MIDAS@IIITD MULTIMODAL DIGITAL MEDIA ANALYSIS LAB

SUMMER INTERNSHIP 2021

Task 3 NLP

README

Submitted by : -

VANSH GUPTA

Problem Statement:-

Use a given dataset to build a model to predict the category using description. Write code in python. Using Jupyter notebook is encouraged.

1. Show how you would clean and process the data
2. Show how you would visualize this data
3. Show how you would measure the accuracy of the model
4. What ideas do you have to improve the accuracy of the model? What other algorithms would you try?

Approach

After taking a glance at the problem statement, I began to consider the approaches that could be used to solve it.

I came across a method called **Hierarchical Classification**[1] that was ideal for our problem statement. Since, it was clearly stated that we have to predict primary category, we'll discuss about this in Future Improvements section.

I've split the definition into subsections to forecast the primary category:-

1. **Getting started**
   We'll take a look at the data and analyse it in this segment.

   Insides discovered:
   • Group tree width
   • Useful columns
   • Additional columns with each depth stage

2. **Label's handling**
   We'll be messing around with category labels in this section, mostly cat level1.

   • There were 266 primary categories at the outset.
   • We discovered that certain categories in the table have only one occurrence.
   • After deleting the categories with less than ten occurrences, we were left with 28 categories.

3. Text Preprossing

We'll do data preprossing steps in this segment, which include things like text cleaning and vector conversion.

Cleaning procedures include the following:

- Lemmatization and stemming
- Remove stop terms
- Remove punctuation
- Remove extra space, special characters, and numbers

Vectorization methods:

- Count Vectorization
- Tf-IDF
- Glove embeddings

4. Trying different models

In this part, we'll compare and contrast various classification models:
The following models have been compared:-

Model	**Accuracy**	**Presion**	**Recall**	**F1 score**
KNN	94.93	94.95	94.93	94.84
SVM	97.27	97.28	97.27	97.23
Naive bayes	81.13	80.94	81.13	77.47
Random Forest	95.57	95.56	95.57	95.34
XgBoost	96.86	96.89	96.86	96.83
LSTM	95.1	95.1	95.1	95.1
Bidirectional LSTM	93.75	93.75	93.75	93.75

SVM is the best model based on the above parameters.

5. Future Improvements

In the future, I'd like to use this information to construct a model that can forecast product classification all the way down to the root of the classification tree.

Submitted by :- Vansh Gupta
vgvanshg25@gmail.com

We have two solutions for this:

1. A complete tree is a class, but this would result in thousands of classes, which would be inaccurate.

2. To divide the tree into tiers and then predict level by level.

This would be a precise solution, but it will include the training of tens of models.

I would personally prefer second method, and would like to discuss it further with MIDAS.

References

1. https://medium.com/rate-engineering/hierarchical-text-classification-for-rates-404c6c399f6b#:~:text=A%20classifier%20assigns%20labels%20to,products%20are%20chosen%20as%20input.&text=That%20is%20why%20we%20make%20the%20classifier%20hierarchical%20as%20well.

2. https://paperswithcode.com/search?q_type=papers&q=Product+Classification+in+E-Commerce+using+Distributional+Semantics

3. https://towardsdatascience.com/multi-label-multi-class-text-classification-with-bert-transformer-and-keras-c6355eccb63a

Submitted by :- Vansh Gupta
vgvanshg25@gmail.com