Financial Statement Classification

Objective

This project aims to automatically categorize HTML files containing financial statements (balance sheet, income statement, cash flow statement, Notes and Others) using a machine learning model. This project involves the processing and classification of text data extracted from HTML files.

Processes Involved

1.Parsing HTML Files:

- Using BeautifulSoup, HTML files are parsed to extract features.
- Extracted features are stored in a structured format.

2. DataFrame Creation:

- Features and class names are used to create a pandas DataFrame.

3. Label Encoding:

- Class names are encoded using LabelEncoder to convert them into numerical values.

4. Word Embedding with BERT:

- Extracted feature text is transformed using a BERT vectorizer.
- BERT model is used to perform word embedding, converting text into numerical vectors.

5. Handling Class Imbalance:

- SMOTE (Synthetic Minority Over-sampling Technique) is applied to eliminate imbalance in the dataset.

6. Dataset Splitting:

- The dataset is split into training and testing sets.

7. Model Training:

- Logistic Regression model is trained using the training dataset.
- The trained model achieved 99% accuracy on the training dataset.

8. Model Prediction:

- The trained model is used to predict the class labels on the test dataset.
- The model achieved 92% accuracy on the test dataset.

9. Model Saving:

- The trained model is saved using pickle for future use.

10. Visualization:

- Word embeddings are plotted on a 2D plane using K-Means clustering for visualization.

Reference:

Table Classification: An Application of Machine Learning to Web-hosted Financial Documents by Marc Vilain, John Gibson, Benjamin Wellner, and Rob Quimby