CS472 Final Code Description File

My final project is implemented in Jupyter Notebook (IPython). It focuses on exploring various machine learning methods for stress classification based on text data. The code performs the following main functionalities:

1. Importing necessary libraries:
   - nltk: Natural Language Toolkit for text preprocessing and tokenization.
   - sklearn: Scikit-learn library for machine learning models and evaluation metrics.
   - pandas: Library for data manipulation and analysis.
   - gensim: Library for Word2Vec implementation.
   - tensorflow: Library for building and training neural networks.
   - matplotlib, seaborn: Libraries for data visualization.

2. Mounting Google Drive:
   - The code mounts Google Drive to access the required data files.

3. Loading and analyzing the data:
   - The code loads a CSV file (`Stress.csv`) into a pandas DataFrame for further processing.
   - Data visualization is performed to understand the distribution of labels.

4. Data preprocessing:
   - Text preprocessing functions are defined to remove punctuation, convert to lowercase, tokenize, and remove stop words.
   - The text data in the DataFrame is preprocessed using the defined functions.

5. Bag of Words (BoW) representation:
   - The preprocessed texts are tokenized, and a vocabulary is created.
   - A BoW matrix is generated using the CountVectorizer from scikit-learn.

6. Machine learning methods based on BoW:
   - Logistic Regression:
     - The data is split into training and testing sets.
     - A logistic regression model is initialized, trained on the training set, and evaluated on the testing set.
   - Support Vector Machine (SVM):
     - An SVM model is initialized, trained, and evaluated using the same procedure as logistic regression.

7. Word2Vec representation:
   - The text data is preprocessed, tokenized, and used to train a Word2Vec model.
   - The trained Word2Vec model is used to obtain word vectors and find similar words.

8. Machine learning methods based on Word2Vec:
   - Logistic Regression:
     - The Word2Vec vectors are averaged to represent each text instance.
     - Data is split into training and testing sets.
     - A logistic regression model is trained and evaluated on the respective sets.

- SVM:
  - An SVM model is initialized, trained, and evaluated using the averaged Word2Vec vectors.
- Convolutional Neural Network (CNN):
  - The text data is converted into sequences of word indices.
  - The data is padded to ensure equal length sequences.
  - A CNN model is defined, compiled, trained, and evaluated on the data.

9. Data Visualization:
- A bar chart is created to compare the accuracies of different machine learning methods.
- The accuracies of the logistic regression, SVM, and CNN models based on both BoW and Word2Vec are displayed.