CEUR-WS.org/Vol-2150/AMI_paper2.pdf

14-ExLab@UniTo for AMI at IberEval2018: Exploiting Lexical Knowledge for Detecting Misogyny in English and Spanish Tweets

Endang Wahyu Pamungkas[1], Alessandra Teresa Cignarella[1,2], Valerio Basile[1], and Viviana Patti[1]

[1] Dipartimento di Informatica, Università degli Studi di Torino
[2] PRHLT Research Center, Universitat Politècnica de València
{pamungka,cigna,basile,patti}@di.unito.it

Abstract We describe our participation to the *Automatic Misogyny Identification (AMI)* shared task at IberEval 2018. The task focused on the detection of misogyny in English and Spanish tweets and was articulated in two sub-tasks addressing the identification of misogyny at different levels of granularity. We describe the final submitted systems for both languages and sub-tasks: Task A is a classical binary classification task to determine whether a tweet is misogynous or not, while Task B is a finer grained classification task devoted to distinguish different types of misogyny, where systems must predict (i) one out of five categories of misogynistic behaviours and (ii) if the abusive content was purposely addressed to a specific target or not. We propose an SVM-based architecture and explore the use of several sets of features, including a wide range of lexical features relying on the use of available and novel lexicons of abusive words, with a special focus on sexist slurs and abusive words targeting women in the two languages at issue. Our systems ranked first in Task A for both English and Spanish (accuracy score of 0.913 for English; 0.815 for Spanish), outperforming the baselines and the other participant systems, and first in Task B on Spanish.

1 Introduction

In the era of mass online communication, more and more episodes of hateful language and harassment against women occur in social media [3]. *Hate Speech* (HS) can be defined as any type of communication that is abusive, insulting, intimidating, harassing, and/or incites to violence or discrimination, and that disparages a person or a group on the basis of some characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics [1]. In particular, when HS is gender-oriented, and it specifically targets women, we refer to it as *misogyny* [2].

Recently, an increasing number of scholars is focusing on the task of automatic detection of abusive or hateful language *online* [3] where hate speech is

[3] https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-3

characterized by some key aspects which distinguish it from offline, face-to-face communication and make it potentially more dangerous and hurtful. In particular, hate speech in the form of racist and misogynist remarks are a common occurrence on social media [4], therefore recent works on the detection of HS focused on HS related to race, religion, and ethnic minorities [5] and on gender-based hate, which is also the focus of the AMI shared task.

Detecting misogynist content and its author is still a difficult task for social media platforms. For instance, the popular social network Facebook is still unable to deal with this issue and it relies on its community to report misogynistic content[4]. The work of Hewitt et al. [6] is a first study that attempts to detect misogyny in Twitter manually, in which the authors used several terms related to slurs against women to gather the data from Twitter. However, the automatic detection of misogynistic content is still an open problem, with few approaches proposed only recently [7].

In this paper, we describe the systems we submitted for detecting misogyny in the context of the *Automatic Misogyny Identification (AMI)* shared task at IberEval 2018 [8], defined as a two-fold task on detecting misogyny in English and Spanish tweets at different levels of granularity. In particular, considering the role of lexical choice in gender stereotypes, we decided to explore the role of lexical knowledge in detecting misogyny, by experimenting with lexical features based on both generic lexicons of slurs and abusive words, and on specific lexicons of sexist slurs and hate words targeting women.

2 The *14-ExLab@UniTo* systems

We built two similar systems for misogyny detection, one for English and one for Spanish. Several sets of features were considered based on a linguistically motivated approach, including *stylistic*, *structural* and *lexical features*. In particular, in order to explore the role of lexical knowledge in this task, we experimented the use of (i) generic lexicons of abusive words and slurs; (ii) specific lexicons of sexist slurs and hate words reflecting specifically gender-based hate and well-known cultural gender bias and stereotypes. In particular, we experimented for the first time in this task the use of a new multilingual lexicon (*HurtLex*), including an inventory of hate words compiled by the Italian linguist Tullio De Mauro [9], which has been semi-automatically translated from Italian into English and Spanish both relying on BabelNet [10].

