Banned Versus Not Banned Book Analysis

Alison Reikher, Samantha White, and Natalia Uruska

IST 664

6/13/2022

**Introduction**

Across the country, thousands of books are banned or challenged every year. Some bans might be justified but most would be considered unreasonable. One banned book that surprises many readers is the book *Looking for Alaska*. According to the American Library Association, back in 2015, *Looking for Alaska* was the most banned book in the country. For perspective, *50 Shades of Grey*was the second most banned. The justification for this ban was for “offensive language, sexually explicit, and unsuited for age group” (Admin, *Top 10 most challenged books lists* 2022).

The author of *Looking for Alaska*, John Green, has written many other books for the same audience of teens and young adults that are not banned and challenged. The popular book *The Fault in Our Stars* is one of Green’s most successful novels to date. While this book also contains content like offensive language and sexually explicit content, it is not on the banned book list. The goal of this report is to analyze both texts to determine what differences exist between the two texts along with understand how much of the text would be considered offensive language. A profanity checker can be used to analyze the text. The checker produces a binary result of whether a given word is considered offensive based on an established dataset of offensive words. The checker also provides the probability any given word could be interpreted as offensive. Additionally, sentiment analysis can be used to compare the sentiment of the two books as well.

**Text Processing**

The first step to process the data is to import the text files. The text was initially converted from a PDF file to a Word document to a text file. It was crucial that nothing important was lost in this process. Initially, when the text was imported, symbols were added in place of certain characters like apostrophes or commas. The text had to be reprocessed and the Word document needed to be converted into a text with utf-8. When this issue was resolved, the text was able to be imported correctly.

After the text was imported, the words needed to be tokenized appropriately. There were several methods used to tokenize and separate the words. The first method was to tokenize the text using the nltk.sent\_tokenize and nltk.word\_tokenize. This creates a list of lists where each list represents each sentence in the document.

For the words to be used in the profanity checker, this list of lists needed to be flatted. A function was created to take all the content from each list and combine it into one list. With this list, all the punctuation needed to be removed to get an accurate word count.

The last text processing method was removing all stopwords. This was done using nltk.corpus.stopwords. The words are converted into all lower case. Then they are filtered for alpha words. These are the words that don't contain any non-alphabetic characters. Then the stopwords are removed. There were no special stopwords that needed to be removed from these documents. Just the standard ones that are imported from the library.

**Exploring the Books**

To further explore the text, profanity checkers were used in determining how much offensive language was used in both texts to determine the differences. The first checker function is called predict. The predict function returns a NumPy array of ones and zeros. If a word in the list is considered profanity or offensive, it will be counted as 1. Otherwise, it is counted as 0. This array can be summed to determine how many words are considered offensive in the entire text.

In *Looking for Alaska*, there are 863 offensive (not unique) words out of 76,753 words total. This means that 0.0112 or 1.12% of the text is offensive. In *The Fault in Our Stars*, there are 1,034 offensive (not unique) words out of 71,720 words total. This means that 0.0144 or 1.14% of the text is offensive. While this checker isn’t perfect, it does demonstrate here that both books have very low levels of offensive language, and The Fault in Our Stars actually has more offensive language than *Looking for Alaska* despite not being banned.

When looking at unique words, there are 196 unique offensive words and 8,576 total unique words in *Looking for Alaska*. The percentage of unique words that are offensive is 0.0228 or 2.28%. There are 167 unique offensive words and 7,982 total unique words in *The Fault in Our Stars*. The percentage of unique words that are offensive is 0.0209 or 2.09%. While this shows that *The Fault in Our Stars* has slightly fewer unique offensive words than *Looking for Alaska*, it is not significant enough to label one book as more offensive than the other.

  Another function used to predict the probability that a word is offensive is called predict probability. This function returns a NumPy array of probabilities based on how each word might be perceived as offensive. For instance, in the *Looking for Alaska* dataset, the word with the highest probability at 0.99999 or 99.99% is the word ‘fuck’. The word with the lowest probability at 0.0009 or 0.009% is the word ‘saint’. And a random word like ‘my’ has a probability of 0.04608 or 4.61%. The average of these probabilities can be determined by summing both arrays and dividing them by their respective length. The probability that any given word would be considered offensive for *Looking for Alaska* is 0.06302 or 6.30%. The probability of any given word would be considered offensive for *The Fault in Our Stars* is 0.0634 or 6.34%. Once again, the profanity checker shows that the difference between these books based on being perceived as offensive is extremely similar. Also, it is again shown that *The Fault in Our Stars* would most likely be perceived as offensive over *Looking for Alaska* despite the latter being the banned book.

