Can We Predict a Bully?

A Maching Learning Approach to Tweet Classification

Patrick Massey

Abstract

Cyber bullying is a rampant problem facing kids today. As long as they have access to their phones the bullying can follow them wherever they go. In an effort to remove these messages from the public sphere I develop two models that will predict the classification of tweets. The tweets I will be analyzing have been hand classified as a type of cyber bullying or not. The first model I used was a simplistic Naive Bayes, where every unique word in the tweets was a feature. In the second model I use Principal Component Analysis to reduce the number of features and use those components in a Random Forest model. The Random Forest model performed significantly better than the Naive Bayes however both models had trouble predicating `other_cyberbullying` and `not_cyberbullying`. This is due to the vague nature of both categories and the models have trouble discerning between both. However, we have strong predictive power on the other categories. This implies that if we can classify a tweet as one of the specific categories then there is a high probability of being correct, however if it does not fit into the specific categories then it is nearly a 50/50 chance of being correct.

Warning: The following paper deals with cyber bullying, specifically in regard to ethnicity, religion, gender, and age. The topics discussed as well as the tables and figures presented may make some feel uncomfortable.

Introduction

Bullying is a problem that children have always faced, however it used to be the case that there was always somewhere they could escape to. If the bullying was happening at school they were safe once they reached their house. In the era of smartphones and social media this is no longer the case. When children are connected to the internet the bullying can follow them wherever they go in the form of cyber bullying. Twitter is one of the major forms of social media that kids today use to ingest their news, share their interests, communicate with their peers, and potentially even communicate with strangers. This allows for a wide form interaction between individuals including cyber bullying. We can combat that by building a model that can accurately flag a tweet for cyber bullying, and remove the tweet. We will never be able to get rid of cyber bullying completely but if we can make the online space marginally better for at least one child then it will have been well worth the effort.

Methods

In this analysis I utilize a data set of 47,692 tweets that have been identified by a human as cyber bullying or not, and if it is cyber bullying a further classification is made. A cyber bullying tweet can be classified as ethnicity, religion, gender, age, or other. This data set has been artificially created, in the sense that we have a near balanced data set as shown below, with approximately 8,000 tweets for each category. I summarize the counts of the tweets by category in Figure 1, below.

