promising result, balanced across the classes. On Spanish, stereotype and derailing were the only two classes which are difficult to be detected. For the classification of target of misogyny, most of our systems only performed well in detecting the active class. We argue that this result is influenced by the label distribution on the gold standard where most of misogyny tweets are labeled as active.

In addition to the analysis of the confusion matrix, we also performed a manual error analysis on the dataset to find other difficulties of this particular task. After a manual inspection of the data, it emerged that there is no clear demarcation line between one category and the others in classification of misogyny behaviour task. The single label introduced for the misogyny behaviour classification task forces tweets to only have one label, representing the dominant category. We argue that it is possible for one tweet to express more than one misogyny behaviour phenomenon. For example, *dominance* and *discredit* are both highly correlated to high presence of swearing, with varying focus (e.g., the agent (man) vs. the wounded part, the target (woman)). Similarly, *stereotype & objectification* is not so conceptually distant from *sexual harassment*, due to a strong use of language referring to sexual body parts or vulgar sexual practices. These insights are reflected in the examples we provide in the following.



(a) EN-AMI EVALITA

(b) EN-AMI IberEval

(c) IT-AMI EVALITA

(d) ES-AMI IberEval

Fig. 4. Confusion Matrix of Target of Misogyny Classification.

STEREOTYPE & OBJECTIFICATION (EN): No girl is even capable of developing morals
until they get the slut fucked out of them. Welcome to my generation
DOMINANCE (EN): Bad girls get spankings
DERAILING: Women want u to automatically believe women who scream rape they
don't understand our position....
SEXUAL HARASSMENT & THREATS OF VIOLENCE (EN): @ 🐦 hey bitch
wassup bitch suck my dick bitch
DISCREDIT (EN): @ 🐦 Fuck that money whore @HillaryClinton Too stupid
to know consensual touching or grabbing is not assault. Only @ChelseaClinton is dumber
STEREOTYPE & OBJECTIFICATION (ES): Que cuza antes la calle, una mujer inteligente
o una tortuga vieja? Una tortuga vieja porque las mujeres inteligentes no existen ⋯
DOMINANCE (ES): Voy a ensearle a esta perra como se trata a un hombre
LMAO IN LOVE WITH EL TITI 😂
SEXUAL HARASSMENT & THREATS OF VIOLENCE (ES): @🐦 ME GUSTARA
abrirte las piernas y clavarte toda mi polla en tu culo.
DISCREDIT (ES): Porque ladra tanto mi perra? La puta madre cllate un poco

7.2. Relationship between Misogyny and other Abusive Phenomena

The overall results of cross-domain classification of misogyny show that deep learning approaches have a better performance in transferring knowledge between different datasets, including the AMI task datasets. The results also show that BERT is the most robust system in cross-domain setting, achieving a stable result in all the experimental settings. This result is in line with the findings in Swamy et al. (2019), which experimentally found that BERT has a better capability than other models to generalize over different abusive language detection tasks. In addition, the use of a lexical resource related to abusive language such as HurtLex is able to improve the system performance, providing domain independent information for the systems. The experimental results also show that training a system on a more general abusive language dataset (i.e., OffensEval), and testing it in more specific dataset (i.e., AMI datasets), still obtains a reasonable performance compared to the in-domain setting. On the contrary, when we train the system on more specific datasets such as the AMI datasets, and test it on more general dataset (OffensEval), were obtain very poor results. This is consistent with the result obtained by Pamungkas and Patti (2019) on cross-domain classification of abusive language. We also found that system trained on OffensEval is more robust than when trainining on other datasets, including WaseemS. The results for the WaseemS dataset, which is related to AMI topic-wise, indicate that having similar topic does not guarantee to get competitive results when tested on AMI datasets. We argue that the performance on cross-domain or cross-dataset classification is not only influenced by the topical focus of the datasets, but also, and heavily so, by data collection approaches and annotation procedures. This finding is also supported by Wiegand et al. (2019), who notices how the WaseemS dataset contains biases, which is problematic when the dataset is used for cross-domain classification, since models are not be able to generalize sufficiently. We also found that when the system is trained on misogyny datasets (both EVALITA and IberEval collection) and tested on WaseemS, they experience a bigger drop than when tested on HatEvalM. This result shows how hate speech toward women has a stronger relation to misogyny than sexism. Interestingly, this result is also in line with recent philosophical accounts of misogyny Manne (2017); Richardson-Self (2018), theorizing *misogyny* and *sexism* as related but *distinct* mechanisms that enforce the norms of patriarchy Manne (2017), and arguing for considering only misogynistic speech as a specific kind of hate speech Richardson-Self (2018).

