# Detecting Misogynous Tweets

Resham Ahluwalia[1], Evgeniia Shcherbinina[1], Edward Callow[1], Anderson Nascimento[1], and Martine De Cock[1,2]

[1] University of Washington, Tacoma WA-98402, USA
[2] Ghent University, 9000 Gent, Belgium
{resh,es43,ecallow,andclay,mdecock}@uw.edu

**Abstract.** Social media companies struggle to control the quality of the content on their platforms. The sheer amount of user-generated content uploaded on a daily basis far exceeds what can be screened by human curators, fuelling the need for intelligent detection algorithms that can automatically flag inappropriate content. In this paper, we present machine learning models that can identify instances of aggression and hate speech towards women in tweets. In particular, we present the system that we submitted for the shared task on automatic misogyny identification at IberEval 2018.

**Keywords:** Automatic misogyny identification · Text classification.

## 1 Introduction

Misogyny is prevalent across all social media platforms and is of increasing concern. In today's hyper-connected social media world, hate speech towards women once limited to a particular place or time can happen anytime anywhere with just a few taps on the keyboard. With millions of contributions added every day by users on large social media platforms, it is impractical to manually police all misogynous posts. Notwithstanding this, Facebook for instance has recently announced plans to hire several thousands of new employees to moderate content [3]. The seriousness of the problem, combined with the sheer amount of user-generated content, necessitates the development of algorithms that can automatically scan user contributions for inappropriate content, to assist human curators.

In this paper, we address the detection of misogyny in tweets, in line with the shared task on automatic misogyny classification organized at IberEval 2018 [7]. Automatically labeling tweets as misogynous vs. non-misogynous is challenging because the language of tweets is full of syntactic and grammatical flaws, making extraction of text-based features difficult. Sometimes tweets consist of only one or a few words, and due to the lack of conversational context, it is difficult to assess whether such very short tweets should be perceived as misogynous. Finally, tweets can be indirectly misogynous, for example through the use of sarcasm, making the intended nature of the tweet difficult to detect.

Below we describe a machine learning approach to automatically identify misogynous comments from Twitter data. We obtain our best results with an

ensemble of classifiers that make inferences based on the presence or absence of word unigrams and bigrams. In addition to classifying tweets as misogynous or not (Task A), in the former case we also identify the kind of the misogyny (e.g. sexual harassment, derailing, ...) and whether the offensive message was targeted at a specific individual vs. being a general comment (Task B). For Task A we were the 6th ranking team (team "resham") at the IberEval 2018 competition, with an accuracy of 78%. For Task B, we were the 3rd ranking team, with an average macro-average F1-score of 0.35.

## 2 Background

Online harassment of women both in the form of personal attacks and in the form of generalized hate speech may be considered as sexual harassment and may have the effect of preventing and discouraging women from participating in social media on an equal footing with men [12]. Although social media sites such as Twitter generally prohibit hate speech and other forms of abuse against women [18], such speech has thrived for several reasons. First, the relative anonymity of the internet has emboldened perpetrators, who might otherwise fear the consequences of such harassment [10]. Second, social media sites primarily rely upon manual screening and reporting of abusive texts, which is not scalable to the amount of data [17]. The presence of hate speech has given rise to various social responses, including attempts to identify and shame such perpetrators through a kind of vigilante justice [6]. However, one cannot always identify the perpetrators, and vigilante justice is not an ideal solution, as it can raise ethical issues of its own [4–6].

Several machine learning based solutions for the automatic detection of online harassment and hate speech have been proposed. Zhang and Luo used a deep neural network partially trained on unlabeled corpora to classify hate speech [17]. They implemented a convolutional neural network (CNN) and a gated recurrent unit (GRU), a kind of recurrent neural network (RNN), to classify social media text as one of four categories: "non-hate", "sexism", "racism", or "both". Razavi et al. proposed a 3-tier Naive Bayes and Decision Tree classifier model, based on a dictionary of 'offensive' words in order to classify text according to the level of offensiveness [1]. Hewitt et al. also used a list of offensive words to collect and then manually label misogynous tweets.[15] However, a purely lexical approach ignores the fact that even words that are often offensive may be used in ways that do not necessarily disparage women. Nobata et al. [2] analyzed character n-grams using a 2-tiered model, by first classifying tweets as "abusive" or not, and then classifying them into one of three categories: "hate speech", "derogatory language", or "profanity". Out of all the related work mentioned above, we followed an approach most similar to Zhang and Luo, but we rely upon both word n-grams and character n-grams as the semantic units for our model, and like Zhang and Luo partially rely upon unclassified tweets to build an embedding layer; however, we also use additional supervised learning models in our analysis, as described infra, and when using neural networks we exclusively use RNNs.

