

Ratings Predictions Project

Submitted by:

Assie Houphouet Olivier

**ACKNOWLEDGMENT**

First I want to say thank you to sir ***Keshav Bansal*** our mentor which gave us this project. The very basic idea behind the gathering the data (Products reviews from e-commerce web sites) has been covered to us, we have been made to work on the projects where we had to scrape the data from websites and then work on it. So with the experience we got from during web scraping projects help us to gather the needed data and finish the first step of the project.

The rest of the project needed text pre-processing and model building; those concept have already been covered in my academy (Data Trained) . I have used NLP method to filter off the odd/extra data from the review texts and the ML modelling is done.

**INTRODUCTION**

* The business problem involved in this project is that we have a client who runs a website where people write different technical review to evaluate the product. Now they want to add a new feature to rate the products. The new feature is for the reviews from now, but they already written ones will be left out. Thus we will have to come up with a model which will predict the ratings based on the product reviews.
* The appropriate domain problem like customer retention projects, sentiments analysis projects are some of the best example to refer and work on it.
* The only objective was to find the right and good model in the end of the project which will predict the ratings for the reviews.

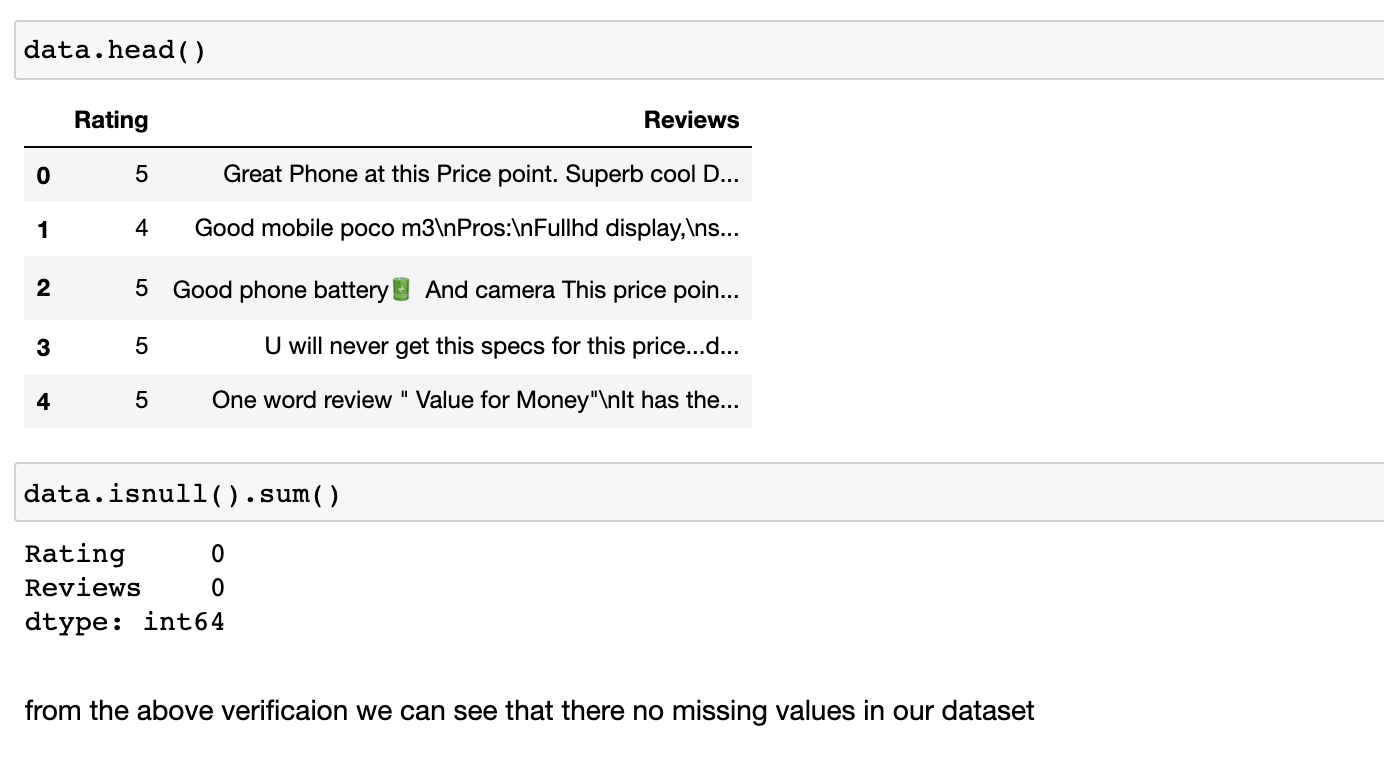
**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

There aren’t much mathematical calculation involved in this project since the data is from reviews texts. The analysis is done during the text conversion to numerical with the help of Tfidf and CountVector.

