JudithJeyafreeda at SemEval-2023 Task 10: Machine Learning for Explainable Detection of Online Sexism

Judith Jeyafreeda Andrew
University of Manchester

Abstract

The rise of the internet and social media platforms has brought about significant changes in how people interact with each another. For a lot of people, the internet have also become the only source of news and information about the world. Thus due to the increase in accessibility of information, online sexism has also increased. Efforts should be made to make the internet a safe space for everyone, irrespective of gender, both from a larger social norms perspective and legal or technical regulations to help alleviate online gender-based violence. As a part of this, this paper explores simple methods that can be easily deployed to automatically detect online sexism in textual statements.

1 Introduction

The Internet is not an equal space. Over the past few years there has been a rise in concerns about the disproportionate levels of abuse experienced by women in social media platforms. Online abuse can take different forms including bullying, stalking, impersonation, non-consensual pornography, revenge porn or image-based sexual abuse/exploitation, and most commonly, hate speech against women or online misogyny ((Sit, a)). A study ((Sit, a)) has shown that women who experience online abuse often adapt their online behaviour, self-censor the content they post and limit interactions on the platform out of fear of violence and abuse However, despite these there still exists a gap to bridge. (Sit, b) explains the need for identifying and stopping online sexism. One of the reason being: *Online gender-based violence can have significant psychological, social, and economic impacts. Most directly, it affects women's freedom of expression.* Thus in this task we aim at automatically identifying online sexism (gender based abusive statements) by taking advantage of Machine Learning methods. In particular, this work explores certain machine learning methods to identify and classify sexist statements in texts into predefined categories. This constitutes a multi-class classification problem.3.

2 Related Work

Sexism detection can be characterized as hate speech detection. Several works have been done in this area of research [(Yin and Zubiaga, 2021), (Chetty and Alathur, 2018), (Gambäck and Sikdar, 2017),(Andrew, 2021b)]. There are several accounts of sexist content on major platforms such as Twitter, motivating the development of models for better detecting and classifying social media posts. Thus, most works focus on developing methods for identifying and classifying text in social media platforms [(Pamungkas et al., 2020),(Chiril et al., 2020),(Rodríguez-Sánchez et al., 2020)]. Most state-of-the-art methodologies can be summarized with the following methods: (i) methods using lexicons (ii) Deep Learning methods, which are more generic but lack domain knowledge (iii) a combination of lexicons and deep learning methods, which is a hybrid method.

Over the years, several research work have been conducted on this topic using transformers based models. BERT (Devlin et al., 2018) RoBERTa (Liu et al., 2019), Electra (Clark et al., 2020) and GPT2 (Radford et al., 2019) are few of such models that have achieved good results in this area. The authors of (Parikh et al., 2021) develop a multi class classification model of sexist content. This is similar to the second and third sub tasks discussed in this paper. The authors propose a model that uses both the outputs of a BERT model and linguistic word embeddings.

The authors of (de Paula and da Silva, 2022), use transformers for multilingual classification of sexist content. The authors build on the work of (de Paula et al., 2021). The models are developed and tested for the datasets provided by (Rodríguez-Sánchez et al., 2022), where the models were developed for

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1324–1328
July 13-14, 2023 ©2023 Association for Computational Linguistics

both English and Spanish languages. The English language models have a high F1 score, while the Spanish model hasn't fared well.

In this work, several machine learning techniques have been tried using the training and development sets. The algorithm that has the highest accuracy on the development test is used for the test set. This has been previously experimented in (Andrew, 2021a), where the authors attempt to classify YouTube comments in the Dravidian Languages of Tamil, Malayalam and Kannada. The authors come to a conclusion SVM models perform well for two out of three languages. In this paper, the 6 algorithms in (Andrew, 2021a) are trained for the task at hand. In this paper, the Stanford Sentiment treebank is incorporated the algorithms in (Andrew, 2021a) . (Klein and Manning, 2003) describes the Stanford Sentiment treebank which includes a total of 215,154 unique phrases from those parse trees, each annotated by 3 human judges. However the labels from this treebank are different from the ones expected in the task. But these labels can help understand the premise of the sentiment. Thus these are incorporated as features for classification. The authors of (Socher et al., 2013) introduce a Sentiment Treebank with Recursive Neural Networks for fine grained labelling tasks. In this work, simple Machine Learning algorithms are used with the sentiment treebank to explore the extent to which simple techniques can help identify online sexism.