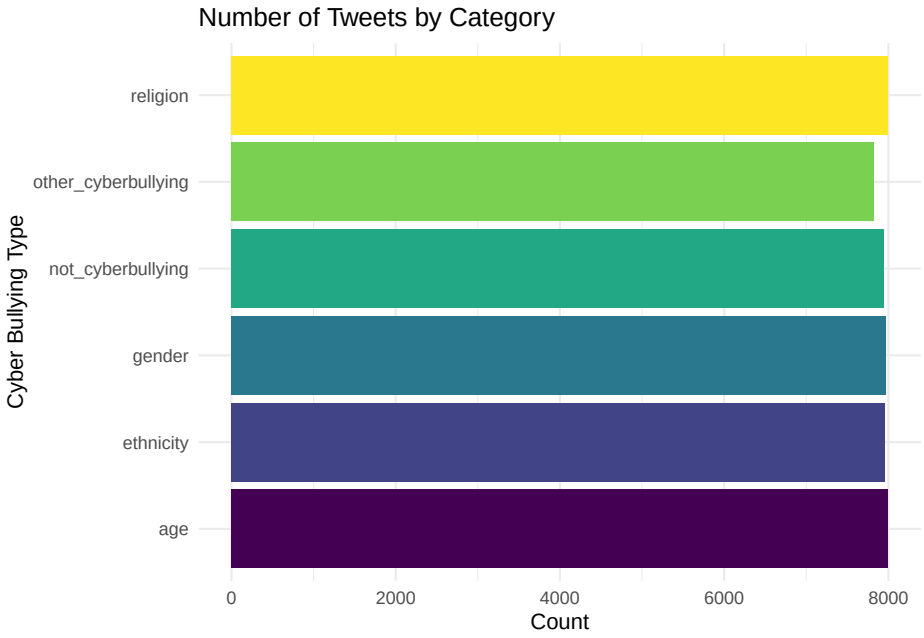

Figure 1: Initial Tweet Distribution

In order to clean the tweets, I remove any punctuation, symbols, emojis, common stop words, as well as additional words that you might see in the cyber world, such as "u" in place of "you". Additionally I also perform a lemmatization of the tweets which will transform words like "sell","selling", or "sold" into one representative word "sell". This will help increase accuracy of the model. After processing the tweets and removing any tweets that may now be completely empty we see in Figure 2 below that most of the removal happens with the `other_cyberbullying` and `not_cyberbullying` categories, meanwhile the others remain relatively unchanged. This makes sense as these two categories are the most likely to have vague and unimportant words.

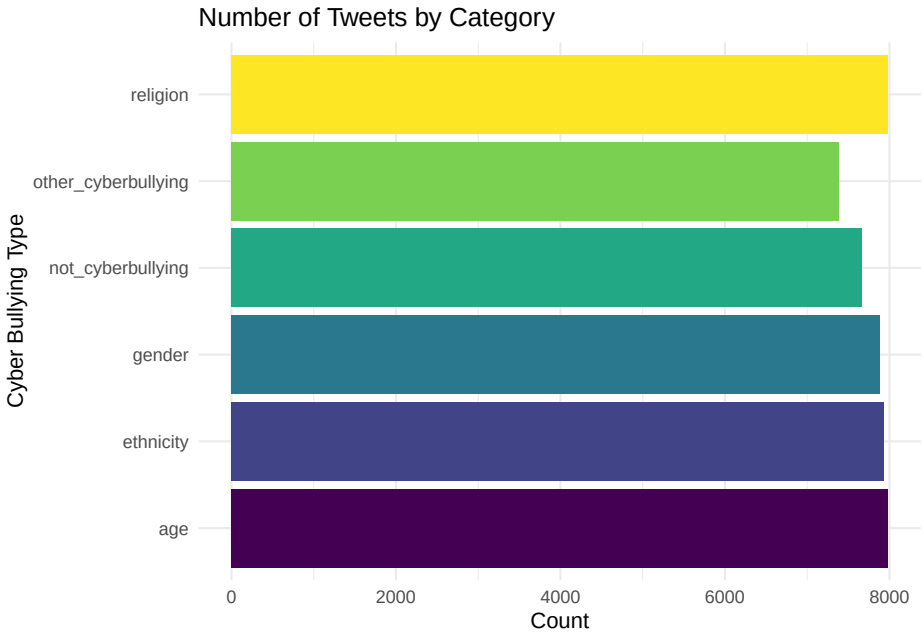

Figure 2: Tweet Distribution After Cleaning

Before getting into the analysis, it is important to know the words that make up the tweets being analyzed. This also helps further understand the motivation. In Figure 3 below I have generated a word cloud based on tweets that have been classified as religious cyber bullying. I have also generated word clouds for the other categories which will be located in the appendix at the end.

