AP3 - Analysis

Model, Accuracy, Confidence Interval:

BERT: 0.78, (0.6988091840319723, 0.8611907587475687)

SVM: 0.67, (0.5778400007999419, 0.7621599992000582)

Random Forests: 0.64, (0.5459217287420775, 0.7340782712579226)

Logistic Regression: 0.65, (0.5565156760890944, 0.7434843239109057)

Balanced Logistic: 0.69 (0.5993529900246857, 0.7806470099753142)

The accuracy, confidence interval, and visualizations for each respective model are printed below their training in our notebook (with the accuracy and confidence intervals shown above for ease of reference). Below we've referenced three relevant visualizations of the ones found there.

Our models were significantly affected by the unequal representation of categories. In our dataset, the majority of our Jeopardy questions were labeled ‘non-fiction’. Two other categories (‘science’ and ‘sports’) are subcategories of ‘non-fiction’ (special instances of ‘non-fiction’). This resulted in the model predicting ‘non-fiction’ more often than not, potentially decreasing the accuracy of what the model could be capable of. In hindsight, it probably would've been better for us to create more subcategories of ‘non-fiction’, thus decreasing class asymmetry and increasing model accuracy. In our model, every category that's not 'non-fiction' is often mistaken as ‘non-fiction’, and occasionally as ‘fiction.’

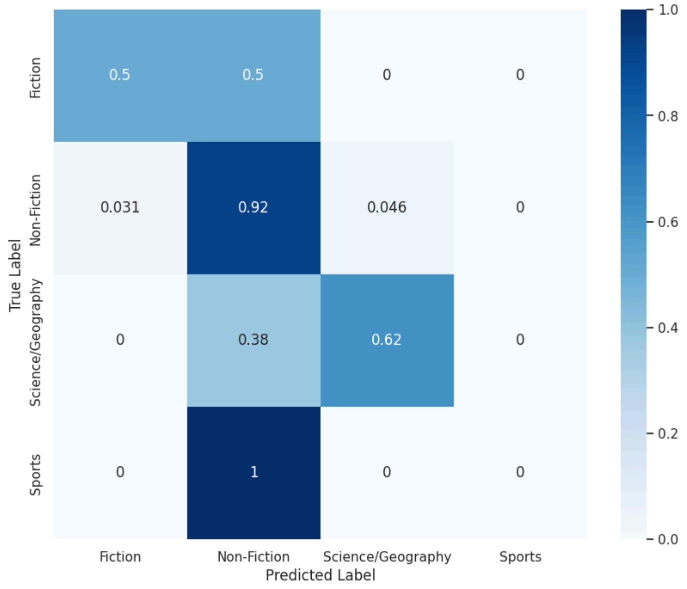
When creating the categories, we didn’t realize that with such a large proportion being classified as 'non-fiction' would leave the algorithms with a minimal amount of data in the other categories, and an extremely large bias towards said category. In hindsight, we should’ve created more subcategories of ‘non-fiction,’ thus enabling a more even class distribution (since many of the Jeopardy questions had multiple ideas that overlapped with the original categories we made).

In our analysis, we choose to use a few different unique models that were not already implemented for us: a SVM, Random Forest, and Balanced Logistic Regression. We chose to add SVMs because we believe that our annotated data was highly accurate, so a more rigid model like SVM might take each of our classifications more seriously and get to the level of detail that separates ‘non-fiction’ from its respective subcategories. We chose to add a Random Forest model because we felt as though our guidelines had a “flowchart”-esque type of reasoning behind them that the Random Forest might be able to pick up on. We chose to use Balanced Logistic Regression because our dataset is very imbalanced (contains a category that makes up a majority of the dataset), and thus it would most likely perform better. In hindsight, most of these models did not perform better than the BERT model despite our reasonings, but it was interesting to look through them nonetheless.

