Amelia Li, Matthew Moon, Francesca Song

Annotation Project Analysis

Model Selection

We used BERT to implement our predictive model and achieved an improved performance of 0.600 for our test accuracy, with a 95% confidence interval of [0.532, 0.668]. We aimed to utilize BERT's sophisticated context-understanding and homonym differentiation abilities to analyze our texts.

Shortcomings

A major difficulty we had during our annotation process was with "slang" terms. In our annotation guidelines, one of the categories that we defined was difficulty of vocabulary. Documents with high-level terms would naturally be seen as being intended for tertiary audiences; similarly, documents with low-level terms would be more attune to primary audiences. Our set of documents contained mostly fiction novels from Project Gutenberg, spanning a wide variety of time periods, therefore, we had to deal with a wide pool of slang terms while classifying our documents.

The nature of slang terms is highly subjective since the words could be quite easy to speakers from that time period but unintelligible to current readers. We also received some feedback highlighting this fact. In order to deal with this complex issue, we chose to annotate documents with the mindset of a reader from our current time i.e. 2022. However, this decision could have impacted our results negatively, especially when considering that children's classics such as *Peter Pan* and *Alice in Wonderland* were written in the early 20th century. Depending on how we choose to view slang terms, the guidelines of our categories could shift.

Similarly, novels often make note of current events or moments in history specific to that time period. While this may be common knowledge even to elementary school students of that time, we cannot expect current elementary school students to be well-versed in events from the early 1900s, which creates a similar problem as with slang described above. Again, we chose to annotate documents with the mindset of a reader now in 2022.

Another difficulty that we experienced was with the length of documents. Since we do not have the capacity to run models on the full novels, we chose to scrape the first few paragraphs of each novel when creating our documents. This technique, admittedly, is subjective. Some novels have longer paragraphs; others have shorter ones. We had to make decisions for each document regarding how much of the text to incorporate and this could have opened up room for bias and error. While we debated simply pulling the first 300 words of each document, this could have cutoff text mid-sentence which we felt would negatively impact our results.

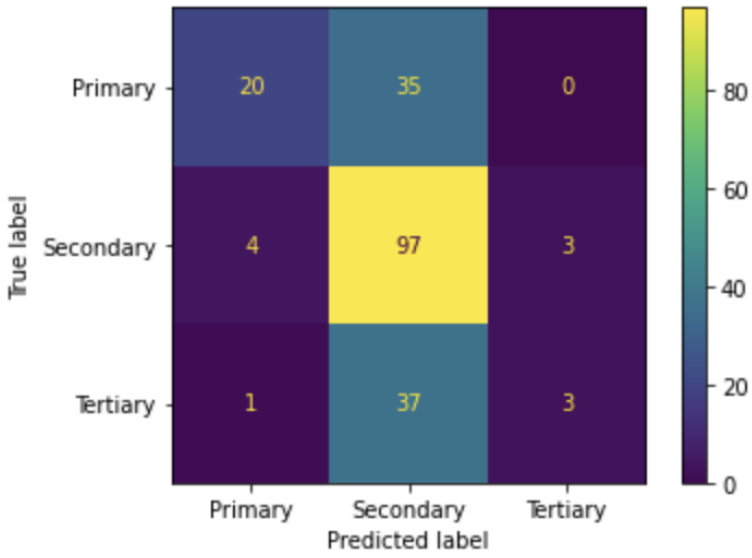
In this sense, we also had difficulties in defining how many difficult terms would serve as the cutoff between the three categories. Since some documents would be longer, we didn't know how to fairly delineate the categories and this remains, in our opinion, a major flaw in our guidelines that we would like to remedy in future iterations of this project.

At the same time, we wonder if it is really possible to generalize the conclusions of the analysis on the first few paragraphs of a novel to the entire novel itself. A quick Google search reveals that the average novel is somewhere between 70,000 to 120,000 words, meaning that there can be quite a significant variation between chapters in the level of difficulty. Throughout this entire process, we make the assumption that the first chapter of a novel is a good approximation for the difficulty of a novel as a whole, but this also opens up room for bias and error.