The list of lexical features includes: **Bag of Words (*BoW*)**: sparse vector encoding the occurrence of unigrams, bigrams and trigrams in a tweet. **Swear Word Count**: this feature represents the number of swear words contained in a tweet. We used the list of swear words from *noswearing* dictionary [5]. **Swear Word Presence**: this feature is a binary value representing the presence of swear words. We used the same dictionary from *noswearing*. **Sexist Slurs Presence**:

[4] https://www.nytimes.com/2013/05/29/business/media/facebook-says-it-failed-to-stop-misogynous-pages.html

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

we use a small set of sexist words aimed towards women from prior work [11]. This feature has a binary value 0 (there is no sexist slur in the tweet) and 1 (there is at least one sexist slur in the tweet). **Woman-related Words Presence**: this feature is used to represent the target of misogyny. Therefore, we manually built a small set of words in English containing synonyms or other words related to the word "woman" [6]. Additionally, we extracted a set of features based on the presence of words from the **HurtLex lexicon** [10]. This lexicon includes a wide inventory of about 1,000 Italian hate words originally compiled in a manual fashion by De Mauro [9] organized in 17 categories grouped in different macro levels:
(a) *Negative stereotypes*: ethnic slurs (PS); locations and demonyms (RCI); professions and occupations (PA); **physical disabilities and diversity (DDF)**; **cognitive disabilities and diversity (DDP)**; moral and behavioral defects (DMC); words related to social and economic disadvantage (IS).
(b) *Hate words and slurs beyond stereotypes*: plants (OR); animals (AN); **male genitalia (ASM)**; **female genitalia (ASF)**; **words related to prostitution (PR)**; words related to homosexuality (OM).
(c) *Other words and insults*: descriptive words with potential negative connotations (QAS); derogatory words (CDS); felonies and words related to crime and immoral behavior (RE); words related to the seven deadly sins of the Christian tradition (SVP). The lexicon has been translated into English and Spanish semi-automatically by extracting all the senses of all the words from BabelNet [12], manually discarding the senses that were not relevant to the context of hate, and finally retrieving all the English and Spanish lemmas for the remaining senses. Thanks a manual inspection we identified five categories as specifically related to gender-based hate: DDF and DDP related to negative stereotypes; PR, ASM and ASF beyond stereotypes (highlighted in bold).

The structural features employed by our systems include: **Bag of Hashtags (*BoH)***: similarly to BoW, we exploit the hashtags. **Bag of Emojis (*BoE)***: we also utilized the Emojis in the tweets as a feature. We used their CLDR short name [7] in our feature matrix. Therefore, we converted the emoji unicode to its CLDR short name by using PyPI library[8]. **Hashtag Presence**: this feature has a binary value 0 (if there is no hashtag in the tweet) or 1 (if there is at least one hashtag in the tweet). **Link Presence**: presence of URLs in the tweets as a binary value: 0 if there is no link, 1 if there is at least one link in the tweet. All the features are encoded as fixed-size numerical or one-hot vector representations, allowing us to experiment extensively with their combination.

3 Experiments and Results

In this section, we report on the result of the evaluation of our system for misogyny detection according to the benchmark established by the AMI task.

[6] For the Spanish system development, we translated all the English word lists described here by using Google Translate: `https://translate.google.com/`.

[7] `https://unicode.org/emoji/charts/full-emoji-list.html`

[8] `https://pypi.org/project/emoji/`

3.1 AMI: Tasks Description and Dataset Composition

The organizers of AMI proposed an automatic detection task of misogynistic content on Twitter, in English (EN) and Spanish (SP). Two different tasks were proposed: Task A is a binary classification task, where every system should determine whether a tweet is *misogynous* or *not misogynous.* Task B is composed of two distinct classification tasks. First, participants were asked to classify the misogynous tweets into five categories of misogynistic behavior including: *"stereotype & objectification"*, *"dominance"*, *"derailing"*, *"sexual harassment & threats of violence"*, and *"discredit"*. Secondly, they were asked to classify the misogynous tweets based on their target, labeling whether it is *active* (i.e. referring to one woman in particular) or *passive* (i.e. referring to a group of women).

