AMI at IberEval2018 Automatic Misogyny Identification in Spanish and English Tweets

Victor Nina-Alcocer

Department of Computer Systems and Computation
Universitat Politècnica de València, Spain
vicnial@inf.upv.es

Abstract. In this paper we describe the submission for the *Automatic Misogyny Identification in Spanish and English Tweets* shared task organized at IberEval[1]. This work proposes an approach based on weights of ngrams, word categories, structural information and lexical analysis to discover whether these components allow us to discriminate between misogynous and no misogynous tweets and their respective categories and targets in case of misogynous tweets. Moreover, we analyze the use of some features created by these components to investigate their impact.

1 Introduction

AMI is the first task on automatic misogyny identification [2]. Its aim was to identify cases of aggressiveness and hate speech towards women in social media [1]. Poland's work [3] was the first attempt to manually classify misogynous tweets. Now this shared task will consider two subtasks for this classification:

- subtask1: Misogyny identification.
- subtask2a: Misogynistic Behaviour.
- subtask2b: Target Classification.

The aim of *subtask1* is to identify whether a tweet is misogynous or not, and the second *subtask2a* aims to identify the category, if a misogynous tweet belongs to: discredit, dominance, sexual_harassment, stereotype, and derailing. Finally, *subtask2b* is in charge to identify whether a misogynous tweet is active or passive i.e. if its target is generic (women in general) or individual. In this work, each of these tasks is approached as a classification task. We will use natural language processing (NLP), machine learning and feature engineering to identify patterns and learn classification models respectively.

2 Approach

This section tries to describe the main approaches that have been used. Generally, misogyny can be expressed written, orally, in a subtle or explicit way, also

[1] https://sites.google.com/view/ibereval-2018

directly or indirectly addressed to someone. In order to investigate how people may express misogyny in tweets, we propose an approach that allows us to discover some aspects about how misogyny is expressed in the corpus provided by the organizers. Hence this approach takes into account some features that we considered important in order to understand if some of them contribute to recognizing misogynous content and its respective category.

Structure (str): Basically, knowing how many words are used in a tweet or if most of those words are written in capital letters, even if some of them use excessively punctuation marks could reveal important information. As we know a tweet is composed of words, punctuations, mentions, URLs, etc. In this approach, we will pay attention to these aspects to see if all of them in some way help to better discriminate between misogynous tweets and not misogynous one. A summary of these features is given below:

- The number of symbols or punctuation marks (!'?,.").
- The number of words written in capital letters.
- The number of words and characters, including stop words.
- Mean of the numbers of words and characters.
- The number of mentions, URLs, and hash-tags.

LIWC categories (lc): Another component that we consider important is the possibility to get features from Linguistic Inquiry and Word Count (LIWC) [2]. We have just taken into account some categories related to misogynous emotions such as: angry, sexual, swear, positive, negative, etc. [4] The idea behind this component is to calculate for instance the percentage of positive or negative emotions, or even if a tweet has sexual content as we can see in Figure 1.

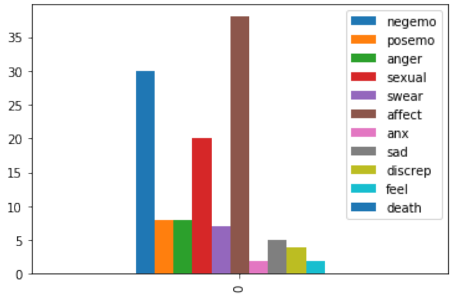

Fig. 1. The content (words) of a tweet belongs to some category: death, anger, etc.

Ngrams (ng): In this component Term frequency - Inverse document frequency based on Words (TFIDFW) or Chars (TFIDFC) schemes are used. For instance in misogynous TFIDFW (see Table 1) the term *bitch* (first place) is more

[2] https://www.receptiviti.ai/liwc-api-get-started

used among the misogynous tweets than among the ones that are not misogynous (fourth place), e.g. in our case the uni-gram *bitch* has a different weight, it means that this word has a specific weight in a misogynous tweet and has another weight in a non misogynous one. The same logic is followed for subtask2a and subtask2b using TFIDFW of their categories and targets respectively.

