ISSN 2007-9737

Automatic Misogyny Detection in Social Media: A Survey

Elena Shushkevich, John Cardiff

Technological University, Dublin,
Ireland

e.shushkevich@yandex.ru, john.cardiff@tudublin.ie

Abstract. This article presents a survey of automated misogyny identification techniques in social media, especially in Twitter. This problem is urgent because of the high speed at which messages on social platforms grow and the widespread use of offensive language (including misogynistic language) in them. In this article we survey approaches proposed in the literature to solve the problem of misogynistic message recognition. These include classical machine learning models like Support Vector Machines, Naive Bayes, Logistic Regression and ensembles of different classical machine learning models, as well as deep neural networks such as Long Short-term memory and Convolutional Neural Networks. We consider results of experiments with these models in different languages: English, Spanish and Italian tweets. The survey describes some features, which help to identify misogynistic tweets and some challenges, which aim was to create misogyny language classifiers. The survey includes not only models, which help to identify misogyny language, but also systems which help to recognize a target of an offense (an individual or a group of persons).

Keywords. Twitter, misogyny detection, machine learning, deep neural networks.

1 Introduction

Nowadays, social networks are becoming an integral part of everyone's life. Along with the expansion of the social networks influence, the number of problems associated with them increases dramatically, including the volume of the offensive language in social media increase.

The challenge of the identifying the various types of hate speech (e.g., racism, aggression, misogyny) in social media is very complicated because of the different meaning of slurs and offenses depend on their context, a gender and type of users and another factors how it was shown in [1,2].

According to [3], it is possible to identify the main features that help to identify hate speech in social media (including, e.g., slurs in messages, problematic hashtags or criticisms of minorities without a well-founded argument). However, it should be noted that different types of offensive messages have different characteristics, for example, the authors of [4] demonstrated that sexist messages are more interactive and more attitudinal than racist ones.

In this article we describe the different approaches for solving the problem of recognizing automatically a specific category of hate speech recognition: misogynistic messages in social networks. We concentrate on the datasets from Twitter, because it is one of the largest and the most popular social media platform, where people have an opportunity to share their opinions.

The paper is organized as follows. Some relevant information about the problem of misogyny detection is described in Section 2. Section 3 presents the main principal approaches to deal with the task of misogynistic message identification. In Section 4, some shared tasks in this area are described and analyzed. In Section 5, we present our conclusions. The problem of misogyny detection in Twitter Misogyny is one of the most urgent problems of the modern world, so the problem of misogyny detection should be identified as a separate group of hate speech.

This group of offensive messages has its own characteristics and patterns. It includes many aspects, such as sexual harassment, the stereotypes associated with "stupid" women's behavior against male, objectification of the female body and a lot of other problems. Also note that

doi: 10.13053/CyS-23-4-3299

despite the fact that the problem of misogyny detection in social media is quite new, the researchers have already enough data to speak about the linguistic features of misogynistic messages. It is necessary to create systems that are able to recognize this kind of aggressive language and to prevent its spread.

Twitter is one of the most popular social media platforms, and therefore the study of its content and the misogyny detection in the messages on this platform is a serious problem. As described in [5], distinctive features of Twitter messages are the high speed of propagation of tweets, and the fact that some of the messages, including misogynistic ones, can stay there forever because of the possibility of retweeting.

Additionally, the problem is that users prefer to create their own social networks, so the misogynistic tweet can be amplified many times by the support and reposting of the user's 'friends'.

It should be noted that in the case of misogyny detection in Twitter it is possible to use some meta-data from users like geolocations, information which users posted at their profiles and links to external platforms (video and audio) from tweets.

2 Principal Approaches for Misogyny Detection

There are two principal approaches for detection offensive language and misogyny in particular in social media. One of them is based on the models created using classical machine learning approach. In [6], a model based on Support Vector Machines and K-fold cross-validation is presented , which achieves quite good result in the task of misogyny identification in Twitter.

The authors of [7] created the ensemble of models using Logistic Regression, Naive Bayes and Support Vector Machines and achieved high results worked on the problem of offensive language connected with African-American people, plus-sized people and misogyny. Authors used data from the communities of active support and groups of haters for each target and achieved good results for all three classifiers (the best results for 'black' with Support Vector Machines: 0.81 accuracy, Logistic Regression for 'plus' target with 0.79 accuracy, Logistic Regression for 'female' target with 0.81 accuracy).

