

DECLARATION

This is to certify that Report entitled “Reviews Analyzer and Summarizer” which is submitted in partial fulfilment of the requirement for the award of degree B.Tech. in Computer Engineering to BPIT, GGSIP University, Dwarka, Delhi comprises only my original work and due acknowledgement has been made in the text to all other material used.

Date: 01/10/2019

Name of Student: Lokesh Aggarwal

Enrollment No. : 03920802716

ACKNOWLEDGEMENT

The success and final outcome of this project required a lot of guidance and assistance; it would not have been possible without the kind support and help of many individuals in the organization. We would like to extend our sincere thanks to all of them.

Kind words do not cost much, yet they accomplish much; therefore we would like to give our sincere and heartfelt thanks to our esteemed guide, **Ms. Kritika Balihar**, for providing us with the right guidance, advice at the crucial junctures & also for her support in completing the project. We are grateful to our coordinator, for her immense support and valuable ideas throughout the work. She made us to get out of our comfort and push the limit.

We would like to express our gratitude towards our parents for their kind co-operation and encouragement which helped us in completion of this project.

Training Coordinator Certificate

This is to certify that Report entitled “ **Reviews Analyzer and Summarizer**” which is submitted by LOKESH AGGARWAL in partial fulfilment of the requirement for the award of degree B.Tech in Computer Science and Engineering to BPIT, GGSIP University, Dwarka, Delhi is a record of the candidate own work and the matter embodied in this report is adhered to the given format.

Date: 01/10/2019

Coordinator: Ms. Kritika Balihar

Table of Contents

Declaration.....	(i)
Acknowledgement.....	(ii)
Training Coordinator Certificate.....	(iii)
Company Certificate.....	(iv)
Table of Contents.....	(v)
List of Figures.....	(vii)
List of Tables.....	(viii)
Abstract.....	(ix)
Chapter 1 –Introduction	1
1.1 Overview	1
1.2 Content Summarization.....	4
1.3 Need of Content Summarization in E Commerce Industry.....	4
1.4 SEO Benefits of Reviews.....	5
1.5 Negative Reviews and Customers.....	6
1.6 Python	7
1.6.1 History of Python.....	8
1.6.2 Python Features.....	8
1.7 Machine Learning.....	9
1.7.1 Machine Learning Techniques.....	10
1.7.2 Evolution of Machine Learning.....	11
1.8 Deep Learning.....	12
1.9 Natural Language Processing.....	13
Chapter 2 – Software Requirement Specification (SRS)	15
2.1 Problem Statement.....	15
2.2 Scope.....	16
2.3 Purpose	17
2.4 Definition and Acronyms	18
2.5 Software Perspective	18
2.6 Software Functioning	19
2.7 User Classes and Characteristicis	19
2.8 Assumptions and Dependencies.....	19
2.9 External Interface Requirements.....	20

2.10 Software Requirements.....	21
2.11 Non-Functional Requirements.....	22
2.12 Software and Hardware Requirements	23
Chapter 3 – Diagrams	24
3.1 Use Case Diagram.....	24
3.2 Data Flow Diagram	25
3.2.1 Zero Level DFD.....	25
3.2.2 First Level DFD.....	26
3.3 ER Diagram	27
Chapter 4 – Process Selection and Implementation	28
4.1 Data Acquisition.....	28
4.1.1 Web Scrapping.....	29
4.1.2 Other Data Sources	32
4.2 Data Pre-processing.....	33
4.2.1 Text Cleaning	34
4.2.2 Summary Cleaning	35
4.2.3 Understanding the distribution of Sequences	36
4.3 Preparing the Tokenizer.....	37
4.4 Model Building.....	38
4.5 Encoder Decoder based Architecture.....	40
4.5.1 Training Phase	40
Chapter 5 – Result	44
Chapter 6 – Comparison and Analysis.....	45
6.1 Brief Study of other Summarization systems.....	45
6.2 Existing vs Proposed System	46
Chapter 7 – Conclusion and Future Scope.....	47
9.1 Conclusion.....	47
9.2 Future Scope	48
References	49

LIST OF FIGURES

Figure 1.1 Amazon Reviews	6
Figure 1.2 Python Logo.....	6
Figure 1.3 The world of Machine Learning	10
Figure 1.4 Classification of Machine Learning	11
Figure 1.5 Evolution of Deep Learning	13
Figure 3.1 Use Case Diagram for Content/Reviews Summarizer	25
Figure 3.2 Level 0 DFD for Content/Reviews Summarizer.....	26
Figure 3.3 Level 1 DFD for Content/Reviews Summarizer.....	26
Figure 3.4 ER Diagram for Content/Reviews Summarizer.....	27
Figure 4.1 Basic Steps of Review Summarizer.....	30
Figure 4.2 Web Scrapping Process.....	33
Figure 4.3 Text and Dataset Cleaning Code.....	34
Figure 4.4 Summary Cleaning Code.....	35
Figure 4.5 Showing Cleaned Code	37
Figure 4.6 Output of Cleaning Code	37
Figure 4.7 Code For Sequence Distribution.....	38
Figure 4.8 Output graph of sequence Distribution.....	38
Figure 4.9 Code for Tokenizer.....	39
Figure 4.10 Model Building	40
Figure 4.11 Basic Encoder.....	40
Figure 4.12 LSTM based encoder architechture.....	41
Figure 4.13 LSTM based decoder architechture.....	42
Figure 4.14 Combined LSTM based encoder decoder architechture.....	42
Figure 4.15 Code for Inference phase	43
Figure 5.1 Individual Reviews Summary	44

LIST OF TABLES

Table 4.1 Scrapy File Structure.....	32
Table 5.1 Review Summary for Products of other category	45
Table 5.2 Review Summary for Product Inputs	46

ABSTRACT

In the recent past, e-commerce sites have made rapid growth. There are thousands of products and various websites sell these products. Massive growth in the number of reviews and their availability along with the advent of opinion-rich review forums for the products sold online, choosing the right one from a large number of products has become difficult for the users. It is thus imminent for the users to verify the genuineness and quality of products. What better way is there to ask people who have already bought the product? This is when customer reviews come into the picture. Customer reviews from customers who have actually purchased and used the product in question can give you more context to the product itself. Each reviewer rates the product from 1 to 5 stars, and provides a text summary of their experiences and opinions about the product. The ratings for each product are averaged together in order to get an overall product rating. The number of reviews on Amazon has grown over the years. Thus, the major problem or hitch that we face today for analysing a product on the basis of its reviews is that the popular product have such a large number of reviews. This number can range from hundreds to thousands and even tens of thousands. In the era of big data, the user itself is also responsible for generating a large amount of data. Also, we as a product buyer or analyzer do not have time to go through this number of reviews. Hence, a tool has been proposed to ease the task by analyzing these product reviews and summarizing it accordingly so that it will greatly help the users to decide what other buyers have experienced on buying this product. We carry out this process by a number of modules that includes feature extraction and opinion extraction that will further improve the process of analysis and provide us with an efficient summary of the product.