The dataset augmentation experiment show that augmenting the data training coverage by adding external data only works when the additional data share similar targets or topics, as observed on the AMI, HatEvalM, and WaseemS datasets. Adding a dataset which has different phenomena and topical focus (in this case OffensEval), failed in enhancing the system performance. We argue that additional training data from the loosely related dataset could not be able to extend the coverage of the dataset, which would help to build a robust system, but instead introduces noise which hurts the systems performance. Again, these results confirm that hate speech toward women (modeled in HatEvalM) provide more valuable additional data than sexism (WaseemS), reaffirming that hate speech targeting women is more strongly related to misogyny than sexism.

Notice that we experimented with the *offensive* label provided for tweets in the OffensEval corpus. However, the annotation of target in this corpus (group vs. individual, see Table 10) may also be a valuable layer to explore in the future to see how well this information can transfer to and from the AMI dataset, where we have an analogous layer of annotation devoted to identify the nature of the target (active vs. passive).

7.3. Cross-lingual Automatic Misogyny Identification

Based on the cross-lingual experimental results, we argue that the performance of LSVC models heavily relies on the translation quality, since they only use the token n-grams as their main feature to estimate the probability of a tweet to contain misogyny. In this case, we observe that the machine translation performance is still not good enough for translating to Italian from other languages. Deep learning models have the capability to update their feature representation (word embedding matrix), optimized on the train set during the training phase, giving more flexibility than only relying on the translation result. Therefore, LSTM has a more stable performance across all cross-lingual settings.

We also observe that the vocabulary size of the pre-trained word embedding is a possible cause for the low performance of multilingual embeddings. Indeed, the pre-trained models from FastText contain 2,000,000 vocabulary items for monolingual embeddings and only 200,000 for multilingual embeddings. The low vocabulary coverage leads to a higher number of out of vocabulary (OOV) words, which causes inaccuracies in the word representation matrix.

Both joint-learning models (LSTM-based and BERT-based) performed better than their standard counterparts (LSTM with multilingual embeddings and Multilingual BERT). The better results obtained by our joint-learning model confirms our idea that allowing the network to learn both the original and translated text is able to reduce some of the noise from the translation, while keeping the original structure of the training set. This in turn enables the system to deal with the issues of low vocabulary coverage of multilingual embeddings and quality of the translation result.

Regarding to the performance improvement when exploiting HurtLex, further investigation proved that there is a significant improvement on the recall side when the model includes the HurtLex features, meaning that the system is able to reduce the number of false negatives in the prediction and to detect more misogynistic instances. We argue that HurtLex has a significant impact to inform the models about specific hurtful words, which are possibly not always translated correctly by the machine translation service, or not covered by multilingual word embeddings. For example, offensive words toward women such as “hoe” and “skank” are mistranslated by machine translation (hoe (English) → azada (Spanish)) and (skank (English) → skank (Italian)).

8. Conclusion

We presented the results of a deep exploration of automatic misogyny identification (AMI). We started from investigating the best approaches to detect misogynistic, exploring state of the art on several AMI benchmark datasets. We also explored the most predictive features for detecting misogyny, by performing an ablation test on the best performing systems on such benchmarks. We performed a manual error analysis to discover the issues and challenges specific to this kind of classification task. Furthermore, we ran experiments in cross-domain classification, involving some of the AMI datasets, in order to investigate the interaction between misogyny and related phenomena, namely sexism, hate speech, and offensive language. Finally, we conducted an experiment on AMI in a cross-lingual setting, building a joint-learning model based on LSTM and BERT, in order to bridge the gap of AMI in low-resource languages.

Our proposed models succeed to outperform the state of the art on all AMI benchmarks, consisting of three different languages: English, Spanish and Italian. We found that traditional models still perform better than more sophisticated, deep learning approaches in English and Spanish AMI IberEval. On the other hand, in English and Italian AMI EVALITA, BERT obtains better performance than other models. We also experimentally proved that lexical features such as sexist slurs and woman words (words which are synonyms or related words to “woman”) are among the most predictive features to detect misogyny. We also observed that treating AMI task B as an independent multi-class classification gives a better performance than a pipeline approach with task A. With this approach, we were able to outperform all of the state of the art results on task B with the exact same system used for task A.

