Further steps for Model Training

In this project, the aim is to develop a predictive model that evaluates the controversiality of comments on Reddit, with a specific focus on discussions about the Russia-Ukraine war. Our goal as a team is to create a tool that assists Reddit in determining whether to enhance or limit the visibility of comments based on their likelihood to spark controversy. By doing so, we aim to foster a healthy online discussion environment.

After completing our ongoing work in Exploratory Data Analysis and Data Cleaning we shall head to Tokenization. We proceed to tokenize and identify the most frequent words by utilizing the NLTK package's tokenizer class and word tokenizer's RegexpTokenizer to extract tokens and character sequences. In the dataset, using those tools, we can identify the "stop words," which are typically common words present in a language that add little to the meaning of the text. Consequently, we can remove them from the NLP tasks and continue tokenization and analysis of the rest.

Techniques of Word Embeddings:

Word2Vec:

- Word2Vec is a popular word embedding technique that learns word representations through shallow neural networks.
- It comprises two models: Continuous Bag of Words (CBOW) and Skip-gram.
- CBOW predicts the current word given the context words, while Skip-gram predicts surrounding context words given the current word.

GloVe (Global Vectors for Word Representation):

- GloVe is another word embedding technique that leverages global word co-occurrence statistics.
- It constructs an explicit co-occurrence matrix and learns word embeddings by minimizing the difference between dot products of word vectors and their corresponding co-occurrence probabilities.

FastText:

- FastText extends Word2Vec by representing words as bags of character n-grams.
- It captures subword information, making it robust for handling out-of-vocabulary words and morphologically rich languages.

Further Suggested models that can be used are and how they can be leveraged

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that can be useful for analyzing the sentiment or emotional tone of text data, including Reddit comments. While VADER is primarily designed to determine the positive, negative, or neutral sentiment of a piece of text, it can also be utilized to gauge the controversiality of comments to some extent.

Here's how VADER can be applied to analyze the controversiality of Reddit comments:

- Sentiment Analysis: VADER can identify the sentiment polarity of a comment, distinguishing between positive, negative, and neutral sentiments. This information alone can be valuable in determining the overall tone of a comment.
- Intensity Scoring: VADER provides intensity scores for sentiment words, which can help gauge the strength of sentiment expressed in a comment. High-intensity scores indicate strong emotions, which are often associated with controversial or divisive statements.
- Contextual Analysis: VADER takes into account the context of words, including negations and modifiers, which allows it to better understand the sentiment expressed. This contextual analysis is crucial for accurately assessing the controversiality of comments, as it can identify intances where seemingly positive or neutral words are used in a sarcastic or ironic manner.
- Aggregating Scores: By analyzing a large number of comments using VADER, you can aggregate sentiment scores across all the comments in a thread or subreddit. This can help identify patterns and trends in controversial comments, allowing you to assess the overall controversiality of the discussion.

It's important to note that while VADER is a useful tool, it has certain limitations. It may struggle with sarcasm, irony, or ambiguous statements, which can affect the accuracy of its results. Additionally, VADER focuses primarily on sentiment analysis and may not capture the full complexity of controversiality, which can involve nuanced arguments and perspectives.

To achieve more accurate results, it's often beneficial to combine VADER's output with other analytical methods and consider human judgment and expertise when interpreting the data.

Second Model which is recommended is

TextBlob is another popular Python library for natural language processing tasks, including sentiment analysis. While it lacks the nuanced features of VADER, TextBlob can still provide useful insights when analyzing the controversiality of Reddit comments. Here's how TextBlob can be applied in this context:

- Sentiment Analysis: TextBlob can determine the sentiment polarity of a comment, categorizing it as positive, negative, or neutral. This information can help in understanding the overall sentiment expressed in a comment and whether it leans towards controversy or agreement.
- Subjectivity Analysis: TextBlob also provides a measure of subjectivity, indicating how subjective or opinionated a comment is. Higher subjectivity scores are often associated with more controversial or polarizing statements, as they reflect personal viewpoints rather than objective facts.
- Noun Phrase Extraction: TextBlob can extract noun phrases from text, which can be helpful in identifying key topics or entities being discussed in a comment. This can aid in understanding the context of controversial statements and the specific issues being debated.
- Sentence Parsing: TextBlob can parse sentences and extract various linguistic features, such as parts of speech and noun-verb relationships. This information can be useful for analyzing the structure and complexity of comments, which can provide insights into the depth of arguments and the potential for controversy.
- Customization and Training: TextBlob allows for customization and training on specific datasets. By providing labeled data on controversial comments, you can train TextBlob to better recognize and classify controversial language, enhancing its effectiveness for your specific analysis.

While TextBlob can provide valuable information regarding sentiment and subjectivity, it's important to note that it may not capture the nuanced aspects of controversiality as effectively as more specialized tools. Understanding and analyzing controversial comments often requires considering factors beyond sentiment, such as argumentation, logical fallacies, and contextual understanding.

To obtain more accurate and comprehensive results, combining TextBlob's output with other techniques, including manual review and analysis by domain experts, is beneficial to gain a deeper understanding of the controversiality of Reddit comments.

Last and most recommended model is

BERT (Bidirectional Encoder Representations from Transformers) is a powerful transformer-based deep learning model that has been widely used for various natural language processing (NLP) tasks, including sentiment analysis and understanding the context of text data. BERT can also be utilized for analyzing the controversiality of Reddit comments. Here's how BERT can be applied in this context:

- Contextual Understanding: BERT excels at capturing the contextual meaning of words and sentences by considering the surrounding context. This is particularly important for analyzing controversial comments on Reddit, as the interpretation of controversiality heavily relies on the context in which statements are made. BERT can capture the dependencies between words and understand the nuanced meaning of text.
- Fine-grained Sentiment Analysis: BERT can provide more fine-grained sentiment analysis by predicting sentiment scores or probabilities for various categories. Instead of simply categorizing a comment as positive, negative, or neutral, BERT can assign sentiment scores on a continuous scale, allowing for a more nuanced assessment of the controversiality of comments.
- Semantic Similarity: BERT can measure the semantic similarity between different comments, which can be useful for identifying similar or related controversial statements across Reddit threads. By comparing the content of comments, BERT can help identify patterns and common themes in controversial discussions.
- Transfer Learning and Pre-training: BERT is typically pretrained on a large corpus of text data, which gives it a strong foundation in understanding language. However, fine-tuning BERT on a specific task, such as analyzing controversiality, can lead to even better performance. By training BERT on labeled data that specifically captures controversial comments, the model can learn to identify and classify controversial language more accurately.
- Multi-task Learning: BERT can be trained on multiple related tasks simultaneously, which can enhance its performance on each individual task. By combining controversiality analysis with other related NLP tasks, such as sentiment analysis or topic classification, BERT can capture a broader understanding of the comments and their controversial nature.

It's important to note that leveraging BERT for analyzing controversiality requires labeled data that specifically captures controversial comments. Collecting and annotating such data can be challenging, but with a well-constructed dataset, BERT can provide valuable insights into the controversial nature of Reddit comments.

Additionally, working with BERT may require significant computational resources due to its large architecture and training requirements. Efficient implementations and hardware acceleration techniques can help mitigate this challenge.