CrotoneMilano for AMI at Evalita2018. A performant, cross-lingual misogyny detection system.

Angelo Basile
Symanto Research
angelo.basile@symanto.net

Chiara Rubagotti
Independent Researcher
chiara.rubagotti@gmail.com

Abstract

We present our systems for misogyny identification on Twitter, for Italian and English. The models are based on a Support Vector Machine and they use n-grams as features. Our solution is very simple and yet we achieve top results on Italian Tweets and excellent results on English Tweets. Furthermore, we experiment with a single model that works across languages by leveraging abstract features. We show that a single multilingual system yields performances comparable to two independently trained systems. We achieve accuracy results ranging from 45% to 85%. Our system is ranked first out of twelve submissions for sub-task B on Italian and second for sub-task A.

In questo articolo presentiamo i nostri modelli per il riconoscimento automatico di testi misogini su Twitter: abbiamo addestrato lo stesso sistema prima su un corpus italiano e poi su uno inglese. Il modello si basa su una macchina a vettori supporto e usa n-grammi come feature. La nostra soluzione è molto semplice e tuttavia ci permette di raggiungere lo stato dell'arte sull'italiano e ottimi risultati sull'inglese. Presentiamo inoltre un sistema che funziona con entrambe le lingue sfruttando una serie di feature astratte. Il nostro livello raggiunge livelli di accuratezza tra il 45% e l'85%: con questi risultati ci piazziamo primi nel task B per l'italiano e secondi nel task A.

1 Introduction

With awareness of violence against women growing in the public discourse and the spread of unfiltered and possibly anonymous communication on social media in our digital culture, the issue of misogyny online has become compelling. Violence against women has been described by the UN as a "Gender-based [..] form of discrimination that seriously inhibits women's ability to enjoy rights and freedoms on a basis of equality with men"[1]. On the web this often takes the form of female-discriminating attacks of different types, which undermine the women's rights of freedom of expression and participation[2]. Following erjavec2012you's understanding of *hate speech*, reported in (Pamungkas et al., 2018) 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, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics", we can define *misogynist speech* as any kind of aggressive discourse which targets women because they are women. Within the larger context of hate speech, online misogyny — or *cybersexism* — stands out as a large and complex phenomenon which reflects other forms of offline abuse on women (Poland, 2016). This holds true for the Italian case as well, where bouts of misogynistic tweets have been linked to episodes of femicides [3]. In recent years the NLP commu-

[1] http://www.un.org/womenwatch/daw/cedaw/recommendations/recomm.htm

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

[3] http://www.voxdiritti.it/wp-content/uploads//2018/06/mappa-intolleranza-3-donne.jpg

nity has addressed the issue of automatic detection of hate speech in general (Schmidt and Wiegand, 2017) and misogyny in particular (Anzovino et al., 2018). This effort to detect and contain verbal violence on social media (or any kind of text) demonstrates how NLP tools can also be used for ethically beneficial purposes and should be considered in the newborn and crucial discourse on Ethics in NLP (Hovy and Spruit, 2016; Hovy et al., 2017; Alfano et al., 2018). We are therefore proud to take up the AMI challenge (Fersini et al., 2018) and present our contribution to the cause of stopping misogynistic speech on Twitter. In this paper we propose a simple linear model using n-grams: we show that such a simple setup can still yield good results. We decided to propose a simple model for three reasons: first, it has been shown that linear SVM can easily outperform more complex deep neural networks (Plank, 2017; Medvedeva et al., 2017); second, training and testing our model does not require expensive hardware but a common laptop is enough to replicate our experiments; third, we experiment with a transformation of the input (i.e. we extract abstract features) and a linear model allows for an easier interpretation of the contribution of this transformation.

To summarise, the following are the contributions of this paper:

- We propose a simple and yet strong misogyny detection system for English and Italian (ranked first out of twelve systems for misogynistic category detection)
- We show how a single system can be trained to work across languages
- We release all the code[4] and our trained systems for reproducibility and for a quick implementation of language technology systems that can help detect and mitigate cybersexism phenomena.

Task Description The AMI task is combined binary and multi-label, short text classification task. Given a Tweet, we have to predict whether it contains or not misogyny (**Task A**) and if it does, we have to classify the misogynistic behaviour and predict who is the subject being targeted (**Task B**). The misogynistic behaviour's space consists of five different labels:

- `Stereotype & Objectification`
- `Dominance`
- `Derailing`
- `Sexual Harassment & Threats of Violence`
- `Discredit`

The target can be either `Active` when the message refers to a specific person or `Passive` when the message expresses generic misogyny. The setup is the same for both Italian and English.

