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.

BERT (Bidirectional Encoder Representations from Transformers):

- BERT is a state-of-the-art language representation model based on the Transformer architecture.
- It learns contextual word embeddings by pre-training on large text corpora with masked language modeling and next sentence prediction tasks.