Identification and Classification of Misogynous Tweets Using Multi-classifier Fusion

Han Liu[1], Fatima Chiroma[2], and Mihaela Cocea[2]

[1] School of Computer Science and Informatics
Cardiff University, Cardiff, United Kingdom
LiuH48@cardiff.ac.uk
[2] School of Computing, University of Portsmouth, Portsmouth, United Kingdom
fatima.chiroma@port.ac.uk, mihaela.cocea@port.ac.uk

Abstract. For this study, we used the Doc2Vec embedding approach for feature extraction, with the context window size of 2, minimum word frequency of 2, sampling rate of 0.001, learning rate of 0.025, minimum learning rate of 1.0E-4, 200 layers, batch size of 10000 and 40 epochs. Distributed Memory (DM) is used as the embedding learning algorithm with the negative sampling rate of 5.0. Before feature extraction, all the tweets were pre-processed by converting the characters to their lower case, removing stop words, numbers, punctuations and words that contain no more than 3 characters as well as stemming all the kept words by Snowball Stemmer. Additionally, three classifiers are trained by using SVM with a linear kernel, random forests (RF) and gradient boosted trees (GBT). In the testing stage, the same way of text pre-processing and feature extraction is applied to test instances separately, and each pair of two out of the three trained classifiers (SVM+RF, SVM+GBT and RF+GBT) are fused by combining the probabilities for each class by averaging.

Keywords: Misogynous · Multi-classifier Fusion · Social Media

1 Introduction

Social media platforms have provided users the ability to freely express themselves, however, it has also resulted to increase in cyberhate such as bullying, threats and abuse. A study has shown how 67% of teenagers between the ages of 15 to 18-year olds have been exposed to hate materials on social media, with 21% becoming victims of such materials [11]. Another type of cyberhate that is increasing and worrying is the use of hateful language specifically misogyny on social media platforms like Twitter [3].

Misogyny is defined as a particular type of hate speech that is targeted towards women [1]. In [10], it is stated that online misogyny or abuse is linked to domestic violence against women offline. For example, 48% of women in the UK that have been victims of domestic violence have also been victims of online abuse. Likewise, [9] stated that misogynist abuse as well as threats that are targeted towards many women are amplified due to other social media users joining in for entertainment or to drive out the targeted user.

Therefore, due to increasing evidence showing that cyberhate is increasingly becoming a threat to the society, it has become necessary to implement techniques that can be automated to classify cyberhate so as to reduce the burden on those responsible for public safety. Hence, the aim of this study is to: 1. Identify and distinguish between misogynous and non-misogynous contents; 2. Classify misogynistic behaviour into several behavioural types; and 3. Identify if the target of the misogynistic behaviour is active or passive.

2 Description

The experiment was carried out using misogynous text collected from Twitter [7] which was made available as a training set (with labels) and a separate test set (without the labels).

Due to the noisy nature of social media data [2, 5, 6, 8], it is necessary to rigorously pre-process the data to improve its quality as well as the performance of the classifiers [2, 8]. Therefore, the data sets were pre-processed and classified using different machine classifiers.

Figure 1 shows the experimental processes for this study, while subsequent sections provide a detailed description of the experiment with the results of the classification.

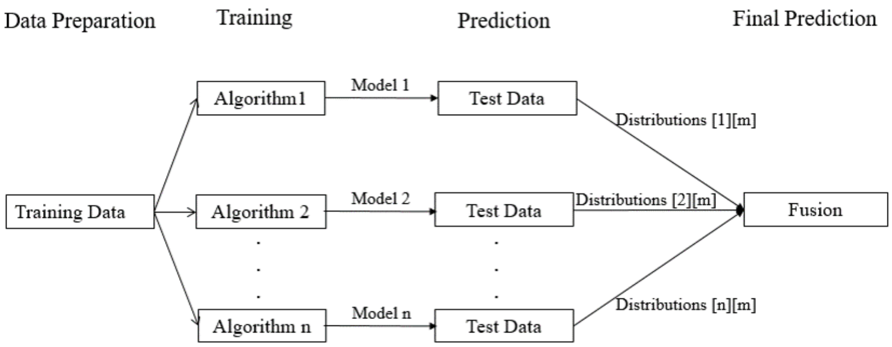

Fig. 1. Experimental Process

2.1 Dataset

As stated in the previous section, the dataset used for this experiment contains misogynous contents extracted from Twitter. Both the training set and the test set contain only English language text with a total of 3, 977 instances as shown in Table 1, which also contains a brief description of the features. Table 2 shows a detailed description of the training set labels with the number of instances for each category.

Table 1. Data-set Description

Dataset	Label	Description	Instances
Training	id	Tweet identifier	
	tweet	Text from twitter	
	misogynous	Misogynous and non-misogynous identifier	3251
	misogynous-category	Misogynistic behaviour types	
	target	Contains the different target classes	
Test	id	Tweet identifier	
	tweet	Text from twitter	726
Total			3977

Table 2. Training Set Instances Per Label

Label	Categories	Instances	Total Instances
Misogynous	0	1683	
	1	1568	3251
Category	stereotype	137	
	dominance	49	
	derailing	29	
	sexual-harassment	410	3251
	discredit	943	
	0	1683	
Target	active	626	
	passive	942	1568

2.2 Pre-processing

The training and test datasets were separately pre-processed and filtered using standard text pre-processing features. User names, URLs and non-ascii characters were removed. The tweets were converted to lower case, and the stop words, numbers and punctuation characters were filtered out. Words that contain less than 3 characters were removed and all the remaining words were stemmed using the Snowball Stemmer in Knime [4].