Our cross-domain classification experiment shows that neural-based models, i.e., LSTM and BERT, facilitate knowledge transfer between different datasets. As expected, our system does not achieve an optimal performance when trained on other abusive phenomena data and tested on AMI data, and vice versa. The experiment with HurtLex showed that the use of a domain independent resource, such as an abusive language lexicon, was able to boost the cross-domain performance, proving how this approach is capable of facilitating domain transfer between datasets. We also found that augmenting the training set only works when the additional data provide a similar topical focus as the original training dataset.Both experiments in Section 5 confirm that hate speech towards women is a more related phenomenon to misogyny than sexism. The overall results show that BERT is the best model for domain transfer between different datasets, able to obtain robust performance in all experimental settings.

Differently from the cross-domain setting, our traditional classifier, i.e. LSVC, still got a better performance than neural architectures in some of our cross-lingual experimental settings. However, further investigations showed that its performance is highly dependant on the quality of the translation result, while deep learning approaches provide a more stable performance across language pairs. Using monolingual word embeddings with translated data with LSTM gives better results than multilingual word embeddings without translating the data. We ascribe this result to the high number of out of vocabulary words resulting from using the multilingual embeddings by FastText. To overcome the translation quality and the out of vocabulary words issues, we proposed a joint-learning model, which was able to outperform all the other systems. Again, the use of additional knowledge form HurtLex in our joint-learning model improved its performance, mainly on the recall side. Similarly to the cross-domain setting, the overall results exhibit that BERT-based model is the best model in cross-lingual setting experiment, even more robust performance is obtained when we build joint-learning model with multilingual BERT.

In future work, we plan to implement a transfer learning approach for improving the task A performance, by propagating information from the task B classification. Transfer learning is also a potential solution for the domain adaptation issue in both cross-domain and cross-lingual settings. We also plan to investigate additional architectures and language models, which may prove beneficial in a domain-specific task such as automatic misogyny identification task.

CRediT authorship contribution statement

Endang Wahyu Pamungkas: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Valerio Basile:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing - original draft, Writing - review & editing. **Viviana Patti:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Writing - original draft, Writing - review & editing.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ipm.2020.102360

References

Ahluwalia, R., Soni, H., Callow, E., Nascimento, A. C. A., & Cock, M. D. (2018). Detecting hate speech against women in english tweets. In T. Caselli, N. Novielli, V. Patti, & P. Rosso (Vol. Eds.), *CEUR Workshop Proceedings: . 2263. Proceedings of the sixth evaluation campaign of natural language processing and speech tools for italian. final workshop (EVALITA 2018) co-located with the fifth italian conference on computational linguistics (clic-it 2018), turin, italy, december 12-13, 2018* (pp. 1–6). CEUR-WS.org.

Anzovino, M., Fersini, E., & Rosso, P. (2018). Automatic identification and classification of misogynistic language on Twitter. In M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, & F. Meziane (Vol. Eds.), *Lecture Notes in Computer Science: . 10859. Natural language processing and information systems - 23rd international conference on applications of natural language to information systems, NLDB 2018, paris, france, june 13-15, 2018, proceedings* (pp. 57–64). Springer. https://doi.org/10.1007/978-3-319-91947-8_6.

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *TACL, 7*, 597–610.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In R. Barrett, R. Cummings, E. Agichtein, & E. Gabrilovich (Eds.). *Proceedings of the 26th international conference on world wide web companion, perth, australia, april 3-7, 2017* (pp. 759–760). ACM. https://doi.org/10.1145/3041021.3054223.

Bakarov, A. (2018). Vector space models for automatic misogyny identification (short paper). In T. Caselli, N. Novielli, V. Patti, & P. Rosso (Vol. Eds.), *CEUR Workshop Proceedings: . 2263. Proceedings of the sixth evaluation campaign of natural language processing and speech tools for italian. final workshop (EVALITA 2018) co-located with the fifth italian conference on computational linguistics (clic-it 2018), turin, italy, december 12-13, 2018* (pp. 1–3). CEUR-WS.org.

