CAPSTONE PROJECT – **REVIEW**

Table of Contents

1. Problem Statement
2. Project Objective
3. Data Description
4. Data Pre-processing Steps and Inspiration
5. Choosing the Algorithm for the Project
6. Motivation and Reasons for Choosing the Algorithm
7. Model Evaluation and Techniques
8. Inferences from the Same
9. Future Possibilities of the Project
10. Conclusion
11. References

Problem Statement

You are working in an e-commerce company, and your company has put forward a task to analyse the customer reviews for various products. You are supposed to create a report that classifies the products based on the customer reviews.

1. Find various trends and patterns in the reviews data, create useful insights that best describe the product quality.
2. Classify each review based on the sentiment associated with the same.

Project Objective

1. To search for products which are highly in demand.
2. To look for products that are constantly getting bad reviews.
3. Performing customer segmentation using LSTM.

Data Description

* The data consists of 10 columns. Each carrying their own characteristics.
* Following are the features in the dataset:

1. ID
2. ProductID
3. UserID
4. ProfileName
5. HelpfullnessNumerator
6. HelpfullnessDenominator
7. Score
8. Time
9. Summary
10. Text
11. Reviews(added by me)

* Data Visualization and Insights:

1. Top 10 Product sold –

'B009WVB40S', 'B0018MWINQ', 'B0018OKA1Q', 'B0018OJK12', 'B0048IACB2', 'B0048IACSK', 'B0048IAE5G', 'B0018N0KXU', 'B0018N0KXA', 'B0048IAIAM'

1. Least Product sold –

'B003HD6NPI', 'B0007WLM9E', 'B004MZ9JLA', 'B000G65238', 'B0007WEP70', 'B0046H8N02', 'B000V1U83W', 'B000V1U82S', 'B003HACP26', 'B003ZIDK5G'

1. Count of different scores –

Count represents the number of times a has score appeared in the dataset.

|  |  |
| --- | --- |
| Score | Count |
| 1 | 36303 |
| 2 | 20800 |
| 3 | 29767 |
| 4 | 56094 |
| 5 | 250955 |

Data Processing Steps:

1. Firstly we check data type of all the columns.
2. We found that all columns have data types as per their characteristic.
3. By combining Text and Summary columns we’ll produce a new columns named ‘Reviews’ which will be helpful for computation.
4. We’ll check for null values. We have null values in ProfileName, Summary, and Reviews column but here it will be dropped due to their presence is 16, 27, and 2 respectively which is minute in comparison to the data.
5. To pass df.describe command to know mean, standard deviation, etc. of all the columns.
6. Let’s check for duplicate values, for this we will pass function df.duplicated and in that function we’ll mention columns names that we are concerned for.
7. Drop the retrieved duplicate values which cutdown our data.
8. Adding column Sentiment and the values will be based on the scores. Suppose the score is more than 3 then it will be labelled as 1 and else 0.

LSTM (Long Short Term Memory)

LSTM is an improved version of Recurrent Neural Network (RNN) designed by Hochreiter & Schmidhuber. LSTM are best at capturing pertinence between last line of a review and the first line of a review. RNN is a good practice but they fail to capture such type of relation within a paragraph/review.

* LSTM has three gates:

1. Input gate:

This decides which information to store inside in the memory cell. It is trained to open when it is important and close when it is not.

1. Forget gate:

This gate decides which information to discard from the memory cell. It is trained to open when the information is no longer required and close when it is.

1. Output gate:

Output gate decides which information to use for the output of the LSTM. It is trained to open when the information is important and close when it is not.

* Working

LSTM network consist of three gates which open and close as per the need of information. The three gates are responsible for flow of information. The structure also consist cell state where all the previous words are memorized and stored in that state.

Model Evaluation and Technique

1. In this model I have used tokenizer and stemming methods for data preparation.
2. Tokenizer: It splits the sentence into words.
3. Stemming: It can be defined as the method to trim the word to its root. Stemming will trim off last few characters of the words without knowing the context of the word.
4. Convert ‘Reviews’ into list and storing it in a variable. And then applying tokenizer and stemming to the variable.
5. Now we’ll create a function where our Reviews data will be split into a word and then those words will be assigned with an index value with help of tokenizer. And the entire individual review will assigned with its sentiment.
6. Then we’ll check if our function has given the desired results or not.
7. Check for number of words present in an individual review. And according to that we’ll do padding. In my model I have taken maxlen as 60.
8. Converting the data into numpy array and splitting the data as train and test data.
9. In my model I have performed Bidirectional and LSTM neural network.

Evaluation Report:

* Here in my case, I am getting loss= 0.2 and val\_loss= 0.27.

Inferences from Project

* As we can see that the model is performing well.
* The parameters chosen were best in my knowledge.
* I tried whereas parameters like changing optimizer and manipulating padding sequences but didn’t receive much difference.

Future Possibilities

* This model is useful for segregating reviews into two parts that is positive and negative.
* By knowing a negative review has been received via feedback following a product.
* We can target specific customer by reviewing their feedback and we can give them discount or coupons as token of apology.
* Classifying reviews can actually help us understand the quality of the product. This can give us a check on sincerity of retailer.
* If we are getting bad reviews on products which are sold by the same retailer it will help us grab our attention and take necessary action against the retailer.
* It can help us analyse how negative reviews are affecting our sales periodically.

Conclusion

This project gave us meaningful insights about how classifying reviews can help us understand the quality that we are providing to our loyal customers. Additionally we cannot compromise in our quality in any manner. This technique will give us a broader picture about our business.

References

* <https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/>
* <https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/>
* <https://intellipaat.com/blog/what-is-lstm/>