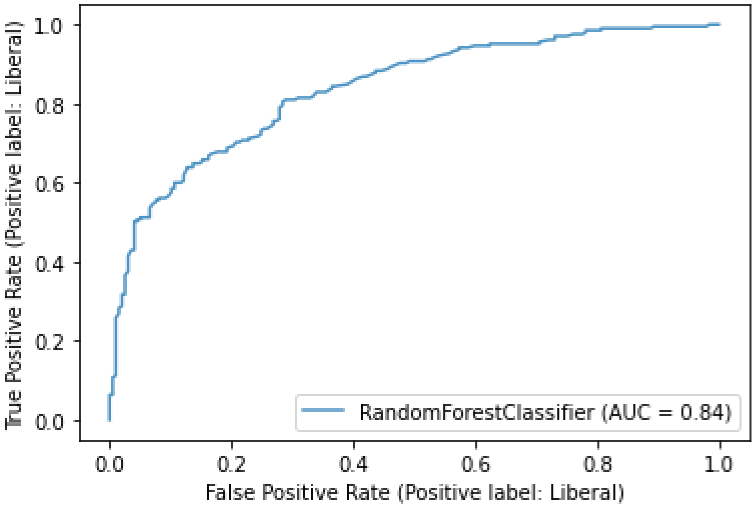
DS 110 Project: Brandon Wong, Eric Gulotty, Michael Che

Introduction

The problem our project is trying to address is to determine whether specific comments made on the political side of Reddit are leaning towards the liberal or conservative side of the political spectrum. The reason we wanted to address this topic is due to our interest in politics and the practicality of this topic in the real world. With social media and politics being heavily involved in our everyday life, we thought this project would be a great opportunity to focus on these concepts. Additionally, Reddit has a reputation for being a unique and popular social platform where people can freely express their political views which could make the results more intriguing.

Methodology

The dataset we utilized consisted of 13000 posts collected from Liberal and Conservative subreddits. To turn the texts of posts into something that can be used by machine learning, we utilized BERT to pass our data through. After downloading the data from its Github location, we returned as a Pandas dataframe.Since 13,000 posts would take a long time to run we took 800 post from each side for a total of 1,600 posts to analyze. From there, we extracted the labels and converted the sentences into lists of tokens to use with BERT. The final step before applying BERT was to convert all the sentences to the same length by padding it with zeros. We then had to create a variable to mask the padding when processing the input. Once the sentences were ready, we utilized a pre-trained DistiliBERT model to analyze the sentences. DistiliBERT has a grasp on the English language making it more effective. Then, we used a basic RandomForestClassfier, we trained the model with a maximum of 80% with an average hovering above 70%. We found the model works especially well with the test cases within the dataset.



The above graph displays the efficiency of our model in predicting the true political leaning of a comment or title. As one might notice, the model is quite accurate up until the 0.6 to 0.8 mark, when false positive rates begin to increase dramatically. Our average accuracy of approximately 0.7 occurs around the middle, while our highest accuracy of 0.8 occurs right before the false positive rate skyrockets.

Results

After running a few test sentences through the predict_from_sentence() function, we decided to scrape forums on Reddit for real comments to run through the algorithm to test its accuracy under real world conditions. Liberal comments are taken from r/PoliticalDiscussion, while conservative comments are taken from r/Conservative. The sample size is relatively small, but each comment was randomly selected, with a certain degree of manual filtering through the data; as all users of Reddit can comment on a particular subreddit, we had to ensure that each comment we selected was both: (1) A proper representation of the political leaning we were selecting for, and (2) To a certain degree of specificity and length to ensure that the model has enough context to make a prediction.

