AIML | MODULE PROJECT

Sequential NLP

TOTAL SCORE 60

General Instructions:

1. Submission of all the parts is expected in 1 notebook only
2. Expected submission format: 1 '.ipynb' notebook and 1 '.html' notebook only
3. 50% marks will be deducted if insights/steps are missing in the corresponding questions.
4. If output for any code cell is missing, 50% marks will be deducted.
5. Any kind of Plagiarism will lead to 0 (zero) Marks.

Submission Format:

1. '.ipynb' (Jupyter Notebook) and
2. '.html' (Jupyter Notebook > File > Download as > HTML)
5 Marks will be deducted if submission in any of the formats is missing.

Part A - 30 Marks

- **DOMAIN:** Digital content and entertainment industry
- **CONTEXT:** The objective of this project is to build a text classification model that analyses the customer's sentiments based on their reviews in the IMDB database. The model uses a complex deep learning model to build an embedding layer followed by a classification algorithm to analyse the sentiment of the customers.
- **DATA DESCRIPTION:** The Dataset of 50,000 movie reviews from IMDB, labelled by sentiment (positive/negative). Reviews have been preprocessed, and each review is encoded as a sequence of word indexes (integers). For convenience, the words are indexed by their frequency in the dataset, meaning the for that has index 1 is the most frequent word. Use the first 20 words from each review to speed up training, using a max vocabulary size of 10,000. As a convention, "0" does not stand for a specific word, but instead is used to encode any unknown word.
- **PROJECT OBJECTIVE:** To Build a sequential NLP classifier which can use input text parameters to determine the customer sentiments.

Steps and tasks: **[Total Score: 30 Marks]**

1. Import and analyse the data set. [5 Marks]
   Hint: - Use `imdb.load_data()` method
   - Get train and test set
   - Take 10000 most frequent words
2. Perform relevant sequence adding on the data. [5 Marks]
3. Perform following data analysis: [5 Marks]
   - Print shape of features and labels
   - Print value of any one feature and it's label
4. Decode the feature value to get original sentence [5 Marks]
5. Design, train, tune and test a sequential model. [5 Marks]
   Hint: The aim here Is to import the text, process it such a way that it can be taken as an inout to the ML/NN classifiers. Be analytical and experimental here in trying new approaches to design the best model.
6. Use the designed model to print the prediction on any one sample. [5 Marks]

Please Note:

Intentionally limited questions/instructions are provided so that learners can explore more and perform more research since learners are comfortable with all the concepts and implementation.

Part B - 30 Marks

- **DOMAIN:** Social media analytics
- **CONTEXT:** Past studies in Sarcasm Detection mostly make use of Twitter datasets collected using hashtag based supervision but such datasets are noisy in terms of labels and language. Furthermore, many tweets are replies to other tweets and detecting sarcasm in these requires the availability of contextual tweets.In this hands-on project, the goal is to build a model to detect whether a sentence is sarcastic or not, using Bidirectional LSTMs.
- **DATA DESCRIPTION:**

The dataset is collected from two news websites, theonion.com and huffingtonpost.com.

This new dataset has the following advantages over the existing Twitter datasets:

Since news headlines are written by professionals in a formal manner, there are no spelling mistakes and informal usage. This reduces the sparsity and also increases the chance of finding pre-trained embeddings.

Furthermore, since the sole purpose of TheOnion is to publish sarcastic news, we get high-quality labels with much less noise as compared to Twitter datasets.

Unlike tweets that reply to other tweets, the news headlines obtained are self-contained. This would help us in teasing apart the real sarcastic elements

Content: Each record consists of three attributes:

is_sarcastic: 1 if the record is sarcastic otherwise 0

headline: the headline of the news article

article_link: link to the original news article. Useful in collecting supplementary data

Reference: https://github.com/rishabhmisra/News-Headlines-Dataset-For-Sarcasm-Detection

- **PROJECT OBJECTIVE:** Build a sequential NLP classifier which can use input text parameters to determine the customer sentiments.

Steps and tasks: **[Total Score: 30 Marks]**

1. Read and explore the data [3 Marks]
2. Retain relevant columns [3 Marks]
3. Get length of each sentence [3 Marks]
4. Define parameters [3 Marks]
5. Get indices for words [3 Marks]
6. Create features and labels [3 Marks]
7. Get vocabulary size [3 Marks]
8. Create a weight matrix using GloVe embeddings [3 Marks]
9. Define and compile a Bidirectional LSTM model. [3 Marks]

   Hint: Be analytical and experimental here in trying new approaches to design the best model.
10. Fit the model and check the validation accuracy [3 Marks]

Please Note:

Intentionally limited questions/instructions are provided so that learners can explore more and perform more research since learners are comfortable with all the concepts and implementation.