Exploration of ready-made projects and articles (relevant reddit comment analysis) to understand the overall gist of what all we have to do

Document Summary:

This project aims to develop a predictive model to assess the controversiality of comments on Reddit, specifically focusing on discussions related to the Russia-Ukraine war. The goal is to create a tool that can help Reddit determine whether to promote or restrict the visibility of comments based on their potential to generate controversy, thus fostering a conducive online discussion environment while respecting diverse viewpoints.

Detailed Explanation of Project Pipeline:

1. Data Cleaning:

Data cleaning is a critical initial step that involves removing duplicates, handling missing values, and eliminating special characters to ensure data integrity. Lowercasing, tokenization, stopword removal, stemming, lemmatization, and addressing contractions, spell checking, and removing URLs and email addresses are essential tasks to standardize and enhance the quality of the dataset. Dealing with imbalanced data, normalization, and standardization ensure consistency and reliability.

2. Exploratory Data Analysis (EDA):

EDA focuses on understanding dataset characteristics, distributions, and relationships through visualizations like plots, histograms, and correlation matrices. It helps identify patterns, outliers, and variables of interest, guiding feature selection and model development. EDA aids in making data-driven decisions, detecting anomalies, and verifying data assumptions.

3. Feature Engineering:

Feature engineering involves techniques for transforming raw data into informative features suitable for model training. In sentiment analysis, methods like NLTK and BERT are used to quantify the emotional tone of comments, enabling the model to discern positive, negative, or neutral sentiments. Topic modeling techniques, such as TF-IDF, NMF, and LDA, identify latent topics within the data, facilitating categorization and understanding of comment content. Other relevant techniques include word embeddings, word frequency analysis, and TF-IDF, which contribute to capturing semantic relationships and contextual information. Feature engineering is essential for enhancing model performance and interpretability, providing valuable insights for accurately predicting comment controversiality.

https://www.kaggle.com/code/andreshg/nlp-glove-bert-tf-idf-lstm-explained *
https://github.com/Abeldewit/RCP/tree/master
https://www.kaggle.com/code/leonwolber/reddit-nlp-topic-modeling-prediction/notebook

https://www.geeksforgeeks.org/sentiment-classification-using-bert/
https://www.geeksforgeeks.org/word-embeddings-in-nlp/
https://www.geeksforgeeks.org/nlp-part-of-speech-default-tagging/

https://www.kaggle.com/datasets/asaniczka/public-opinion-russia-ukraine-war-updated-daily?resource=download

https://github.com/avisionary/reddit-comments-analysis/blob/main/project_eda.ipynb

https://github.com/topics/reddit-comments

https://github.com/SoniSakshi1999/SOCIAL-MEDIA-LISTENING-A-CASE-STUDY-OF-REDDIT-USING-BIG-DATA-ANALYTICS/tree/master

https://github.com/wafer110/Python-NLP-Analyze_TextualData_on_Reddit_Comments

https://medium.com/coinmonks/remaking-of-shortened-sms-tweet-post-slangs-and-word-contraction-into-sentences-nlp-7bd1bbc6fcff

https://github.com/jshiohaha/redditCommentsAndPresidentialElection

https://medium.com/wisio/a-gentle-introduction-to-doc2vec-db3e8c0cce5e

https://github.com/ishalyminov/reddit_tools

https://github.com/syedaminx/Reddit-Sentiment-Analysis

https://github.com/bkent97/reddit-sentiment-analysis

https://monkeylearn.com/blog/what-is-tf-idf/

https://www.kaggle.com/code/prakharrathi25/sentiment-analysis-using-bert

https://www.kaggle.com/code/buyuknacar/comparing-sentiment-classifiers-txblob-vader-bert

https://medium.com/@amanabdulla296/sentiment-analysis-with-vader-and-twitter-roberta-2ede7fb78909

https://francisgichere.medium.com/sentiment-analysis-of-app-reviews-a-comparison-of-bert-spacy-textblob-and-nltk-9016054d54dc

https://itsmariodias.medium.com/leveraging-pre-trained-transformers-for-downstream-tasks-4e234656ca96

https://www.kaggle.com/code/tanulsingh077/twitter-sentiment-extaction-analysis-eda-and-model

https://www.kaggle.com/code/aryantiwari123/reddit-tweets-analysis

https://www.kaggle.com/code/dylanyves/twitter-and-reddit-sentiment-analysis

https://www.kaggle.com/code/kaushalkrishna2000/reddit-post-analysis

https://www.kaggle.com/code/alessiopeluso/reddit-sentiment-analysis

https://www.kaggle.com/code/thomaskonstantin/reddit-wallstreetbets-posts-sentiment-analysis