Basile, A., & Rubagotti, C. (2018). Crotonemilano for AMI at evalita2018. A performant, cross-lingual misogyny detection system. In T. Caselli, N. Novielli, V. Patti, & P. Rosso (Vol. Eds.), *CEUR Workshop Proceedings: . 2263. Proceedings of the sixth evaluation campaign of natural language processing and speech tools for italian. final workshop (EVALITA 2018) co-located with the fifth italian conference on computational linguistics (clic-it 2018), turin, italy, december 12-13, 2018* (pp. 1–5). CEUR-WS.org.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., ... Sanguinetti, M. (2019). *SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. Proceedings of the 13th international workshop on semantic evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics54–63. https://doi.org/10.18653/v1/S19-2007.

Bassignana, E., Basile, V., & Patti, V. (2018). Hurtlex: A multilingual lexicon of words to hurt. In E. Cabrio, A. Mazzei, & F. Tamburini (Vol. Eds.), *CEUR Workshop Proceedings: . 2253. Proceedings of the fifth italian conference on computational linguistics (clic-it 2018), torino, italy, december 10-12, 2018* (pp. 1–6). CEUR-WS.org.

Baziotis, C., Pelekis, N., & Doulkeridis, C. (2017). *Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. Proceedings of the 11th international workshop on semantic evaluation (semeval-2017)*. Vancouver, Canada: Association for Computational Linguistics747–754.

Buscaldi, D. (2018). Tweetaneuse @ AMI EVALITA2018: character-based models for the automatic misogyny identification task (short paper). In T. Caselli, N. Novielli, V. Patti, & P. Rosso (Vol. Eds.), *CEUR Workshop Proceedings: . 2263. Proceedings of the sixth evaluation campaign of natural language processing and speech tools for italian. final workshop (EVALITA 2018) co-located with the fifth italian conference on computational linguistics (clic-it 2018), turin, italy, december 12-13, 2018* (pp. 1–4). CEUR-WS.org.

Cambria, E., Chandra, P., Sharma, A., & Hussain, A. (2010). *Do not feel the trolls. Proceedings of the 3rd international workshop on social data on the webCEUR Workshop Proceedings664. Proceedings of the 3rd international workshop on social data on the web* Shanghai, China: CEUR-WS.org1–12.

Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems, 32*(6), 74–80.

Canós, J. S. (2018). Misogyny identification through SVM at ibereval 2018. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, & J. C. de Albornoz (Vol. Eds.), *CEUR Workshop Proceedings: . 2150. Proceedings of the third workshop on evaluation of human language technologies for iberian languages (ibereval 2018) co-located with 34th conference of the spanish society for natural language processing (SEPLN 2018), sevilla, spain, september 18th, 2018* (pp. 229–233). CEUR-WS.org.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). *Learning phrase representations using RNN encoder–decoder for statistical machine translation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics1724–1734. https://doi.org/10.3115/v1/D14-1179.

Chung, Y.-L., Kuzmenko, E., Tekiroglu, S. S., & Guerini, M. (2019). *CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. Proceedings of the 57th annual meeting of the association for computational linguistics*. Association for Computational Linguistics2819–2829.

Code, L. (2002). *Encyclopedia of feminist theories*. Routledge.

Daumé III, H. (2007). *Frustratingly easy domain adaptation. Proceedings of the 45th annual meeting of the association of computational linguistics*. Prague, Czech Republic: Association for Computational Linguistics256–263.

Davidson, T., Bhattacharya, D., & Weber, I. (2019). *Racial bias in hate speech and abusive language detection datasets. Proceedings of the third workshop on abusive language online*. Florence, Italy: Association for Computational Linguistics25–35. https://doi.org/10.18653/v1/W19-3504.

De Mauro, T. (2016). Le parole per ferire. *Internazionale* 27 settembre 2016

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*. Minneapolis, Minnesota: Association for Computational Linguistics4171–4186. https://doi.org/10.18653/v1/N19-1423.

Fasoli, F., Carnaghi, A., & Paladino, M. P. (2015). Social acceptability of sexist derogatory and sexist objectifying slurs across contexts. *Language Sciences, 52*, 98–107.

Fehn Unsvåg, E., & Gambäck, B. (2018). *The effects of user features on Twitter hate speech detection. Proceedings of the 2nd workshop on abusive language online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics75–85. https://doi.org/10.18653/v1/W18-5110.

