K-Means Clustering for News Classification (Fake or Not)

Jagannadham Bharath

23M0038

Aerospace Propulsion, Aerospace Engineering Department, IIT Bombay

Introduction

In this report, we present an analysis of a text clustering project aimed at categorizing news articles as either fake or real. The project involves preprocessing textual data, utilizing Word2Vec for feature extraction, applying KMeans clustering, and evaluating the clustering accuracy. The goal is to enhance the understanding of the text analysis process and assess the efficacy of the applied approach.

Data Preprocessing

The initial steps of the project include data preprocessing to prepare the news articles for analysis. URL links, lowercase letters, punctuation, and other irrelevant elements are removed using custom filtering techniques. The preprocessed text data is then stored for further processing.

Feature Extraction using Word2Vec

Word2Vec, a popular word embedding model, is employed to transform the preprocessed text data into numerical vectors. Each word in the articles is represented by a vector, capturing semantic relationships between words. This conversion enables the comparison and analysis of textual data in a numerical form.

KMeans Clustering

To categorize the news articles, a KMeans clustering algorithm is utilized. The preprocessed and transformed data is clustered into two distinct categories: fake and real news. The algorithm aims to group similar articles together based on the extracted features from Word2Vec.

Evaluation and Results

The accuracy of the clustering process is evaluated by comparing the predicted labels from the KMeans algorithm with the actual labels of the articles. The accuracy metric indicates the percentage of correctly clustered news articles. The evaluation provides insights into the effectiveness of the clustering approach in differentiating between fake and real news.

Conclusion

The presented text clustering project showcases the application of data preprocessing, feature extraction with Word2Vec, and KMeans clustering in categorizing news articles as fake or real. The results of the clustering process are informative in terms of evaluating the accuracy of the model and understanding the text analysis workflow. Further enhancements could involve exploring different algorithms, incorporating additional features, and fine-tuning the model for improved accuracy.