Our project was to develop a predictive model to classify slogans based on their persuasiveness on a scale from 1 to 5, with 1 being the least persuasive and 5 the most. Utilizing a dataset annotated with these persuasiveness ratings, we employed ordinal regression to predict the effectiveness of unseen slogans, splitting our data into a training, dev, and test set.

For our predictive model, we chose to use the ordinal regression method, as its properties aligned with our numerical persuasiveness scale. After running the baseline model, we arrived at a baseline accuracy of 0.2 (blindly guessing one of 5 categories). Subsequently, after running our baseline ordinal regression model, we arrived at a test accuracy of 0.430, with our 95% confidence interval consisting of [0.333 0.527].

To improve upon our model, we chose to apply feature engineering, adding 9 features respectively:

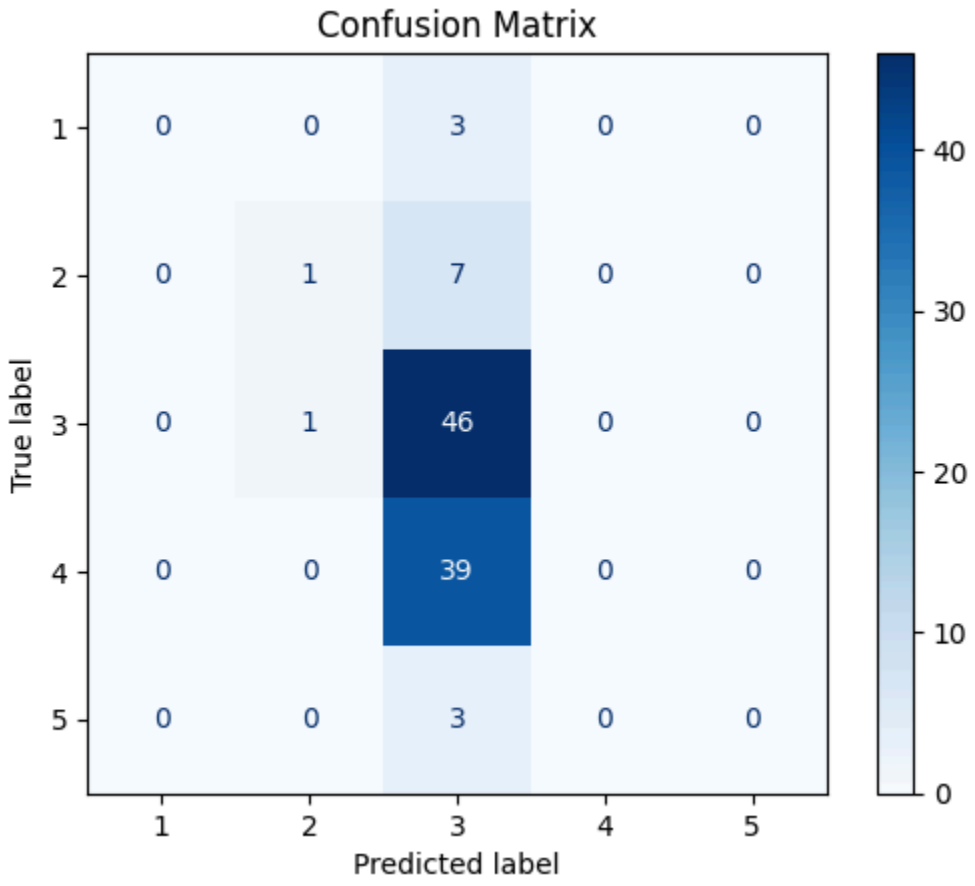
- Slogan length
- Punctuation
- Readability
- English
- Rhyme
- Alliteration
- Action words
- Compelling words
- Stop words

In this section, we manually conducted feature selection in order to see which features were helping the model accurately classify models. Through this section, we found that features such as punctuation (whether a slogan had a ? or !) and readability (if all characters were numbers or letters) didn't help the model increase its accuracy. We also found that certain features such as rhyme and length were crucial in any iteration to help the model assess a slogan's persuasiveness.