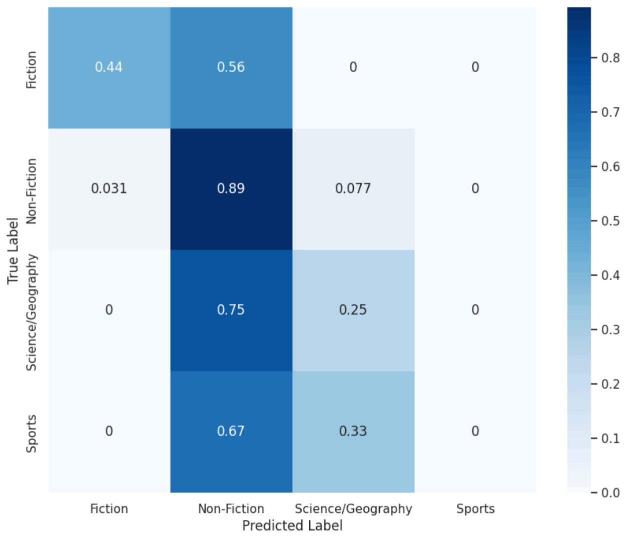
Our BERT model had by far the best accuracy with the Balanced Logistic Regression model second behind. Looking between both, their Confusion Matrices paint a clear picture of where this extra accuracy is coming from: the BERT model seems to have a much richer understanding of our guidelines and annotation methodology, correctly predicting ‘science’ (as opposed to mis-classifying it as ‘non-fiction) a much larger proportion of the time than the

Balanced Logistic model. Both of these models were significantly better than all the other models we tried, most likely because they were able to account for the class imbalance much more so than other models (each of which got stuck at nearly always classifying every data point as 'non-fiction,' even more so than the balanced logistic model). Since BERT understood that everything should not always be classified as 'non-fiction' we did not explore any other models further, as even the ones that were supposed to account for class-imbalance (e.g., balanced logistic regression) still struggled.

