MIDAS Evaluation Task 4:NLP Report

- Yash Mishra

Task: To predict product category using product description

- The Task is Requirements are NLP(Natural Language Processing) and classification models.
- I have done this task in the Jupyter Notebook on Google colab as it provides good computations power(Higher RAM,High Space,GPU)

Experiments during text preprocessing

Model or technique used	*Results*
Word2vec()	don't give fair results(fail)
Countvectorizer()	give fair results(Succeed)
Tfidfvectorizer()	RAM crashed(fail)

Experiment done while doing data visualization

Model or technique used	*Results*
Pandas visualization failed	failed
Matplotlib	failed
Seaborn	able to plot distplot(pass)

Experiment done while making ML models

Model used	*Results*
Multinomial Naïve bayes	Succeed(92% accuracy)
Logistic regression	Succeed(96% accuracy)
Random Forest	Failed to fit(RAM crashed)
Xgboost classifier	Failed to fit(RAM crashed)

Results

1. Data cleaning and processing is done by removing stopwords ,symbols numbers , converting text to lowercase and stemming of data.
2. Visualization is done by seaborn as it is simple to use and give nice results.
3. Multinomial Naive bayes is used classification product category.
4. Accuracy is determined using classification report and confusion matrix.
5. I have used other models like logistic regression ,random forest etc, but nothing worked every model has some flaws.
6. I would like to try LGBM classifier if I got the chance further to work on this project.

References

1. https://towardsdatascience.com/tagged/nlp
2. https://stackoverflow.com/
3. https://www.kaggle.com/abhishek/approaching-almost-any-nlp-problem-on-kaggle
4. https://www.udemy.com/course/nlp-natural-language-processing-with-python/