Additionally, tagging each document using NLTK’s Sentiment Intensity Analyzer, both *Looking for Alaska*and *The Fault in Our Stars* had similar percentages of positive, neutral, and negative words showing that there is not a big difference in sentiment between the books. Specifically, *Looking for Alaska*was tagged as 9.2% positive, 79.4% neutral, and 11.4% negative. *The Fault in Our Stars* was tagged as 9.2% positive, 77.9% neutral, and 12.9% negative. *The Fault in Our Stars* has slightly more negative keywords than the book that is banned, *Looking for Alaska*.

Word Clouds were created for both *Looking for Alaska* and *The Fault in Our Stars* with the most frequently used words in each text. The purpose of these word clouds is to help visually see which words are used most frequently. From these word clouds, it is shown that the most common words in both books are the word ‘said’ and the name of the main characters.

A picture containing graphical user interface

Description automatically generated

Word Cloud for Looking for Alaska

Text

Description automatically generated

Word Cloud for The Fault in Our Stars

**Feature Engineering and Naïve Bayes**

The processed corpus was then used to train two Naïve Bayes models. The first model aims to predict whether a word is considered profanity, while the second model predicts sentiment. Both models use a feature set that consists of the most common words and most common profanity words. In this feature set, the value for each word is a Boolean value. True means that the word is included in the corpus, while false means that it isn’t.

After creating the feature sets, the documents were randomized and combined. Train and test groups were then created to train and test the models. After testing the models on the test set, we found that the profanity model had an accuracy score of about 0.99, and the sentiment model had an accuracy score of about 0.84. After some deeper digging, it seems that the Naïve Bayes model was not efficient at predicting profanity or sentiment on these corpora because it tagged every word as “profanity” and “neutral.”

Lastly, a 5-fold cross-validation technique was used to gain a better accuracy score for the sentiment model. Cross-validation is a useful method to train the model on 5 different, equally sized, pieces of the document, and every part of the document is accounted for in the training of the model. In this case, accuracy stayed the same at around 0.84 because it most likely is still predicting all neutral sentiment. Because of this, when calculating precision, recall, and f1 for negative and positive predictions true positive and false positive will always be 0. This means that when calculating these metrics, the code had to account for the possibility of division by 0. To combat this issue, recall, precision and f1 will equal 0 if the result cannot be calculated.

After implementing the 5-fold cross-validation, it was found that the average accuracy was 0.84, as stated above. Average recall, precision and f1 were 0.28, 0.33, and 0.92 respectively. Recall means that there were on average about 0.28 true predictions out of the “positive” predictions (not to be confused with positive sentiment). Precision means that on average 0.33 of the “positive” predictions were true. Lastly, the F1 of 0.92 represents a combination of the two. F1 is considered high, however, this is most likely because most words in the corpus were neutral, to begin with.

**Additional Experiments**

To further the analysis, additional experiments were performed and focused on the overlapping profanity in both *Looking for Alaska* and *The Fault in Our Stars*. Since the books share the same author, John Green, it was interesting to find how many profane words and “clean” words overlap between the two books. After processing the tokens further by removing punctuation, ensuring everything is lowercase, and removing all duplicates, it was found that there are 79 unique profane words that are the same in both books. The same steps were taken to process the overlapping “clean” words, but all profane words from both *Looking for Alaska* and *The Fault in Our Stars* were removed from that list. After the processing, it was found that 2,916 “clean” words overlapped between the two books.

To further explore sentiment among these filtered lists, another one of NLTK’s Sentiment Analysis Lexicon was used called SentiWordNet. SentiWordNet is an opinion lexicon that is derived from the WordNet database. Each term is associated with a numerical score that indicates positive, negative, and neutral (objective) sentiment. The ratings provided are on a scale between 0 and 1. It is another valuable resource for performing opinion mining tasks.

