Detecting Hate Speech Against Women in English Tweets

Resham Ahluwalia, Himani Soni, Edward Callow, Anderson Nascimento, Martine De Cock*
School of Engineering and Technology
University of Washington Tacoma
{resh,himanis7,ecallow,andclay,mdecock}@uw.edu

Abstract

English. Hate speech is prevalent in social media platforms. Systems that can automatically detect offensive content are of great value to assist human curators with removal of hateful language. In this paper, we present machine learning models developed at UW Tacoma for detection of misogyny, i.e. hate speech against women, in English tweets, and the results obtained with these models in the shared task for Automatic Misogyny Identification (AMI) at EVALITA2018.

Italiano. *Commenti offensivi nei confronti di persone con diversa orientazione sessuale o provenienza sociale sono oggigiorno prevalenti nelle piattaforme di social media. A tale fine, sistemi automatici in grado di rilevare contenuti offensivi nei confronti di alcuni gruppi sociali sono importanti per facilitare il lavoro dei moderatori di queste piattaforme a rimuovere ogni commento offensivo usato nei social media. In questo articolo, vi presentiamo sia dei modelli di apprendimento automatico sviluppati all'Università di Washington in Tacoma per il rilevamento della misoginia, ovvero discorsi offensivi usati nei tweet in lingua inglese contro le donne, sia i risultati ottenuti con questi modelli nel processo per l'identificazione automatica della misoginia in EVALITA2018.*

1 Introduction

Inappropriate user generated content is of great concern to social media platforms. Although social media sites such as Twitter generally prohibit hate speech[1], it thrives online due to lack of accountability and insufficient supervision. Although social media companies hire employees to moderate content (Gershgorn and Murphy, 2017), the number of social media posts exceeds the capacity of humans to monitor without the assistance of automated detection systems.

In this paper, we focus on the automatic detection of misogyny, i.e. hate speech against women, in tweets that are written in English. We present machine learning (ML) models trained for the tasks posed in the competition for Automatic Misogyny Identification (AMI) at EVALITA2018 (Fersini et al., 2018b). Within this competition, Task A was the binary classification problem of labeling a tweet as misogynous or not. As becomes clear from Table 1, Task B consisted of two parts: the multiclass classification problem of assigning a misogynous tweet to the correct category of misogyny (e.g. sexual harassment, stereotype, ...), and the binary classification problem of determining whether a tweet is actively targeted against a specific person or not.

Interest in the use of ML for automatic detection of online harassment and hate speech is fairly recent (Razavi et al., 2010; Nobata et al., 2016; Anzovino et al., 2018; Zhang and Luo, 2018). Most relevant to our work are approaches published in the context of a recent competition on automatic misogyny identification organized at IberEval2018 (Fersini et al., 2018a), which posed the same binary classification and multiclass classification tasks addressed in this paper. The AMI-baseline system for each task in the AMI@IberEval competition was an SVM trained on a unigram representation of the tweets, where each tweet was represented as a bag of words (BOW) composed of 1000 terms. We participated in the AMI@IberEval competition with an Ensem-

*Guest Professor at Dept. of Applied Mathematics, Computer Science and Statistics, Ghent University

[1] https://help.twitter.com/en/rules-and-policies/twitter-rules

Task A: Misogyny	Train	Test	**Task B: Category**	Train	Test	**Task B: Target**	Train	Test
Non-misogynous	2215	540	0	2215	540	0	2215	540
Misogynous	1785	460	Discredit	1014	141	Active	1058	401
			Sexual harassment	352	44	Passive	727	59
			Stereotype	179	140			
			Dominance	148	124			
			Derailing	92	11			

Table 1: Distribution of tweets in the dataset

ble of Classifiers (EoC) containing a Logistic Regression model, an SVM, a Random Forest, a Gradient Boosting model, and a Stochastic Gradient Descent model, all trained on a BOW representation of the tweets (composed of both word unigrams and word bigrams) (Ahluwalia et al., 2018). In AMI@IberEval, our team *resham* was the 7th best team (out of 11) for Task A, and the 3rd best team (out of 9) for Task B. The winning system for Task A in AMI@IberEval was an SVM trained on vectors with lexical features extracted from the tweets, such as the number of swear words in the tweet, whether the tweet contains any words from a lexicon with sexist words, etc. (Pamungkas et al., 2018). Very similarly, the winning system for the English tweets in Task B in AMI@IberEval was also an SVM trained on lexical features derived from the tweets, using lexicons that the authors built specifically for the competition (Frenda et al., 2018).