After adding these features, we found an increase in our test accuracy, as running the improved model returned a test accuracy of 0.48 with a 95% confidence interval of [0.382 0.578].

While we did increase our accuracy by an amount, we must acknowledge that it is hard to achieve a near perfect accuracy when rating slogans with a predictive model, as perceiving slogans and persuasiveness is a highly subjective topic amongst humans.

To conduct our error analysis of the data, we first decided to implement a confusion matrix that takes the true labels (manually annotated persuasiveness) , predicted labels (predicted persuasiveness), and class labels (ordinal values like "1", "2", etc.) as input, where each cell in the matrix indicates the number of predictions for a pair of true and predicted classes. The darker cells indicate which persuasiveness level is more accurately predicted. We also decided to add a function called “feature importance” which identifies which aspects of a slogan most influence its persuasiveness rating according to the model. It does this by looking at the weight or 'coefficient' each feature has in the regression equation.



Feature Importance Results:

Most important features for classifying above vs below 2:
english: 0.1580
compelling_words: 0.0607
punctuation: 0.0520
action_words: 0.0069
length: -0.0000
alliteration: -0.0598
rhyme: -0.1796

Most important features for classifying above vs below 3:
english: 0.2801
rhyme: 0.1617
action_words: 0.0264
length: 0.0000
alliteration: -0.0038
compelling_words: -0.0339
punctuation: -0.3401

Most important features for classifying above vs below 4:
rhyme: 0.1808
english: 0.0881
alliteration: 0.0774
action_words: 0.0585
length: -0.0000
compelling_words: -0.0001
punctuation: -0.1748

Most important features for classifying above vs below 5:
alliteration: 0.3567
compelling_words: 0.1410
rhyme: 0.0775
english: 0.0615
length: 0.0000
action_words: -0.0095
punctuation: -0.0653

When analyzing our confusion matrix and feature importance outputs, we can see that the model mostly gets moderate persuasion ratings (scores of 3 or 4) correct compared to extreme persuasion ratings; this could signify a couple of things:

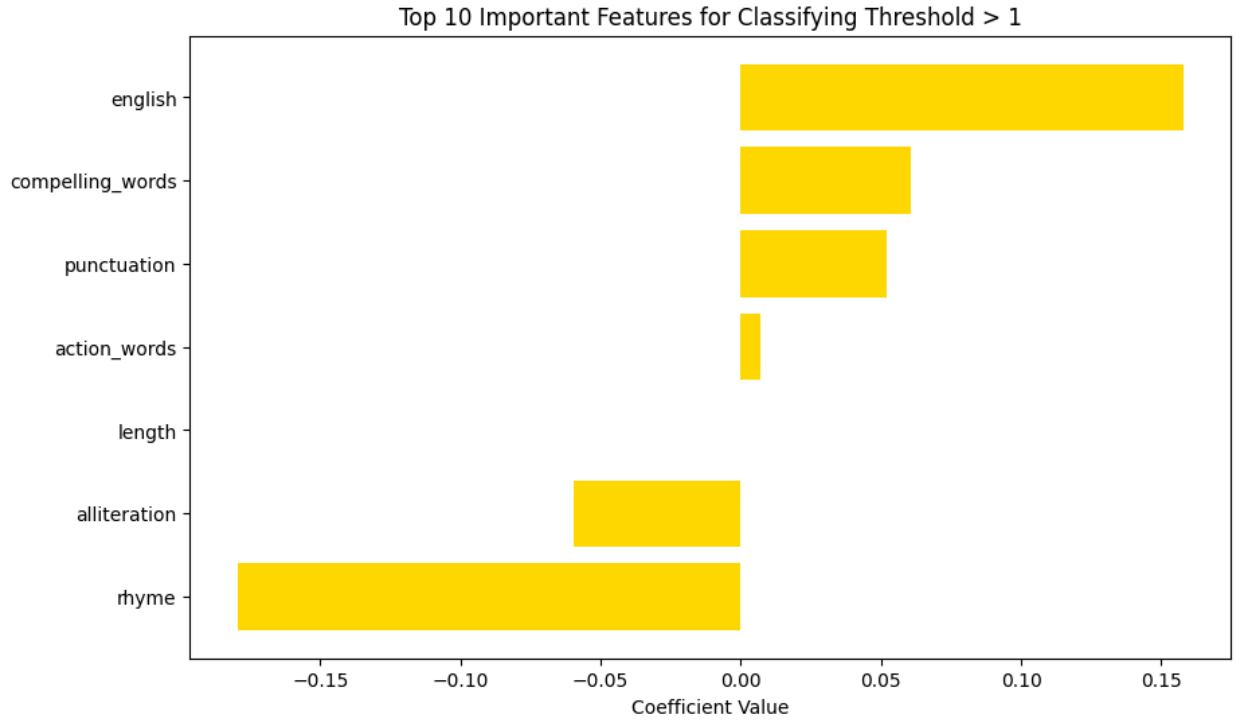
- Complexity of Sentiment: Slogans that are very persuasive might use subtle sentiment cues or complex rhetorical devices that influence human emotions. Such subtleties might be lost on a model that relies on more surface-level textual features. Human emotions are complex, enough to the point where machines are still unable to fully comprehend them
- Central Tendency Bias: There might have been bias toward the middle categories when unsure, leading to over-representation of moderate persuasiveness in the training data. This could have also been due to the fact that we had only a small number of annotators who all exhibited this bias. More on this later.
- Inconsistencies in Annotations: There may be inconsistencies or subjectivity in how the extreme persuasiveness levels were annotated, leading to noisy labels that confuse the model: especially given how we had only 3 annotators.