Sentence (Liberal)	**Prediction Output**	**True/False**
“Trump and his supporters are the ones who turned politics into a weird cult.”	Liberal	True
“The Republican Party is QAnon. MAGA is QAnon.”	Liberal	True
“Tax cuts for the rich doesn't work. We've known that for decades. But they just keep chasing the same failed policies for no reason. This entire thing makes no sense.”	Conservative	False
“What do you think would happen if Obama tried to pull out? The country would have collapsed to the Taliban, and the Republicans would have shredded Obama and accused him of failing their war.”	Liberal	True
“It's just crazy that Trump still has this much power over the party. He lost them the House, he lost them the Senate and he lost them the White House. Heck even when he "won" he lost by 3 million votes. Normally with that much failure a political party would not be willing to be held hostage yet instead they've gone to a full cult.”	Liberal	True
“I wouldn’t call them “racial undertones”. The racism was very overt and clear cut to me.”	Liberal	True
“Nope. The south is full of racists that live by a failed ideology of forced integration and the thoughts of losing their heritage causes them to embolden.”	Conservative	False
“Yes, the police were at fault, thank you.”	Liberal	True
“Well, it's not that simple. Many of Trump's policies, if you can call them that, don't really align with the GOP in general.”	Liberal	True
“The majority is against Trump.”	Liberal	True
“Raise the minimum wage.”	Liberal	True

Sentence (Conservative)	**Prediction Output**	**True/False**
“Harboring and assisting direct threats to the United States makes you a direct threat to the United States in my book.”	Conservative	True
"That sounds like a populist lie you have been told."	Conservative	True
"Trump actually wanted to reduce our dependence on China, unfortunately it's a lot longer process than a few years. Biden wants us dependent on them again."	Liberal	False
"The Chinacrats from the Chinacratic Party are an open book..."	Conservative	True
"Rules for thee, but not for me! - dems "	Conservative	True
"Didnt WHO also say that the Floyd protests did’t present a risk for virus spreading but the stay at home order protests did? Anyone who doesn't realize they were being played about the seriousness of this whole thing needs to wake up."	Conservative	True
It’s the media. If you look up “media” in the dictionary you’ll see the definition of hypocrisy.	Conservative	True
"WHO and the NPR are hypocrites"	Conservative	True
"This programming is made possible by listeners like you...and the Chinese government."	Conservative	True
"Well in Biden's defense he probably won't remember having rallies."	Liberal	False

Results Analysis

The model was able to accurately predict the political leaning of 8 out of 10 total sentences for each political ideology. Since our model defaults to returning a “liberal” prediction when unsure of the true political leaning, it is interesting to see that it classified some sentences as being conservative. For the first sentence regarding tax cuts, we believe that the prevalence of “tax cuts” in conservative literature could be the reason for such a result, as tax cuts are a popular conservative macroeconomic strategy to boost economic productivity in times of recession. As for the second sentence, we believe that the sentence, “Nope. The south is full of racists that live by a failed ideology of forced integration and the thoughts of losing their heritage causes them to embolden”, might be too abstract for the model, and we cannot be sure which words triggered the model to produce a conservative prediction.

For conservative sentences, we can observe some of the more traditional shortcomings of a simple BERT model. The sentence regarding Trump and Biden’s respective stances on China does not provide the model with the assumed context which would indicate to a human reader that the sentence was likely written by a conservative; for a BERT model taking the sentence at face value, there doesn’t appear to be enough evidence for the model to overturn its default “liberal” output. The second sentence, "well in Biden’s defense he probably won't remember having rallies”, relies on sarcasm to resonate with its human audience, which would be able to correctly identify the comment as conservative. Once again, our simple BERT model is unable to identify the sarcasm in the statement, and thus defaults to a “liberal” output as in the previous sentence.

Conclusions

Overall, from our model, we concluded that in general, BERT is able to successfully determine whether most cases is liberal or conservative generally with 80% accuracy. However, we learned that BERT is sometimes confused by long sentences or tricky word-phrasings causing inaccuracies. Also, Reddit posts tend to lean towards the neutral side, challenging the model to correctly separate liberal and conservative posts. The greatest challenge that we faced was that even though BERT could understand proper english it doesn't understand things like sarcasm and instead takes stuff very literally. In conclusion, we were successful in predicting the political lean of reddit comments with the majority of our model prediction being correct. We also learned BERT is a great way to turn text into something usable by machine learning but has its limitations.