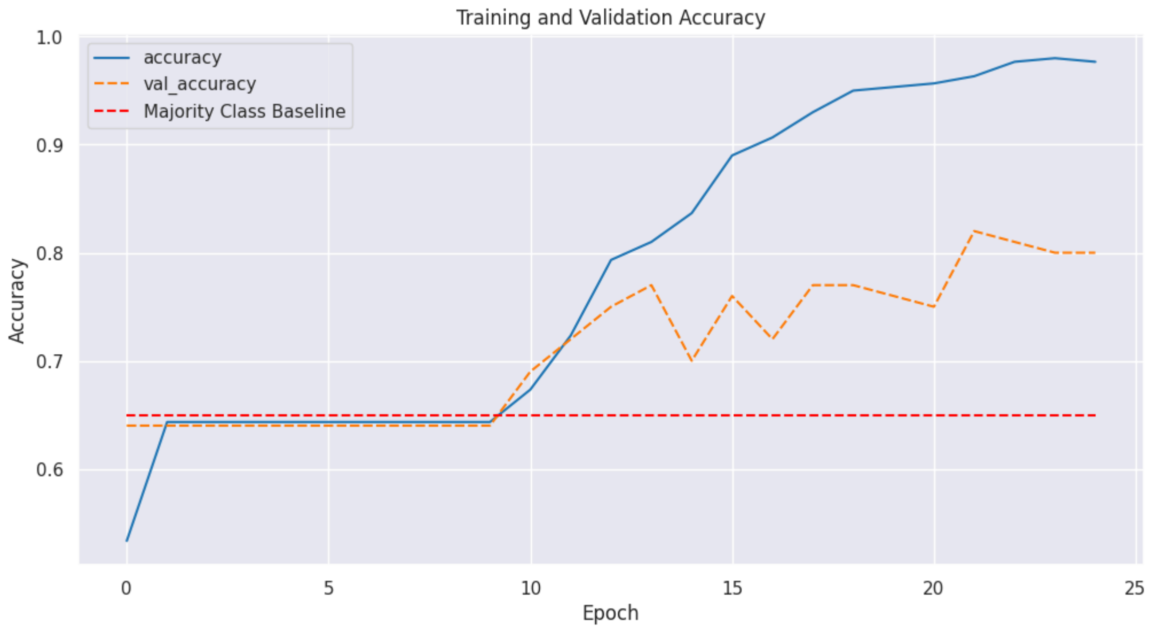
BERT Confusion Matrix

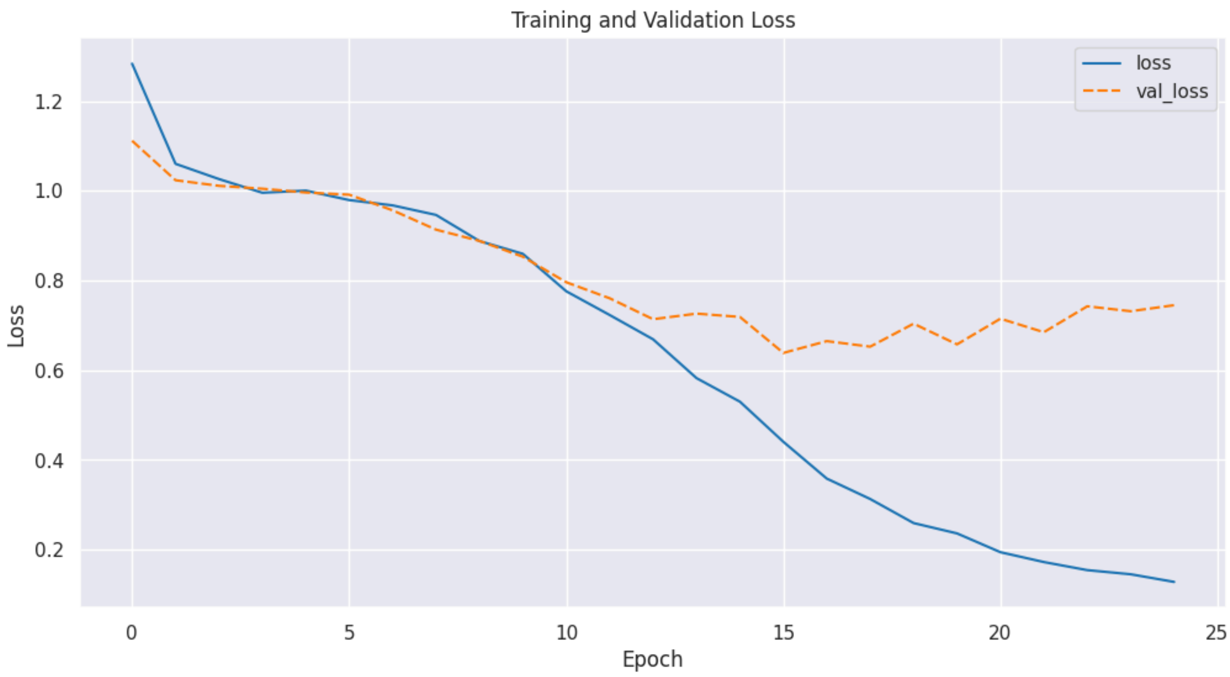


Balanced Logistic Regression Confusion Matrix



Looking through the BERT model, it is very interesting to note that it often (a large minority of the time) confuses ‘science’ with ‘non-fiction’ and always classifies ‘sports’ as ‘non-fiction’ (which makes sense according to our guidelines, both are subclasses of that parent class). Also, any ‘fiction’ label is equally likely to be considered as ‘non-fiction’ (aka the model was not able to accurately distinguish our idea of what makes a fiction Jeopardy clue ‘fiction’). It seems that when in doubt the model will guess ‘non-fiction’, for which it is not penalized since it is ‘non-fiction’ such a large proportion of the time. In addition to these more direct classification considerations, it is also interesting to look at more broad notions, like if the model is overfitting and to what degree. Looking over the Training/Validation Accuracy and Loss graphs, it seems that the model began to overfit and lose validation accuracy around the 10th epoch (as expected by a BERT style model). Despite that, the validation accuracy seemed to hold around constant, most likely because overfitting in a class-imbalance scenario is generally beneficial.





The imbalances in the categories make our dataset a good candidate for oversampling techniques (for the small categories: ‘sports’, ‘science’, and ‘fiction’) and undersampling techniques (for the large category: ‘non-fiction’). Balancing out the categories in our dataset could possibly allow the models to learn more about the other categories (and not always guess ‘non-fiction’) and increase accuracy. In addition to this, future work could also include refining the guidelines we made earlier to have more granularity in the ‘non-fiction’ category (i.e., creating more subcategories to balance the class imbalance), which has the potential to significantly increase various model’s accuracy.