**Online Detection of Misogyny**

| **Anjali Shahi**  School of Information Studies,  Syracuse University  Syracuse, NY 13210  ashahi@syr.edu | **Nimisha Kadam**  School of Information Studies,  Syracuse University  Syracuse, NY 13210  nikadam@syr.edu | **Ayyappa Bhuma**  School of Information Studies,  Syracuse University  Syracuse, NY 13210  abhuma@syr.edu | |
| --- | --- | --- | --- |
|  |  | |

**Abstract**

The purpose of this project is to detect online misogyny through analyzing sexist content from Gab and Reddit. We have performed binary classification using a baseline Support Vector Classifier and tried different vectorization techniques. Further, we used different sampling methods on the best baseline SVC model, to resolve the issue of imbalanced data. Later, we have used BERT classifier to perform binary classification, misogyny category classification and for fine-grained vector misogyny classification we performed topic modeling to identify if there are any similarities between the categories defined and the topics derived. Through error analysis and evaluation methods we have explored scope of improvements in our classification models.

**1. Introduction**

The nature of social media means that anyone can post anything they desire, putting forward any position, whether it is enlightening, repugnant or anywhere between. However, a small percentage of users utilize those platforms to promote hate speech, increasing the impacts of racism, sexism. This project seeks to identify specific instances of sexism and misogyny in sentences.

“Misogyny” can be defined as hatred or prejudice against women based on their gender or their gender combined with another identity. Online misogyny is being identified as one of the most emerging issues on the web. Hatred against women on online platforms can cause harm to targeted women making these platforms unwelcoming and inaccessible and can lead to gender inequality and injustice against women.

Many automated tools have been created to detect inappropriate sexist content on online spaces, but most of them are modeled just to deploy a generic binary classification with no further explanation. This can cause misleading interpretations of those comments who are wrongly interpreted by the model which can lead to banning of wrong people from these online spaces. The capacity to identify sexist content and to explain why it is sexist increases the interpretability, trust, and comprehension of the choices made by automated technologies, giving users and moderators more control.

Our project aims to provide an in-depth classification within the misogyny category to understand the context behind a comment and identify what type of misogyny it is. The level 1 classification of our model is generic binary classification identifying whether a comment is misogynistic or non-misogynistic. The level 2 classification dives deeper into the misogyny category and identifies the type of misogyny: Threats, Derogation, Animosity, Prejudiced discussion. The level 3 classification provides a fine-grained vector classification of misogyny.

