Misogyny Detection and Classification in English Tweets: The Experience of the ITT Team

Elena Shushkevich
Social Media Research Group
Institute of Technology Tallaght
Dublin, Ireland
e.shushkevich@yandex.ru

John Cardiff
Social Media Research Group
Institute of Technology Tallaght
Dublin, Ireland
john.cardiff@it-tallaght.ie

Abstract

English. The problem of online misogyny and women-based offending has become increasingly widespread, and the automatic detection of such messages is an urgent priority. In this paper, we present an approach based on an ensemble of Logistic Regression, Support Vector Machines, and Naïve Bayes models for the detection of misogyny in texts extracted from the Twitter platform. Our method has been presented in the framework of the participation in the Automatic Misogyny Identification (AMI) Shared Task in the EVALITA 2018 evaluation campaign.

Italiano. *Il problema della misoginia online e dell'odio diretto verso le donne si sta diffondendo sempre più, e così il riconoscimento automatico di tali messaggi è una priorità importante.*
In questo articolo, presentiamo un approccio basato sui classificatori Logistic Regression, SVM e Naive Bayes per il riconoscimento automatico della misoginia in testi estratti da Twitter. Il nostro metodo è stato presentato attraverso la nostra partecipazione allo shared task AMI presso la campagna di valutazione EVALITA 2018.

1 Introduction

It is hard to miss the fact that an intensive growth of social networking has led not only to the rise of personal communication opportunities, but also to an increase in aggression on social media. Hate speech can be aimed at sexual orientation, race, religion as gender as a whole. In particular, when the target of hate speech is women, we could say that this is misogyny. Nowadays, more and more attention is paid to this problem, and one of the directions for the hate speech recognition is the women-oriented aggression detection in social networks.

It is important to work with hate speech and misogyny detection now, because over the course of time the data from social networks will grow and this problem will become more and more serious. It is necessary to create a range of systems which allow us to detect and control the number of hate speech messages, and we need to understand how to classify this type of information and how we could reduce the number of it. So, it is a big challenge to find the way of misogyny data detection and processing.

This paper describes our participation in the Automatic Misogyny Identification (AMI) Shared Task, in EVALITA 2018 (Fersini, Nozza and Rosso, 2018). The aim of the task is to identify misogynistic text in tweets. The task contained two different subtasks:
Subtask A - Misogyny Identification: the main goal of the task was to separate misogynous tweets from non-misogynous.
Subtask B - Misogynistic Behavior and Target Classification: the idea of the target classification was to define misogynous tweet which offends a specific person (Active) and tweets which insult a group of people (Passive).
Misogynistic behavior task was intended to divide misogynous tweets into different groups:
- Stereotype & Objectification: a widely held but fixed and oversimplified image or idea of a woman, description of women's physical and/or comparisons to narrow standards.
- Dominance: to assert the superiority of men over women or to highlight gender inequality.
- Derailing: to justify abuse of women, rejecting male responsibility and an attempt to disrupt the conversation in order to redirect women's conversations on something more comfortable for men.
- Sexual Harassment & Threats of Violence: to describe actions as sexual advances, requests for sexual favours, harassment of a sexual nature, intent to physically assert power over women through threats of violence.
- Discredit: slurring of women with no other larger intention.

There were two datasets for the task, one of which contained tweets in the English language and another containing Italian tweets. Our team worked with English dataset only. The English dataset was composed of 4,000 tweets for training and 1,000 tweets for testing. The results were evaluated using the accuracy performance for Task A and macro F-measure performance for Task B.

This paper presents our approach to solve the above problems. The main thrust of our approach is to build a model that allows us to assess the classification of any tweet to its assigned group.

The paper is organized as follows. Some relevant related works in the area are described in Section 2. Section 3 presents the way we conducted data preprocessing and the approach we chose for building the desired model. In Section 4 the results are described and analyzed. In Section 5 we summarize our work.

2 Related work

There are a number of approaches in the area of text processing by machine learning methods which allow us to deal with misogyny and harassment in texts. Some of these were presented in the AMI@IBEREVAL-2018 shared task (Fersini, Anzovino and Rosso, 2018). The aim of this challenge was to detect misogynistic tweets and to create the model which was able to classify misogynistic tweets for different groups depending on the type of misogyny. In particular, it was demonstrated that, using models based on Support Vector Machines (Pamungkas et al., 2018) and ensembles of models (Frenda et al., 2018), it is possible and quite successful in cases where the aim is to make a classification of tweets for different types and functions of misogyny. In our work we apply several of the same techniques - Support Vectors Machines and ensembles of models - to the task of misogyny tweets detection.

Some works which could help us to understand the way to hate speech messages classification were published in recent years. In (Schmidt and Wiegand, 2017) the authors demonstrated methodologies of hate speech data processing. In another work (Waseem and Hovy, 2016) there were presented useful approaches to detect racial and sexist offenses. It should be noted that there was a classification for 3 different groups (hate speech, derogatory, profanity) with the understanding that hate speech is a kind of abusive language.