Task A is evaluated in terms of accuracy, while for Task B the evaluation consists in the macro-average of the F_1-scores on the positive classes. Each participating team could submit a maximum of 5 runs, pertaining to two different scenarios: constrained and unconstrained.

Dataset As summarized in Table 1, the organizers provided 3,251 tweets for the English training set and 3,307 tweets for the Spanish training set. Each tweet, in both languages, was annotated at three levels: 1) presence of misogynous content, 2) categories of misogynistic behavior, as described in Section 3.1, and 3) target of misogyny (*active* or *passive*). The organizers provided a balanced label distribution for Task A (*misogynous* vs. *not misogynous*), while the distribution of data for Task B was highly unbalanced, reflecting the natural distribution of misogynistic behaviours and targets in the corpus.

Task A			**Task B**		
	English	**Spanish**		**English**	**Spanish**
Misogynistic	1,568	1,649	Stereotype	137	151
			Dominance	49	302
			Derailing	29	20
			Sexual Harassment	410	198
			Discredit	943	978
			Active	942	1455
			Passive	626	194
Not misogynistic	1,683	1,658	No class	1,683	1,658
Total				3,251	3,307

Table 1. Dataset label distribution.

3.2 Experimental Setup

We built two variants of our system and trained them on the available training sets. We tuned the system on the basis of the results of a 10-fold cross validation, using accuracy as an evaluation metric for Task A. The system for English is based on SVM with Radial Basis Function (RBF) kernel, while the system built for Spanish is based on SVM with a linear kernel. Both systems were built by using *scikit-learn* Python library[9]. Additionally, we performed an ablation test on our feature sets to study the impact of the different features on the system performance. Table 2 shows the features selected for each of our submissions and accuracy scores from cross-validation on training sets for English and Spanish. For what concerns features based on HurtLex, in S4 (EN) and S3 (SP) we explored the impact of hate words belonging to categories specifically related to gender-based hate (see Sec. 2). In addition, we tested the performance of the best-performing sets of features of one language applied to the other language, to gauge the multilingual potential of the best systems: the English submission 5 is based on the best-performing (in cross-validation) combination of features for Spanish, and the Spanish submission 5 is based on the best-performing com-

Languages	**English**					**Spanish**				
Systems	S1	**S2**	S3	S4	S5	**S1**	S2	S3	S4	S5
Accuracy	0.748	0.75	0.75	0.737	0.73	0.791	0.789	0.787	0.789	0.73
Bag of Word	-	-	-	-	✓	✓	✓	✓	✓	-
Bag of Hashtags	-	-	-	-	✓	✓	✓	✓	✓	-
Bag of Emojis	-	-	-	-	✓	✓	✓	✓	✓	-
Hashtag Presence	✓	✓	✓	✓	-	-	-	-	-	✓
Link Presence	✓	✓	✓	✓	-	-	-	-	-	✓
Swear Word Count	✓	✓	✓	✓	-	-	-	-	-	✓
Swear Word Presence	✓	✓	✓	✓	-	-	-	-	-	✓
Sexist Slurs Pres.	✓	✓	✓	✓	-	-	✓	✓	✓	✓
Woman Word Pres.	✓	✓	✓	✓	-	-	✓	✓	✓	✓
ASF Count	-	✓	✓	✓	-	-	-	✓	✓	✓
PR Count	-	✓	✓	✓	-	-	-	✓	✓	✓
OM Count	-	✓	-	-	-	-	-	-	✓	✓
DDF Count	-	-	✓	✓	-	-	-	✓	✓	-
CDS Count	-	-	-	-	-	-	-	-	✓	-
DDP Count	-	-	-	✓	-	-	-	✓	✓	-
AN Count	-	-	-	-	-	-	-	-	✓	-
ASM Count	-	-	-	✓	-	-	-	✓	✓	-
DMC Count	-	-	-	-	-	-	-	-	✓	-
IS Count	-	-	-	-	-	-	-	-	✓	-
OR Count	-	-	-	-	-	-	-	-	✓	-
PA Count	-	-	-	-	-	-	-	-	✓	-
PS Count	-	-	-	-	-	-	-	-	✓	-
QAS Count	-	-	-	-	-	-	-	-	✓	-
RCI Count	-	-	-	-	-	-	-	-	✓	-
RE Count	-	-	-	-	-	-	-	-	✓	-
SVP Count	-	-	-	-	-	-	-	-	✓	-