Fersini, E., Nozza, D., & Rosso, P. (Nozza, Rosso, 2018a). Overview of the evalita 2018 task on automatic misogyny identification (AMI). In T. Caselli, N. Novielli, V. Patti, & P. Rosso (Vol. Eds.), *CEUR Workshop Proceedings: . 2263. Proceedings of the sixth evaluation campaign of natural language processing and speech tools for italian. final workshop (EVALITA 2018) co-located with the fifth italian conference on computational linguistics (clic-it 2018), turin, italy, december 12-13, 2018* (pp. 1–9). CEUR-WS.org.

Fersini, E., Rosso, P., & Anzovino, M. (Rosso, Anzovino, 2018b). Overview of the task on automatic misogyny identification at ibereval 2018. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, & J. C. de Albornoz (Vol. Eds.), *CEUR Workshop Proceedings: . 2150. Proceedings of the third workshop on evaluation of human language technologies for iberian languages (ibereval 2018) co-located with 34th conference of the spanish society for natural language processing (SEPLN 2018), sevilla, spain, september 18th, 2018* (pp. 214–228). CEUR-WS.org.

Frenda, S., Ghanem, B., Guzmán-Falcón, E., Montes-y-Gómez, M., & Pineda, L. V. (Ghanem, Guzmán-Falcón, Montes-y-Gómez, Pineda, 2018a). Automatic expansion of lexicons for multilingual misogyny detection. In T. Caselli, N. Novielli, V. Patti, & P. Rosso (Vol. Eds.), *CEUR Workshop Proceedings: . 2263. Proceedings of the sixth evaluation campaign of natural language processing and speech tools for italian. final workshop (EVALITA 2018) co-located with the fifth italian conference on computational linguistics (clic-it 2018), turin, italy, december 12-13, 2018* (pp. 1–6). CEUR-WS.org.

Frenda, S., Ghanem, B., & Montes-y-Gómez, M. (Ghanem, Montes-y-Gómez, 2018b). Exploration of misogyny in spanish and english tweets. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, & J. C. de Albornoz (Vol. Eds.), *CEUR Workshop Proceedings: . 2150. Proceedings of the third workshop on evaluation of human language technologies for iberian languages (ibereval 2018) co-located with 34th conference of the spanish society for natural language processing (SEPLN 2018), sevilla, spain, september 18th, 2018* (pp. 260–267). CEUR-WS.org.

Frenda, S., Ghanem, B., Montes-y-Gómez, M., & Rosso, P. (2019). Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent and Fuzzy Systems, 36*(5), 4743–4752. https://doi.org/10.3233/JIFS-179023.

Fulper, R., Ciampaglia, G. L., Ferrara, E., Ahn, Y., Flammini, A., Menczer, F., ... Rowe, K. (2014). *Misogynistic language on Twitter and sexual violence. Proceedings of the acm web science workshop on computational approaches to social modeling (chasm)*1–4.

Gaur, M., Alambo, A., Sain, J. P., Kursuncu, U., Thirunarayan, K., Kavuluru, R., ... Pathak, J. (2019). *Knowledge-aware assessment of severity of suicide risk for early intervention. The world wide web conferenceWWW '19*New York, NY, USA: Association for Computing Machinery514–525. https://doi.org/10.1145/3308558.3313698.

Goenaga, I., Atutxa, A., Gojenola, K., Casillas, A., de Ilarraza, A. D., Ezeiza, N., ... Perez-de-Viñaspre, O. (2018). Automatic misogyny identification using neural networks. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, & J. C. de Albornoz (Vol. Eds.), *CEUR Workshop Proceedings: . 2150. Proceedings of the third workshop on evaluation of human language technologies for iberian languages (ibereval 2018) co-located with 34th conference of the spanish society for natural language processing (SEPLN 2018), sevilla, spain, september 18th, 2018* (pp. 249–254). CEUR-WS.org.

van der Goot, R., Ljubešić, N., Matroos, I., Nissim, M., & Plank, B. (2018). *Bleaching text: Abstract features for cross-lingual gender prediction. Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers)*. Melbourne, Australia: Association for Computational Linguistics383–389. https://doi.org/10.18653/v1/P18-2061.