For the AMI@EVALITA competition, which is the focus of the current paper, we experimented with the extraction of lexical features based on dedicated lexicons as in (Pamungkas et al., 2018; Frenda et al., 2018). For Task A, we were the 2nd best team (*resham.c.run3*), with an EoC approach based on BOW features, lexical features, and sentiment features. For Task B, we were the winning team (*himani.c.run3*) with a two-step approach: for the first step, we trained an LSTM (Long Short-Term Memory) neural network to classify a tweet as misogynous or not; tweets that are labeled as misogynous in step 1 are subsequently assigned a category and target label in step 2 with an EoC approach trained on bags of words, bigrams, and trigrams. In Section 2 we provide more details about our methods for Task A and Task B. In Section 3 we present and analyze the results.

2 Description of the System

The training data consists of 4,000 labeled tweets that were made available to participants in the AMI@EVALITA competition. As Table 1 shows, the distribution of the tweets over the various labels is imbalanced; the large majority of misogynistic tweets in the training data for instance belong to the category "Discredit". In addition, the distribution of tweets in the test data differs from that in the training data. As the ground truth labels for the test data were only revealed after the competition, we constructed and evaluated the ML models described below using 5-fold cross-validation on the training data.

2.1 Task A: Misogyny

Text Preprocessing. We used NLTK[2] to tokenize the tweets and to remove English stopwords.

Feature Extraction. We extracted three kinds of features from the tweets:

- *Bag of Word Features.* We turned the preprocessed tweets into BOW vectors by counting the occurrences of token unigrams in tweets, normalizing the counts and using them as weights.
- *Lexical Features.* Inspired by the work of (Pamungkas et al., 2018; Frenda et al., 2018), we extracted the following features from the tweets:
  - Link Presence: 1 if there is a link or URL present in the tweet; 0 otherwise.
  - Hashtag Presence: 1 if there is a Hashtag present; 0 otherwise.
  - Swear Word Count: the number of swear words from the *noswearing dictionary*[3] that appear in the tweet.
  - Swear Word Presence: 1 if there is a swear word from the *noswearing dictionary* present in the tweet; 0 otherwise.
  - Sexist Slur Presence: 1 if there is a sexist word from the list in (Fasoli et al., 2015) present in the tweet; 0 otherwise.
  - Women Word Presence: The feature value is 1 if there is a *woman synonym word* [4] present in the tweet; 0 otherwise.

[2] https://www.nltk.org/, TweetTokenizer

[3] https://www.noswearing.com/dictionary

[4] https://www.thesaurus.com/browse/woman

- *Sentiment scores.* We used SentiWordNet (Baccianella et al., 2010) to retrieve a positive and a negative sentiment score for each word occurring in the tweet, and computed the average of those numbers to obtain an aggregated positive score and an aggregated negative score for the tweet.

Model Training. We trained 3 EoC models for designating a tweet as misogynous or not (Task A). The EoC models differ in the kind of features they consume as well as in the kinds of classifiers that they contain internally.

- *EoC with BOW (resham.c.run2)*[5]*:* an ensemble consisting of a Random Forest classifier (RF), a Logistic Regression classifier (LR), a Stochastic Gradient Descent (SGD) classifier, and a Gradient Boosting (GB) classifier, each of them trained on the BOW features.
- *EoC with BOW and sentiment scores (resham.c.run1):* an ensemble consisting of the same 4 kinds of classifiers as above, each of them trained on the BOW and sentiment score features.
- *EoC with BOW, sentiment scores, and lexical features (resham.c.run3):* an ensemble consisting of
  - RF on the BOW and sentiment score features
  - SVM on the lexical features
  - GB on the lexical features
  - LR on the lexical features.
  - GB on the BOW and sentiment features

All the ensembles use hard voting. For training the classifiers we used scikit-learn (Pedregosa et al., 2011) with the default choices for all parameters.

2.2 Task B: Category And Target

For Task B, our winning system *himani.c.run3* consists of a pipeline of two classifiers: the first classifier (step 1) in the pipeline labels a tweet as misogynous or not, while the second classifier (step 2) assigns the tweets that were labeled misogynous to their proper category and target.