* Data Sources and their formats

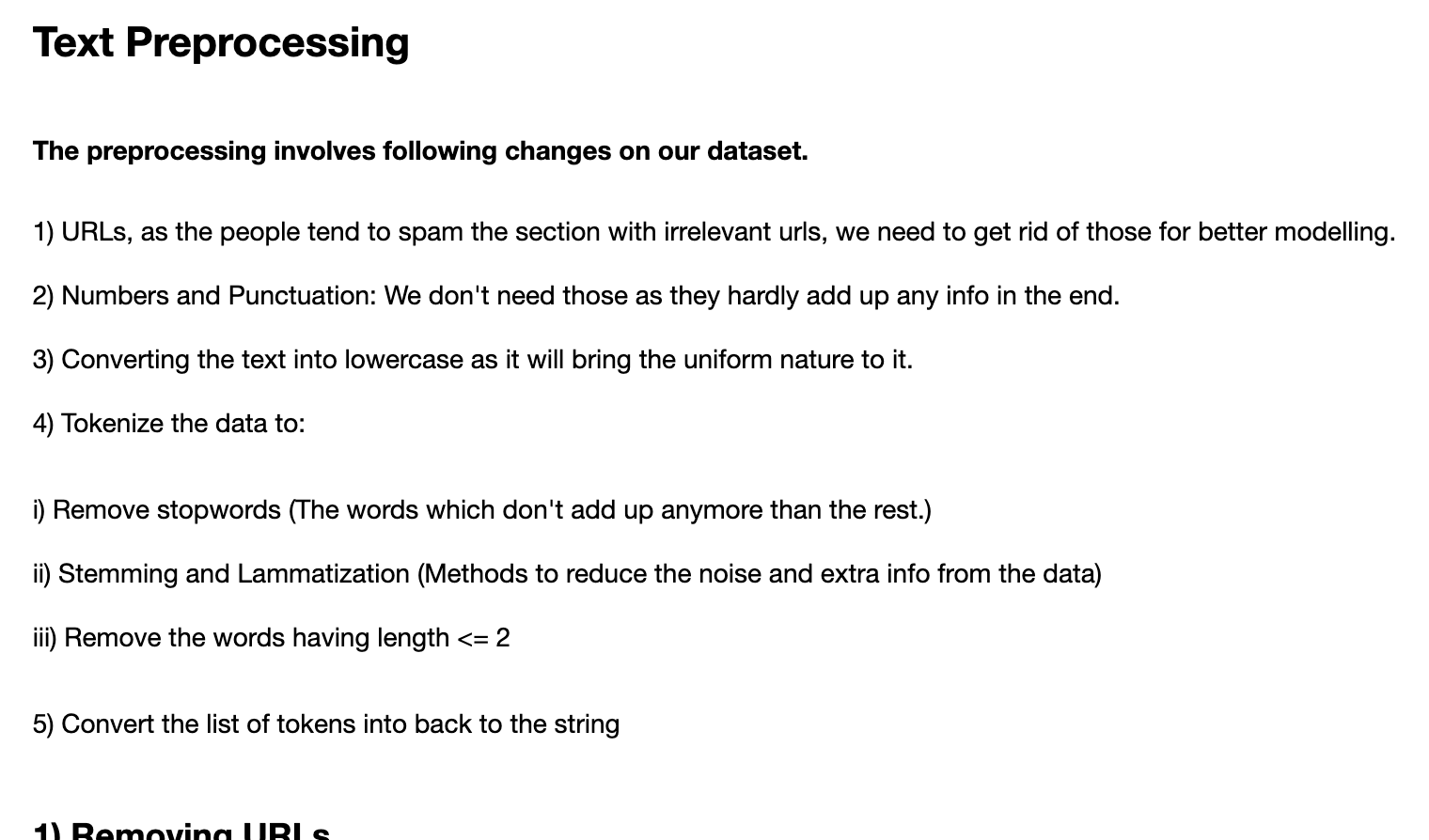
The raw data is basically the reviews from e-commerce websites. The customers express their opinions on their site, we as data scientist scrape such data to do the sentiment analysis on it and model it eventually. The format of the data is raw and object datatype in nature after converted it into comma separated values (csv).