Hewitt, S., Tiropanis, T., & Bokhove, C. (2016). *The problem of identifying misogynist language on Twitter (and other online social spaces). Proceedings of the 8th acm conference on web science*. ACM333–335.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput. 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Jha, A., & Mamidi, R. (2017). *When does a compliment become sexist? analysis and classification of ambivalent sexism using Twitter data. Proceedings of the second workshop* *on NLP and computational social science*. Vancouver, Canada: Association for Computational Linguistics7–16. https://doi.org/10.18653/v1/W17-2902.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. Nedellec, & C. Rouveirol (Vol. Eds.), *Lecture Notes in Computer Science: . 1398*. *Machine learning: Ecml-98, 10th european conference on machine learning, chemnitz, germany, april 21-23, 1998, proceedings* (pp. 137–142). Springer. https://doi.org/10.1007/BFb0026683.

Jurgens, D., Chandrasekharan, E., & Hemphill, L. (2019). *A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. Proceedings of the 57th annual meeting of the association for computational linguistics*. Association for Computational Linguistics (ACL)3658–3666.

Karan, M., & Snajder, J. (2018). Cross-domain detection of abusive language online. In D. Fiser, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, & J. Wernimont (Eds.). *Proceedings of the 2nd workshop on abusive language online, alw@emnlp 2018, brussels, belgium, october 31, 2018* (pp. 132–137). Association for Computational Linguistics.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.). *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, 4-9 december 2017, long beach, ca, USA* (pp. 3146–3154). .

Khatua, A., Cambria, E., Ghosh, K., Chaki, N., & Khatua, A. (2019). *Tweeting in support of lgbt? a deep learning approach. Proceedings of the acm india joint international conference on data science and management of dataCoDS-COMAD '19*New York, NY, USA: Association for Computing Machinery342–345. https://doi.org/10.1145/3297001.3297057.

Khatua, A., E., C., & Khatua, A. (2018). *Sounds of silence breakers: Exploring sexual violence on twitter. 2018 ieee/acm international conference on advances in social networks analysis and mining (asonam)*397–400.

Kramerae, C., & Spender, D. (2000). *Routledge international encyclopedia of women.* New York, London: Routledge.

Kshirsagar, R., Cukuvac, T., McKeown, K., & McGregor, S. (2018). *Predictive embeddings for hate speech detection on Twitter. Proceedings of the 2nd workshop on abusive language online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics26–32. https://doi.org/10.18653/v1/W18-5104.

Lample, G., Conneau, A., Ranzato, M., Denoyer, L., & Jégou, H. (2018). *Word translation without parallel data. 6th international conference on learning representations, ICLR 2018, vancouver, bc, canada, april 30 - may 3, 2018, conference track proceedings*. OpenReview.net1–14.

Liu, H., Chiroma, F., & Cocea, M. (2018). Identification and classification of misogynous tweets using multi-classifier fusion. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, & J. C. de Albornoz (Vol. Eds.), *CEUR Workshop Proceedings: . 2150*. *Proceedings of the third workshop on evaluation of human language technologies for iberian languages (ibereval 2018) co-located with 34th conference of the spanish society for natural language processing (SEPLN 2018), sevilla, spain, september 18th, 2018* (pp. 268–273). CEUR-WS.org.

Majumder, N., Poria, S., Peng, H., Chhaya, N., Cambria, E., & Gelbukh, A. (2019). Sentiment and sarcasm classification with multitask learning. *IEEE Intelligent Systems, 34*(3), 38–43.

Malmasi, S., & Zampieri, M. (2018). Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence, 30*(2), 187–202.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019). *Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. Proceedings of the 11th forum for information retrieval evaluation.* ACM14–17.

Manne, K. (2017). *Down girl: The logic of misogyny.* Oxford University Press.

Menini, S., Moretti, G., Corazza, M., Cabrio, E., Tonelli, S., & Villata, S. (2019). *A system to monitor cyberbullying based on message classification and social network analysis. Proceedings of the third workshop on abusive language online*. Florence, Italy: Association for Computational Linguistics105–110. https://doi.org/10.18653/v1/W19-3511.

Mishra, P., Tredici, M. D., Yannakoudakis, H., & Shutova, E. (2018). Author profiling for abuse detection. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.). *Proceedings of the 27th international conference on computational linguistics, COLING 2018, santa fe, new mexico, usa, august 20-26, 2018* (pp. 1088–1098). Association for Computational Linguistics.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). A bert-based transfer learning approach for hate speech detection in online social media. In H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, & L. M. Rocha (Vol. Eds.), *Studies in Computational Intelligence: . 881*. *Complex networks and their applications VIII - volume 1 proceedings of the eighth international conference on complex networks and their applications COMPLEX NETWORKS 2019, lisbon, portugal, december 10-12, 2019* (pp. 928–940). Springer. https://doi.org/10.1007/978-3-030-36687-2_77.

