Analysis

For our model, we decided to improve upon the logistic regression model that uses a simple bag of words as its only feature. Our final model has a test accuracy of 0.705 with a 95% confidence interval of [0.642, 0.768]. We experimented with different syntactic features, like different kinds of punctuation, along with average word length and number of words. We also tried to use an external hate speech detection library that predicts the likelihood of a text being offensive and hate speech, but it didn’t increase the accuracy. We also manually created small dictionaries that contain words we would associate with being in posts for each type of sexism. We also tried to use a dictionary that does sentiment analysis on specific words based on how positive or negative they are, but this also ended up decreasing test accuracy so we didn’t include it in our final model.

The features that most define the neutral class are words that aren’t related to sexism or include various irrelevant topics that are common in reddit posts but not for sexist posts. For ambivalent sexism, the features most learned to define the class include words that indicate a woman should be docile and reinforce traditional stereotypes like being a stay at home wife. For benevolent sexism, the features most learned to define the class include words that are often used in conjunction to when men objectify women like “body” and “butt”. For hostile sexism, the features most learned to define the class include hate-filled speech often used against women like “toxic”, “feminazi”, and “inferior”. For institutional sexism, the features most learned to define the class include institutions like “workplace” and “law”. For interpersonal sexism, the features most learned to define the class include words that often are used to indicate one is talking about another like “her” and “she”. For internalized sexism, the features most learned to define the class include “I” and “personally” to indicate that the poster is talking about sexist ideals that they hold toward themselves.

Our model makes the systematic mistake of having several sexist posts being misclassified as neutral. This can be due to the fact that at times reddit posts were not blatantly sexist, but had undertones of sexism that makes it difficult for a model to interpret. Predictions for Neutral and Hostile classes are the most accurate, and this can be attributed to the fact that there is a data imbalance. Most of the reddit posts were classified as neutral and the leading category was hostile sexism. Interpersonal sexism is at times classified as internalized sexism, because both categories often involve personal stories in reddit posts. It is difficult at times to distinguish which category a post belongs to, because at times, there are posts that contain several instances of sexism and we have to manually weigh which one is more prevalent.

In terms of label mistakes, neutral often gets mistaken for Hostile Sexism (22 in the confusion matrix) and less frequently vice versa (5). Additionally, Interpersonal Sexism gets mistaken as Neutral (9), Benevolent Sexism as Neutral (5) and Institutional Sexism as Neutral (5). Internalized Sexism gets mixed up for Neutral (4) and Ambivalent Sexism for Neutral (3) and Ambivalent for Hostile Sexism (2).

Looking at some of the mistakes of the model displayed by the analyze function, it seems to have more difficulty correctly identifying posts that have slang words, shortened versions of words, or acronyms in them. For example, one post that uses slang like replacing “business” with “bizz” and “BTW” for “by the way” is incorrectly classified as neutral rather than hostile sexism. Some other acronyms used in posts that are misclassified are “wth”, “BS”, and “LVF” which stands for “low value female.” A lot of posts are also misclassified if they have a question in the text. For instance, one post asks “How do you find out if a woman is hot without makeup before sleeping with them?” and is classified as neutral rather than benevolent sexism. The model may be misclassifying these posts with questions because it may have “learned” that posts with questions are neutral because they’re asking questions about sexism rather than being sexist themselves. This also reveals issues in our guidelines since we want to highlight not just posts where the poster is sexist but also posts that discuss others being sexist or sexism in general as a sexist post. This suggests that the model performs better on posts with a more conventional and “proper” use of language and posts that make more statements rather than asking questions.

The most prevalent label is neutral by far: 133 out of 200 true labels are neutral. This impacts the model because it makes it difficult to make improvements to the simple bag of words model with such a large number of labels being attributed to one class. Many features that we tried like analyzing a text for the probability that it is hate speech and a sentiment analysis that classifies a text as positive or negative lowered the accuracy of our model, possibly because of the unbalanced dataset we used. Oversampling or undersampling may help solve this problem as both would reduce the proportion of neutral labels to the other labels. This would help by increasing the focus on the rest of the minority labels and making the predictions for the different types of sexism more accurate. This would allow for better real world applications since we want to differentiate the different kinds of sexism, not just whether a post is sexist or neutral. Another possibility is to change class weights in the model from None to balanced. This would change the weight of every class from 1 to inversely proportional to their frequencies, which can also help make the predictions for labels other than neutral more accurate.

Furthermore, the model learns features of the phenomenon that we did not consider in the guidelines, such as if it contains a link, it is often labeled as neutral rather than a sexist label. For example, “Feminists don't know what respect means: [[https://twitter.com/Saviour12380391/status/1499859675667337216](https://twitter.com/Saviour12380391/status/1499859675667337216)\n](https://twitter.com/Saviour12380391/status/1499859675667337216%5D(https://twitter.com/Saviour12380391/status/1499859675667337216)%5Cn)” is labeled as neutral even though it attacks feminists. The model may be misclassifying these posts because it may have “learned” that posts that contain links are neutral because the poster can simply share information rather than also having a sexist message. This caused us to rethink the category boundaries because the post may be viewed as neutral because it may be simply summarizing the information in a link rather than choosing a side. Therefore, it may make sense to make any post with links be classified as neutral. Additionally, as mentioned earlier, questions in the text are often mislabeled. For example, “Where do some women get this idea that the men they date are doing “the bare minimum”?: Is it not the men who usually drive, pay for the dates, and try to be chivalrous?.” While the true label was interpersonal sexism, the model labeled it as neutral. This made us begin to rethink the category boundaries because potentially people who ask questions are trying to learn or unlearn sexist beliefs and are therefore not being sexist, but neutral. Therefore, posts with questions might simply be gathering more knowledge and can be neutral.