In the research reported in (Nobata et al., 2016), it was shown (Bartlett et al., 2014) how to use NLP to analyse English-language misogynistic tweets to find the frequencies of abusive words and the users who used this type of words more often. In other works (Alexandrov et al., 2013; Kaurova et al., 2010) the authors focused on creating models which could allow the evaluation of the tone of the text on a scale from very negative to very positive. They constructed a model for the groups of 3, 5 and 8 different categories and were able to achieve the results with a high accuracy using additional tools like GMDH Shell and Semantic Orientation Calculation (So-CAL), which demonstrates the very high potential of using inductive modelling for text-mining tasks. We are planning to use techniques which were mentioned above to improve the results of our model in future.

3 System

In our approach we perform a number of sequential actions including preprocessing, model design, and finally embedding the constructed models in one ensemble.

3.1 Preprocessing

In the first step, we prepared the data for the classification. To clean the data we removed the string punctuation and converted words to lower case. For the vectorization we used the tf-idf (term frequency–inverse document frequency) method which allows us to reduce the weight of frequently occurring in many documents words and to increase the weight of frequently occurring words in the documents. These were carried out for the first run. For the subsequent two runs, we added some extra preprocessing steps:

- the replacement of all links with the string "URL"
- the replacement of all references to Twitter users (i.e, terms starting with the "@" symbol) with the term "USER".
- we marked some combinations of symbols which were used often in messages such as "!!! ", "??? " and other emotional expressions, and replaced them with the term "emoji".

3.2 Models

The main idea of the modeling was to create an ensemble of different models which could complement each other to achieve the best results. The final blended model assigns the tweet to a specific class by majority voting. We used a number of simple models which include:

- Logistic regression model. Logistic regression involves the construction of a discriminant model, which calculates the probability from a function of a weighted set of observation features and assigns a class to each observation. The classifier based on logistic regression applies an exponential function to a linear combination of objects obtained from the input data (Wang et al., 2012; Wright, 1995).

- Support Vector Machines classifier. As it was shown in (Joachims et al., 2002), this method is very useful in work with texts. The idea of this method is to translate the source vectors into a higher dimension space and search for such a separating hyperplane so that the gap in this space is maximal. There are two parallel hyperplanes on both sides of the hyperplane that are constructed to separate the classes, and one hyperplane that will maximize the distance to two parallel ones is sought.

- Naive Bayes classifier. One of the advantages of this method is the high speed of calculations (Zhang and Di Li, 2007), and another one is the number of the data which is needed to train the model - in this case it is not necessary to have a big training dataset to achieve a high level of classification parameter estimation.

In the next step we combined the Naive Bayes approach and Logistic regression approach in one model, as presented in the work

(Genkin et al., 2007),which produced quite good results.

In the final step we combined the models we have mentioned, Logistic regression (LR), Support Vector Machines (SVM), Naive Bayes and Logistic Regression (NB+LR), into one ensemble. In this blended model the probabilities of belonging to different classes from the simple models were summed and averaged. We marked as a final choice the class which had the highest average probability.

4 Results

We chose three different runs for the evaluation: one of them was implemented by using the simplest type of preprocessing (we just deleted punctuation symbols and changed all letters to the low case) and this variant supposed that we marked a tweet as misogynistic one in case that two of three types of classification marked this tweet as misogynous (Misogyny+Target or Misogyny+Misogynistic Behavior or Target+Misogynistic Behavior).

In the next step, we carried out a more intricate preprocessing as described in Section 3.1 and applied the type of tweets labeling such a way as we detected a tweet as misogynistic each time when at least one classifier worked.

The last run was implemented by using the most complicated preprocessing and the type of tweets labeling such as at the first run.

Table 1 shows the results of all three classification types. As can be seen, the fourth type of selection was the most successful. It could be concluded that the blended model which contained more simple models (Logistic Regression, Naive Bayes + Logistic Regression and Support Vector Machines) allows us to achieve the best results for all classification types: Misogyny Identification, Target Classification and Misogynistic Behavior classification.

It should be noted that we used the F-Measure for the results' evaluation because this assessment allows bringing together both recall and precision and because of the imbalance within both the Misogynistic Category Classification and the Target Classification.

Task	Classifier	F1-score
Misogyny Identification	LR	0.78
	NB+LR	0.72
	SVM	0.71
	Blend	**0.78**
Target Classification	LR	0.60
	NB+LR	0.66
	SVM	0.76
	Blend	**0.76**
Misogynistic Behavior	LR	0.50
	NB+LR	0.52
	SVM	0.57
	Blend	**0.64**

Table 1.Performance on the validation set.

Also note that the results of our model increase when the number of different classes decreases, thus an efficiency of the blended model is reduced from the Misogyny Identification classification results to the Misogynistic Behavior classification ones.