For Step 1 we trained a deep neural network that consists of a word embedding layer, followed by a bi-directional LSTM layer with 50 cells, a hidden dense layer with 50 cells with relu activation, and an output layer with sigmoid activation. For the embedding layer we used the pretrained Twitter Embedding from the GloVe package (Pennington et al., 2014), which maps each word to a 100-dimensional numerical vector. The LSTM network is trained to classify tweets as misogynous or not. We participated with this trained network in Task A of the competition as well (*himani.c.run3*). The results were not as good as those obtained with the models described in Section 2.1, so we do not go into further detail.

Next we describe how we trained the models used in Step 2 in *himani.c.run3*.

Text Preprocessing. We used the same text preprocessing as in Section 2.1. In addition we removed words occurring in more than 60 percent of the tweets along with those that had a word frequency less than 4.

Feature Extraction. We turned the preprocessed tweets into *Bag of N-Gram* vectors by counting the occurrences of token unigrams, bigrams and trigrams in tweets, normalizing the counts and using them as weights. For simplicity, we keep referring to this as a BOW representation.

Model Training. For category and target identification, *himani.c.run3* uses an EoC approach where all classifiers are trained on the BOW features mentioned above. The EoC models for category identification on one hand, and target detection on the other hand, differ in the classifiers they contain internally, and in the values of the hyperparameters. Below we list parameter values that differ from the default values in scikit-learn (Pedregosa et al., 2011).

- *EoC for Category Identification:*
  - LR: inverse of regularization strength C is 0.7; norm used in the penalization is L1; optimization algorithm is 'saga'.
  - RF: number of trees is 250; splitting attributes are chosen based on information gain.
  - SGD: loss function is 'modified huber'; constant that multiplies the regularization term is 0.01; maximum number of passes over the training data is 5.
  - Multinomial Naive Bayes: all set to defaults.
  - XGBoost: maximum depth of tree is 25; number of trees is 200.
- *EoC for Target Identification:*

[5]Here 'resham.c.run2' refers to the second run of the data submitted by the author in connection with the competition. Similar citations that follow have a corresponding meaning.

Approach	5-fold CV on Train	Test
majority baseline	0.553	0.540
resham.c.run1	0.790	0.648
resham.c.run2	0.787	0.647
resham.c.run3	**0.795**	**0.651**
himani.c.run3	0.785	0.614

Table 2: Accuracy results for Task A: Misogyny detection on English tweets.

- LR: inverse of regularization strength C is 0.5; norm used in the penalization is L1; optimization algorithm is 'saga'.
- RF: number of trees is 200; splitting attributes are chosen based on information gain.

For completeness we mention that *himani.c.run2* consisted of a two-step approach very similar to the one outlined above. In Step 1 of *himani.c.run2* tweets are labeled as misogynous or not with an EoC model (RF, XGBoost) trained on the Bag of N-Gram features. In Step 2, a category and target label are assigned with respectively an LR, XGBoost-EoC model and an LR, RF-EoC model in which all classifiers are trained on the Bag of N-Gram features as well. Since this approach is highly similar to the *himani.c.run3* approach described above and did not give better results, we do not go into further detail.

3 Results and Discussion

3.1 Results for Task A

Table 2 presents accuracy results for Task A, i.e. classifying tweets as misogynous or not, evaluated with 5-fold cross-validation (CV) on the 4,000 tweets in the training data from Table 1. In addition, the last column of Table 2 contains the accuracy when the models are trained on all 4,000 tweets and subsequently applied to the test data. We include a simple majority baseline algorithm that labels all tweets as non-misogynous, which is the most common class in the training data.

The accuracy on the test data is noticeably lower than the accuracy obtained with 5-fold CV on the training data. At first sight, this is surprising because the label distributions are very similar: 45% of the training tweets are misogynous, and 46% of the testing tweets are misogynous. Looking more carefully at the distribution across the different categories of misogyny in Table 1, one can observe that the training and test datasets do vary quite a lot in the kind (category) of misogyny. It is plausible that tweets in different misogyny categories are characterized by their own, particular language, and that during training our binary classifiers have simply become good at flagging misogynous tweets from categories that occur most often in the training data, leaving them under-prepared to detect tweets from other categories.

