MINI PROJECT 3
SENTIMENT ANALYSIS FOR MOVIE REVIEWS - NLP

By Divendra Raj
18/03/2023

AGENDA

- Business Background
- Business Main Objective
- Data Pipeline
- Data Summary
- Data Cleaning
- Modelling
- Model Performance
- Interpretations
- Conclusion

Business Background

In modern world, most of the business customers relay heavily on review and star ratings before approaching any new supplier or business for product and services. The better the reviews and ratings, the more confidence customers get. Especially “Positive” comments. Also, if they are heading in the right direction.

According to govt survey in New Zealand **60 per cent** of consumers **read reviews** before **making a purchase**.

Stuff link about online reviews:

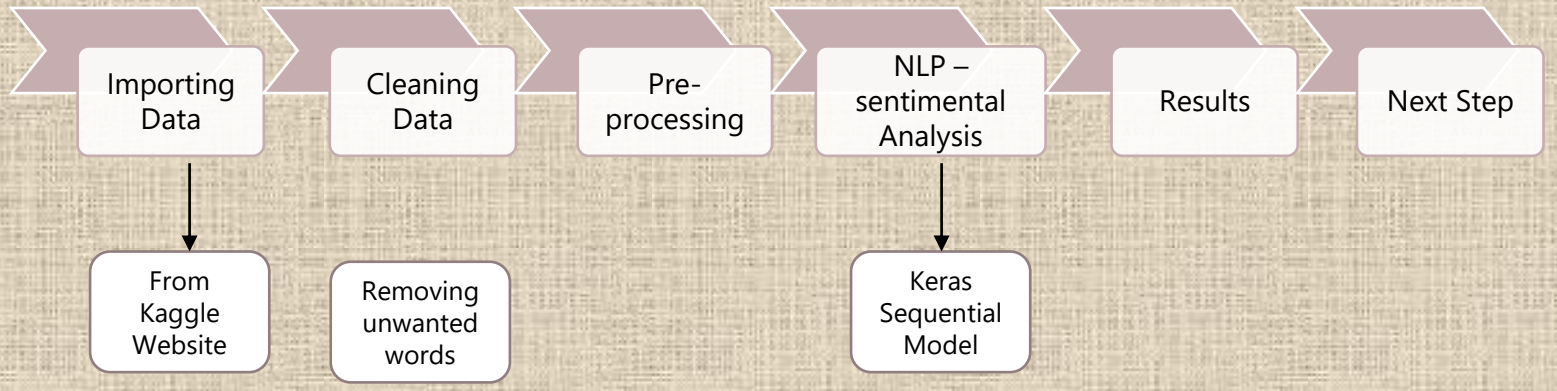
https://www.stuff.co.nz/business/113125913/online-reviews-may-be-fake-but-we-still-put-our-trust-in-digital-word-of-mouth

Business Main Objective

The main objective is to build a model which can tell whether a review are Positive or Negative correct

Also, which bucket has got the most.

Data Pipeline



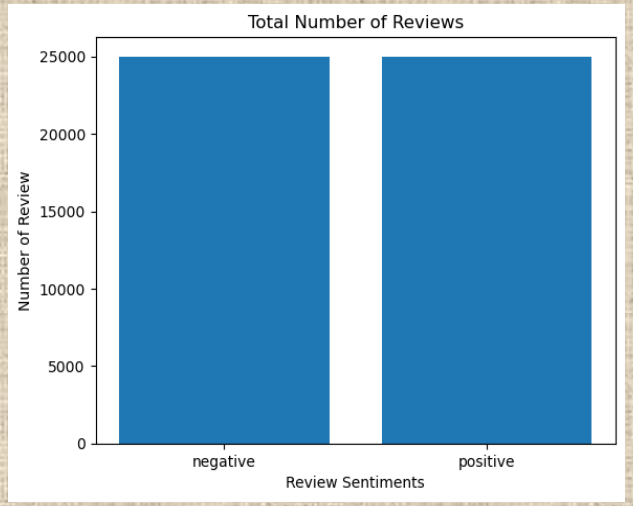
Data Summary/Cleaning

IMDB dataset having 50000 movie reviews based of 25000 popular movies for Natural Language Processing