Confusion Matrix

Below in our confusion matrix, we can see that the true label 'Primary' is most often mislabeled as 'Secondary' and the true label 'Tertiary' is also most often mislabeled as 'Secondary'. However, the majority (i.e. 93%) of true 'Secondary' texts were correctly labeled as 'Secondary'. This suggests that our categories and guidelines were not adequately differentiating between the categories (e.g. taking into account text length).

However, this result can be expected, to an extent. Primary and Secondary novels and Secondary and Tertiary novels will be more similar than Primary and Tertiary novels. The fact that our model, at the very least, rarely mislabels Tertiary novels as Primary and vice versa indicates some success.



Patterns

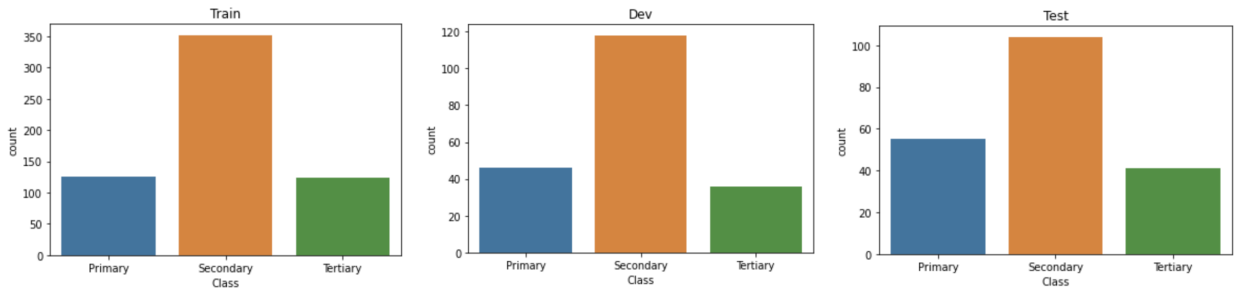
Our model tended to over classify Primary and Tertiary novels as Secondary. For novels that were Tertiary (but misclassified as Primary or Secondary), we found that a majority of the time, the number of difficult terms in the text was more spread out compared to the novels that our model correctly identified as Tertiary. In other words, the density of difficulty terms was lower. Interestingly, we also found that when Tertiary texts mentioned medieval themes such as “princess”, “king”, or “knight”, these were also often misclassified as Secondary or Primary. This was because many of the Primary texts in our dataset were fairy tales and mentioned these themes as well.

We found that our model seemed to have formed a “resistance” against words specifying characters, names, and locations. Some of the texts that we classified as Tertiary contained a dense network of locations and references within a single paragraph, which would be difficult for primary or even secondary readers to follow. However, our model almost always classified these text examples incorrectly leading us to form this conclusion. We agree that to some extent, devaluing proper nouns can help to form more accurate conclusions, but in the case of classifying text by readability, it leads to inaccurate results.

The last systematic error that we noticed was with texts containing the characters “Äù” i.e. character encoding errors from copying and pasting the novel text from Project Gutenberg. Because there were relatively few instances of this error, we chose not to remedy it as resolving that would not severely affect predictions over the text as a whole. However, we found that texts containing these encoding errors were frequently classified as Secondary. It is possible our model viewed “Äù” as a difficult term.

In terms of biases, most of our novels were in English and we purposely filtered out novels that were in different languages. There were also very few novels written in different dialects so we concluded that bias had relatively little effect on our results.

Dataset & Balance



While annotating our data, we did discover that the “Secondary” label was assigned more frequently than “Primary” and “Tertiary”. This can be seen in our training set – the “Secondary” label was associated with 348 novels while “Primary” and “Tertiary” were associated with 123 and 122 novels, respectively. This has the potential to obscure the results of our model. Simply predicting the most common class and applying it to all novels in our training set produces an accuracy rate of 0.587.

In terms of utilizing oversampling, our dataset would not be a good candidate. Although “Secondary” labels are more prevalent than the other two options, a large enough skew is not present. The linked article discusses datasets with a 1:100 skew, whereas our dataset is skewed around 1:2. Therefore, adjusting class weights could be a better alternative in accounting for the skew in our data.