Table 2. Feature Selection for all the submitted systems.

[9] http://scikit-learn.org/

bination of features for English. For Task B, we used exactly the same features as Task A in each submission. We only submitted constrained runs.

3.3 Official Results and Analysis

Table 3 until Table 6 shows our submission ranking based on the competition official results [10]. The submission name is based on the submission numbering on Table 2 (run 1 is result of S1 and so on). Our systems ranked first in Subtask A for both English (accuracy 0.913 by run 1) and Spanish (accuracy 0.815 by run 3). Meanwhile, for Subtask B (Table 5 and Table 6), one of our systems was the best result on Spanish (average Macro F-measure 0.446 by run 2) and the 6^{th} on English (average Macro F-measure 0.370 by run 5).

Our experiment in testing the multilingual setting proved to be a challenge. Not surprisingly, both submissions 5 were the worst-performing compared to other submissions. However, the English S5 shows a comparatively good performance in absolute terms. On Table 3, we can see that all of our submissions in English were above the competition baseline. However as we can see on Table 4, with the same system applied to the Spanish dataset, we obtained a very low accuracy score in Spanish (ranked 24^{th}, accuracy 0.537). This asymmetry indicates that the combination of BoW, BoH and BoE is a better representation of tweets in a multilingual setting than more ad-hoc, task-specific features.

RANK	SUBMISSIONS	ACCURACY
1	14-exlab.c.run1	**0.913**
2	14-exlab.c.run2	0.902
3	14-exlab.c.run4	0.898
4	14-exlab.c.run3	0.879
...	...	...
10	14-exlab.c.run5	0.824
...	...	...
15	AMI-BASELINE	0.784

Table 3. Task A rankings (English)

RANK	SUBMISSIONS	ACCURACY
1	14-exlab.c.run3	**0.815**
4	14-exlab.c.run1	0.812
5	14-exlab.c.run2	0.812
6	14-exlab.c.run4	0.809
...	...	...
18	AMI-BASELINE	0.767
...	...	...
24	14-exlab.c.run5	0.536

Table 4. Task A rankings (Spanish)

On Task B, most participants achieved relatively low results, showing the difficulty of this task, especially in classifying misogynistic behavior categories. We found the datasets' unbalanced distribution of labels to be the main issue. Based on the detailed result provided by the organizers, we note that most of the submitted system are not able to detect the less represented classes including *derailing* (29), *dominance* (49), and *stereotype & objectification* (137). Also classifying the target of misogyny (active and passive) has not been an easy task, which can be seen looking at the F_1-score of the result on official results.
Features including Swear Words Count, Swear Words Presence, Hashtag Presence, Link Presence, Sexist Slurs, and Woman-related Words outperformed all

[10] https://amiibereval2018.wordpress.com/important-dates/results/

RANK	SUBMISSIONS	F_1-SCORE
...	...	...
6	14-exlab.c.run5	0.369
8	14-exlab.c.run3	0.351
10	14-exlab.c.run4	0.343
12	14-exlab.c.run2	0.342
15	14-exlab.c.run1	0.338
...	...	...
16	AMI-BASELINE	0.337

Table 5. Task B rankings (English)

RANK	SUBMISSIONS	F_1-SCORE
1	14-exlab.c.run2	**0.446**
2	14-exlab.c.run3	0.445
3	14-exlab.c.run4	0.444
5	14-exlab.c.run1	0.441
...	...	...
14	AMI-BASELINE	0.410
...	...	...
20	14-exlab.c.run5	0.279

Table 6. Task B rankings (Spanish)

other submissions in English. In Spanish, the use of terms from the HurtLex lexicon, which were selected as related to gender-based hate, improves system performance in submission 3. However, not all the lexicon categories have been shown useful on this task, as indicated by the result of submission 4.

