Automatic Expansion of Lexicons for Multilingual Misogyny Detection

Simona Frenda
Università degli Studi di Torino, Italy
Universitat Politècnica de València, Spain
simona.frenda@unito.it

Bilal Ghanem
Universitat Politècnica de València, Spain
bigha@doctor.upv.es

Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez and Luis Villaseñor-Pineda
Instituto Nacional de Astrofísica
Óptica y Electrónica (INAOE), Mexico.
{fany.guzman, mmontesg, villasen}@inaoep.mx

Abstract

English. The automatic misogyny identification (AMI) task proposed at IberEval and EVALITA 2018 is an example of the active involvement of scientific Research to face up the online spread of hate contents against women. Considering the encouraging results obtained for Spanish and English in the precedent edition of AMI, in the EVALITA framework we tested the robustness of a similar approach based on topic and stylistic information on a new collection of Italian and English tweets. Moreover, to deal with the dynamism of the language on social platforms, we also propose an approach based on automatically-enriched lexica. Despite resources like the lexica prove to be useful for a specific domain like misogyny, the analysis of the results reveals the limitations of the proposed approaches.

Italiano. *Il task AMI circa l'identificatione automatica della misoginia proposto a IberEval e a EVALITA 2018 è un chiaro esempio dell'attivo coinvolgimento della Ricerca per fronteggiare la diffusione online di contenuti di odio contro le donne. Considerando i promettenti risultati ottenuti per spagnolo e inglese nella precedente edizione di AMI, nel contesto di EVALITA abbiamo testato la robustezza di un approccio simile, basato su informationi stilistiche e di dominio, su una nuova collezione di tweet in inglese e in italiano. Tenendo conto dei repentini cambiamenti del linguaggio nei social network, proponiamo anche un approccio basato su lessici automaticamente estesi. Nonostante risorse come i lessici risultano utili per domini specifici come quello della misoginia, analizzando i risultati emergono i limiti degli approcci proposti.*

1 Introduction

The anonymity and the interactivity, typical of computer-mediated communication, facilitate the spread of hate messages and the perpetuated presence of hate contents online. As investigated by Fox et al. (2015), these factors increase and influence social misbehaviors also offline. In order to foster scientific research to find optimal solutions that could help to monitor the spread of hate speech contents, different tasks have been proposed in various campaigns of evaluation. An example is the AMI shared task proposed at IberEval 2018[1] and later at EVALITA 2018[2]. This task focuses on the automatic identification of misogyny in different languages. In particular, the first edition focuses on Spanish and English languages, and the second one on a new English corpus and Italian language. The multilingual context allows to observe the analogies and differences between different languages. The AMI's organizers (Fersini et al., 2018a; Fersini et al., 2018b) asked participants to detect firstly misogynistic tweets and then classify the misogynistic categories and the kind of target (individuals or groups). In the first edition, we proposed an approach based on stylistic and topic information captured respectively by means of character n-grams and a set of modeled lexica (Frenda et al., 2018). Considering the encouraging results obtained with the lexicon-based approach in Spanish and English languages, we re-proposed a similar approach for Italian language and a new collection of English tweets in

[1] http://amiibereval2018.wordpress.com/
[2] http://amievalita2018.wordpress.com/

order to test the performance and robustness of this approach. Actually, in this paper we propose two approaches. The first one, similar to previous work (Frenda et al., 2018), involves topic, linguistic and stylistic information. The second one focuses mainly on the automatic extension of the original lexica. Indeed, to deal with the continuous variation of the language on social platforms, the modeled lexica are enriched considering the contextual similarity of lexica by the use of pre-trained word embeddings. This technique helps the system to consider also new terms relative to the topic information of the original lexica. It could be considered as a good methodology to upgrade automatically the existing list of words used to block offensive contents in real applications of Internet companies. Indeed, a comparison between the two approaches reveals that the automatic enrichment of the lexica improves the results especially for English language. However, comparing the results obtained in both competitions and observing the error analyses, we notice that lexica represent a good resource for a specific domain like misogyny, but they are not sufficient to detect misogyny online.

Following, Section 2 describes the studies that inspired our work. Section 3 explains the approaches employed in both languages. Section 4 discusses the obtained results and delineates some conclusions.