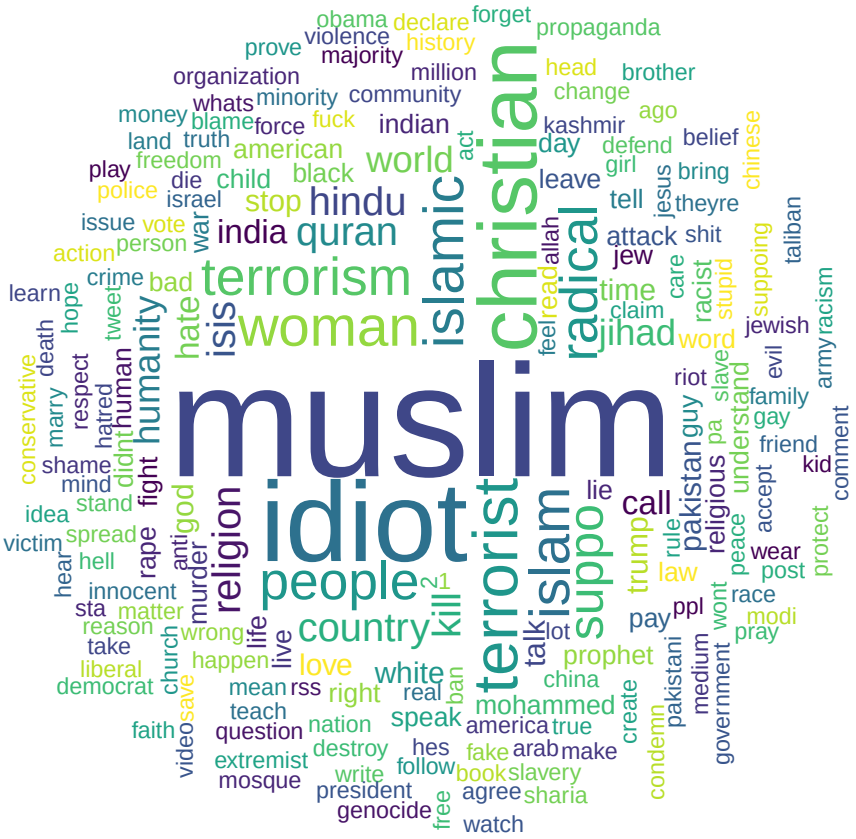

Figure 3: Cyber Bullying Tweets: Religion

Now that we have an idea of what the distribution and contents of the tweets look like, we can now move on to the analysis. In order to know what methodology we need to use to analyze the data, we need to clarify the problem. Here we will be doing text analysis on a multinomial classification problem. Which is to say we are going to analyze the contents of the tweets and attempt to predict one of six categories the tweet could be classified as.

When considering the model to use for this analysis, I simply asked the question of what the goal was. Here our goal is "given a combination of words, what is the classification of this tweet?". This points directly towards using a Naive Bayes model. The Naive Bayes model is an extremely popular model for doing text analysis, because of its simplicity, but also still being fairly accurate. It is especially powerful in the case of multinomial classification. Its simplicity comes from the main assumption that the features being used are independent of each other. This is probably a bit too simplistic, given a tweet is somewhat of a sentence, the words are likely not independent of each other. With the removal of stop words this may help with the independence assumption as we are now only considering words outside of filler words. In addition to the Naive Bayes model since I am not convinced that words are independent of each other, I will also be using Principal Component Analysis (PCA) in combination with a Random Forest model. Here the PCA is being

used to reduce the number of features into a select number of components, I will be discussing more on the features being used later on. The Random Forest is also extremely powerful when it comes to classification problems, but does not require such strong assumptions like the Naive Bayes.

In order to use either of these models the tweets must be in a format that is usable. I first turn the tweets into a corpus and then transform the corpus into a Document Term Matrix. The tweets have now been broken apart and for every word that appeared in a tweet it is now a feature of the data set. Every row represents a tweet and every column is a word and the values are the frequencies that each word appears in a tweet. This gives a total of 40,166 features to work with. However, there are likely some words that only appear in a handful of tweets and are likely not very useful in the analysis. After inspecting the documents, it is an extremely sparse matrix. I remove terms that are sparse in 99% of the tweets which is to say that a word must appear in about 400 tweets in order to remain. After removing sparse terms we are left with 102 features. Now that we have done the feature engineering, I split the data set into training and testing sets. Using the same training set I will train both the Naive Bayes, and Random Forest model and then test on the testing set. As stated earlier for the Random Forest model I will be using PCA to reduce the number of features used. To determine the number of components to use I create a scree plot in Figure X, from the plot I see that past 15 components there is not much gain in variance explanation. Because I am not gaining much past 15 components, that is the number of components I will use for this analysis. This will reduce the number of features from 102 to 15.

Results

After training the models and then testing them against a separate testing set, I summarize the results in confusion matrices in Figures 4 and 5 below. On the Y axis I have plotted the observed classification, and on the X axis I have plotted the predicted classification using the relative model. Along the diagonal is the true positive matches between the actual and predicted values. I also summarize the overall accuracy of both models in Table 1 below.