2 Data

We use only the data released by the task organisers: they consist of Italian and English Tweets. The organisers report that the corpus has been manually labelled by several annotators. We provide an overview of the data set in Table 1. As it can be seen from the table, the data for Task A is more or less balanced, while the data for Task B is highly skewed.

3 Experiments

In this section we describe the feature extraction process and the model that we built.

3.1 Pre-processing

We decide not to pre-process the data in any way, since we do not have linguistic (or non-linguistic) reasons for doing so. To tokenize the text we simply split at every white space.

3.2 Model and Features

We built a sparse linear model for approaching this task.

We use n-grams extracted at the word level as well as at the character level. We use 3-10 n-grams and binary tf-idf. We feed these features to a Support Vector Machine (SVM) model with a linear kernel; we use the implementation included in `scikit-learn` (Pedregosa et al., 2011). Furthermore, we experiment with feature abstraction: we follow the bleaching approach recently proposed by (van der Goot et al., 2018). First, we transform each word in a list of symbols that 1) represents the shape of the individual characters

[4]The code can be found at https://github.com/anbasile/AMI/.

	ITALIAN					ENGLISH				
	SO	DO	DE	ST	DI	SO	DO	DE	ST	DI
Active	625	61	21	428	586	54	78	24	207	695
Passive	40	9	3	2	43	125	70	68	145	319
Non-Misogynous	2172					2215				

Table 1: Data Set Overview, showing the label distribution across the five misogynistic behaviours: Stereotype & Objectification (**SO**), Dominance (**DO**), Derailing (**DE**), Sexual Harassment & Threats of Violence (**ST**) and Discredit (**DI**).

	SHAPE	FREQ	LEN	ALPHA
This	Ccvc	46	04	True
is	vc	650	02	True
an	vc	116	02	True
example	vcvcccv	1	07	True
.	.	60	01	False
😎	😎	1	01	False

Table 2: An illustration of the bleaching process.

and 2) abstracts from meaning by still approximating the vowels and characters that compose the word; then, we compute the length of the word and its frequency (while taking care of padding the first one with a zero in order to avoid feature collision); finally, we use a Boolean label for explicitly distinguishing words from non-alphanumeric token (e.g. emojis). Table 2 shows an example of this feature abstraction process.

(van der Goot et al., 2018) proposed this bleaching approach for modelling gender across languages, by leveraging the language-independent nature of these features: here, we try to re-use the technique for classifying misogynist text across languages. We slightly modify the representation proposed by (van der Goot et al., 2018) by merging the shape feature (e.g. `Xxx`) with the vowel-consonant approximation feature (e.g. `CVC`) into one single feature (e.g. `Cvc`).

We propose three different multi-lingual experiments:

- `TRAIN` Italian → `TEST` English
- `TRAIN` English → `TEST` Italian
- `TRAIN` Ita. & Eng. → `TEST` Ita. & Eng.

For the last experiment, we use half the data set for each language. We report scores obtained by training on the whole training set and testing on the official test set, using the gold labels released by the organisers after the evaluation period.

4 Evaluation and Results

Since the data set labels for the sub-task B are not evenly distributed across the classes, we use f1-score to evaluate our model. First we report results obtained via a 10-fold cross-validation on the training set; then, we report results from the official test set, whose labels have been released. The official evaluation does not take into account the joint prediction of the labels, however here we report results considering the 0 label: since we train different models for the different label sets, we make sure that the models trained on Task B are able to detect if a message is misogynistic in the first place.

4.1 Development Results

We report the development results obtained by using different text representations. Table 3 presents an overview of these results. We note that all four representations — words, characters, a combination of these two and the bleached representation — all yield comparable results. The combination of words and characters seems to be the best format. Overall, we note that the system performs better on the Italian corpus than on the English corpus.

4.1.1 Cross-lingual Results

In Table 4 we present the results of our cross-lingual experiments. We train and test different systems using lexical and abstract features. We note that the abstract model trained on Italian outperforms the fully lexicalized model when tested on English, but the opposite is not true. The English data set seems particularly hard for both the abstract and the lexicalized model. Interestingly, the abstract model trained on both corpora shows good results.