* Data Pre-processing Done

Data pre-processing is an integral part of the projects like these.

The basic idea behind this is to filter the raw data first while applying the NLP methods to accomplish cleaning.



The above image text contains all the steps to pre-process the data before send it to the modelling stage.

* Data inputs- Logic- Output Relationships

The input was nothing that a raw data scraped from the e-retailors websites which has reviews (text) and ratings count(1 star, 2 stars ,etc..). The relationship between the input and the output has been made to be more clear post text as it has many outcasts which the model couldn’t read, thus getting rid of those made the job easy in modelling stage.

The final model is now able to predict the rating if the input are given as reviews.

* State the set of assumptions (if any) related to the problem under consideration

The assumption had to be taken while the text pre-processing stage Since I wasn’t sure how well the scraped data is from the outcast’s presence. Thus, I had to take an assumption that there might to be presence of outcasts like Unnecessary single texts, number, punctuations etc. I can’t go through each and every piece of data since it was around 36k in length.

Hardware and Software Requirements and Tools Used

I used **Jupyter notebook , NumPy, Pandas** libraries to this project.

Since NLP and Neural Network concepts were used in this project, there are some package and libraries :

**NLTK** (Natural Language Toolkit), Stemming ,Lamenting and also to get rid of the Stopwords, **NLTK** package was useful.

To convert the txt data into numerical data, we used **Sklearn.feature\_extraction.text import CountVectorizer, TfidfTransformer**

For data modelling we used **RandomForestClassifier, LSTM of ANN, DecissionTreeClassifer.**

**Model/s Development and Evaluation**

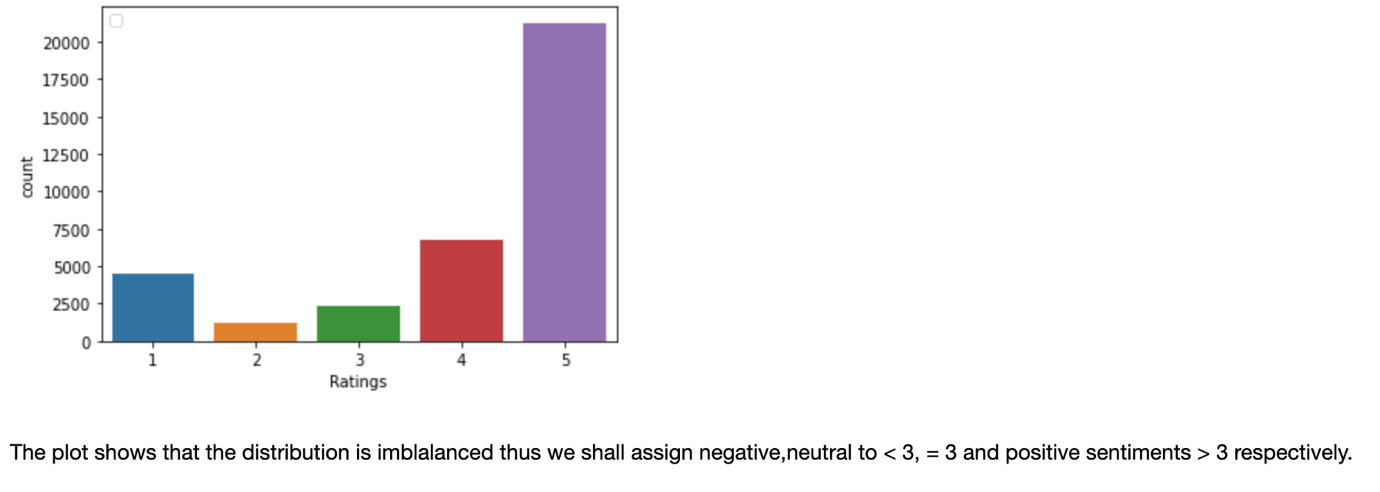
* Identification of possible problem-solving approaches (methods)

For the problem-solving, I had to first look at the data properly in order to see what are the tools or libraries required for it. Then, I had to used NLP methods like getting the data off the stopwords, punctuations, single letter words, URL links and so on. Text pre- processing method has done a half job, the other half of the job is to find the right models for it predict the data correctly.

.

* Testing of Identified Approaches (Algorithms)

i)   First, the ratings column of the data is Imbalanced thus, we’ve reduced to sentiments.