Nina-Alcocer, V. (2018). AMI at ibereval2018 automatic misogyny identification in spanish and english tweets. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, & J. C. de Albornoz (Vol. Eds.), *CEUR Workshop Proceedings: . 2150*. *Proceedings of the third workshop on evaluation of human language technologies for iberian languages (ibereval 2018) co-located with 34th conference of the spanish society for natural language processing (SEPLN 2018), sevilla, spain, september 18th, 2018* (pp. 274–279). CEUR-WS.org.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). *Abusive language detection in online user content. Proceedings of the 25th international conference on world wide web*145–153.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., & Yeung, D.-Y. (2019). *Multilingual and multi-aspect hate speech analysis. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*. Hong Kong, China: Association for Computational Linguistics4675–4684. https://doi.org/10.18653/v1/D19-1474.

Pamungkas, E. W., Basile, V., & Patti, V. (Basile, Patti, 2018a). Stance classification for rumour analysis in twitter: Exploiting affective information and conversation structure. In A. Cuzzocrea, F. Bonchi, & D. Gunopulos (Vol. Eds.), *CEUR Workshop Proceedings: . 2482*. *Proceedings of the CIKM 2018 workshops co-located with 27th ACM international conference on information and knowledge management (CIKM 2018), torino, italy, october 22, 2018* (pp. 1–7). CEUR-WS.org.

Pamungkas, E. W., Cignarella, A. T., Basile, V., & Patti, V. (Cignarella, Basile, Patti, 2018b). Automatic identification of misogyny in english and italian tweets at EVALITA 2018 with a multilingual hate lexicon. In T. Caselli, N. Novielli, V. Patti, & P. Rosso (Vol. Eds.), *CEUR Workshop Proceedings: . 2263*. *Proceedings of the sixth evaluation campaign of natural language processing and speech tools for italian. final workshop (EVALITA 2018) co-located with the fifth italian conference on computational linguistics (clic-it 2018), turin, italy, december 12-13, 2018* (pp. 1–9). CEUR-WS.org.

Pamungkas, E. W., Cignarella, A. T., Basile, V., & Patti, V. (Cignarella, Basile, Patti, 2018c). *14-ExLab@UniTo* for AMI at ibereval2018: Exploiting lexical knowledge for detecting misogyny in english and spanish tweets. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, & J. C. de Albornoz (Vol. Eds.), *CEUR Workshop Proceedings: . 2150*. *Proceedings of the third workshop on evaluation of human language technologies for iberian languages (ibereval 2018) co-located with 34th conference of the spanish society for natural language processing (SEPLN 2018), sevilla, spain, september 18th, 2018* (pp. 234–241). CEUR-WS.org.

Pamungkas, E. W., & Patti, V. (2019). Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In F. Alva-Manchego, E. Choi, & D. Khashabi (Eds.). *Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, florence, italy, july 28 - august 2, 2019, volume 2: Student research workshop* (pp. 363–370). Association for Computational Linguistics.

Park, J. H., Shin, J., & Fung, P. (2018). *Reducing gender bias in abusive language detection. Proceedings of the 2018 conference on empirical methods in natural language processing*. Brussels, Belgium: Association for Computational Linguistics2799–2804. https://doi.org/10.18653/v1/D18-1302.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature.* Penguin.

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

Poria, S., Majumder, N., Hazarika, D., Cambria, E., Gelbukh, A., & Hussain, A. (2018). Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems, 33*(6), 17–25.

Qian, J., ElSherief, M., Belding, E., & Wang, W. Y. (2018). *Leveraging intra-user and inter-user representation learning for automated hate speech detection. Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 2 (short papers)*. New Orleans, Louisiana: Association for Computational Linguistics118–123. https://doi.org/10.18653/v1/N18-2019.

Richardson-Self, L. (2018). Woman-hating: On misogyny, sexism, and hate speech. *Hypatia, 33*(2), 256–272. https://doi.org/10.1111/hypa.12398.

Ruder12, S., Bingel, J., Augenstein, I., & Søgaard, A. (2017). Sluice networks: Learning what to share between loosely related tasks. *stat, 1050*, 23.