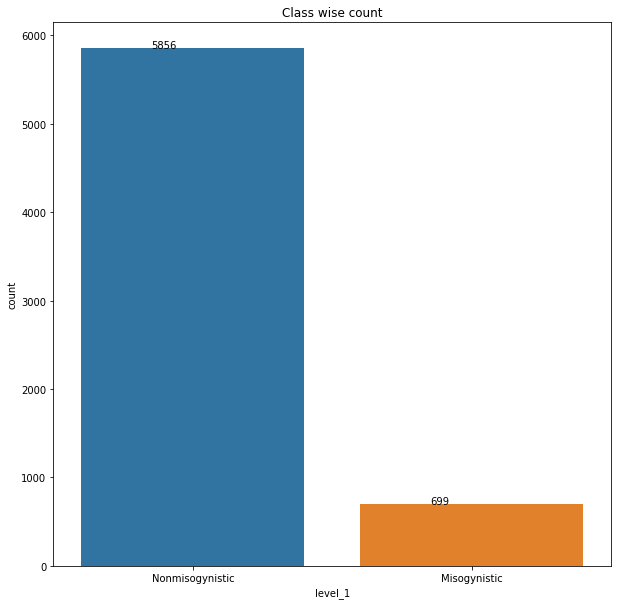
**2. Methodology**

**2.1 Data Cleaning**

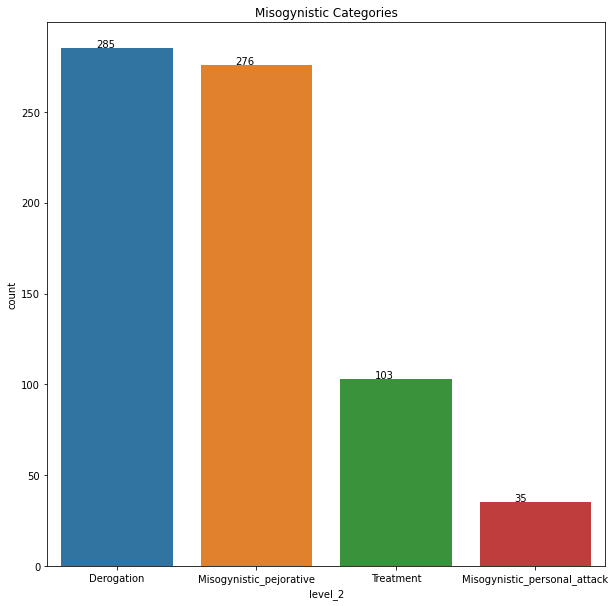
The data is scanned for null values and rows with missing values are dropped. Similarly, punctuation marks and stop words are removed. As we are using topic modeling in level 3 of our analysis hence, stop words are removed as they help in more defined topics.

**2.2 Data exploration**

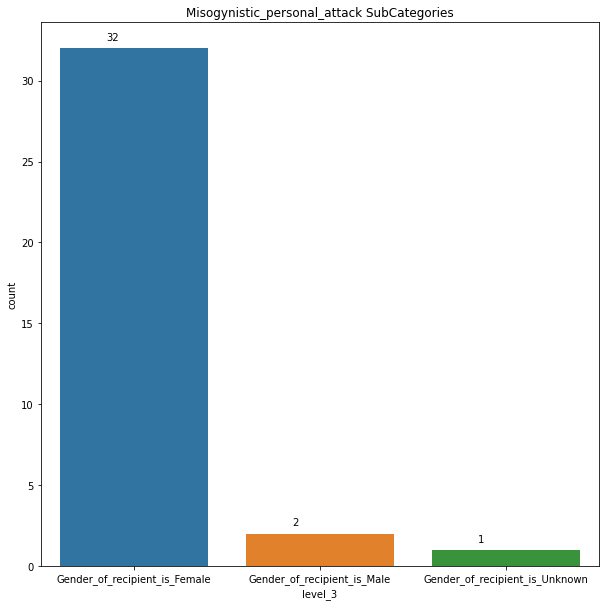
Visualizations are created to understand the distribution of the data on the various levels of analysis to be performed. In level 1, we find the distribution of misogynistic versus non-misogynistic annotated comment threads.



As can be seen in the above figure, the data is imbalanced and hence to avoid bias in modeling results sampling techniques will have to be used.



Further deep diving into misogynistic annotated data we see more levels of classification.



In the above visualization we can see that the majority of the misogynistic threats which are further categorized as personal attacks are directed towards women.

**2.3 Vectorization**

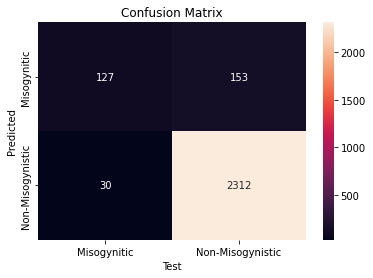
To identify the best vectorization method for the given dataset we compared 3 vectorization techniques by creating a baseline Support Vector Classification (SVC) model.

The three vectorization techniques we compared were:

1. Count vectorizer
2. Bigram vectorizer
3. TF-IDF vectorizer

The vectorizer settings for all the three techniques were, minimum document frequency set to 5 and English stop word removal.

Of these three TF-IDF vectorizer gave the highest cross-validation score of 0.91.



The above confusion matrix for the result of using TF-IDF vectorizer with SVC modeling helped us identify and reiterate two things:

* The imbalance between the 2 classes of data.
* The number of instances when non-misogynistic comments are misclassified as misogynistic is higher than misogynistic comments being classified as non-misogynistic.

Our goal is to further reduce and improve the performance of the classification model in those two aspects.

**2.4 Sampling**

The 2 major categories of sampling techniques are : Oversampling and Under-sampling. To identify the best technique to be implemented for the given data we applied both these techniques and compared the results.

Of the above techniques, Oversampling gave us the best results with a higher precision value compared to the under-sampling.

**3. Data Modeling**

**3.1 Level 1 Classification**

For the first task, determining whether a sentence is misogynistic or not requires predictive modeling. We figured a neural network model would do better because this is a 2-class classification task. We choose to use the pre-trained BERT model because the data we have is limited in size.

**3.1.1 BERT model**

We used the pre-trained BERT uncased model with a sklearn wrapper for this purpose. Link to model: (1)

In order to test the model's performance and maintain all of the data we had, we initially trained it using the original, unsampled data. The model performed well even with skewed data, yielding an accuracy of 92.94%. and F1- macro of 0.775

Later, using the knowledge we had gained from the sampling phase, we sampled the data and utilized it to train the BERT model.The model performed well, resulting in an accuracy of 82.38% and an F1-macro of 0.823.