Regardless, one can see that the ensembles benefit from having more features available. Recall that *resham.c.run2* was trained on BOW features, *resham.c.run1* on BOW features and sentiment scores, and *resham.c.run3* on BOW features, sentiment scores, and lexical features. As is clear from Table 2, the addition of each feature set increases the accuracy. As already mentioned in Section 2.2, the accuracy of *himani.c.run3*, which is a bidirectional LSTM that takes tweets as strings of words as its input, is lower than that of the *resham* models, which involve explicit feature extraction.

3.2 Results for Task B

Table 3 contains the results of our models for Task B in terms of F1-scores. Following the approach used on the AMI@EVALITA scoreboard, both subtasks are evaluated as multiclass classification problems. For Category detection, there are 6 possible class labels, namely the label 'non-misogynous' and each of the 5 category labels. Similarly, for Target detection, there are 3 possible class labels, namely 'non-misogynous', 'Active', and 'Passive'.

When singling out a specific class c as the "positive" class, the corresponding F1-score for that class is defined as usual as the harmonic mean of the precision and recall for that class. These values are computed treating all tweets with ground truth label c as positive examples, and all other tweets as negative examples. For example, when computing the F1-score for the label "Sexual harassment" in the task of Category detection, all tweets with ground truth label "Sexual harassment" are treated as positive examples, while the tweets from the other 4 categories of misogyny *and* the non-misogynous tweets are considered to be negative examples. The average of the F1-scores computed in this way for the 5 categories of misogyny is reported in the columns F1 (Category) in Table 3, while the average of the F1-scores for 'Active' and 'Passive' is reported in the columns F1 (Target) in Table 3. The first columns contain results obtained

	5-fold CV on Train			Test		
Approach	**F1 (Category)**	**F1 (Target)**	**Average F1**	**F1 (Category)**	**F1 (Target)**	**Average F1**
majority baseline	0.079	0.209	0.135	0.049	0.286	0.167
himani.c.run2	0.283	0.622	0.452	0.323	0.431	0.377
himani.c.run3	**0.313**	**0.626**	**0.469**	**0.361**	**0.451**	**0.406**
Step 1 from resham.c.run3 & Step 2 from himani.c.run3	0.278	0.515	0.396	0.246	0.361	0.303

Table 3: F1-score results for Task B on English tweets

	5-fold CV on Train						Test					
Approach	**Pr (A)**	**Re (A)**	**F1 (A)**	**Pr (P)**	**Re (P)**	**F1 (P)**	**Pr (A)**	**Re (A)**	**F1 (A)**	**Pr (P)**	**Re (P)**	**F1 (P)**
himani.c.run3	0.61	0.79	0.69	0.53	0.56	0.54	0.61	0.75	0.67	0.14	0.61	0.23
Step 1 from resham.c.run3 & Step 2 from himani.c.run3	0.70	0.70	0.70	0.51	0.31	0.39	0.67	0.45	0.54	0.17	0.19	0.18

Table 4: Detailed precision (Pr), recall Re), and F1-score (F1) results for Task B: Target Identification on English tweets; 'A' and 'P' refer to 'Active' and 'Passive' respectively.

		Predicted value		
		N	**A**	**P**
Actual value	**N**	202	176	162
	A	40	301	60
	P	8	15	36

Table 5: Confusion matrix for Task B: Target Identification with *himani.c.run3* on the test data; 'N', 'A', and 'P' refer to 'Non-misogynous', 'Active' and 'Passive' respectively.

		Predicted value		
		N	**A**	**P**
Actual value	**N**	428	78	34
	A	201	182	18
	P	38	10	11

Table 6: Confusion matrix for Task B: Target Identification with Step 1 from *resham.c.run3* and Step 2 from *himani.c.run3* on the test data; 'N', 'A', and 'P' refer to 'Non-misogynous', 'Active' and 'Passive' respectively.

with 5-fold CV over the training data with 4,000 tweets from Table 1, while the last columns contain results for models trained on the entire training data of 4,000 tweets and subsequently applied to the test data. The latter correspond to the results on the competition scoreboard.

As a simple baseline model, we include an algorithm that labels every tweet as misogynous and subsequently assigns it to the most frequently occurring Category and Target from the training data, i.e. 'Discredit' and 'Active'. This model has a very low precision, which explains why its F1-scores are so low. The best results on the test data are obtained with *himani.c.run3*, which is an EoC approach using a BOW representation of extracted word unigrams, bigrams, and trigrams as features. This was the best performing model for Task B in the AMI@EVALITA competition.