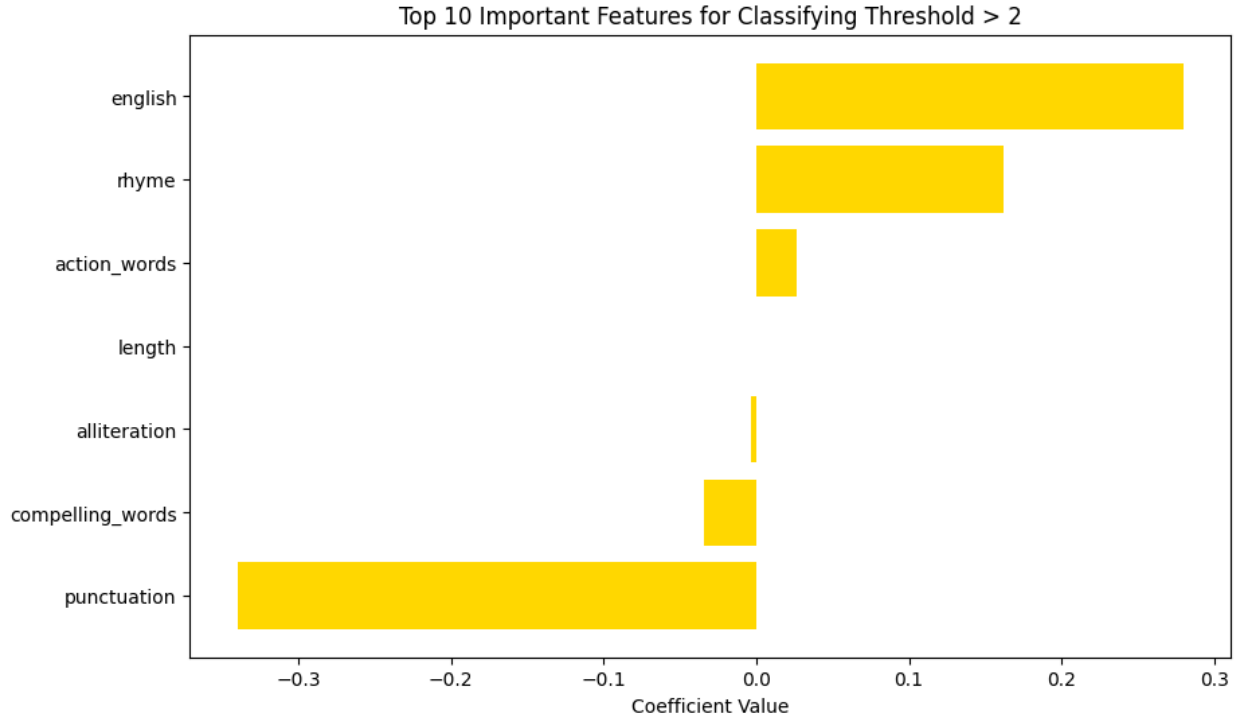
Analyzing our specific results, which will be in our python file we submit as well as pictures later, we can also see how the important features change when dealing with different persuasion scores. For example, we can see that for scoring above or below a 2, English possesses the highest coefficient in feature importance (desirable for persuasion) and rhyming the lowest (less desirable). However, for scoring above or below a 5, alliteration possesses the highest coefficient and punctuation the lowest, perhaps signifying that some features become more important as considerations of persuasion levels increase; this could signify what it takes to go from a moderate level, say a 3 in persuasion to a 5 in high persuasion.