The decrease in accuracy is expected with sampled data, as the train and test sets have almost the same number of sentences from each class but increase in F1-macro is what we are trying to get from the model. In order to better evaluate the model, we tested it with the test set of original biased data and got an accuracy of 87.23% and F1-macro of 0.763. Though there is a decrease in accuracy and F1-macro, compared to the first model, the second model gave good F1-scores for both the misogynistic and non-misogynistic classes

For both classes, the BERT model with sampled data showed better results, thus we chose to refine its hyperparameters in order to make it even better. The learning rate and epoch number are the parameters for the BERT model.

We fine-tuned the BERT hyperparameters using a grid search, and scored the models with accuracy using a validation set. Our best model has a learning rate of 3e-5 and a total of 6 epochs.

Using the grid search parameters and a sampled data set, we trained the BERT uncased model. When tested on a test data set the model resulted with an accuracy of 85.5% and F1-macro of 0.855. On evaluating the model with a biased test data set we got an accuracy of 88.56% and F1-macro of 0.786.

**3.1.2 Level 1 model results:**

**Accuracy:**

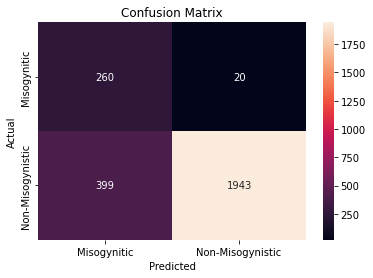
| **Model** | **Accuracy** |
| --- | --- |
| (1)BERT without sampling | 92.94% |
| (2)BERT with sampling | 82.38% |
| (3)optimized BERT | 85.47% |

**F-1 score:**

| **Model** | **F-1 Positive** | **F-1 Negative** | **F-1 Macro** |
| --- | --- | --- | --- |
| 1 | 0.58 | 0.96 | 0.77 |
| 2 | 0.81 | 0.83 | 0.82 |
| 3 | 0.85 | 0.85 | 0.85 |

**3.1.2 Level 1 Error Analysis:**

To understand possible improvements in our classification model, we have performed error analysis over our classification results. The confusion matrix for level 1 classification can be shown as below:



We can infer from the confusion matrix that there were 20 cases where the comment was Misogynistic but was predicted as non-misogynistic. Whereas there were 399 non-misogynistic cases which were predicted as Misogynistic. According to the ethics statement of our project, we have aimed to minimize cases where people are banned from online platforms for those comments which are non-misogynistic but have been predicted as misogynistic. Hence, we have performed error analysis for this use case below:

Actual: Non-misogynistic

Predicted: Misogynistic

* “I was in 2 abusive relationships had ZERO friends and my life consisted of nothing that really made me happy I didn't think any women would ever truly like me my girlfriend at the time hated me…”

* “Ive met some fucking amazing women over the years since then including my current girlfriend of 2 years Shes the sweetest coolest girl Ive met she adds a hell of a lot to my life…”

It can be observed that frequency of profanity and words describing “women” in a sentence is causing the model to infer a non-misogynistic sentence as misogynistic. In the second example it is observed that the use of profanity and “women” is used in a good sense to show appreciation towards women but since our model is terming all comments consisting of slang language and “women” words as misogynistic hence are the wrong predictions.

**3.2 Level 2 Classification**

The next level of classification is to identify the type of misogyny used in a sentence. It has four categories: "Derogation", "Misogynistic pejorative", "Misogynistic personal attack" and "Treatment." As non-misogynistic sentences are not needed for this task, we filtered them out of the data set. The final data for level 2 task is of 699 sentences

