Product Browse Node Classification

1. Problem Statement :-

- Amazon catalog consists of billions of products that belong to thousands of browse nodes (each browse node represents a collection of items for sale). Browse nodes are used to help customer navigate through our website and classify products to product type groups. Hence, it is important to predict the node assignment at the time of listing of the product or when the browse node information is absent.
- Our goal is to create a model that accurately classifies BROWSE_NODE_ID.

2. Introduction:-

- As we all know, Amazon already has 310 million active Amazon customer accounts. Day-by-day, millions are people uploading their products, so it is hard to classify products to correct id at that time. For solving this problem, we have to develop a model that will classify those products.

- We can use various machine learning classification algorithms like random forest classifier, K-Nearest Neighbor to solve the problem. But for this problem, we will use a neural network, and we know that for extensive data neural network is the best option to solve the problem efficiently.

3. Dataset:-

- The dataset contains columns like product title, description, bullet points and brand with the label for 3M Products with some noise.

  Here is a link to the dataset:- https://bit.ly/3lun2R3

Data Description

- Key column – PRODUCT_ID
- Input features – TITLE, DESCRIPTION, BULLET_POINTS, BRAND
- Target column – BROWSE_NODE_ID
- Train dataset size – 2,903,024
- Number of classes in Train – 9,919
- Overall Test dataset size – 110,775
- These Datasets contain 9919 unique classes.

4. Preprocessing of Data:-

- To process the data, first, we have to handle the null values. We have two choices either delete the null rows or fill the null values with meaningful values. For better performance, we would fill the values with some meaningful text. And the dataset has two main features columns. So we will combine the TITLE and BRAND columns and drop other columns except for BROWSE_NODE_ID, fill the null values with 'unknown' for better performance.

- Now we have a column with valuable features. For the process, the text data NLTK is the best option. After text processing, we'll use a tokenizer to convert data into an integer. Now data is prepared to feed the neural network.

  (Note:-Dataset is huge, so I suggest you guys use a chunks of the dataset.)

5. Model:-

LSTM neural network

- Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems.

- LSTMs are a complex area of deep learning. It can be hard to get your hands around what LSTMs are, and how terms like bidirectional and sequence-to-sequence relate to the field.

6. Conclusion:-

- In this report, we learned how efficiently solve the Product Browse Node Classification problem with a Neural network. We can manipulate the model architecture for better performance.