**Table 1.** Distribution of tweets in the dataset

| **Misogyny** | | **Category** | | **Target** | |
|---|---|---|---|---|---|
| Non-misogynous | 51.8% | | 0% | | 0% |
| Misogynous | 48.2% | Discredit | 60.0% | Active | 60% |
| | | Sexual harassment | 26.1% | Passive | 40% |
| | | Stereotype | 8.8% | | |
| | | Dominance | 3.1% | | |
| | | Derailing | 2.0% | | |

In Anzovino et al., unlike the previous work, the problem of misogyny classification is more clearly defined and isolated as a distinct category of hate speech [13]. In particular, misogynist tweets are considered as distinct subcategories, consisting of 'discredit', 'stereotype', 'objectification', 'sexual harassment', 'threats of violence', 'dominance', or 'derailing'. Our model attempts to classify tweets according to this classification system and generally to follow the approach of Anzovino et al to the classification of misogynist tweets [13].

## 3 System Description

*Description of Dataset and Classification Tasks.* We use the labeled English tweet corpus from the shared task on Automatic Misogyny Classification at IberEval 2018 [7]. The corpus is composed of 3,251 tweets manually labeled by human annotators as either misogynous or non-misogynous, and if misogynous, with their respective category and target (see Table 1). Some examples of labeled tweets are provided in Table 2. We split this data into a *training set* with 90% of the instances and a *validation set* with 10% of the instances, using a stratified split to ensure that relative class frequencies are approximately preserved in each set. We address the following classification tasks:

- Task A: Given a tweet, classify it as either misogynous or not misogynous (binary classification).
- Task B: Classify the misogynous tweets according to the *category* of misogyny (multi-class classification; see Table 1 for the 5 different labels) and the *target* of the message (binary classification). For the latter, "Active" means that the text includes offensive messages purposely sent to a specific target, while "Passive" (generic) refers to messages posted to many potential receivers.

**Table 2.** Examples of tweets in the dataset

| **Tweet** | **Misogyny** | **Category** | **Target** |
|---|---|---|---|
| All of us from dealing with feminists #ThingsTimHowardCouldSave | 0 | NA | NA |
| You a pussy bitch watch when I catch you Im beat tf outta you | 1 | Sexual harassment | Active |
| jul_hamomelett sorry sweetie, but women are the causes of all problems | 1 | Stereotype | Passive |

*Text Preprocessing.* As a preprocessing step, we remove links, periods, commas, and usernames in @-mentions. Then, we tokenize tweets keeping all emoticons and emoji, which can convey the emotions of the entire tweet.

*Bag of Words and Ensemble Classification Model.* The most common way to turn raw text into numerical features is a bag of words approach that includes tokenizing, counting the occurrences of the tokens, normalizing the counts and using them as weights. In addition to individual words, we also extract bigrams, where occurrences of pairs of consecutive words are counted.

For the binary classification task of designating the tweets as misogynous or not (Task A), we train an ensemble of 5 classifiers, namely Logistic Regression, Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting, and Stochastic Gradient Descent (SGD). The ensemble selects the class that has the highest-class probability averaged over all the individual classifiers. For training the classifiers we use scikit-learn [9] with the default choices for all parameters except for the following:

- Logistic Regression: inverse of regularization strength $C$ is 0.1.
- SVM: penalty parameter $C$ is 5, kernel coefficient gamma is 0.01.
- RF: maximum depth of a tree is 145, number of trees is 200.
- Gradient Boosting: maximum depth of the individual regression estimators is 25, number of boosting stages to perform is 150.
- SGD: log loss, the constant that multiplies the regularization term is 0.001.

For Task B, i.e. category detection and target detection, we train similar ensembles as above. The main difference is that the classifiers for Task A are trained on the entire training set, while the classifiers for Task B are trained using only those tweets from the training set that have been labeled as misogynous by the human annotators. Finally, since some of the categories for the category detection problem have only a small number of instances, we train 3-class classifiers that label a misogynous tweet as "Discredit", "Sexual harassment", or "Other" instead of distinguishing between all 5 categories.