**3.2.1 BERT Model**

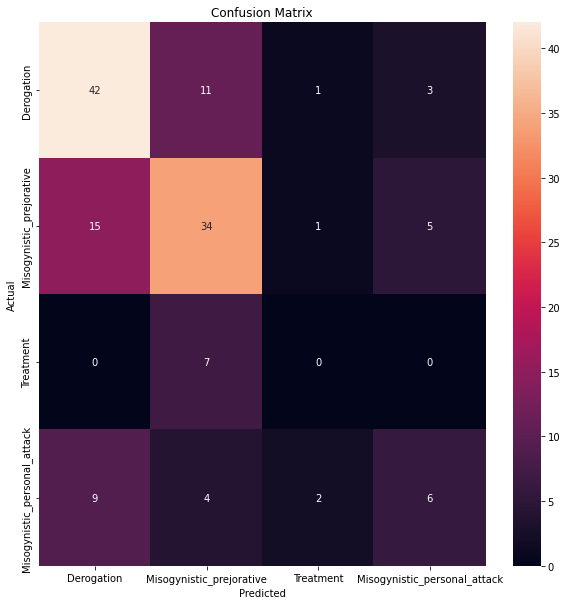
We utilized the best model from the level 1 task to classify the level 2 variable because the BERT model performed well in classifying the level 1 task. The model had an accuracy of 58.57% and an F-1 macro of 0.405 when tested on test data.

Given the small size of the train data, the poor accuracy and F-1 score values are to be expected. Details on where the errors are occuring is explained in the Error analysis section below.

**Results:**

| **Metric** | **Optimized BERT** |
| --- | --- |
| Accuracy | 58.57% |
| F1 - class-wise | Derogation : 68.29%  Pejorative : 61.26%  Treatment : 0  PersonalAttack:34.28% |
| F1 - Macro | 40.95% |

**3.2.2 Level 2 Error Analysis:**

****

The above image describes a confusion matrix for the level 2 classification. The highest number of errors have been computed as Pejorative being the actual case but being predicted as Derogation. We have performed error analysis for this use case below:

Actual : Derogation (disrespectful)

Prediction : Pejorative (expressing contempt / disapproval)

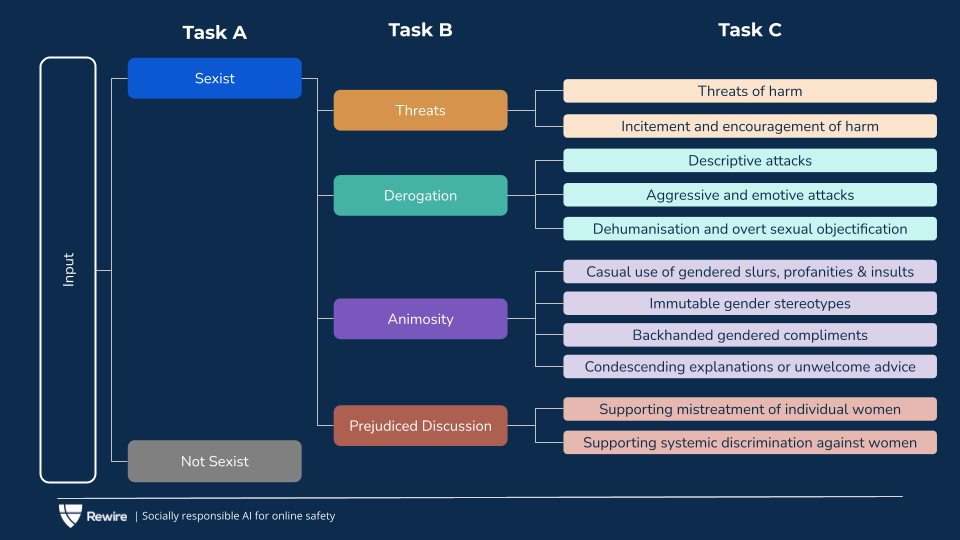
1. “Men are such simple minded humans and dont always Choose friends based on discrimination Females are the exact opposite”.
2. “Online gives room for people voice their true opinion but theres also bias towards the opinions of those who spend more time talking to strangers than they do socializing with friends”

Level 2 classification deals with classifying categories within the misogynistic category. Since the similarity in the definition for these categories is high, even we as humans, were not able to distinguish. Also, lack of data for the misogyny category could be a problem to our classification model.

**3.3 Level 3 Topic Modeling**

Considering the limitations of level 2 classification, for the third level classification we decided to focus on topic modeling inorder to see how well clusters can be formed and analyze the clusters created to identify any possible annotation mistakes in the data.