2 Related Work

A first work about misogyny detection is proposed in Anzovino et al. (2018). In this study, the authors compared the performance of different supervised approaches using word embeddings, stylistic and syntactic features. In particular, their results reveal that the best machine learning approach for identification of misogyny is the linear Support Vector Machine (SVM) classifier. In general machine learning techniques are the most used in hate speech detection (Escalante et al., 2017; Nobata et al., 2016), because they allow researchers for exploring closely the issue exploiting different features, such as textual (Chen et al., 2012) and syntactical aspects (Burnap and Williams, 2014) or semantic and sentiment information (Samghabadi et al., 2017; Nobata et al., 2016; Gitari et al., 2015). Finally, some recent works have investigated also the potential of deep learning techniques (Mehdad and Tetreault, 2016; Del Vigna et al., 2017). Considering the specific domain concerning the hate against women, this work exploits stylistic, linguistic and topic information about the misogynistic speech. In particular, differently from previous studies, we use specific lexica relative to offensiveness and discredit of women for English and Italian languages, and we extend them with new words relative to the issues of the considered lexica. Considering the fact that commercial methods rely currently on the use of blacklists to monitor or block offensive contents, the proposed approach could help to upgrade their blacklists automatizing the process of the lexicon building.

3 Proposed Approaches

The AMI shared task proposed at EVALITA 2018 aims to detect misogyny in English and Italian collections of tweets. The organizers asked participants to detect misogynistic texts (Task A), and then, if the tweet is predicted as misogynistic, to distinguish the nature of target (individuals or groups labeled respectively “active” and “passive”), and identify the type of misogyny (Task B), according to the following classes proposed by Poland (2016): (a) stereotype and objectification, (b) dominance, (c) derailing, (d) sexual harassment and threats of violence, and (e) discredit. Actually, these classes represent the different manifestations and the various aspects of this social misbehavior. Table 1 shows the composition of the datasets.

Considering the promising results obtained at the IberEval campaign, in this work we use two approaches mainly based on lexica. The first one (Section 3.1) is similar to the approach used in Frenda et al. (2018), based on topic, linguistic and stylistic information captured by means of modeled lexica and n-grams of characters and words. The second one (Section 3.2) principally involves the automatically extended versions of the original lexica (Guzmán Falcón, 2018). In particular, we aim: 1) to test the robustness of lexicon based approaches in the new collections of tweets and in a new language, and 2) to understand the impact of automatically enriched lexica to face up the variation of the language in the multilingual computer-mediated communication.

	Misogynistic							Non-misogynistic
	(a)	(b)	(c)	(d)	(e)	active	passive	
Italian								
Training set	668	71	24	431	634	1721	97	2172
Test set	175	61	2	170	104	446	66	488
English								
Training set	179	148	92	352	1014	1058	727	2215
Test set	140	124	11	44	141	401	59	540

Table 1: Composition of AMI's datasets at EVALITA 2018.

3.1 Approach 1: using manually-modeled lexica (MML)

The first proposed approach aims to capture topic, linguistic and stylistic information by means of manually-modeled lexica and n-grams of words and characters. Below the features description for each language.

English Features. For the detection of misogyny in English tweets, we employed the manually-modeled lexica proposed in Frenda et al. (2018). These lexica concerns sexuality, profanity, femininity and human body as described in Table 2.

These lexica contain also slang expressions. Moreover, we take into account hashtags and abbreviations collected in Frenda et al. (2018): 40 misogynistic hashtags, such as: $\#ihatefemales$ or $\#bitchesstink$; and a list of 50 negative abbreviations, such as wtf or $stfu$. Considering the most relevant n-grams of words, we employ the bigrams for the first task and the combination of unigrams, bigrams and trigrams (hence defined as UBT) for the second task. Moreover, the bag of characters (BoC) in a range from 1 to 7 grams is employed to manage misspellings and to capture stylistic aspects of digital writing. In order to perform the experiments, each tweet is represented as a vector. The presence of words in each lexicon is pondered with Information Gain, and character and word n-grams are weighted with Term Frequency-Inverse Document Frequency (TF-IDF) measure. In addition, considering the fact that in Frenda et al. (2018) several misclassified misogynistic tweets were ironic or sarcastic, we try to analyze the impact of irony in misogyny detection in English. Indeed, Ford and Boxer (2011) reveal that sexist jokes that in general are considered innocent, truthfully they are experienced by women as sexual harassment. In particular, inspired by Barbieri and Saggion (2014), we calculate the imbalance of the sentiment polarities (positive and negative) in each tweet using SentiWordNet provided by Baccianella et al. (2010). For each degree of imbalance, we associate a weight used in the vectorial representation of the tweets. Despite our hypothesis is well funded, we obtained lower results for the runs that contain sentiment imbalance among the features (see Table 4).

