How can doing Principal Component Analysis with Text Embeddings/Univariate Feature Selection help in determining the controversiality of the reddit post

Text embedding is a representation of words, phrases, or documents in a high-dimensional vector space. These embeddings will result in a high-dimensional feature vector, especially when dealing with large vocabularies or complex embedding models which is very much prevalent in our dataset. Each word or token in the vocabulary will be represented by a vector with hundreds or even thousands of dimensions.

Seeing these issues we can use PCS for dimensionality reduction. As we know PCA (Principal Component Analysis) is a technique used to reduce the number of features (dimensions) in a dataset while preserving most of its variability or information content. In context of text embeddings, PCA can be applied to these high-dimensional feature vectors to reduce their dimensionality. PCA will work by identifying the directions, or principal components, along which the data which varies the most. It will then project the data onto a lower-dimensional subspace spanned by these principal components.

Next step shall be Preserving Variance and Capturing Relevant Information. Since PCA aims to retain as much variance as possible in the original data while reducing its dimensionality. By selecting a lower number of principal components (dimensions), PCA effectively captures the most relevant information in the data. The retained principal components represent combinations of the original features that explain the maximum amount of variance in the dataset.

Identification of Important Features: Univariate feature selection techniques, such as SelectKBest or chi-square tests, can be applied to the text embeddings to identify the most informative features or words that contribute the most to the controversiality of Reddit posts. These techniques evaluate each feature independently of others, making them suitable for high-dimensional data. By selecting only the top k features based on their scores, you can reduce noise and focus on the most relevant information for determining controversiality.

Steps to achieve Principal Component Analysis with Text Embeddings/Univariate Feature Selection help in determining the controversiality of the reddit post

1. Extract Text Embeddings: Text embeddings represent words or documents in a high-dimensional vector space, capturing semantic similarities and relationships between words. You need to first extract text embeddings from the textual data in your dataset. This can be done using pre-trained word embeddings models such as Word2Vec, GloVe, or fastText, or using more advanced language models like BERT or GPT.
2. Vectorize Text Data: Convert the textual data (post titles, self-text, comments) into numerical representations that can be input to PCA. This often involves tokenization, where each word is mapped to its corresponding embedding vector.
3. Aggregate Embeddings: Since each post or comment may contain multiple words, you need to aggregate the embeddings to obtain a single vector representation for each post or comment. Common aggregation methods include averaging the embeddings of individual words or using more sophisticated techniques such as weighted averaging or concatenation.
4. Apply PCA: Once you have obtained the aggregated embeddings for each post or comment, you can apply PCA to reduce the dimensionality of the data. PCA will identify

the principal components (linear combinations of the original features) that capture the most variance in the data. By projecting the data onto a lower-dimensional subspace spanned by these principal components, you can reduce noise and focus on the most informative features.

5. Select Number of Components: Decide on the number of principal components to retain based on the desired level of dimensionality reduction and the amount of variance explained by the components. You can use techniques like scree plots, cumulative explained variance, or cross-validation to determine the optimal number of components.
6. Transform Data: Transform the original data into the reduced-dimensional space using the selected principal components.

To achieve univariate feature selection for the given dataset, which contains Reddit posts and associated attributes, including text data. Post Data Preprocessing, Feature Extraction is the next process where in following tasks need to be performed

- Generate text embeddings by Use techniques like Word2Vec, GloVe, or BERT to convert the text data into dense numerical vectors.
- Alternatively, create a document-term matrix or TF-IDF matrix to represent the text data as a numerical feature matrix.

Apply univariate feature selection techniques to select the most informative features (words or tokens) from the dataset based on their relationship with the target variable (e.g., controversiality of Reddit posts). Common univariate feature selection methods include

- SelectKBest: Select the top k features based on a scoring function (e.g., chi-square, ANOVA F-test).
- Chi-square tests: Assess the independence between each feature and the target variable (controversiality).
- Information gain: Measure the reduction in entropy or uncertainty in the target variable given the presence of a feature.
- Mutual information: Measure the amount of information that one feature provides about another.

Model Building and Evaluation:

Split the dataset into training and testing sets. Train machine learning models (e.g., logistic regression, random forest, support vector machine) using the selected features as input variables. Evaluate the performance of the models using appropriate metrics (e.g., accuracy, precision, recall, F1-score) on the testing set. Compare the performance of models built with and without feature selection to assess the impact of feature selection on model performance and interpretability.

Model Training and Prediction: Once the dimensionality is reduced and important features are selected, you can use the processed data to train machine learning models for predicting the controversiality of Reddit posts. You can experiment with various classification algorithms, such as logistic regression, random forests, or support vector machines, to build predictive models. These models can learn patterns in the selected features and predict whether a post is controversial or not based on its textual content.

Evaluation and Interpretation: After training the models, it's essential to evaluate their performance using appropriate metrics such as accuracy, precision, recall, or F1-score.

Additionally, you can analyze the coefficients or feature importances of the models to understand which words or features contribute the most to predicting controversiality. This analysis can provide insights into the characteristics of controversial Reddit posts and help in understanding the factors driving their controversial nature.

Overall, applying PCA with text embeddings and univariate feature selection techniques can help in preprocessing and extracting meaningful information from textual data, enabling the development of predictive models for determining the controversiality of Reddit posts.

How Can Ridge and Lasso Regression be leveraged to gain insights whether the post if controversial or not.

Ridge Regression and Lasso Regression are two regularization techniques used in linear regression models to prevent overfitting and improve model generalization.

Ridge Regression:

Ridge Regression adds a penalty term to the ordinary least squares (OLS) objective function, which is proportional to the square of the magnitude of the coefficients. This penalty term, also known as the L2 regularization term, helps in shrinking the coefficients towards zero, but not exactly to zero. Ridge Regression is particularly useful when dealing with multicollinearity in the data, as it tends to distribute the coefficients more evenly across correlated features.

Lasso Regression:

Lasso Regression (Least Absolute Shrinkage and Selection Operator) also adds a penalty term to the OLS objective function, but this penalty term is proportional to the absolute value of the coefficients (L1 regularization term). Lasso Regression has the property of inducing sparsity in the coefficient estimates, meaning it can force some coefficients to be exactly zero. Due to this property, Lasso Regression can be used for feature selection, as it effectively selects only the most relevant features while setting the coefficients of irrelevant features to zero.

How we can determine the controversiality of posts in the given dataset using Ridge and Lasso Regression models

Feature Engineering:

- Preprocess the text data to extract features that can be used for prediction, such as word frequencies, TF-IDF scores, or text embeddings.
- Transform the text data into a numerical format that can be fed into the regression models.

Model Training:

- Split the dataset into training and testing sets.
- Train Ridge Regression and Lasso Regression models using the training data, with the controversiality of posts as the target variable and the extracted features as input variables.

Model Evaluation:

- Evaluate the performance of the trained models on the testing set using appropriate evaluation metrics, such as mean squared error (MSE) or R-squared.
- Compare the performance of the Ridge Regression and Lasso Regression models to assess their ability to predict the controversiality of posts.

Interpretation:

- Analyze the coefficients of the trained models to understand the importance of different features in predicting the controversiality of posts.
- In Lasso Regression, features with non-zero coefficients are considered important predictors of controversiality, while features with zero coefficients are considered irrelevant.

Iterative Refinement:

- Experiment with different hyperparameters (e.g., regularization strength) to optimize the performance of the models.
- Consider incorporating additional features or refining the feature extraction process based on the insights gained from model interpretation.