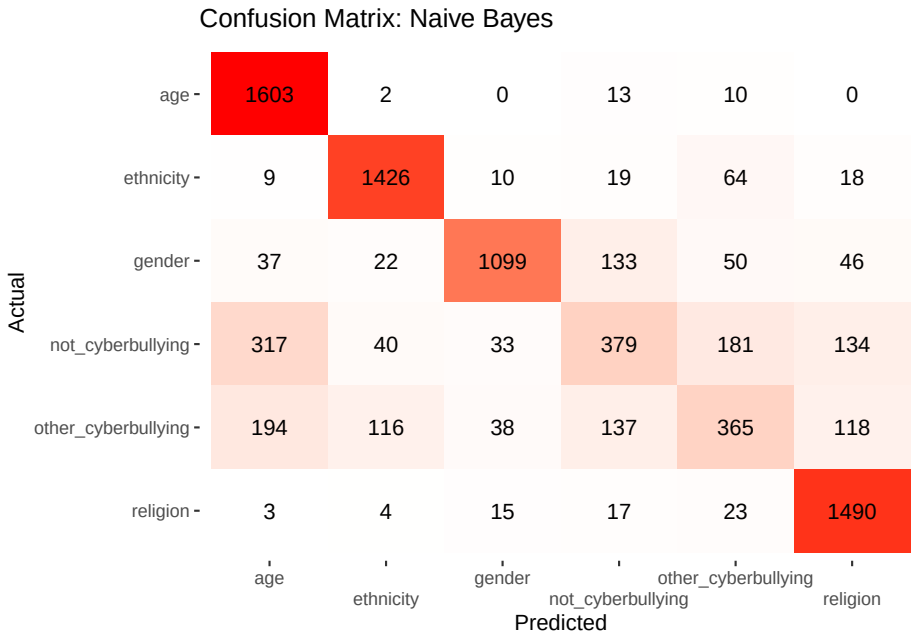

Figure 4: Confusion Matrix Heatmap for Naive Bayes

From Figure 4 above we see strong initial results from the Naive Bayes. There is an overall accuracy of 0.779, and I break down the performance results by class in Table 2 located in the appendix. What we see is

that in both the confusion matrix and Table 2, the model is not a strong predictor for `not_cyberbullying` and `other_cyberbullying`. The true positivity rate for those is 0.543 and 0.527 respectively. Lets see if we can improve upon this using a Random Forest with PCA. Figure 5 below summarizes the performance in the same confusion matrix format as before and Table 3 summarizes the performance results by class.