The results of all 3 runs for the blended model with the testing dataset are presented in Table 2.

Subtask A - English		
Rank	Team	Accuracy
8	ITT.c.run2.tsv	0.638
9	ITT.c.run3.tsv	0.636
10	ITT.c.run1.tsv	0.636

Table 2. Results of the classification.

It can be concluded by the results on the test data, the best run is the one with the most complicated preprocessing and the type of labelling, when we mark tweet as misogynistic every time when at least one of classifiers worked.

5 Conclusion

A negative aspect of the increased usage of platforms like Twitter is that incidents of aggression and related activities like harassment and misogyny have increased significantly. Nowadays it is an urgent problem to deal with such type of text information and messages, and there are a lot of challenges that have a connection with this task. In this article we have described our approach to misogyny detection and classification of tweets. The method was presented for evaluation in the framework of the Automatic Misogyny Identification (AMI) Shared Task at EVALITA 2018. We built an ensemble of models that includes Logistic regression, Naive Bayes and Support Vector Machines approaches, which classified the data taking into account the probabilities of belonging to classes calculated by simpler models. It was shown that it is possible to achieve quite good results using the final blended model and our model showed the best results for the binary classification of misogynistic tweets and non-misogynistic ones.

We observed preprocessing to be a very important part of the data handling and it has a high impact on the results of all models. From our results it could be concluded that the highest accuracy has been produced with maximum additional work at the preprocessing stage. It was important to pay attention to the replacement of links and references with special symbols, because the run with this type of alteration demonstrated the best results. Also, the best type of labelling misogynistic tweets was to mark the message as misogyny if any one of the type of classification worked. At first we had an idea that it could be more reliably if we mark tweet when 2 of 3 classifications mark it, but the real results disproved that hypothesis. We are currently investigating the addition of more features and models for the blended model to improve our results in the future.

References

Alexandrov M., Danilova V., Koshulko A., Tejada, J. 2013. *Models for opinion classication of blogs taken from Peruvian Facebook.* Proceedings of 4th International Conference on Inductive Modeling (ICIM-2013), pp. 241–246 .

Bartlett J., Norrie R., Patel S., Rumpel R., Wibberley S. 2014. *Misogyny on twitter*, http://www.demos.co.uk/, 05.

Fersini, E., Anzovino, M., Rosso. P. 2018. *Overview of the Task on Automatic Misogyny Identification at IberEval*. 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. CEUR-WS.org

Fersini E., Nozza D., Rosso P. 2018. *Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI).* Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18). Caselli, Tommaso and Novielli, Nicole and Patti, Viviana and Rosso, Paolo CEUR.org, Turin, Italy

Frenda S., Ghanem B. 2018. *Montes-y-Gómez M. Exploration of Misogyny in Spanish and English tweets.* CEUR Workshop Proceedings. CEUR-WS.org.

Genkin A., Lewis D., Madigan D. 2007. *Large-scale bayesian logistic regression for text categorization*. Technometrics, 49(3):291–304.

Joachims, T. 2002. *Learning to classify text using support vector machines: Methods, theory and algorithms.* Kluwer Academic Publishers.

Kaurova O., Alexandrov M., Ponomareva N. 2010. *The Study of Sentiment Word Granularity for Opinion Analysis (a Comparison with Maite Taboada Works*). International Journal on Social Media. MMM: Monitoring, Measurement, and Mining 1(1), 45–57.

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, pp. 145–153. International World Wide Web Conferences Steering Committee.

Pamungkas E.W., Cignarella A.T., Basile V., Patti V. 2018. *14-ExLab@UniTo for AMI at IberEval2018: Exploiting Lexical Knowledge for Detecting Misogyny in English and Spanish Tweets.* CEUR Workshop Proceedings. CEUR-WS.org.

Schmidt, A., Wiegand, M. 2017. *A survey on hate speech detection using natural language processing*. Proceedings of the Fifth International

Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics, Valencia, Spain, pp. 1–10.

Shushkevich E., Cardiff J. 2018. *Classifying Misogynistic Tweets Using a Blended Model: The AMI Shared Task in IBEREVAL 2018*. CEUR Workshop Proceedings. CEUR-WS.org.

Wang S., Manning C.D. 2012. *Baselines and bigrams: simple, good sentiment and topic classification.* Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, ACL 2012, vol. 2, pp. 90–94.

Waseem, Z., Hovy, D. 2016. *Hateful symbols or hateful people? predictive features for hate speech detection on Twitter.* SRW@ HLT-NAACL, pp. 88–93.

Wright R. 1995. *Logistic regression*. L.C. Grimm & P.R. Yarnold (Eds.) Reading and understanding multivariate statistics. Washington, DC: American Psychological Association, 217-244

Zhang H. and Di Li. 2007. *Naıve bayes text classifier.* Granular Computing. GRC 2007. IEEE International Conference on, pages 708–708. IEEE.