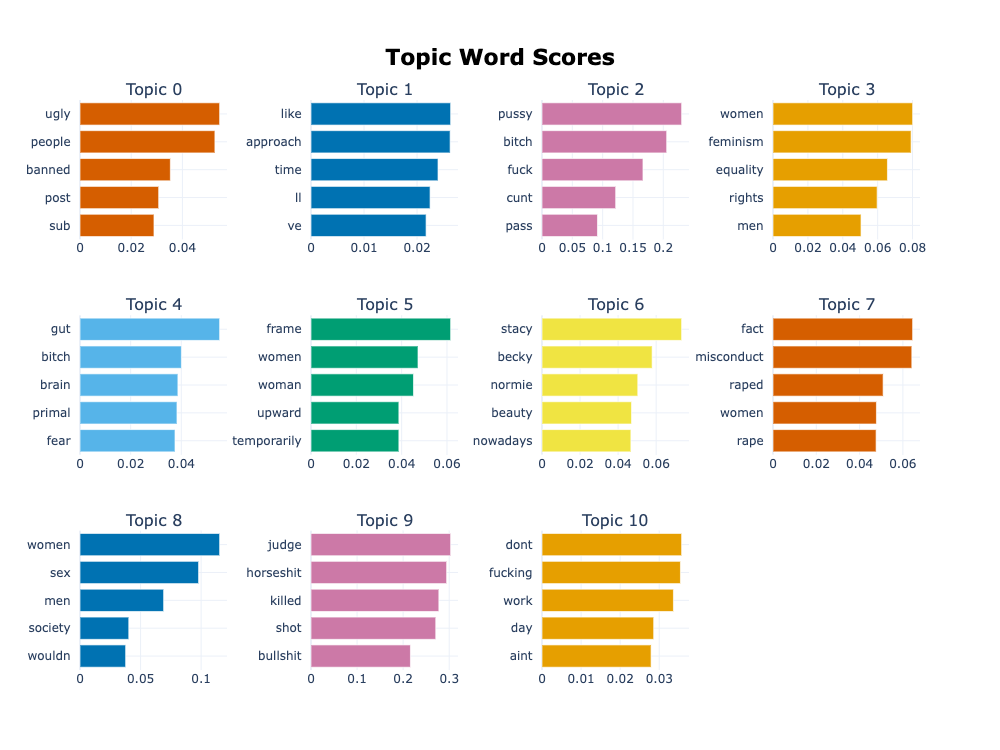
The original dataset consisted of 11 categories on third level of classification that are represented in the figure below:



For the purpose of comparison we also created 11 topic clusters using LDA and BERT topic modeling.

The results of LDA were vague and hard to analyze hence, we went ahead with applying BERT topic modeling as well along with visualizing the cluster results.

The results were as follows:



Comparing the original categories and the computed clusters we see that certain topics can be easily categorized for example; topic 6 denotes the category - ‘Supporting mistreatment of individual women’ and topic 3 denotes category - ‘ Backhanded gendered compliments’. But there are other clusters that are slightly more ambiguous like topic 9, topic 4 which can be classified in more than one categories.

**4. Ethics Statement**

* This NLP model was designed for sexism detection that is more accurate as well as explainable, with fine-grained classifications for sexist content from Gab and Reddit. However, the model cannot ascertain that it is classifying these claims with absolute accuracy.
* There are various ways of phrasing a sentence and sarcastic statements could hinder accurate classification by the model, this could have implications in a real world setting where individuals get flagged or blocked for making such remarks.

**5. Conclusion**

The goal of our project is to understand and narrow the gap between a generic automated classification model and a fine-grained classification model which would consider the context behind a comment and provide explainable misogyny. As mentioned in our ethics statement, throughout our project we have concentrated on minimizing the errors of non-misogynistic comments being wrongly predicted as misogynistic comments to avoid erroneous flagging on social media platforms. Hence, we have looked for a good precision value in each of our model comparisons and have tried to provide possible model improvements based on our analysis.

It is easier for a model to distinguish between misogynistic and non-misogynistic comments but further classification becomes difficult as it is ambiguous even for the humans to distinguish between. Future scope for this project would be to collect data from other online spaces to resolve the issue of imbalanced data and use contextual classification methods to provide better fine-grained classifications.

Developing an algorithm that precisely identifies misogyny in comments will further help owners of various social media platforms create more specific guidelines and policies to deal with the various kinds of harassment and mistreatment.

Additionally, in the current social media climate we see a lot of social media platforms imposing bans on individuals due to propagation of hate speech. While the sentiment is right, greater transparency over the factors enabling these decisions will help provide understanding of the working of existing algorithms.

**6. References**

SemEval :<https://codalab.lisn.upsaclay.fr/competitions/7124#learn_the_details-overview>

Katarya, Rahul. "Analysis of Online Toxicity Detection Using Machine Learning Approaches." *International Conference on Artificial Intelligence and Sustainable Engineering*. Springer, Singapore, 2022.

Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of EMNLP 2002*, 79-86.