As a first step in prepping for the SentiWordNet analysis, part of speech (POS) tagging on the word lists using PennTreebank tags was performed. Each word from the profane and clean lists was tagged as ‘NN’ (noun), ‘JJ’ (adjective), ‘VBG’ (verb), or ‘RB’ (adverb). Since SentiWordNet works with WordNet synsets, the score is assigned based on word meaning. The tags were then converted from PennTreebank tags to simple WordNet tags. WordNet Lemmatizer function was then called to bring context to the lists of words, and used within the function we created to return the list of positive, negative, and neutral (objective) scores for each word. There is a caveat though that if the word is not present in SentiWordNet, an empty list will be returned instead of a score. After running the function on both word lists and checking for any overlap, two lists of scores were created, one for the profanity and one for the clean words.

To get a better idea of what the sentiment in the overlapping lists looks like after running SentiWordNet, two Pandas dataframes were created for easier manipulation. Columns were created for the word, and its positive, negative, and objective scores. After an initial inspection, a few NAs were found. As previously mentioned, if the word is not present in SentiWordNet, blank list will be returned, hence why multiple NA values appeared. For this analysis, all the aforementioned words were removed which had no scores associated with them. Next, a function was created that automatically assigned a sentiment to the word based on its score. The sentiment which had the highest score was respectively labeled in a separate column as “Positive”, “Negative”, or “Neutral”.

After words in both dataframes were properly labeled with their corresponding sentiments, a high-level overview of what the sentiment in each of them looks like was created through Seaborn. Seaborn was imported and used to create a stacked bar chart showcasing the sentiment. Figure 1 below shows the overall sentiment for the data frame with the profane words, while Figure 2 shows the overall sentiment for the data frame with the clean words. Figure 3 shows the sentiment with both of the data frames combined.

Chart, bar chart

Description automatically generated

Figure 1. Overall sentiment for profane word data frame

Chart, bar chart

Description automatically generated

Figure 2. Overall sentiment for clean word data frame

Chart, bar chart

Description automatically generated

Figure 3. Overall sentiment for both profane and clean word data frames combined

As pictured in the charts above, the data frame with the profane words surprisingly had more neutral sentiment than it did negative. Unsurprisingly though, there is no positive sentiment at all. One thing to be mindful of is that the data set is only focused on words that overlap, it has been processed, and therefore it is much smaller than the profane word lists that was found at the very beginning of the analysis. In the clean word dataframe, mostly neutral sentiment was found as well, with amounts of both negative and positive sentiment being much smaller in comparison. This data set was much larger, as unsurprisingly there were more clean words than profane words overlapping between the two books.

**Conclusion**

Out of the thousands of banned and challenged books, many vary greatly in terms of the content and reasoning for the ban. Some may be considered extremely inappropriate for just about anyone, whereas others may only be offensive to certain audiences. When looking at John Green’s *Looking for Alaska* and *The Fault in Our Stars*, it is apparent that both books have offensive language. In fact, both books have similar levels of sentiment as well. Despite both books being similar, only one is consistently being banned more than the other. All analyses indicated that the overall sentiment towards the majority of the words was mostly neutral. Since some profanity was present in both books, the ban can be partially justified as these books may be deemed inappropriate for younger age groups, but likely not for adults. It could be seen as hypocritical, however, to ban one and not the other at the same degree. Both books have the same target audience. No justification was found in the analysis for one text to be more or less banned than the other.

**Group Breakdown**

The process of completing this project was a group effort. All group members agreed to the data set and the allocation of work. The process for how each step would be complete was also agreed upon by all members. Alison did the text processing, initial text exploration using the profanity checker, and created the word clouds. Samantha calculated sentiment using Sentiment Intensity Analyzer. She also trained two Naïve Bayes models. Natalia performed additional experiments using the overlapping words from both texts. She also did parts of speech tagging and furthered the analysis of sentiment analysis. Writing the report was evenly divided among the group members.

**References**

Admin. (2022, April 4). *Top 10 most challenged books lists*. Advocacy, Legislation & Issues. Retrieved June 2, 2022, from <https://www.ala.org/advocacy/bbooks/frequentlychallengedbooks/top10>