Amazon ML Challenge

Team : Elementals

Prateek | Manish | Adhesh | Shanmukh

Details of Competition

Competition : Multi Class Text Classification
Host : Hacker-Earth
Metric : Accuracy
Time of Competition : 2days:23hrs:59min

Details of Data

Full Train/Test dataset details:

- **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

Data Preprocessing

Libraries used:

- re
- langdetect
- deep-translator

Steps followed:

- Removed special characters and emojis using re.
- Translated non-english text to english using langdetect and deep-translator.
- Removed stop-words.
- Decontracted some of the words.

Our Approach

Libraries used:

- Sentence_transoformer
- RAPIDS

Steps followed:

- First the text is converted to embeddings using pre-trained models such as **paraphrase-mpnet-base-v2 , paraphrase-MiniLM-L6-v2 paraphrase-MiniLM-L3-v2**
- Dimension of Embeddings : 384
- Embeddings of training data are sent into **KNNClassifier** present in **CuML** library
- Then the trained KNNClassifier is used to predict on test embeddings.
- We also used **mode** technique i.e. using most frequently predicted label obtained from different experiments
- **Cross Validation** is also used to train **KNNClassifier**

Experiments

- Used **NearestNeighbour, SVM, RandomForest Classifier** techniques but results are not better compared to **KNNClassifier**.
- Different size embeddings (384,768)
- Combined TITLE,DESCRIPTION and TITLE,DESCRIPTION, BULLET_POINTS and TITLE, DESCRIPTION, BULLET_POINTS

Details of Files

1. **amazon_ml_preprocessing.ipynb** : code to preprocess text
2. **amazon_ml_translation_csv.ipynb** : code to translate non-english text
3. **amazon_ml_mode.ipynb** : code to create submission file from multiple submission files using mode technique
4. **amazon_ml_training.ipynb** : code to train embeddings and predict on test embeddings
5. **amazon_ml_embeddings.ipynb** : code to generate embeddings from csv files