Italian Features. For the Italian language, we selected some specific issue groups, described in Bassignana et al. (2018), from the Italian lexicon "Le parole per ferire" provided by Tullio De Mauro[3]. In particular, we consider the lists of words described in Table 3. Differently from English, the experiments reveal that: the UBT is useful for both tasks and the best range for BoC is from 3 to 5 grams[4]. Indeed, in a morphological complex language like Italian the desinences of the words (such as the extracted n-grams "tona" or "ana ") contain relevant linguistic information. Diversely, in English, longer sequences of characters could help to capture multi-word expressions containing also pronouns, adjectives or prepositions, such as "ing at" or "ss bitc".

To extract the features correctly, in order to train our models, we pre-process the data deleting emoticons, emojis and URLs. Indeed, from our experiments, the emoticons and emojis do not prove to be relevant for these tasks. In order to perform a correct match between the dictionaries of the corpora and the single lexicon, we use the lemmatizer provided by the Natural Language Toolkit (NLTK[5]) for English, and the Snowball Stemmer for Italian. Differently from English, the use of lemmatizer for Italian tweets hinders the match.

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

[4] The experiments are carried out using the Grid Search.

[5] http://www.nltk.org/

Lexicons	*Words*	*Definition*
Sexuality	290	contains words relative to sexual subject ($orgasm$, $orgy$, $pussy$) and especially male domination on women ($rape$, $pimp$, $slave$)
Profanity	170	is a collection of vulgar words such as $motherfucker$, $slut$ and $scum$
Femininity	90	is a list of terms used to identify the women as target. It contains personal pronouns or possessive adjectives (such as she, her, $herself$), common words used to refer to women ($girl$, $mother$) and also offensive words towards women (such as $barbie$, $hooker$ or $non-male$)
Human body	50	is a lexicon strongly connected with sexuality collecting words referred especially to feminine body also with negative connotations (such as $holes$, $throat$ or $boobs$)

Table 2: Composition of English lexica.

Lexicons	*Words*	*Definition*
AN	111	collects words relative to animals, such as $sanguisuga$ or $pecora$
ASF	31	contains terms referred to female genitalia, such as $fessa$
ASM	76	contains terms referred to male genitalia, such as $verga$
CDS	298	is a list of derogatory words, such as $bastardo$ or $spazzatura$
OR	17	contains words derived from plants but that are used as offensive words, such as $finocchio$ or $rapa$
PA	83	is a list of professions or jobs that have also a negative connotations, such as $portinaia$ or $impiegato$
PR	54	contains terms about prostitution, such as $bagascia$ or $zoccolona$
PS	42	is a list of words relative to stereotypes, such as $negro$ or $ostrogoto$
QAS	82	collects words that have in general negative connotations, such as $parassita$ or $dilettante$
RE	37	contains terms relative to criminal acts or immoral actions, such as $stupro$ or $violento$

Table 3: Composition of Italian lexica.

3.2 Approach 2: using automatically-enriched lexica (AEL)

The second approach aims to deal with the dynamism of the informal language online trying to capture new words relative to contexts defined in each lexicon. Therefore, we use enriched versions of the original lexica (described above), and stylistic and linguistic information captured by means of n-grams of words and characters as in the first approach. The method for the expansion of a given lexicon shares the idea of identifying new words by considering their contextual similarity with known words, as defined by some pre-trained word embeddings. For its description, let assume that $\mathcal{L} = \{l_1, \ldots, l_m\}$ is the initial lexicon of m words, and $\mathcal{W} = \{(w_1, e(w_1)), \ldots, (w_n, e(w_n))\}$ is the set of pre-trained word embeddings, where each pair represents a word and its corresponding embedding vector. This method aims to enrich the lexicon with words strongly related to the context from the original lexicon without being necessarily associated to any particular word. Its idea is to search for words having similar contexts to the entire lexicon. This method has two main steps, described below.

