Analysis of Pretrained Word Embedding Models for Detecting AI-Generated Text

Moroti Sonde - R11904541
Sruthi Mandalapu - R11906160
Charishma Rathan Bala - R11908154
Chandrika Kancheti - R11908347
Gowtham Edumudi - R11912904
Raga Poojitha Kinthada - R11902142

Summary

This research work aims to address the challenges faced in the education sector particularly with the use of AI in writing essays by leveraging Natural Language Processing (NLP) techniques and Bidirectional Long Short-Term Memory (Bi-LSTM) model in identifying AI-generated text. The dataset to be used comprises a total of 28827 text containing human written essays and AI-generated essays. The primary focus of this work is to evaluate the effectiveness of two popular word pre-trained embeddings, namely Word2Vec and Glove in training a Bi-LSTM model to identify AI-generated and human-generated text.

First is the data collection and preprocessing of the text by removing stop-words and special characters before splitting it into training and test sets. Subsequently, we will be using the pretrained word embeddings (word2Vec and Glove) to convert the textual data into vector representation. The embedding created will then serve as input features to our Bi-LSTM model which is known for capturing textual information in sequential data.

We will then train the model on the training set, with subsequent evaluation on the testing dataset to assess its performance in identifying AI-generated text and human-generated text. Furthermore, we will be using metrics such as accuracy and precision to measure the effectiveness of our model.

Finally, a comprehensive analysis of the experimental results of word2vec and Glove embeddings model will be done to provide more insights into the classification task as well as discussion about real-world implications, such as how to improve the performance of each model to reduce the rate of false positives.