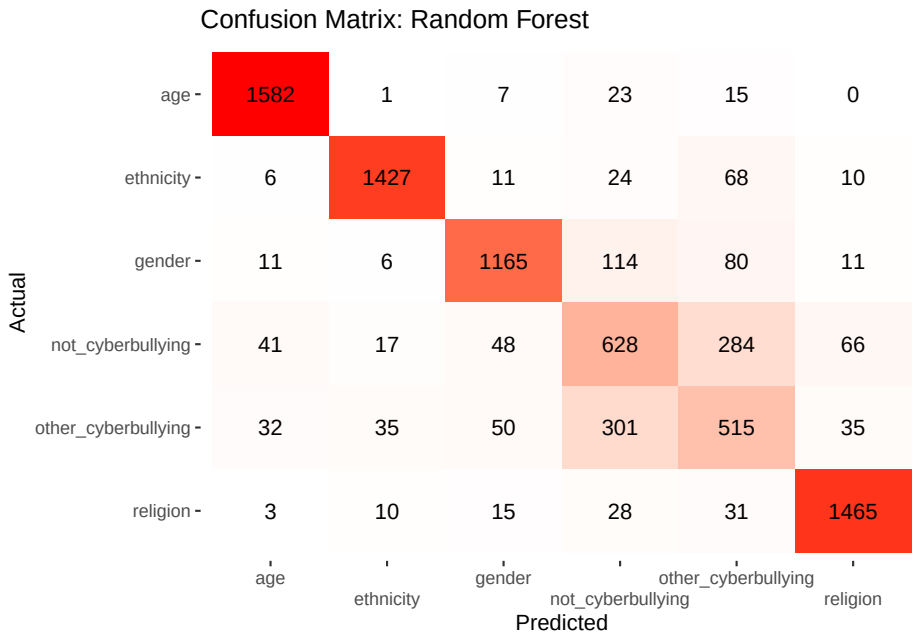

Figure 5: Confusion Matrix Heatmap for Random Forest using PCA

Table 1: Overall Accuracy

	Naive Bayes	Random Forest
Accuracy	0.779	0.831

From Figure 5 and Table 3, we can see there is an improvement in the performance, with an overall accuracy of 0.831. From Table 1 what we see is that the Random Forest provides a 6.6 percent increase in accuracy. This is a sizable increase in performance. However once again we see that the Random Forest is also not a strong predictor for `not_cyberbullying` and `other_cyberbullying`. The true positivity rate for those is 0.562 and 0.519 respectively. So although we see a large increase in other classes of tweets and improvement overall, there is still not much change in the problematic categories.

Conclusion

In this analysis I have used two models to predict the cyber bullying classification of a set of tweets. The first model was a simplistic Naive Bayes model where each word was a feature. The second model used PCA to reduce the number of features to 15 components and then I used those components in a Random Forest model. The results show strong predictive power for `age`, `ethnicity`, `gender`, and `religion`. It should be noted that due to the balanced nature of the data set used in this analysis we would very likely not see such strong results when using live scraped tweets. The Random Forest performed significantly better than the Naive Bayes, albeit the Naive Bayes is a simpler and easier to implement model. The improvement in overall accuracy is significant enough to warrant the more complicated Random Forest model. However, when it comes to `other_cyberbullying` and `not_cyberbullying` both models performed roughly the same,

and it essentially is a 50/50 chance of being correct. This is understandable especially once Figures 9 and 10 are considered. Figures 9 and 10 show a word cloud of `other_cyberbullying` and `not_cyberbullying` respectively. The same words appear in both figures and at a similar frequency. This will make it difficult for any Machine Learning model to discern between the two, as it seems the only thing that makes them different is the context in which they appear. In a simplistic example we may see two tweets "Fuck yeah!" and "Fuck you!", the former is an obvious positive encounter, and the latter is a negative one. After processing a model would only see "Fuck" and "Fuck", it would be impossible for a machine to discern between the two. That's not to say that it is pointless to try to predict a tweets classification. We see strong predictive powers for the specific cyber bullying classifiers, so if we can classify a tweet as a specific category then there is a high probability of being correct. However, if a tweet can not be classified as a specific category, then it would require a human to understand the context.

