Word Embeddings and text Classification

Problem Statement:

Develop an approach to effectively train a feedforward neural network for text classification while simultaneously deriving meaningful word embeddings. The objective is to build a neural network model that not only classifies text accurately but also learns high-quality word representations, optimizing both classification performance and embedding utility.

Dataset:
ScienctTopics.csv uploaded along with problem statement.

Part 1 : The goal in this part is to prepare the text data. (3 Marks)

a. Read the text data and lowercasing and EDA
b. Tokenization, encoding and Text vectorization
c. Extract features and labels

Part 2 (3 marks):

Build a simple feedforward Neural Network with the below architecture:

1 Embedding layer of shape (,10)

3 Dense layers with 100 neurons and relu activation functions

1 appropriate output dense layer with appropriate activation function

Compile the model and train for 50 epocs (or accuracy score of atleast 75%)

Part 3:

Extract the embedding vector from the model – 1 marks

Visualize the word embeddings in a scatter plot – 1 marks (use PCA)

Predict the class for the below given description (1 marks):

a. “For the last one, neutrinos have mass. This is well established at many many sigma. There is good agreement in the so-called solar parameters between solar oscillations measured by SNO, Borexino, SuperK, and others with long-baseline reactor measurements by KamLand. The former measurement will improve considerably with DUNE and the latter with JUNO.”

b. “Polypropylene is used specifically for its chemical inert-ness. You're not going to have much luck with a chemical solvent.\n\nYeah, I read something like that. To be honest, I can't be arsed to melt in a metal mesh”

Calculate Cosine Similarity of the below pair of words from the extracted word embeddings: (1 marks)

a. Concentration, Purity
b. Safety, Precaution
c. Toxic, Harmful

Expected Output :
Jupyter notebook with clear mention of all the members of your team including Name and Bits ID.

There should not be any error cell, and all the cells should be executable.