Dictionary modeling. Firstly, we extract the embedding $e(l_i)$ for each word $l_i \in \mathcal{L}$; then, we compute the average of these vectors to obtain a vector describing the entire lexicon, $e(\mathcal{L})$. We name this vector the context embedding.

Dictionary expansion. Using the cosine similarity, we compare $e(\mathcal{L})$ against the embedding $e(w_i)$ of each $w_i \in \mathcal{W}$; then, we extract the k most similar words to $e(\mathcal{L})$, defining the set $E_L = (w_1, \ldots, w_k)$. Finally, we insert the extracted words into the original lexicon to build the new lexicon, i.e., $\mathcal{L}_\mathcal{E} = \mathcal{L} \cup E_L$.

Therefore, we carry out the experiments using different pre-trained word embeddings for each language: *GloVe* embeddings trained on 2 billion tweets (Pennington et al., 2014) for English, and word embeddings built on TWITA corpus[6] for Italian (Basile and Novielli, 2014). Finally, the proposed expansion method is parametric and requires a value for k, the number of words that are going to extend the lexica. In particular, we use $k = 1000$, 500 and 100.

3.3 Experiments and Results

To carry out the experiments, a SVM classifier is employed with the radial basis function kernel (RBF) using the following parameters: $C = 5$ and $\gamma = 0.1$ for English and $\gamma = 0.01$ for Italian. Considering the complexity of the target classification for the Italian language due to imbalanced training set (see Table 1), we used a Random Forest (RF) classifier that aggregates the votes from different

[6] http://valeriobasile.github.io/twita/about.html

decision trees to decide the final class of the tweet.

The evaluation is performed using the test set provided by the organizers of the AMI shared task. For the competition, they use as evaluation measures the Accuracy for Task A and the average of F-score of both classes for Task B.

English			
Run	Approach	Accuracy	Rank
run 2[7]	AEL	**0.613**	**17**
baseline_AMI		*0.605*	*19*
run 1	AEL	0.592	21
run 3	MML	0.584	25
Italian			
Run	Approach	Accuracy	Rank
baseline_AMI		*0.830*	*7*
run 1	AEL	0.824	9
run 3[8]	AEL	0.823	11
run 2	MML	0.822	12

Table 4: Results obtained in Task A.

Table 4 and Table 5 show the results obtained in the competition compared with the baselines provided by the organizers for each task. Comparing the two approaches, in general AEL seems to work better than MML. However, the improvement of the results is very slight, especially for Italian language. This soft variation is unexpected considered the results obtained during the experiments employing 10-fold cross validations. In fact, AEL with enriched lexica using k equal 100 performed an Accuracy of 0.880. Moreover, looking at Table 4, reporting the official results of the AMI Task, only run 2 overcomes the baseline for the detection of misogyny in English, and for this run we used AEL approach excluding the sentiment imbalance as feature. About the identification of misogyny in Italian, the obtained results are lower than provided baselines as well as the values of F-score obtained in Task B for both languages (see Table 5). Despite the usefulness of lexica for a specific domain like misogyny, a lexicon-based approach proves to be insufficient for this task. Indeed, as the error analysis will confirm, misogyny, as well as general hate speech, involves linguistic devices such as humour, exclamations typical of orality and contextual information that completes the meaning transmitted by the tweet. Moreover, the low values obtained also in Task B suggest the necessity to implement dedicated approach for each misogynistic category.