Recall that *himani.c.run3* uses a two step approach where tweets are initially labeled as misogynous or not (Step 1) and then assigned to a Category and Target (Step 2). Given that for the task in Step 1, the binary classifier of *himani.c.run3* was outperformed in terms of accuracy by the binary classifier of *resham.c.run3* (see Table 2), an obvious question is whether higher F1-scores for Task B could be obtained by combining the binary classifier for misogyny detection from *resham.c.run3* with the EoC models for Category and Target identification from *himani.c.run3*. As the last row in Table 3 shows, this is not the case. To give more insight into where the differences in predictive performance in the last two rows of Table 3 stem from, Table 4 contains more detailed results about the precision, recall, and F1-scores for Task B: Target Identification on the train as well as the test data, while Table 5 and 6 contain corresponding confusion matrices on the test data. These tables reveal that the drop in F1-scores in the last row in Table 3 is due to a substantial drop

in recall. As can be seen in Table 4, replacing the binary classifier in Step 1 by the method from *resham.c.run3*, causes the recall for 'Active' tweets in the test data to drop from 0.75 to 0.45, and for 'Passive' tweets from 0.61 to 0.19. The slight increase in precision is not sufficient to compensate for the loss in recall. As can be inferred from Table 5 and 6, the recall of misogynous tweets overall with *himani.c.run3* is $(301 + 60 + 15 + 36)/460 \approx 0.896$ while with *resham.c.run3* it is only $(182 + 18 + 10 + 11)/460 \approx 0.480$.

4 Conclusion

In this paper we presented machine learning models developed at UW Tacoma for detection of hate speech against women in English language tweets, and the results obtained with these models in the shared task for Automatic Misogyny Identification (AMI) at EVALITA2018. For the binary classification task of distinguishing between misogynous and non-misogynous tweets, we obtained our best results (2nd best team) with an Ensemble of Classifiers (EoC) approach trained on 3 kinds of features: bag of words, sentiment scores, and lexical features. For the multiclass classification tasks of Category and Target Identification, we obtained our best results (winning team) with an EoC approach trained on a bag of words representation containing unigrams, bigrams, and trigrams. All EoC models contain traditional machine learning classifiers, such as logistic regression and tree ensemble models.

Thus far, the success of our deep learning models has been modest. This could be due to the limited size of the dataset and/or the limited length of tweets. Regarding the latter, an interesting direction to explore next is training neural networks that can consume the tweets at character level instead of at word level, as we did in this paper.

References

Resham Ahluwalia, Evgeniia Shcherbinina, Edward Callow, Anderson Nascimento, and Martine De Cock. 2018. Detecting misogynous tweets. In *Proc. of IberEval 2018*, volume 2150 of *CEUR-WS*, pages 242–248.

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.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204.

Fabio Fasoli, Andrea Carnaghi, and Maria Paola Paladino. 2015. Social acceptability of sexist derogatory and sexist objectifying slurs across contexts. *Language Sciences*, 52:98–107.

Elisabetta Fersini, M Anzovino, and P Rosso. 2018a. Overview of the task on automatic misogyny identification at IberEval. In *Proc. of IberEval 2018*, volume 2150 of *CEUR-WS*, pages 214–228.

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

Simona Frenda, Bilal Ghanem, and Manuel Montes-y Gómez. 2018. Exploration of misogyny in Spanish and English tweets. In *Proc. of IberEval 2018*, volume 2150 of *CEUR-WS*, pages 260–267.

Dave Gershgorn and Mike Murphy. 2017. Facebook is hiring more people to moderate content than Twitter has at its entire company. `https://qz.com/1101455/facebook-fb-is-hiring-more-people-to-moderate-content-than-twitter-twtr-has-at-its-entire-company/`.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proc. of the 25th International Conference on World Wide Web*, pages 145–153.

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 *Proc. of IberEval 2018*, volume 2150 of *CEUR-WS*, pages 234–241.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proc. of EMNLP*, pages 1532–1543.

Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer.

Ziqi Zhang and Lei Luo. 2018. Hate speech detection: A solved problem? The challenging case of long tail on Twitter. *arXiv preprint arXiv:1803.03662*.