In addition to the standard pre-processing, features containing these four labels: Id, Misogynous, Misogynous-category and Target were extracted individually using the Doc2Vec Learner. The learner had a context window size and minimum word frequency of 2, while the sampling rate of 0.001, learning rate of 0.025, minimum learning rate of 1.0E-4, 200 layers, batch size of 10000 and 40 epochs. Also, the Distributed Memory (DM) is used as the embedding learning algorithm with the negative sampling rate of 5.0. The extracted labels were exported as tables for classification.

2.3 Classification

The pre-processed training set was used to train five classifiers using the following machine learning algorithms: SVM with a linear kernel, random forests (RF), gradient boosted trees (GBT), Decision Tree (DT) and Naïve Bayes (NB). The 10-fold cross validation approach was used for evaluation and the results of this experiment was used to determine the algorithms that had the best performance among the five machine learning algorithms used.

To obtain the labels for the test set, the highest three performing trained classifiers (determined as described in the previous paragraph) were used: SVM with a linear kernal, random forest (RF) and gradient boosted trees (GBT). These classifiers were paired and each pair of two out of the three trained classifiers, i.e. SVM+RF, SVM+GBT and RF+GBT, were fused using algebraic fusion to combine the probabilities for each class by averaging them. This experiment was executed three times.

3 Results and Discussion

In this section, the results of the experiment for the three runs will be discussed and compared to the results published in [7]. Additionally, the three machine classifiers used were selected based on the results achieved when training the classifiers. Table 3, shows the results achieved for the Misogyny Identification, Misogynistic Behaviour Classification and Misogynistic Target Classification in this study. Table 4 shows the best results obtained on the test set published in [7] and Table 5 shows the results on our approach on the test set.

Table 3. Experiment Results using cross-validation on the training set

	Identification			Behaviour Classification			Target Classification		
Run	*libSVM*	*RF*	*GBT*	*libSVM*	*RF*	*GBT*	*libSVM*	*RF*	*GBT*
1	0.627	0.597	0.606	0.199	0.214	0.238	0.619	0.613	0.618
2	0.623	0.608	0.604	0.186	0.218	0.247	0.614	0.611	0.618
3	0.628	0.606	0.605	0.187	0.230	0.233	0.617	0.623	0.608

The results have shown that an accuracy up to 0.627, 0.247 and 0.623 were achieved for the misogyny identification, misogynistic behaviour classification and misogynistic target classification, respectively. Also, it has been observed that the results in [7] has better accuracy for the misogyny identification and

Table 4. Highest results reported in [7]

Label	Performance
Identification	0.913 (accuracy)
Behaviour Classification	0.292 (macro F-measure)
Target Classification	0.599 (macro F-measure)

Table 5. AMI Experiment Results on the test set: GrCML2016

Run	Identification	Behaviour Classification	Target Classification
1	N/A	0.086	0.270
2	0.525	0.053	0.113
3	0.527	0.065	0.119

misogynistic behaviour classification. We assume that the incompatibility of the features in the training set with the features in the test set had an effect on the performance in this experimental study.

4 Conclusions

In this study, text containing both misogynous and non-misogynous contents were extracted from Twitter. The extracted training set was pre-processed and used to train three machine classifiers. With this text, we were able to achieve an accuracy of 0.624 (misogyny identification), 0.247 (misogynistic behaviour classification) and 0.623 (misogynistic target classification). On the test set, the performance was lower, which we believe is due to the incompatibility between the features extracted from the training set and the ones extracted from the test set - we will further investigate this issue. Additionally, we strongly believe it is empirical to further improve the identification and classification performance for misogynous contents on social media, specifically Twitter, as this can potentially safe lives.

References

1. Anzovino, M., Fersini, E., Rosso, P.: Automatic identification and classification of misogynistic language on twitter. In: International Conference on Applications of Natural Language to Information Systems. pp. 57–64. Springer (2018)
2. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd international conference on computational linguistics: posters. pp. 36–44. Association for Computational Linguistics (2010)
3. Bartlett, J., Norrie, R., Patel, S., Rumpel, R., Wibberley, S.: Misogyny on twitter. Demos (2014)
4. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B.: KNIME: The Konstanz Information Miner. In: Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007). Springer (2007)

5. Burnap, P., Colombo, W., Scourfield, J.: Machine classification and analysis of suicide-related communication on twitter. In: Proceedings of the 26th ACM conference on hypertext & social media. pp. 75–84. ACM (2015)
6. Colombo, G.B., Burnap, P., Hodorog, A., Scourfield, J.: Analysing the connectivity and communication of suicidal users on twitter. Computer communications **73**, 291–300 (2016)
7. Fersini, E., Anzovino, M., Rosso, P.: Overview of the task 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). pp. 57–64. CEUR-WS.org (2018)
8. Haddi, E., Liu, X., Shi, Y.: The role of text pre-processing in sentiment analysis. Procedia Computer Science **17**, 26–32 (2013)
9. 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. ACM (2016)
10. Jane, E.A.: Online misogyny and feminist digilantism. Continuum **30**(3), 284–297 (2016)
11. Perry, B., Olsson, P.: Cyberhate: the globalization of hate. Information & Communications Technology Law **18**(2), 185–199 (2009)