3 Task

The task in SemEval 2023 - Task 10 - Explainable Detection of Online Sexism (EDOS) (Kirk et al., 2023) is to indentify and classify sexist statements. Sub Task A aims at classifying statements into two categories: sexist and non sexist. Sub Task B aims at classifying the sexist comments from sub task A into 4 classes - (1) threats, (2) derogation, (3) animosity, (4) prejudiced discussions. Sub Task C aims at further classification of statements within the class "Threats" from sub task B into Threats of harm and Incitement, encouragement of harm; statements within the class "derogation" from sub task B into Descriptive attacks, Aggressive and emotive attacks, Dehumanisation and overt sexual objectification; statements within the class "animosity" from sub task B into Casual use of gendered slurs, profanities insults, Immutable gender stereotypes, Backhanded gendered compliments, Condescending explanations or unwelcome advice; statements within the class "prejudiced discussions" from sub task B into Supporting mistreatment of individual women, Supporting systemic discrimination against women. Figure 1 shows the classes for each sub task.

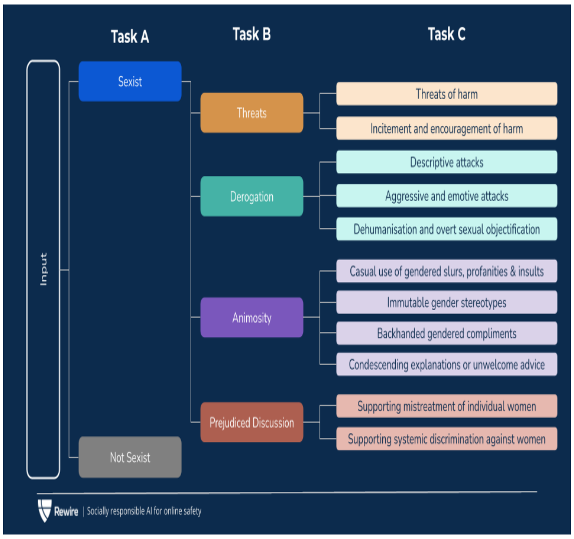

Figure 1: Task Description

4 Data

The data contains contains hateful and sexist statements. The following tables shows some statistics on the number of statements for each sub tasks. Every statement that is classified as "non-sexist" is eliminated from sub tasks B and C.

Train	Dev	Test
14000	2000	4000

Table 1: Sub Task A

Train	Dev	Test
3398	486	970

Table 2: Sub Task B and C

5 Data representation

The bag of words approach is used for text representation. The Term Frequency, Inverse Document Frequency (tf-idf) measure is calculated for each term in the dataset. However, in adiition to this representation, the Stanford sentiment tree bank is used to generate an extra feature. (Klein and Manning, 2003) describes the Stanford Sentiment treebank which includes a total of 215,154 unique phrases from those parse trees, each annotated by 3 human judges. The Stanford Sentiment treebank includes labels for every syntactically plausible phrase. This allows to identify intricate sentiments in texts. In this task, the sentiment output for each statement from the treebank is considered as a feature. Firstly, each statement is run through the Stanford Sentiment treebank, the classification given by the treebank is then converted to a vector representation and added as a feature to the tf-idf representation of the text. This helps to understand the premise of the statement before classifying into the fine grained classes of the task by the following machine learning algorithms.

6 Machine Learning Models

In this section, several machine learning methods are designed for the task at hand. Theses are explained in (Andrew, 2020). The models are Logistic Regression, Naïve Bayes, Support Vector Machines and Random Forests. The algorithms are used on the training and Development sets. The accuracy on the development sets are taken into account. The algorithm with the highest accuracy for

6.1 Logistic Regression

The well established multi-class logistic regression model is implemented for the task at hand (LR, 2017). The model of logistic regression for a multi-class classification problem forces the output layer to have discrete probability distributions over the possible k classes. This is accomplished by using the softmax function. Given the input vector(z), the softmax function works as follows:

$$softmax(z) = \frac{e^z}{\sum_{i=1}^{k} e^{z_i}} \quad (1)$$

At this point, there are k outputs and thus there is a necessity to impose weights connecting each input to each output. The model thus is as follows:

$$\hat{y} = softmax(xW + b) \quad (2)$$

where, W is the weight matrix between the input and output, x being the input and b is the bias.

6.2 Random Forest

Random Forest is a collection of large number of individual decision trees. Decision Trees for samples from the training data sets are constructed. Following this, each decision tree predicts a class. A vote is performed on all predicted result. The class with the maximum vote is decided on to be the output class. For the training process, the random subspace method is used. (i.e) if one or a few features are very strong predictors for the target output, these features will be selected in many of the decision trees. This makes them correlated.