[7]This run does not involve the sentiment imbalance

[8]This run involves the expansions of lexica with $k = 100$

4 Discussion and Conclusions

This paper reports our participation in the AMI shared task. The organizers provide also the gold test set that helps us to understand better what are the misclassified cases and the aspects that should be considered in the next experiments. Carrying out the error analysis, we notice that in both datasets the content of URL affects the transmitted information in the tweet (such as *Right! As they rape and butcher women and children !!!!!! https://t.co/maEhwuYQ8B*). The swear words are often used also as exclamation without the aim to offend (such as *Volevo dire alla Yamamay che tettona non sinonimo di curvy dato che di vita ha una 40, quindi confidence sta minchia.*). Moreover, despite the actual English corpus does not contain several jokes, Italian misclassified tweets involve humourous utterances (such as *@GrianneOhmsfor1 @BarbaraRaval A parte il fatto poi che culona inchiavabile" è il miglior giudizio politico sentito sulla Merkel negli ultimi anni??"*). In fact, in general, humour, irony and sarcasm hinder the correct classification of the texts, as we noticed in English and Spanish corpora provided in the IberEval framework. Participating in this shared task gave us the opportunity to analyze and compare multilingual datasets, and thus, to discover and infer general aspects typical of hate speech against women.

Acknowledgments

The work of Simona Frenda was partially funded by the Spanish research project SomEMBED TIN2015-71147-C2-1-P (MINECO/FEDER). We also thank the support of CONACYT-Mexico (projects FC-2410, CB-2015-01-257383).

References

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204.

Francesco Barbieri and Horacio Saggion. 2014. Modelling irony in twitter. In *Proceedings of the Stu-*

English						
Run	Categories	F-score	Target	F-score	total	ranks
baseline_AMI		*0.342*		*0.399*	*0.370*	*3*
run 2	UBT	0.282	UBT+BoC	**0.407**	0.344	6
run 1	UBT	0.282	UBT+BoC	0.389	0.335	8
run 3	UBT	0.269	UBT+BoC	0.387	0.328	10
Italian						
Run	Categories	F-score	Target	F-score	Total	ranks
baseline_AMI		*0.534*		*0.440*	*0.487*	*2*
run 3	UBT+BoC	0.485	UBT+BoC	0.414	0.449	7
run 1	UBT+BoC	0.483	UBT+BoC	0.414	0.448	8
run 2	UBT+BoC	0.480	UBT+BoC	0.411	0.446	10

Table 5: Results obtained in Task B.

dent Research Workshop at the 14th Conference of the European Chapter of the ACL.

Pierpaolo Basile and Nicole Novielli. 2014. Uniba at evalita 2014-sentipolc task: Predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features. In *Proceedings of EVALITA 2014*.

Elisa Bassignana, Valerio Basile, and Patti Viviana. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of CLiC-it, Turin, 10-12 December 2018, CEUR*.

Peter Burnap and Matthew Leighton Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making. *Internet, Policy & Politics*.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT)*, pages 71–80. IEEE.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of ITASEC17*.

Hugo Jair Escalante, Esaú Villatoro-Tello, Sara E Garza, A Pastor López-Monroy, Manuel Montes-y Gómez, and Luis Villaseñor-Pineda. 2017. Early detection of deception and aggressiveness using profile-based representations. *Expert Systems with Applications*, 89:99–111.

Elisabetta Fersini, Maria Anzovino, and Paolo Rosso. 2018a. Overview of the task on automatic misogyny identification at ibereval. In *Proceedings of Workshop IBEREVAL at 3rd SEPLN*.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018b. Overview of the evalita 2018 task on automatic misogyny identification (ami). In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Thomas E Ford and Christie Fitzgerald Boxer. 2011. Sexist humor in the workplace: A case of subtle harassment. In *Insidious Workplace Behavior*, pages 203–234. Routledge.

Jesse Fox, Carlos Cruz, and Ji Young Lee. 2015. Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Computers in Human Behavior*, 52:436–442.

Simona Frenda, Bilal Ghanem, and Manuel Montes-y Gómez. 2018. Exploration of misogyny in spanish and english tweets. In *Proceedings of Workshop IBEREVAL at 3rd SEPLN*.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Estefanía Guzmán Falcón. 2018. *Detección de lenguaje ofensivo en Twitter basada en expansión automática de lexicones (tesis de maestría)*. Instituto Nacional de Astrofísica, Óptica y Electrónica. Puebla, México.

Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on WWW*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*.

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

Niloofar Safi Samghabadi, Suraj Maharjan, Alan Sprague, Raquel Diaz-Sprague, and Thamar Solorio. 2017. Detecting nastiness in social media. In *Proceedings of ALW1*.