Another approach involves the use of neural networks, and these models show competitive results too. The authors of [8] took standard Convolutional Neural Networks (CNNs) with three convolutional layers and added a new GRU layer. GRU can be compared to the popular Long Short-term memory networks (LSTMs), but these have three gates (input, output and forget gates), whereas GRU have only two gates (reset and update gates). This simpler structure allows us to train and generalize better on small datasets and they achieved very high results in case of classification of Twitter data for three different groups: sexist posts, racist messages and neither sexist nor racist tweets.

In the work [9], the authors presented a model named HybridCNN , which combined a character-level and a word-level convolutional networks and it achieved the F1-score equals 0.827 in the task of classification tweets for sexist, racist and neutral ones. Also, the authors of [10] presented the model based on Long Short-Term Memory Networks with which they achieved 0.930 of F1-score with the same type of tweet classification.

Another interesting approach based on Neural Networks was shown in [11]. The authors used a Recurrent Neural Networks (RNN) and its subclass Bi-LSTM (bidirectional Long Short-term Memory Networks) for misogyny language identification. The main difference between the bidirectional LSTM and the classic one is the fact that the latter creates a representation of each word of the sentence using only the left context, and Bi-LSTM using right representation too, so this representation is more complete.

The authors of [12] created models based on Lineal Support Vector Machines, Random Forest, Naive Bayes and Multi-layer Perception Neural Network to classify misogynistic tweets in two ways: the first experiment was a binary classification and the second one was a multi-classification for five different types of misogyny in Twitter. The result shown by Support Vector Machine was the best for the binary classification with 0.7739 of macro F1-score and in case of multi-classification the best was Multi-layer Perception Neural Network with 0.3697 of macro F1-score.

3 Shared Tasks Focusing on Automated Misogyny Detection

One of the interesting approaches dedicated to the problem of misogyny detection in social media was Automatic Misogyny Identification (AMI) shared tasks. Two such tasks were held in 2018, at the IberEval 2018[1] and Evalita 2018[2] evaluation campaigns. The aim of these tasks was to indicate misogyny behavior in tweets, and the task contained two different subtasks:

- Subtask A - Misogyny Identification: to separate misogynistic tweets from non-misogynous using binary classification.
- Subtask B - Misogynistic Behavior and Target Classification.

The main idea of the target classification was to identify tweets in which a misogynous tweet offends a specific person or group of people. It was binary classification tasks for two groups: one of which included texts with active (or individual) offenses which were sent to a specific person, and the other consisted of texts with passive (or generic) offenses, which were posted with the aim not to offend a specific person, but a group of people.

The misogynistic behavior task was intended to divide misogynous tweets to different groups, included: Stereotype & Objectification (a description of women's physical and/or comparisons to narrow standards), Dominance (an assertion of the superiority of men over women), Derailing (abuse of a woman), Sexual Harassment & Treats of Violence (actions as sexual advances, requests for sexual favors, harassment), Discredit (slurring over women with no other larger intention).

3.1 AMI@IBERALEVAL 2018

There were two datasets for this Task: one of them contained tweets in English language and another one was in Spanish language. The English dataset was composed of 3,251 tweets for training and 726 tweets for testing, and for the Spanish corpus there were 3,307 tweets in training dataset and 831 in testing dataset. Table 1 presents achieved by the top 3 participating teams, showing the results for the English dataset for Subtask A.

Table 1. The best results for the Subtask A (IberEval)

Team	Accuracy
14-exlab	0.913
SB	0.902
AnotherOne	0.871

For the Task A the best results were achieved by 14-exlab [13] team. The team used SVM models: with Radial Basis Function (RBF) kernel for the English dataset and SVM with a linear kernel for the Spanish dataset. AnotherOne used just SVM for modelling.

Also, the ensemble of models presented by ITT team [14] showed quite good results in case of misogyny detection on the Twitter dataset achieving 0.79 score. The proposed technique is based on combining of several simpler classifiers into one more complex blended model, which classified the data taking into account the probabilities of belonging to classes calculated by simpler models. They used the Logistic Regression, Naive Bayes, and SVM classifiers.

The best lexical features for the English corpus were shown by 14-exlab team and included Swear Word Count (a representation of the number of swear words contained in a tweet), Swear Word Presence (a binary value representing the presence of swear words), Sexist Slurs Presence (it was used a small set of sexist words aimed towards women), Hashtag Presence (a binary value equals 0 if there is no hashtag in the tweet or equals 1 if there is at least one hashtag in the tweet).

Table 2 presents the best results in Subtask B for the English dataset.