	ENGLISH			ITALIAN		
	MIS.	CAT.	TGT.	MIS.	CAT.	TGT.
Words (W)	0.68	0.29	0.57	0.88	0.60	0.59
Chars (C)	0.71	0.30	0.61	0.88	0.59	0.58
W+C	0.70	0.31	0.59	0.88	0.62	0.59
Bleaching	0.68	0.27	0.57	0.85	0.55	0.56

Table 3: An overview of the development f1-macro scores obtained via cross-validation.

	TEST →	IT	EN
TRAIN	IT lex	0.85	0.51
	IT abs	0.83	0.52
	EN lex	0.47	0.62
	EN abs	0.45	0.52
	IT + EN lex	0.83	0.60
	IT + EN abs	0.81	0.58

Table 4: Pair-wise accuracy results for Task A. We compare lexicalized vs. abstract models. The combined IT+EN data set is built by randomly sampling 50% of instances from both corpora.

4.2 Test Results

In Table 5 we present official test results (Fersini et al., 2018). We submitted only one, constrained run; a run is considered *constrained* when only the data released by the organisers are used. We submitted the model using the combined representation with word- and character-ngrams, trained once on the English corpus and once on the Italian corpus. We achieve the top and the second position for the tasks B and A respectively on the Italian data set. On the English data set our system is ranked 15th and 4th on the tasks A and B respectively.

	TASK A	TASK B		
		CATEGORY	TARGET	AVG.
IT	0.843	0.579	0.423	0.501
EN	0.617	0.293	0.444	0.369

Table 5: Official test results. Task A is measured using accuracy and Task B is measured using f1-score. We reach the first position on Task B for Italian.

5 Discussion and Conclusions

A warning to the reader: this section contains explicit language. In the attempt to understand better the big difference in performance between the English and Italian models, we show the importance of words as learned by the model: we print the ten most important words, ranked by their learned weights. The result is shown in Table 6. From the output we see that the model trained on Italian learned meaningful words for identifying a misogynist message, such as *zitta* [shut up], *tua* [your] and *muori* [die!]: these words stand out from the rest of the profanity for directly referring to someone, while the rest of the words and almost all the most important English words could be used as interjections or could be more generic insults.

RANK	ITA	ENG
1	zitta	woman
2	bel	hoe
3	pompinara	she
4	puttanona	hoes
5	tua	women
6	muori	whore
7	baldracca	her
8	troie	bitches
9	culona	womensuck
10	tettona	bitch

Table 6: Top ten words ranked by their positive weights learned during training.

The results of the abstract system are satisfactory for eventually building a light, portable model that could be adapted to different language. In the future we will try training on English and Italian and testing on a third corpus (such as the Spanish version of the AMI data set).

In this paper we described our participation to the AMI - Automatic Misogyny Identification for Italian and English. We proposed a very simple solution that can be implemented quickly and we scored a state-of-the-art result for classification of misogynistic behaviours in five classes.

Acknowledgements

The authors would like to thank the two anonymous reviewers who helped improve the quality of this paper. The first author has conducted this research as he was still part of the Erasmus Mundus master in Language and Communication Technology, a shared master program between the University of Groningen (NL) and the University of Malta (MT).

References

Mark Alfano, Dirk Hovy, Margaret Mitchell, and Michael Strube. 2018. Proceedings of the second acl workshop on ethics in natural language processing. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*.

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

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. 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.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 591–598.

Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M Bender, Michael Strube, and Hanna Wallach. 2017. Proceedings of the first acl workshop on ethics in natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*.

Maria Medvedeva, Martin Kroon, and Barbara Plank. 2017. When sparse traditional models outperform dense neural networks: the curious case of discriminating between similar languages. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 156–163.

Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, and Viviana Patti. 2018. 14-exlab@unito for ami at ibereval2018: Exploiting lexical knowledge for detecting misogyny in english and spanish tweets. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, pages 234–241.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Barbara Plank. 2017. All-in-1 at ijcnlp-2017 task 4: Short text classification with one model for all languages. *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 143–148.

Bailey Poland. 2016. *Haters: Harassment, Abuse, and Violence Online*. University of Nebraska Press.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Rob van der Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank. 2018. Bleaching text: Abstract features for cross-lingual gender prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 383–389.