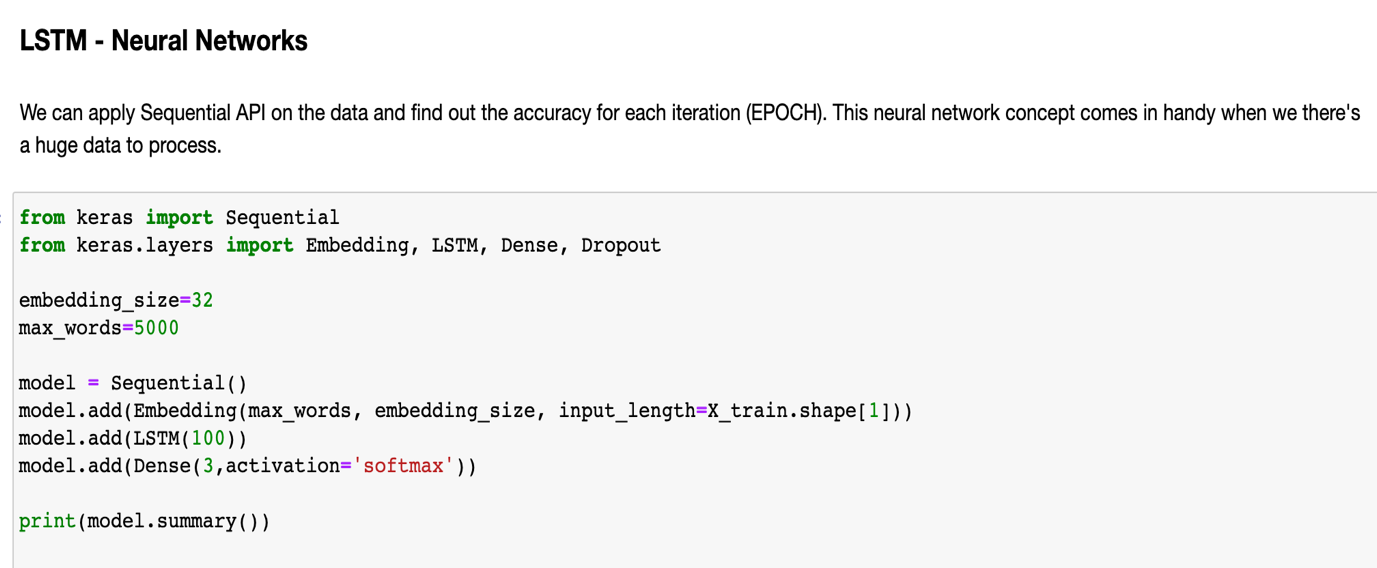
ii)   Second, the reviews column of the data is converted to string datatype and then X and Y data split is done using traintestsplit library.

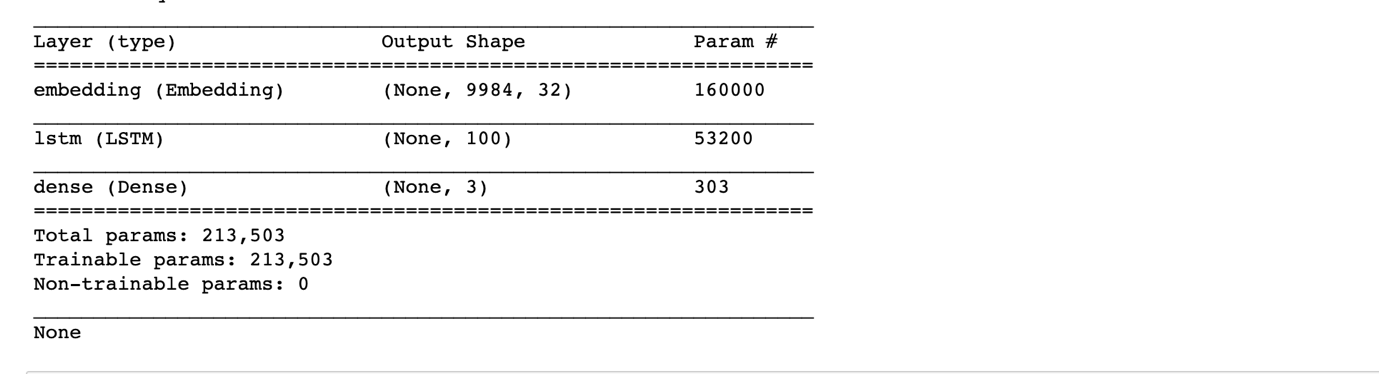
iii)  Further training and testing is done in the model selection phase. Every model is made to train the data and test it simultaneously.