For Task B (tweet classification by different types of misogyny and target classification: active or passive types) the best results were achieved by the SB team [15] with 0.44 average F-Measure. The teams created the best models using SVM with lineal kernel and an ensemble model which combined SVM, Random Forest and Gradient Boosting classifiers. Also, the team created lists of specific lexicons concerning sexuality (p*ssy,

[1] http://ceur-ws.org/Vol-2150/

[2] http://ceur-ws.org/Vol-2263/

c*ncha), profanity, femininity (some words which could be used in negative sense like *gallina (chicken)*, *blonde*) and the human body (having a strong connection with sexuality) and Abbreviations and Hashtags lists which included typical for the Internet slang words like 'smh'. It should be noted that the best results of evaluation were achieved for different datasets (English and Spanish ones) using different approaches.

For example, the JoseSebastian team [16] showed the 10th result for the English dataset, but the top result for the Spanish one with 0.81 accuracy for the binary classification in Subtask A using the SVM model. They replaced all hashtags with keyword HASHTAG and some of them which are known as misogynistic ones with keyword MISO_HASHTAG. The authors note that the large difference between the results for the English and Spanish datasets could have a close connection with the choice of misogynistic hashtags in different languages.

The Resham team [17] used both an ensemble of models and neural networks for their modeling. They presented two approaches to deal with the challenge, the first of them was to create an ensemble of models including Logistic Regression, Support Vectors Machine, Random Forest, Gradient Boosting and Stochastic Gradient Descent models. Their second idea for modeling was to apply Word-level and Document-level Embedding and Recurrent Neural Network.

The authors applied the Continuous Bag of Words (CBOW) approach to create words vectors, so they collected 20 words which are potentially misogynistic (like b*tch, sl*t) and download 20,000 tweets which contained these words with the aim of finding the closest connection words. As the result, they got 100-dimensional word vectors and 300-dimensional word vectors for modeling.

Also, they did the same for Document-level Embedding presented the whole tweet as a word. The result of this unsupervised type of modeling are promising and the authors note that the accuracy could be higher in condition of using extended labeled dataset.

3.2 AMI@EVALITA 2018

The second shared task named Evalita -2018 has the same Subtasks A and B and there were 2 datasets in English and Italian languages [18]. The training English dataset consisted of 1,785 misogynistic tweets and 2,215 non-misogynistic messages. There were 460 misogynous and 540 non-misogynous tweets for testing. The Italian dataset included 4,000 tweets for training (with 46% misogynistic posts) and 1000 tweets for testing (with 51% misogynistic messages). The best results of the Subtask A for the English dataset are presented in Table 3.

Table 2. The best results for the Subtask B (IberEval)

Team	Accuracy
SB	0.442
14-exlab	0.369
Resham	0.351

The models with the highest accuracy were presented by models of Logistic regression from Bakarov team [19], and the ensemble of models from Resham team [20], so we can conclude that the best models in this case are based on the classical machine learning approach. The best results for the Subtask B using the English dataset are presented in Table 4.

In this case, the best models were also created using classical machine learning: ensemble of models by Himani team, Support Vector Machines by CrotoneMilano team [21] and Logistic Regression by Hateminers team. Also, it should be noted that the Bakarov team made the text classification based on using semantic features obtained from vector space models of texts. They used a factorization of the term-document matrix (the method of singular value decomposition) and normalization of factorized values.

As an interesting feature, the CrotoneMilano team calculated the length of words and took it into accounting during the experiments.

4 Conclusion

The problem of misogyny identification in social media is complicated and has a lot of different aspects. There are some approaches based on classical machine learning like Support Vector Machines, Naive Bayes, Logistic Regression and

Table 3. The best results for the Subtask A (Evalita)

Team	Accuracy
Hateminers	0.704
Resham	0.651
Bakarov	0.649

Table 4. The best results for the Subtask B (Evalita)

Team	Accuracy
Himani	0.406
CrotoneMilano	0.369
Hateminers	0.369

ensembles of these models which have achieved quite good results, as shown in Evalita-2018 and IberEval-2018 shared tasks.

Also, models based on neural networks like Convolutional Neural Networks and Long Short-Term Memory Networks work good in case of misogyny recognition and combinations of these models have a high potential in this area of study.

To compare the classical machine learning models, especially ensembles of this models, allow to achieve higher results than the models based on neural networks in case of misogyny identification in Twitter. But these experiments were carried out on relatively small datasets, and we cannot say that the results will be the same with an expanded dataset. It is well known that neural networks work well on large amounts of data, so we assume that their use is also necessary in the further work to make the most approximate to the "real life" experiments to detect misogyny in social media.

Acknowledgements

The research was completed by with the partial support of the grant PPRA1707 of the Technological University of Dublin.