Table 1. Weight of uni-grams and bi-grams

	uni-grams						**bi-grams**				
	misogynous			**no-misogynous**			**misogynous**			**no-misogynous**	
N	term	weights	N	term	weights	N	term	weights	N	term	weights
1	**bitch**	**0.054913**	1	rape	0.021782	1	stupid bitch	0.010204	1	stupid cunt	0.006159
2	dick	0.027398	2	dick	0.019902	2	ass bitch	0.006658	2	son bitch	0.002429
3	stupid	0.024436	3	cunt	0.019422	3	suck dick	0.004807	3	men rights	0.002079
4	like	0.024388	4	**bitch**	**0.018755**						
5	woman	0.023752	5	hoe	0.017120						

Part of Speech (pos): The last component of our approach takes into account part of speech information, which has the task of tagging each word in a sentence with its appropriate part of speech. We decide whether each word is a noun, adjective, verbs, etc. Using this component we can identify some patterns, for instance in our corpus some nouns are followed by punctuation marks *e.g. bitch!!!!!!*.

3 Experiments and Results

Thanks to the organizers we count with a dataset of 3307 Spanish and 3251 English tweets respectively. Each tweet is labeled as misogynous (1) or no-misogynous (0) and both datasets are balanced. Regarding the type of misogyny and target, each tweet is labeled as: discredit, dominance, sexual_harassment, stereotype, derailing and active or passive in case of the target. With respect to the *category* and *target* information, the corpus is unbalanced. The first one is biased in favor of discredit (60%) and regarding the *target* is biased in favor of active (almost 75%). Moreover, to evaluate our system, a test dataset with 831 and 726 unlabeled tweets in Spanish and English respectively was provided.

For the experiments, we employed a set of feature combinations which has been used to feed some classifiers: Support Vector Machine (SVM), Multi-layer Perceptron (MLP) and MultinomialNB (MNB).
SVM and 10 K-fold cross-validation were used. The first one was chosen because its performance was good enough with thousands of features, and the second one allows us to avoid over-fitting in all the experiments.

Firstly, the main goal was to face the classification of misogynous tweets in Spanish in order to apply the best performing approach to the rest of subtasks in English or Spanish.

Table 2. Configuration of the main experiments

Name	**Set up**
ap1.	TFIDFW + TFIDFC + BOW + BOC
ap2.	SVD30(TFIDFW) + SVD30(TFIDFC) + BOW
ap3.	MNB(PREDICTED) + SVD20(TFIDFW) + str + lc
ap4.	BOW + str + lc + ng + pos
ap5.	TFIDFW + str + lc + ng + pos

Table 2 shows how the experiments were set up. Approaches *ap1* and *ap2* had the aim to find out whether features created by TFIDFW, TFIDFC, Bag of word-grams (BOW) or Bag of char-grams (BOC) are useful. *ap1* uses the whole group of features (thousands of them) created by TFIDFW or TFIDFC, while *ap2* obtains the 60 best features using truncated singular value decomposition (SVD) on TFIDFW and TFIDFC then combines with BOW. Unfortunately, those approaches were interesting but we did not obtain results over our baselines with any of the classifiers (MLP, SVM, MNB). *ap3* tries to reduce the number of features: firstly we classified a tweet using MNB and then we obtained their respective probabilities to use them as features (2), additionally we got the best 20 features using SVD on TFIDFW and lastly, we added the features *str* (5) and *lc* (10). Unfortunately, with these 37 features we did not achieve results over our baselines in subtask1 and subtask2ab respectively.

Now we proceed to analyze the results that we got with the approach proposed in Section 2. *ap4* and *ap5* follow the same logic, but ap5 obtains better results than ap4 because it uses TFIDFW. Tables 3 and 4 show the best values that we achieved: run4 in Table 3 uses *TFIDFW* plus *structure*, *category* and *weight of ngrams(unigrams+bigrams+trigrams)* as features and we obtained 0.782 of accuracy applying linear SVM on subtask1 . While with respect to the subtask2a, we added *part of speech* as feature and we obtained 0.370 of F1-macro.

