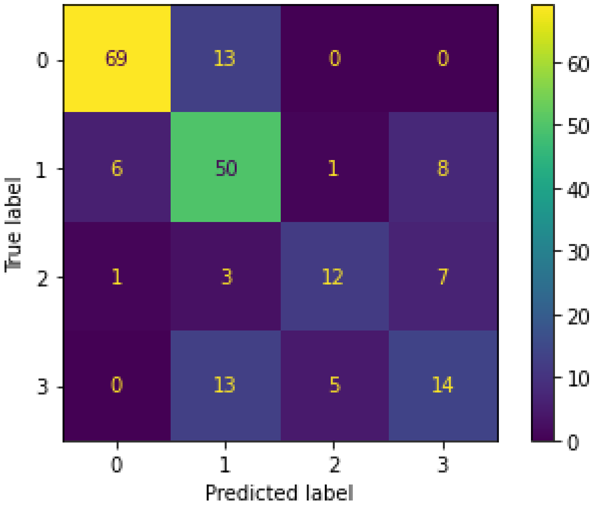
We annotate a custom dataset of book reviews and based on how professional sounding they are on a scale of [0,1,2,3] where 0 (1 in our annotation data) is the least professional and 3 (4 in our annotated data) is the most professional.

We trained a multiclass logistic regression model and a BERT model on our data. The logistic regression model had a test accuracy of 0.673 and with a 95% confidence interval between 0.609 and 0.738. The BERT model had a test accuracy of 0.718 with a 95% confidence interval between 0.656 and 0.780. Not surprisingly, the BERT model performs better since it was trained on a large amount of data and fine tuned with our training set. BERT also uses contextual embeddings which allows it to embed more information for each word. Both models outperform the bag of words logistic regression, which has a test accuracy of 0.65.



For our analysis, we will focus on the BERT model. The distribution of ratings in our test data is listed in the table below. Ratings 1 and 2 are more prevalent, as is reflected in the training data.

Rating	# of Reviews
1	82
2	65
3	32
4	23

The accuracy of the BERT model predictions for each of the ratings is in the table below. It is more accurate in predicting rating 1 and least accurate in predicting rating 3. Overall, it is more accurate in predicting ratings 1 and 2, which correlates to the fact that we have more training data for those two ratings. Future attempts at this task should prioritize collecting more high-rating reviews.

Rating	# of Ratings Correctly Predicted	Percentage of Ratings Correctly Predicted
1	69	84.14%
2	50	76.92%
3	14	43.75%
4	12	52.17%

The table below provides a breakdown of what labels the model was confused with. It most commonly mixed up 1 and 2, and 2 and 3.

Rating Confusion	# of Ratings Mixed Up	Percentage of Ratings Mixed Up
1 and 2	19	33.33%
1 and 3	0	0.00%
1 and 4	1	1.75%
2 and 3	21	36.84%
2 and 4	4	7.01%
3 and 4	12	21.05%

The model had the most difficulty distinguishing ratings 2 and 3 from each other. In other words, more than a third of the errors (21/57, which is 36.84%) consisted of the model being confused between 2 and 3. This is reasonable since those two ratings are in the middle of our rating scale and essentially only differed in terms of meeting “most” vs. “some” of the professionality criteria, which is hard to quantify and ambiguous. These two ratings need to be revisited and clarified, as was suggested in the reviews we got of our annotation guidelines. Once we have more clear criteria as to what should be labeled as a 2 vs. a 3, then the data will need to be re-annotated to reflect those changes. And then hopefully the model will be less confused about those ratings and will perform better.

Another prevalent error consists of the model predicting a review to have a professionality of 2, when in fact it was labeled as a 1. This occurred for nearly a quarter of the test data (13/57, which is 22.80%). This might be because these reviews had a bit more length to them than what was typically labeled as 1 in the training data. There was generally a positive correlation between length of review and predicted professionality rating by the model, which may explain this error. This issue can, again, be addressed by balancing out the labels among the annotated data.

To look at some of the less common errors, we see that 6/57 (or 10.52%) of model errors were falsely predicted as a 1, instead of a 2. This might be due to the brevity of the review, the odd capitalization in it, or the plethora of or the non-standard punctuation present in it. The model falsely predicts a review that was labeled as a 3 as a 4 five times (5/57, which is 8.77%), and this is because these were lengthy reviews, and so the model mistakenly thought they were a 4. Lastly, model errors falsely predicting a rating of 4 as a rating of 3 occurs seven times (7/57, which is 12.28%). Several of these reviewers mention how they wrote the review in exchange for a free copy – so maybe the model picked up on those reviews that have such a disclaimer are generally not as professional and so shouldn't be labeled as a 4.