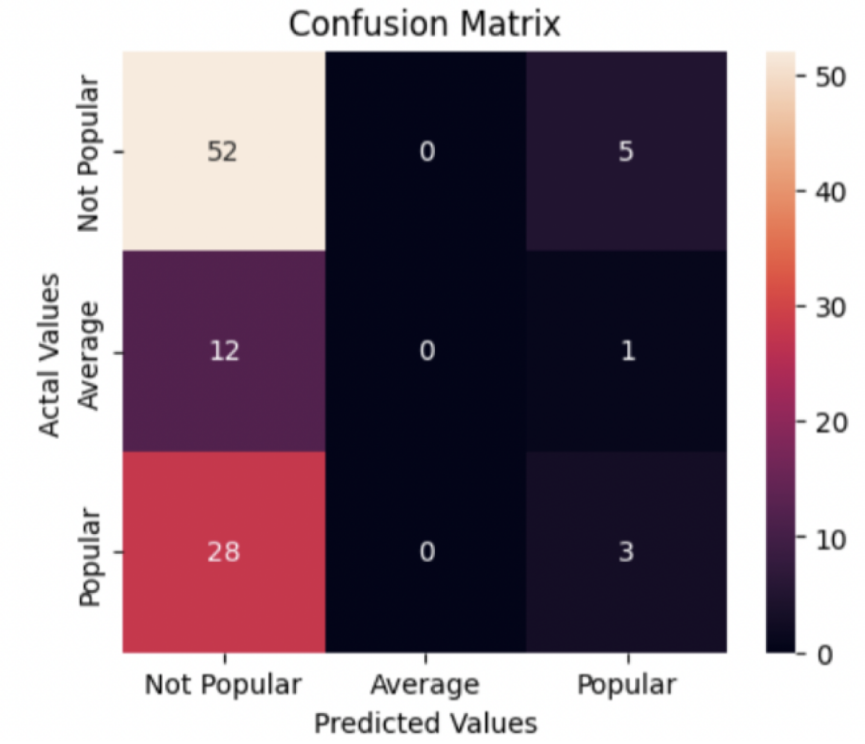
We tested our data using five different models. One of these models was a logistic regression model, which we implemented by eliminating the bag-of-words feature and adding other features, such as ‘clarity’ and ‘sarcasm’ that are included in the AP2 guidelines. Based on four categories, including 1. the length of the sentence, 2. the use of specific words, 3. the presence/number of positive/negative words from the positive/negative set of words from the nltk library, 4. and typical positive/negative words, the model computed clarity, presence of sarcasm, and tone of the data points.

For both logistic regression and ordinal classification, we found that the bag-of-words feature was not a good fit and actually resulted in lower accuracy when included because the presence of a particular set of words did not necessarily indicate the popularity of posts. Instead, we found that features such as capitalization, exclamation/questions, and length of the sentence were more effective in determining the label. Thus, we decided to only include the features that we introduced through our guidelines (clarity, sarcasm, capitalization, tone, Q vs Statement) into the binary features. As a result, we obtained an accuracy of 0.535 for logistic regression and 0.53 for ordinal classification. It may seem odd that our model performed worse in ordinal classification when our labels seem to be in an ordinal scale, but as explained through the AP2 guidelines, the level features such as sarcasm, clarity, and Question vs. Statements are in fact binary features. Even for multiclass features like sentence length, the length of the sentence did not determine the popularity solely based on its length, but rather through intervals.

The majority classification method generated the poorest performance on our data, resulting in an accuracy below 40 percent. With this result, we believed that logistic/ordinal regression/classification methods were better fits to test our data.

Finally, two of the models, TF-IDF (Term frequency, inverse document frequency) (implemented by ourselves) and BERT (which was given to us) resulted in the highest and equal accuracy: 0.564. These highest performances in accuracy may stem from the TF-IDF model’s ability to capture the significance/correlation of certain single words (unlike the bag of words model) and the BERT model’s ability to clearly understand the context of sentences through its usage of bidirectional encoding. Our labels for data (not popular, average, and popular) place a huge emphasis on understanding the context of a sentence, so the two models that were better able to understand the context of Reddit post titles generated the highest accuracy among the other models.

From the confusion matrix, we noticed that the logistic regression model tended to output 'Not Popular' as the end result. Below is the result:



A noticeable finding was that although 52% of the actual data points were labeled as “not popular,” the model predicted the label 'not popular' for 92% of the data points. This difference between prediction and the actual data for “not popular” labels may result from the class imbalance between the ‘popular,’ ‘average,’ and ‘not popular’ categories in our actual data. With ‘not popular’ data representing 52% of data points and thus being more prevalent than the other categories, the model may be biased towards predicting the ‘not popular’ class. In order to overcome class imbalances, one improvement we could have made was adjusting the class weights to give more importance to the minority class during the training process. Also, our logistic regression model may not include all relevant features to determine whether the data point is ‘popular,’ ‘average,’ or ‘not popular,’ which can lead to inaccurate and biased predictions in our model. Additionally, as the data labels are based on subjective categories, and there may be disagreement among labelers, this inconsistency can introduce noise and thus lead to a lower number of accurate predictions.

Below is a distribution of weights among features in determining the label of posts:

Average	0.162	QS
Average	0.152	length:20+
Average	0.127	Clarity
Average	0.124	neutral
Average	0.113	allcap
Average	-0.030	Sarcasm
Average	-0.033	length:11-20
Average	-0.034	positive
Average	-0.089	negative
Average	-0.119	length:0-10

Popular	0.104	positive
Popular	0.048	length:20+
Popular	0.024	length:0-10
Popular	0.009	Sarcasm
Popular	-0.003	neutral
Popular	-0.071	length:11-20
Popular	-0.087	allcap
Popular	-0.101	negative
Popular	-0.116	QS
Popular	-0.150	Clarity

Not popular	0.210	negative
Not popular	0.142	length:11-20
Not popular	0.054	Sarcasm
Not popular	0.053	length:0-10
Not popular	0.005	allcap
Not popular	-0.001	Clarity
Not popular	-0.013	QS
Not popular	-0.052	positive
Not popular	-0.158	neutral
Not popular	-0.195	length:20+

label	0.042	length:0-10
label	0.038	neutral
label	0.024	Clarity
label	-0.005	length:20+
label	-0.018	positive
label	-0.020	negative
label	-0.031	allcap
label	-0.033	QS
label	-0.033	Sarcasm
label	-0.038	length:11-20

From the distribution above, we have found that weights across features are well-spread, suggesting that the model is able to capture a diverse range of relevant features in determining the label. 'Question vs. Statement' represented the highest weight (0.126) for predicting labels as 'average,' 'negative' represented the highest weight (0.21) for predicting labels as ‘not popular,' and ‘positive’ represented the highest weight (0.104) for predicting labels as ‘popular.'