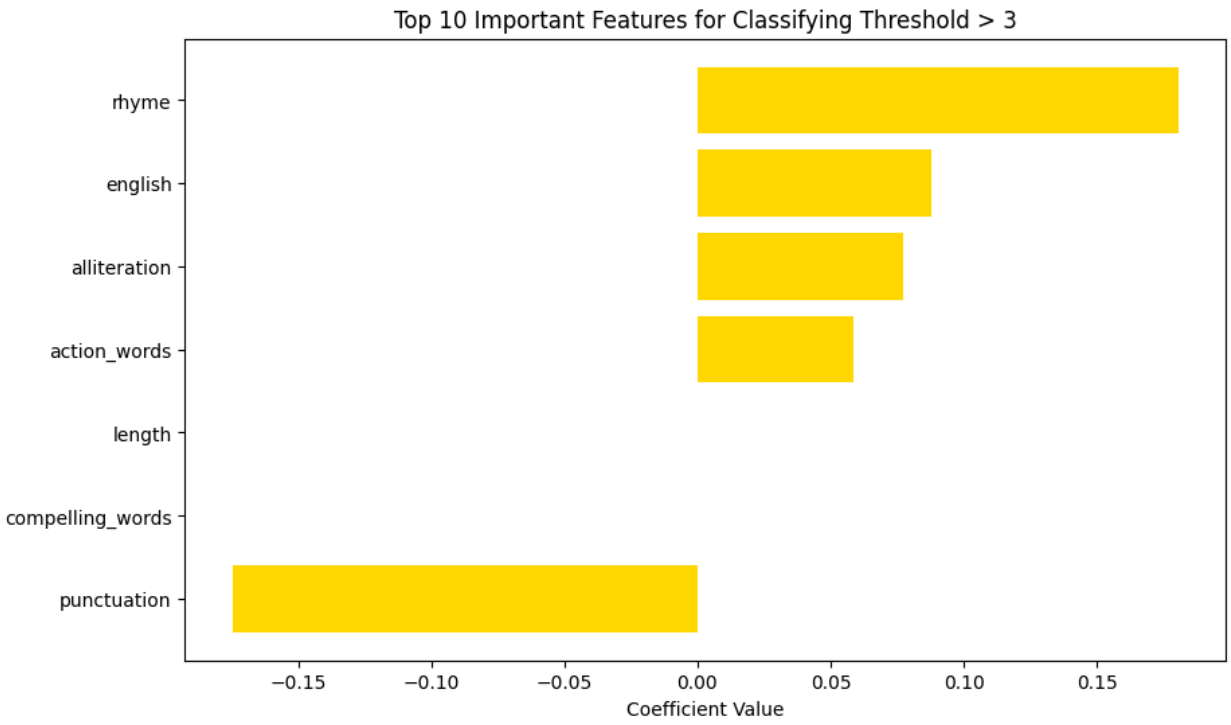
In addition, we similarly performed a regression analysis again between each feature and persuasiveness and visualized this to make it easier to recognize; the results are the same as the feature importance function outputs. All the graphs will be outputted at the end of the document.