4 Discussion and Conclusion

In this paper we described the *14-ExLab@UniTO* submission for the *Automatic Misogyny Identification (AMI)* shared task at IberEval 2018. Our approach based on lexical knowledge was successful and our systems turned out to be the best-performing out of the ones participating in the Task A for both English and Spanish. We also introduced a novel hate-specific lexical resource which helped to improve the performance on the misogyny identification task.

For what concerns Task B, it was hard for all systems to classify misogynous tweets into the 5 categories proposed. After a manual inspection of the data, it emerged that there is no clear demarcation line between one category and the other and that the high presence of swearing in categories such as *dominance* and/or *discredit* just depends on the focus (e.g. the agent (man) vs. the wounded part, the target (woman)). At the same time, *stereotype & objectification* is not so conceptually distant from the *sexual harassment* category, due to a strong use of language referring to sexual body parts or vulgar sexual practices. Some examples from the English and Spanish datasets:

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): @[redacted] hey bitch wassup bitch suck my dick bitch
DISCREDIT (EN): @[redacted] 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 enseñarle a esta perra como se trata a un hombre" LMAO IN LOVE WITH EL TITI 😂
sexual harassment & threats of violence (ES): @[redacted] Me gustaría abrirte las piernas y clavarte toda mi polla en tu culo.
discredit (ES): Porque ladra tanto mi perra? La puta madre cállate un poco

We are planning to participate to the upcoming AMI shared task at EVALITA 2018, in order to validate our approach also for the Italian language.

Acknowledgments

V. Basile and V. Patti were partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01).

References

1. Erjavec, K., Kovačič, M.P.: "You don't understand, this is a new war!" Analysis of hate speech in news web sites' comments. Mass Communication and Society **15** (2012) 899–920
2. Manne, K.: Down Girl: The Logic of Misogyny. Oxford University Press (2017)
3. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. (2017) 1–10
4. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In: Proceedings of the NAACL student research workshop. (2016) 88–93
5. Sanguinetti, M., Poletto, F., Bosco, C., Patti, Stranisci, M.: An Italian Twitter Corpus of Hate Speech against Immigrants. In: Proc. of the 11th International Conference on Language Resources and Evaluation (LREC 2018), ELRA (2018)
6. Hewitt, S., Tiropanis, T., Bokhove, C.: The problem of identifying misogynist language on Twitter (and other online social spaces). In: Proceedings of the 8th ACM Conference on Web Science, ACM (2016) 333–335
7. Anzovino, M., Fersini, E., Rosso, P.: Automatic Identification and Classification of Misogynistic Language on Twitter. In: Proc. of the 23rd Int. Conf. on Applications of Natural Language & Information Systems, Springer (2018) 57–64
8. Fersini, E., Anzovino, M., Rosso, P.: Overview of the Task on Automatic Misogyny Identification at IberEval. In: Proc. of 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with SEPLN 2018), CEUR-WS.org (2018) 57–64
9. De Mauro, T.: Le parole per ferire. Internazionale (2016) 27 settembre 2016.
10. Bassignana, E.: HurtLex: Developing a multilingual computational lexicon of words to hurt (2018) Bachelor's thesis. Supervisor: V. Patti, Co-supervisor: V. Basile.
11. Fasoli, F., Carnaghi, A., Paladino, M.P.: Social acceptability of sexist derogatory and sexist objectifying slurs across contexts. Language Sciences **52** (2015) 98–107
12. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193** (2012) 217–250