References

1. **Fasoli, F., Carnaghi, A., & Paladino, M.P. (2015).** Social acceptability of sexist derogatory and sexist objectifying slurs across contexts. *Language Sciences*, Vol. 52, pp. 98–107. DOI: 10.1016/j.langsci.2015.03.003.
2. **Hardaker, C. & McGlashan, M. (2015).** Real men don't hate women: Twitter rape threats and group identity. *Journal of Pragmatics*, Vol. 91, pp.80–93. DOI: 10.1016/j.pragma.2015.11.005.
3. **Waseem, Z. & Hovy, D. (2016).** Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. *Proceedings of the NAACL Student Research Workshop*, pp. 88–93. DOI: 10.18653/v1/N16-2013.
4. **Clarke, I. & Grieve, J. (2017).** Dimensions of Abusive Language on Twitter. *Proceedings of the First Workshop on Abusive Language Online,* pp. 1–10. DOI: 0.18653/v1/W17-3001.
5. **Hewitt, S., Tiropanis, T., & Bokhove, C. (2016).** The problem of identifying misogynist language on Twitter (and other online social spaces). *Proceedings of the 8th ACM Conference on Web Science*, pp. 333–335. DOI: 10.1145/2908131. 2908183.
6. **Nina-Alcocer, V. (2018).** AMI at IberEval2018 Automatic Misogyny Identification in Spanish and English Tweets. *CEUR Workshop Proceedings.* pp. 274–279.
7. **Saleem, H.M., Dillon, K.P., Benesch, S., & Ruths, D. (2017).** A Web of Hate: Tackling Hateful Speech in Online Social Spaces. *CoRR abs/1709.10159*.
8. **Ziki, Z. & Lei, L. (2018).** Hate speech detection: A solved problem? The challenging case of long tail on Twitter. *arXiv preprint arXiv:1803.03662*. Vol. 10, No. 5, pp. 925–945. DOI: 10.3233/SW-180338.
9. **Park, J.H. & Fung, P. (2017).** One-step and two-step classification for abusive language detection on Twitter. *ArXiv preprint aeXiv:1706.01206*.
10. **Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017).** Deep learning for hate speech detection in tweets. *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 759–760. DOI: 10.1145/3041021.3054223.
11. **Anzovino, M., Fersini, E., & Rosso, P. (2018).** Automatic Identification and Classification of Misogynistic Language on Twitter. *Proc. 23rd Int. Conf. on Applications of Natural Language to Information Systems, NLDB-2018, LNCS* 10859, pp. 57–64. DOI: 10.1007/978-3-319-91947-8_6.
12. **Goenaga, I., Atutxa, A., Gojenola, K., Casilas, A., Dıaz de Ilarraza, A., Ezeiza, N., Oronoz, M., Perez, A., & Perez de Vinaspre, O. (2018).** Automatic Misogyny Identification Using Neural Networks. *CEUR Workshop Proceedings.* pp. 249–254 .

13. **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.* pp. 234–241.
14. **Shushkevich, E. & Cardiff, J. (2018).** Classifying misogynistic tweets using a blended model: The AMI shared task in IBEREVAL 2018. *CEUR Workshop Proceedings.* pp. 255–259.
15. **Frenda, S., Ghanem, B., & Montes-y-Gomez, M. (2018).** Exploration of Misogyny in Spanish and English tweets. *CEUR Workshop Proceedings.* Vol 2150, pp.260–267.
16. **Canós, J.S. (2018).** Misogyny identification through SVM at IberEval 2018. *CEUR Workshop Proceedings.* pp. 229–233.
17. **Ahluwalia, R., Shcherbinina, E., Callow, E., Nascimento, A., & De Cock, M. (2018).** Detecting Misogynous Tweets. *CEUR Workshop Proceedings.* pp. 242–248.
18. **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*. Turin, Italy, pp. 214–228.
19. **Bakarov, A. (2018).** Vector Space Models for Automatic Misogyny Identification*. Proceedings of Sixth Evaluation Campaign of Natural Language. Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018),* Turin, Italy. pp. 211–213.
20. **Ahluwalia, R., Soni, H., Callow, E., Nascimento, A., & De Cock., M. (2018).** Detecting Hate Speech Against Women in English Tweets. *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018),* Turin, Italy.
21. **Pamungkas, E.W., Cignarella, A,T., Basile, V., & Patti, V. (2018).** Automatic Identification of Misogyny in English and Italian Tweets at EVALITA 2018 with a Multilingual Hate Lexicon. *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018),* Turin, Italy.

Article received on 11/07/2019; accepted on 28/09/2019. Corresponding author is John Cardiff.