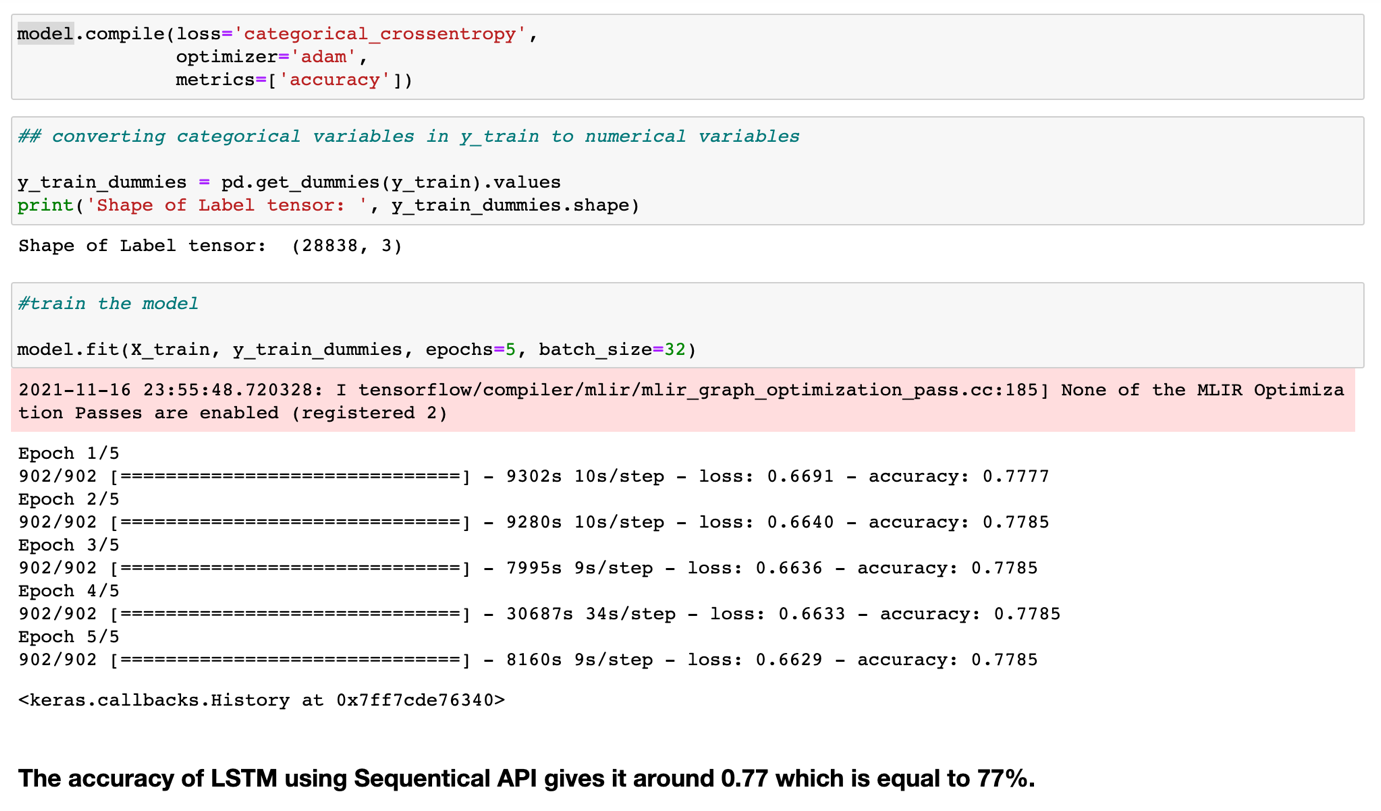
* iv)  The tools like: Accuracy score, f1 score, Classification report, Confusion Matrix are used to find the score of the test done.

testing.

* Run and Evaluate selected models





. 

.

**CONCLUSION**

* Key Findings and Conclusions of the Study

1)  I’ve learned that the type of data can play a major part in the final prediction. More refined the data is, better will be the prediction by the best-chosen model.

* Limitations of this work and Scope for Future Work

What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.

* Learning Outcomes of the Study in respect of Data Science

According to what I’ve gathered from doing this project is that, the role NLP – Natural Language Processing is superior than the rest. It not only helps in data cleaning but also helps in finding the sentiments out of it.

Neural Networks methods can come in handy here, the ability to work on the large datasets is something that almost no other models can do it .

• Limitations of this work and Scope for Future Work  
I wish there was more time given as the data is huge, the processing

time will take longer time to process the data.

For future work, I expect the data requirement for the work should be accounted for the time being given to finish the project.