In deployment, classification is accomplished in two steps. First, the ensemble model for misogyny detection infers whether the given tweet is misogynous. Then, for the misogynous tweets, we apply the ensemble models for category and target classifications trained on misogynous tweets only. In case the inferred category is "Other", we map it to "Stereotype".

*Word-level Embedding and Recurrent Neural Network.* Word-embedding is a natural-language processing technique that maps words into a vector space, where vectors representing words with similar meanings are located close to each other. This embedding is automatically learned from text using a neural network in an unsupervised fashion. There are two popular approaches to creating word vectors: Skip-gram and Continuous Bag of Words (CBOW). In Skip-gram, the input of the neural network is the target word, while the outputs are the surrounding words. CBOW is the opposite of Skip-gram [16]. The hidden layer of the neural network encodes the word representation.

**Table 3.** Examples of words relatively close in the vector space according to the fastText model

| Word | Semantically close words according to the fastText model |
|---|---|
| bitch | bitches, hoes, @, heabitch, lmfaoooo |
| hoes | hoe, bitches, bitch, hella, different emoji |
| slut | horny, cum, hotwife, cumwhore, sluthub |
| suck | dick, tittie, dicks, chyna, sucks |
| whore | presstitute, paula, slut, suppoon, stormy |

Successfully training neural networks requires a substantial amount of data. To collect such data, we compiled a list of 20 words $w$ for which the conditional probability that a tweet is misogynous given that it contains the word $w$, is high. We estimated these conditional probabilities based on the training set, for all words occurring in it. The highest scoring words are, in decreasing order of conditional probability, *bitches, whore, suck, bitch, hoes, slut, dumb, ...*. Next, we downloaded 200,000 tweets that contain any of the 20 top ranking words. The fact that we do not have ground truth labels for these 200,000 tweets is not a problem since we use them only for unsupervised learning, namely to train a fastText model with the CBOW option from the Gensim package.[3] The trained model maps every word to a 100-dimensional word vector. Table 3 shows words that are close together according to the learned representation.

Next, we use the tweets from the labeled training set to train an RNN classifier for Task A as well as for category detection and for target detection in Task B. The RNN classifiers have the following configuration: an embedding layer with dropout rate 0.5, two LSTM layers with 100 cells and dropout rate 0.5, and a densely-connected layer with sigmoid activation function. For the embedding layer we use the 100-dimensional word vectors computed with our fastText model. For comparison, we trained the same LSTM-networks using a fastText model trained on Common Crawl and Wikipedia text that maps words to 300-dimensional vectors [8].

Document-level embedding is an extension of word-level embedding that involves learning a vector representation not only for separate words or character n-grams but also for word n-grams, sentences, paragraphs, and entire documents [11]. Every paragraph (or in our case: every tweet) is mapped to a unique vector [14]. We used the Gensim package to train a 100-dimensional doc2vec model over the 200,000 unlabeled tweets. Next, we used the trained doc2vec model to convert each tweet in the labeled training set into a vector of 100 numerical features, which we subsequently used as input to train an SVM classifier for all three classification problems. The SVM parameters are the same as mentioned above.

## 4 Results and Discussion

As explained in Section 3, we train our models on 90% of the data, and hold out the remaining 10% for validation. Results for Task A are reported in terms of

[3] https://radimrehurek.com/gensim/

**Table 4.** Accuracy and macro-average F1-scores for different models trained on the training set and evaluated on the validation set with our model in bold. The results on the competition's test data (with hidden labels) were 78.51% accuracy, macro-F1=0.15 for category detection, and macro-F1=0.55 for target detection.

| | **Task A** | **Task B** | |
|---|---|---|---|
| **Model** | **Misogyny Accuracy %** | **Category macro-F1** | **Target macro-F1** |
| Baseline (majority baseline algorithm) | 51.77 | 0.15 | 0.37 |
| Bag of Words and Logistic Regression | 76.69 | 0.23 | 0.71 |
| Bag of Words and SVM | 77.30 | 0.23 | 0.69 |
| Bag of Words and Random Forest | 77.61 | 0.21 | 0.68 |
| Bag of Words and Gradient Boosting | 75.15 | 0.24 | 0.70 |
| Bag of Words and SGD | 77.30 | 0.24 | 0.73 |
| **Bag of Words and the Ensemble Model** | **79.14** | **0.23** | **0.73** |
| Word-level Embedding (100 dim) and LSTM | 68.40 | 0.15 | 0.67 |
| Word-level Embedding (300 dim) and LSTM | 70.03 | 0.21 | 0.70 |
| Document-level Embedding (100 dim) and SVM | 64.54 | 0.15 | 0.37 |

accuracy. Labeling all tweets as non-misogynous results in a baseline accuracy of 51.77%, as indicated in the Table 4. Because of the class unbalances, Task B is evaluated in terms of macro-average F1-score.