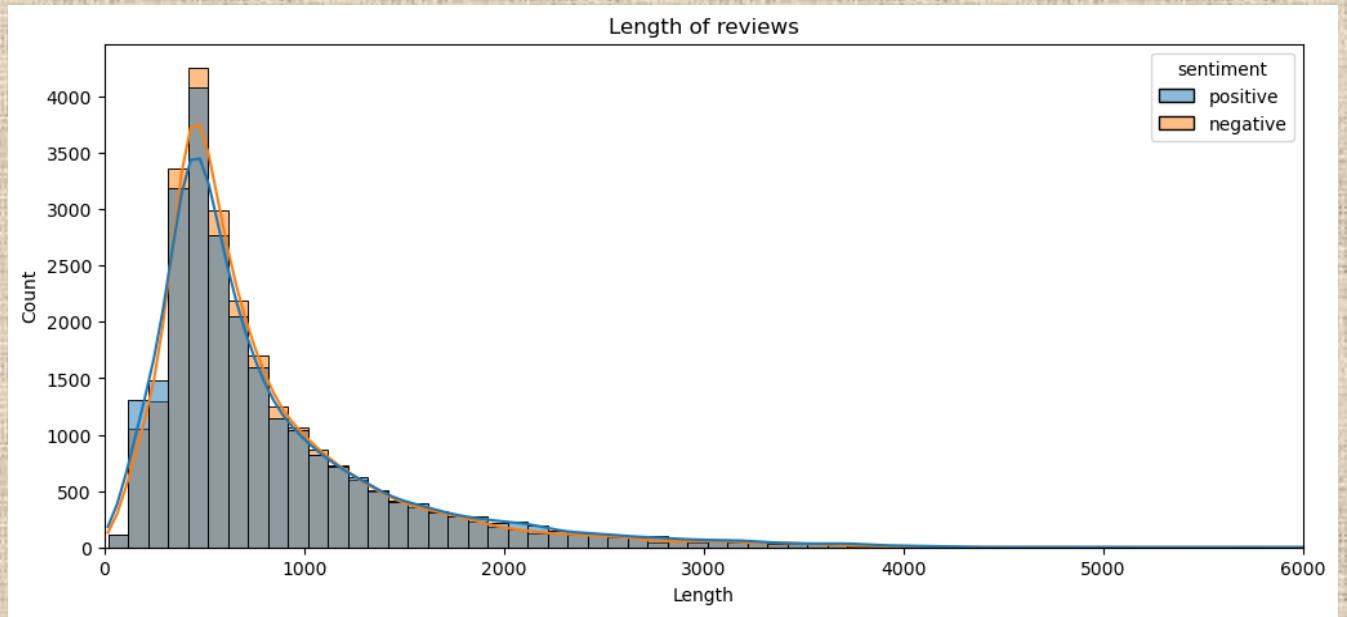
- Regular Expression, Stop words, Lemmetization and Tokenizer were used to standardise data for building models

Exploratory Data Analysis

The Graph on the right show the totals number of Positive and Negative received.
- According to the graph it is equal



The graph on the right shows the length of reviews for both Positive and Negative.
- According to the graph Negative reviews are more lengthy than Positive



Modelling

After data cleaning and standardizing, Model was built on Keras Sequential - NLP
Length - 1823s

- The running time is not as good

Model Performance

The model accuracy score was **86.23 Per Cent**

The accuracy of prediction on the test set comes out to be 86.23%! We can improve the accuracy further by playing around with the model hyperparameters, further tuning the model architecture or changing the train-test split ratio. We can also train the model for a more significant number of epochs, and we stopped at five epochs because of the computational time. Ideally, it would help prepare the model until the train and test losses converge.

• An embedding layer of dimension 100 converts each word in the sentence into a fixed-length dense vector of size 100. The input dimension is set as the vocabulary size, and the output dimension is 100. Each word in the input will hence get represented by a vector of size 100.
•A bidirectional LSTM layer of 64 units.
•A dense (fully connected) layer of 24 units with relu activation.
•A dense layer of 1 unit and sigmoid activation outputs the probability of the review is positive, i.e., if the label is 1.

Model: "sequential"

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 200, 100)	300000
bidirectional (Bidirectional)	(None, 128)	84480
dense (Dense)	(None, 24)	3096
dense_1 (Dense)	(None, 1)	25

Total params: 387,601
Trainable params: 387,601
Non-trainable params: 0

Interpretations

The result on the right represents the outcome of the model testing.
- According to the model testing it classified the correct review.

Conclusion

Keras Sequential Natural language Processing can be a good Model for classifying Movie reviews however my next step would be to use classification models and other application to improve the run time of the Model. Also, more graphical approach to illustrate better.

```
1/1 [==============================] - 0s 91ms/step
The movie was very touching and heart whelming
Predicted sentiment :  Positive
I have never seen a terrible movie like this
Predicted sentiment :  Negative
the movie plot is terrible, but it had good acting
Predicted sentiment :  Negative
```