Saha, P., Mathew, B., Goyal, P., & Mukherjee, A. (2018). Hateminers: Detecting hate speech against women. *arXiv preprint arXiv:1812.06700*.

Saha, P., Mathew, B., Goyal, P., & Mukherjee, A. (2019). *Hatemonitors: Language agnostic abuse detection in social media. Working notes of fire 2019 - forum for information retrieval evaluation*246–253 Kolkata, India

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018). An italian Twitter corpus of hate speech against immigrants. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, ... T. Tokunaga (Eds.). *Proceedings of the eleventh international conference on language resources and evaluation, LREC 2018, miyazaki, japan, may 7-12, 2018* (pp. 2798–2805). European Language Resources Association (ELRA).

Schneider, J. M., Roller, R., Bourgonje, P., Hegele, S., & Rehm, G. (2018). *Towards the automatic classification of offensive language and related phenomena in german tweets. 14th conference on natural language processing konvens 2018*95.

Sharifirad, S., & Jacovi, A. (2019). *Learning and understanding different categories of sexism using convolutional neural network's filters. Proceedings of the 2019 workshop on widening nlp*. Florence, Italy: Association for Computational Linguistics21–23.

Shushkevich, E., & Cardiff, J. (Cardiff, 2018a). Classifying misogynistic tweets using a blended model: The AMI shared task in IBEREVAL 2018. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, & J. C. de Albornoz (Vol. Eds.), *CEUR Workshop Proceedings: . 2150. Proceedings of the third workshop on evaluation of human language technologies for iberian languages (ibereval 2018) co-located with 34th conference of the spanish society for natural language processing (SEPLN 2018), sevilla, spain, september 18th, 2018* (pp. 255–259). CEUR-WS.org.

Shushkevich, E., & Cardiff, J. (Cardiff, 2018b). Misogyny detection and classification in english tweets: The experience of the ITT team. In T. Caselli, N. Novielli, V. Patti, & P. Rosso (Vol. Eds.), *CEUR Workshop Proceedings: . 2263. Proceedings of the sixth evaluation campaign of natural language processing and speech tools for italian. final workshop (EVALITA 2018) co-located with the fifth italian conference on computational linguistics (clic-it 2018), turin, italy, december 12-13, 2018* (pp. 1–6). CEUR-WS.org.

Smith, S. L., Turban, D. H., Hamblin, S., & Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*.

Sulis, E., Farías, D. I. H., Rosso, P., Patti, V., & Ruffo, G. (2016). Figurative messages and affect in twitter: Differences between #irony, #sarcasm and #not. *Knowledge Based Systems, 108*, 132–143. https://doi.org/10.1016/j.knosys.2016.05.035.

Swamy, S. D., Jamatia, A., & Gambäck, B. (2019). *Studying generalisability across abusive language detection datasets. Proceedings of the 23rd conference on computational natural language learning (conll)*. Hong Kong, China: Association for Computational Linguistics940–950. https://doi.org/10.18653/v1/K19-1088.

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., ... Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PloS one, 13*(10).

Wallace, B. C., Kertz, L., Charniak, E., et al. (2014). *Humans require context to infer ironic intent (so computers probably do, too). Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)2. Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)* 512–516.

Waseem, Z., & Hovy, D. (2016). *Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. Proceedings of the NAACL student research workshop*. San Diego, California: Association for Computational Linguistics88–93. https://doi.org/10.18653/v1/N16-2013.

Waseem, Z., Thorne, J., & Bingel, J. (2018). *Bridging the gaps: Multi task learning for domain transfer of hate speech detection. Online harassment*. Springer29–55.

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). *Detection of Abusive Language: the Problem of Biased Datasets. Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*. Minneapolis, Minnesota: Association for Computational Linguistics602–608. https://doi.org/10.18653/v1/N19-1060.

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). *Overview of the germeval 2018 shared task on the identification of offensive language. 14th conference on natural language processing konvens 2018*1.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (Malmasi, Nakov, Rosenthal, Farra, Kumar, 2019a). *Predicting the type and target of offensive posts in social media. Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*. Minneapolis, Minnesota: Association for Computational Linguistics1415–1420. https://doi.org/10.18653/v1/N19-1144.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (Malmasi, Nakov, Rosenthal, Farra, Kumar, 2019b). *SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). Proceedings of the 13th international workshop on semantic evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics75–86. https://doi.org/10.18653/v1/S19-2010.