Table 4 contains the results from all models from Section 3 when trained on the training set and evaluated on the validation set. For Task A, the evaluation was performed over the entire validation set; for Task B, the evaluation was carried out for the tweets from the validation set that were labeled as misogynous by the human annotators. A majority baseline algorithm that labels all tweets for Task B as category "Discredit" and target "Active" achieves a macro-average F1-score of 0.15 for the category detection task, and a macro-average F1-score of 0.37 for the target detection task. As category detection is a multi-class classification problem, F1 scores are lower than for the binary classification tasks.

## 5 Conclusion

We evaluated supervised and unsupervised approaches to misogyny detection in tweets. An ensemble of 5 classifiers trained in a supervised manner on a bag of words (consisting of word unigrams and bigrams) performs best overall. The word-level and document-level embedding approaches look promising considering that they obtain very reasonable results despite the limited size of the training dataset. Increasing the number of tweets for training word and document vectors may improve the accuracy of mapping words into the vector space while preserving the uniqueness of the language used on Twitter. Using an extended labeled data set may improve the performance of the neural networks resulting in higher accuracy and macro-average F1-scores. An option to obtain a larger, noisily labeled dataset, is to apply semi-supervised machine learning.

## References


1. A. H. Razavi, D. Inkpen, S. Uritsky, S. Matwin. Offensive Language Detection Using Multi-level Classification. Proceedings of 23rd Canadian Conference on Artificial Intelligence, LNCS 6085, p. 16-27, 2010.
2. C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Yi Chang: Abusive Language Detection in Online User Content. Proceedings of the 25th International Conference on World Wide Web, p.145-153, 2016.
3. D. Gershgorn, M. Murphy. Facebook is hiring more people to moderate content than Twitter has at its entire company, 2017. https://qz.com/1101455/facebook-fb-is-hiring-more-people-to-moderate-content-than-twitter-twtr-has-at-its-entire-company/
4. E. A. Jane. "Your a Ugly, Whorish, Slut": Understanding E-bile. Feminist Media Studies 14(4), p.531-546, 2012.
5. E. A. Jane. Flaming? What flaming? The pitfalls and potentials of researching online hostility. Ethics and Information Technology 17(1), p.65-87, 2015.
6. E. A. Jane. Online misogyny and feminist digilantism. Continuum 30(3), p.284-297, 2016.
7. E. Fersini, M. Anzovino, P. Rosso. 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. CEUR-WS.org, Seville, Spain, September 18, 2018.
8. E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov. Learning Word Vectors for 157 Languages, 2018. arXiv:1802.06893
9. 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, E. Duchesnay. Scikit-learn: Machine Learning in Python.Journal of Machine Learning Research 12, p.2825-2830, 2011.
10. J. Fox, C. Cruz, and J. Y. Lee. Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. Computers In Human Behaviour 52, p.436-442, 2015.
11. J. H. Lau, T. Baldwin. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation, 2016. arXiv:1607.05368
12. J. Megarry. Online incivility or sexual harassment? Conceptualising women's experiences in the digital age. In Women's Studies International Forum Vol. 47, pp. 46-55. Pergamon, 2014.
13. M.Anzovino, E. Fersini, P. Rosso. Automatic Identification and Classification of Misogynistic Language on Twitter. In Proceedings of the 23rd International Conference on Natural Language  Information Systems, 2018.
14. Q. V. Le, T. Mikolov. Distributed Representations of Sentences and Documents. Proceedings of International Conference on Machine Learning, p.1188-1196, 2014.
15. S. Hewitt, T. Tiropanis, C. Bokhove, The Problem of Identifying Misogynist Language on Twitter. Proceedings of the 8th ACM Conference on Web Science, p.333-335, 2016.
16. X. Rong. word2vec Parameter Learning Explained, 2014. arXiv:1411.2738
17. Z. Zhang, L.Luo. Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter, 2018. arXiv:1803.03662
18. https://help.twitter.com/en/rules-and-policies/twitter-rules