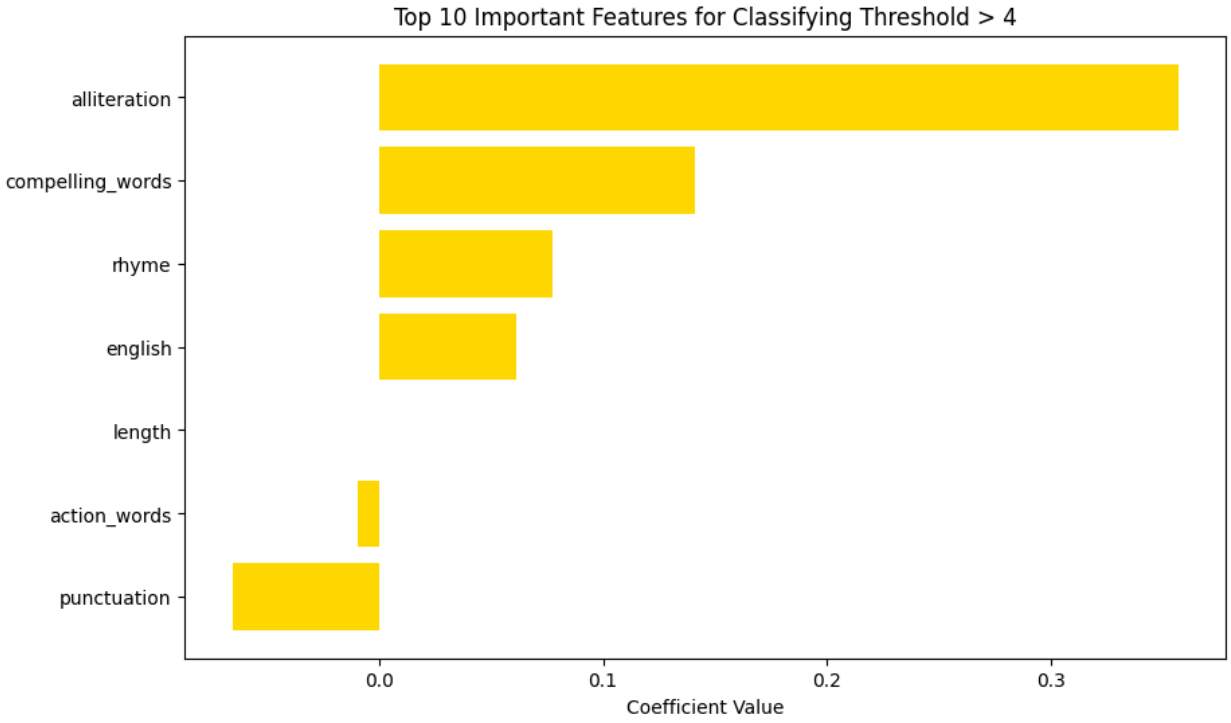
To address the issue of **bias**, there will always be some level of subjectivity when annotating. Given that we only had 3 annotators, this is nowhere close to the amount of annotators needed to provide some sort of consistency to our annotated data. With less annotators, it is harder to settle on a persuasion score that is more reflective of the general public, thus leading to more inconsistency. In addition, the majority of our data was in English, so if the slogan contained a foreign language, it does not have enough training data that was in that language to train on, thus leading to further inaccuracies.

Graphs:


Top 10 Important Features for Classifying Threshold > 1
english
compelling_words
punctuation
action_words
length
alliteration
rhyme
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
Coefficient Value


Top 10 Important Features for Classifying Threshold > 2
english
rhyme
action_words
length
alliteration
compelling_words
punctuation
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
Coefficient Value


Top 10 Important Features for Classifying Threshold > 3
rhyme
english
alliteration
action_words
length
compelling_words
punctuation
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
Coefficient Value


Top 10 Important Features for Classifying Threshold > 4
alliteration
compelling_words
rhyme
english
length
action_words
punctuation
0.0
0.1
0.2
0.3
Coefficient Value