Appendix

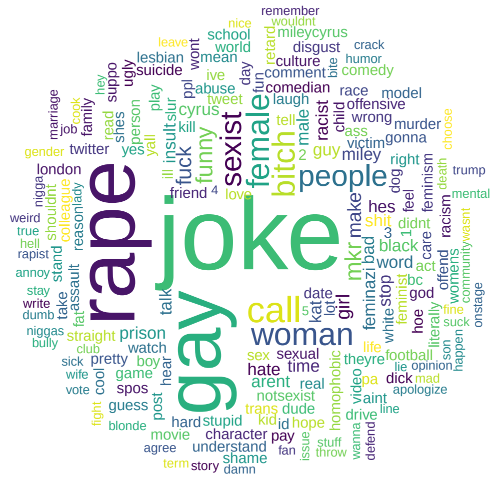

Figure 6: Cyber Bullying Tweets: Gender

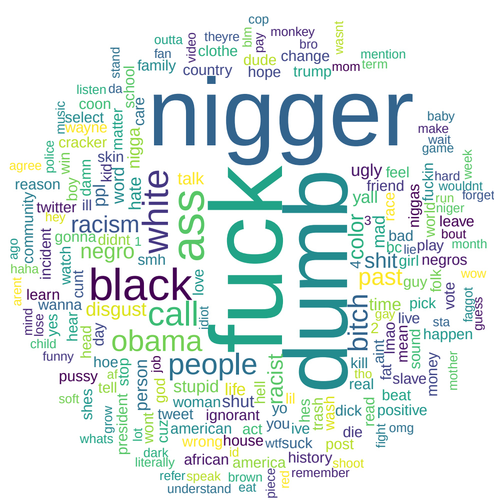

Figure 7: Cyber Bullying Tweets: Ethnicity

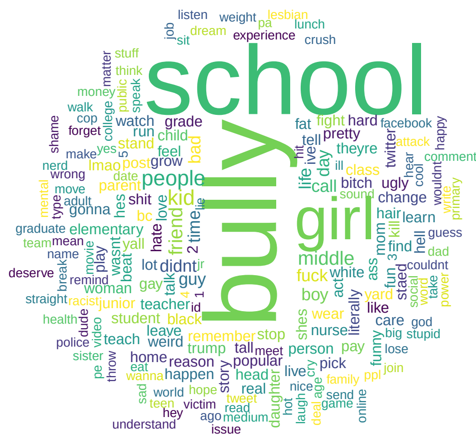

Figure 8: Cyber Bullying Tweets: Age

Figure 9: Cyber Bullying Tweets: Other Cyber Bullying

Figure 10: Cyber Bullying Tweets: Not Cyber Bullying

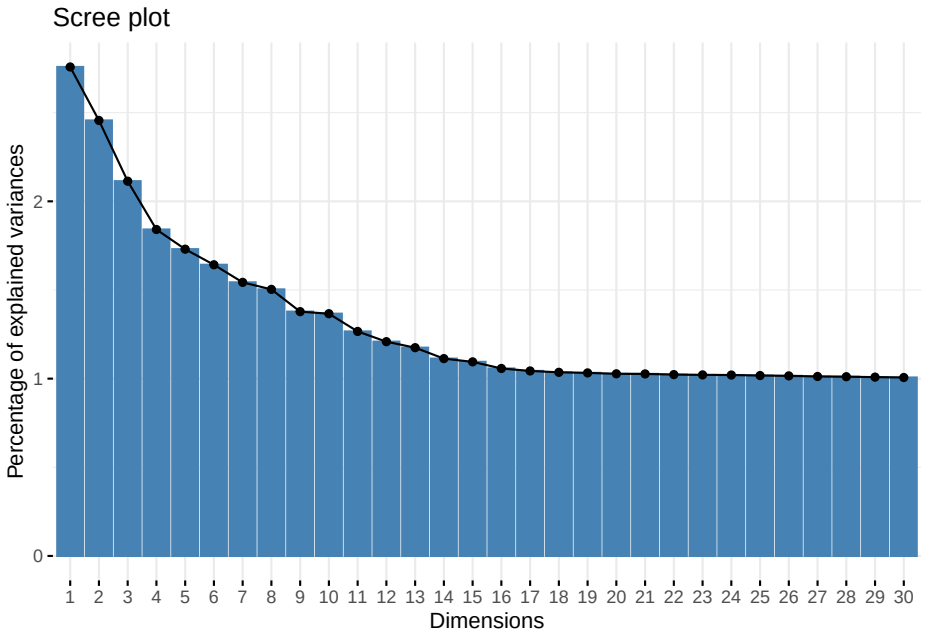

Figure 11: Scree Plot for PCA

Table 2: Naive Bayes Performance Measures by Class

	Sensitivity	Specificity	Precision
Class: age	0.741	0.996	0.985
Class: ethnicity	0.886	0.982	0.922
Class: gender	0.920	0.959	0.792
Class: not_cyberbullying	0.543	0.906	0.350
Class: other_cyberbullying	0.527	0.919	0.377
Class: religion	0.825	0.990	0.960

Table 3: Random Forest Performance Measures by Class

	Sensitivity	Specificity	Precision
Class: age	0.944	0.993	0.972
Class: ethnicity	0.954	0.982	0.923
Class: gender	0.899	0.968	0.840
Class: not_cyberbullying	0.562	0.935	0.579
Class: other_cyberbullying	0.519	0.937	0.532
Class: religion	0.923	0.987	0.944