Table 3. Results with ap5 on English training tweets

run	subtask1		subtask2a		subtask2b
		Accuracy		F1-macro	F1-macro
run1→SVM on TFIDFW	+str+lc	0.733	+pos	0.299	0.721
run2→SVM on TFIDFW	+str+lc+ng(u)	0.781	+pos	0.302	0.762
run3→SVM on TFIDFW	+str+lc+ng(u+b)	0.781	+pos	0.343	0.763
run4→SVM on TFIDFW	**+str+lc+ng(u+b+t)**	**0.782**	**+pos**	**0.370**	**0.764**

Table 4. Results with ap5 on Spanish training tweets

run	subtask1		subtask2a		subtask2b	
		Accuracy		F1-macro		F1-macro
run1→SVM on TFIDFW	+str+lc	0.804		0.472	-str	0.781
run2→SVM on TFIDFW	**+str+lc+ng(b)**	**0.860**	**-lc**	**0.503**	**-str-ng(b)**	**0.780**

Looking at Table 4, we may observe that in run2 we obtained 0.780 of F1-macro in subtask2b just using *lc* as feature. Also, that just using *str* and *ng(bigram)* we obtained 0.503 of F1-macro on subtask2a.

3.1 Official ranking

We did not expect good results in English (see Table 5), but we obtained scores slightly above the average macro F1-baseline (0.3374) in subtask2a and subtask2b (see run3 and run4). While in subtask1 we were below the accuracy baseline (0.7837). These results can be due to a bad combination of our features.

Table 5. Official results for English subtask1, subtask2a and subtask2b

	subtask1			subtask2ab
Rank	**Run**	**Accuracy**	**Rank**	**Average F1-macro**
16	Our approach.run2	0.7809	17	0.336433966
17	Our approach.run3	0.7809	14	0.33914113
18	Our approach.run4	0.7809	13	0.339590051
26	Our approach.run1	0.7094	23	0.316368399

Table 6 shows the better results we obtained in Spanish (between the first five teams). However, we think that classifying misogynous tweets in this corpus was quite difficult because the performance of the teams was approximately 80% in terms of accuracy. Similarly, in subtask2a and subtask2b, mostly the teams were not far from the baseline.

Table 6. Official results for Spanish subtask1, subtask2a and subtask2b

	subtask1			subtask2ab
Rank	**Run**	**Accuracy**	**Rank**	**Average Macro F1**
9	Our approach.run1	0.805054152	8	0.42722476
20	Our approach.run2	0.76654633	13	0.41174962
22	Our approach.run3	0.65944645	21	0.27271983

4 Conclusions

In this work, we proposed an approach that takes into account some aspects: weights of ngrams, LIWC categories, structural information and lexical analysis. We observed that each aspect contributes in some way to the different subtasks. Moreover, we notice that the four aspects contributed to obtaining a better accuracy and F1-macro in the corpus of English tweets. However, only the first three aspects were useful for the Spanish tweets.
As future work, it is interesting to use some techniques to face unbalanced dataset and explore other features. Moreover, we plan to use deep learning to see what performance this technique could achieve.

References

1. Bailey, M.: Haters: Harassment, abuse, and violence online by bailey poland. Signs: Journal of Women in Culture and Society **43**(2), 495–497 (2018). https://doi.org/10.1086/693771
2. Fersini, E., Anzovino, M., Rosso, P.: Overview of the task on automatic misogyny identification at ibereval. In: 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) CEUR Workshop Proceedings. Seville, Spain, September 18, 2018, CEUR-WS.org
3. 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. pp. 333–335. WebSci '16 (2016), http://doi.acm.org/10.1145/2908131.2908183
4. Poland, B.: Haters: Harassment, Abuse, and Violence Online. University of Nebraska Press (2016), http://www.jstor.org/stable/j.ctt1fq9wdp