6.3 Support Vector Machines

SVMs are very good classification algorithm. The idea is to identify hyper-planes that will separate the various features. A linar SVM classification decision is performed as follows:

$$f(x) = sign(W^*.x + b^*) \quad (3)$$

where x represents the input feature, W represents the model weight and b represents the bias. For the multi-class classification problem, a one-vs-rest (also known as one-vs-all) approach is used. It involves splitting the dataset into multiple binary classification problems. Thus a binary classification boundary are constructed to train each binary SVMs and the one with the highest confidence is used to solve the multi-class classification problem. As the task at hand in this paper is a multi-class classification problem, the one-vs-rest approach is used.

6.4 Naïve Bayes

Naïve Bayes (Ng and Jordan, 2002) is based on the Bayes theorem. For a given training dataset, the joint probability distribution (P(X,Y)) is learned. When using Naïve Bayes for classification for an input x, the posterior probability is calculated by the classification model. The class with the highest posterior probability is the predicted class.

6.5 Model Selection

The implementation of the models were done using scikit-learn[1] (same as in (Andrew, 2020)).

Table 3 shows the accuracy achieved by each algorithm on the development sets for each sub task. For the test set, the model with the highest accuracy on the development set in each sub task is chosen. However, it can be seen from table 3 that Logistic Regression performs well for all three tasks. Thus this method is used on the test sets.

7 Results

From table 4, it can be seen that the F1 scores for the tasks is not as good as expected. Although the

[1]https://scikit-learn.org/

Model	Accuracy	Sub Task
Support Vector Machine	0.818	A
LogisticRegression	0.821	A
MultinomialNB	0.791	A
RandomForestClassifier	0.757	A
Support Vector Machine	0.780	B
LogisticRegression	0.783	B
MultinomialNB	0.760	B
RandomForestClassifier	0.757	B
Support Vector Machine	0.779	C
LogisticRegression	0.774	C
MultinomialNB	0.758	C
RandomForestClassifier	0.757	C

Table 3: Accuracy of the different models on the development sets.

Sub-Task	F1 score
A	0.5191
B	0.4200
C	0.2128

Table 4: Results on test set

accuracy of the algorithms on the development set seems to be very good (Table 3). The results for binary classification (sub task A) seems decent with a simple Logistic Regression model. Classification into more fine grained classes needs more effort. In this work, the idea was to use simple techniques to study their effectiveness in detecting online sexism. Although, the models fail for fine grained classification, a binary classification can still help spot online sexism. This implies that simple techniques can be put in place by social media organisation to help prevent sexist comments/statements on the web.

References

a. Toxic twitter – the silencing effect. https://www.amnesty.org/en/latest/news/2018/03/online-violence-against-women-chapter-5-5/#topanchor.

b. When social media is sexist: A call to action against online gender-based violence. https://gehweb.ucsd.edu/social-media-sexist-online-gender-based-violence/.

2017. Multiclass logistic regression from scratch.

Judith Andrew. 2020. Judithjeyafreeda@ dravidian-codemix-fire2020:: Sentiment analysis of youtube comments for dravidian languages. In *Forum for Information Retrieval Evaluation*.

Judith Jeyafreeda Andrew. 2021a. Judithjeyafreedaandrew@ dravidianlangtech-eacl2021: offensive language detection for dravidian code-mixed youtube comments. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 169–174.

Judith Jeyafreeda Andrew. 2021b. JudithJeyafreedaAndrew@DravidianLangTech-EACL2021:offensive language detection for Dravidian code-mixed YouTube comments. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 169–174, Kyiv. Association for Computational Linguistics.

Naganna Chetty and Sreejith Alathur. 2018. Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, 40:108–118.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. An annotated corpus for sexism detection in French tweets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1397–1403, Marseille, France. European Language Resources Association.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Angel Felipe Magnossão de Paula and Roberto Fray da Silva. 2022. Detection and classification of sexism on social media using multiple languages, transformers, and ensemble models. In *CEUR Workshop Proceedings, IberLEF 2022*, Coruña, Spain. CEUR.

Angel Felipe Magnossão de Paula, Roberto Fray da Silva, and Ipek Baris Schlicht. 2021. Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada. Association for Computational Linguistics.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the*

41st Annual Meeting of the Association for Computational Linguistics, pages 423–430, Sapporo, Japan. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrew Y Ng and Michael I Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Inf. Process. Manag.*, 57:102360.

Pulkit Parikh, Harika Abburi, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2021. Categorizing sexism and misogyny through neural approaches. *ACM Trans. Web*, 15(4).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Adrián Mendieta-Aragón, Guillermo Marco-Remón, Maryna Makeienko, María Plaza, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2022. Overview of exist 2022: sexism identification in social networks. *Procesamiento de Lenguaje Natural*, 69:229–240.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions.