Report On

Text Summarization on Amazon Reviews

Submitted in partial fulfillment of the requirements of the Course project in
Semester VII of fourth year Computer Science Engineering (Data Science)

by
Preet Raut   (51)
Dipanshu Vartak(62)
Shubham Warik (66)
Naman Mehta   (34)

Supervisor

Dr. Tatwadarshi P.N.

University of Mumbai

Vidyavardhini's College of Engineering & Technology

Department of Computer Science Engineering (Data Science)

(2023-24)

Vidyavardhini's College of Engineering & Technology Department of Computer Science Engineering (Data Science)

CERTIFICATE

This is to certify that the project entitled "Text Summarization on Amazon Reviews " is a bonafide work of "Preet Raut (51), Dipanshu Vartak(62), Shubham Warik(66), Naman Mehta(34)" submitted to the University of Mumbai in partial fulfillment of the requirement for the Course project in Semester VII of fourth year Computer Science Engineering (Data Science).

Supervisor

Dr. Tatwadarshi P. N.

Dr.Vikas Gupta
Head of Department

Table of Contents

Chapter 1: Introduction

1.1 Introduction

Text summarization is a crucial natural language processing (NLP) task that plays a significant role in distilling vast amounts of text into concise and informative summaries. When applied to Amazon reviews, text summarization can be a valuable tool for both consumers and businesses. Amazon, one of the world's largest e-commerce platforms, hosts an extensive array of product reviews from millions of customers. These reviews contain valuable insights and opinions that can guide potential buyers in making informed decisions. However, sifting through these reviews can be a time-consuming process. Text summarization techniques aim to automate and simplify this task by extracting the most important and relevant information from these reviews, providing users with a quick overview of the key aspects, sentiments, and considerations associated with a particular product. In this context, text summarization on Amazon reviews can enhance the shopping experience, assist businesses in understanding customer feedback, and contribute to a more efficient and insightful online marketplace.

1.2 Problem Statement & Objectives

The central challenge in this context is the need to efficiently extract valuable insights from the vast pool of Amazon product reviews. Given Amazon's status as a colossal e-commerce platform, consumers often find it arduous to manually navigate through the extensive array of reviews to make well-informed purchasing decisions. Simultaneously, businesses operating on Amazon seek effective mechanisms for analyzing and comprehending customer feedback on a large scale, with an emphasis on recognizing patterns, sentiments, and suggestions for product enhancements. Consequently, the primary objective is to craft a sophisticated text summarization system tailored specifically to Amazon reviews. This system's purpose is twofold: it should automatically generate concise and coherent summaries of reviews, thereby aiding consumers in their decision-making process and empowering businesses to glean actionable insights from the wealth of available customer feedback.

Chapter 2: Proposed system

2.1 Introduction

Our proposed text summarization system for Amazon reviews dataset will leverage a comprehensive approach. We will first collect a diverse dataset of Amazon reviews and perform thorough preprocessing, including text cleaning, tokenization, stopword removal, and sentiment analysis. We will employ summarization techniques, utilizing state-of-the-art NLP models like NLTK for optimal results. PreProcessing would include getting rid of stop words and replacing contractions with their full form.A user-friendly interface will allow users to input review data and receive concise summaries. Continuous improvement will be a priority, incorporating feedback loops for enhanced summarization quality, ultimately aiding consumers in making informed purchase decisions. This system trained on Amazon.csv dataset hosted on kaggle accounting for 364mb.
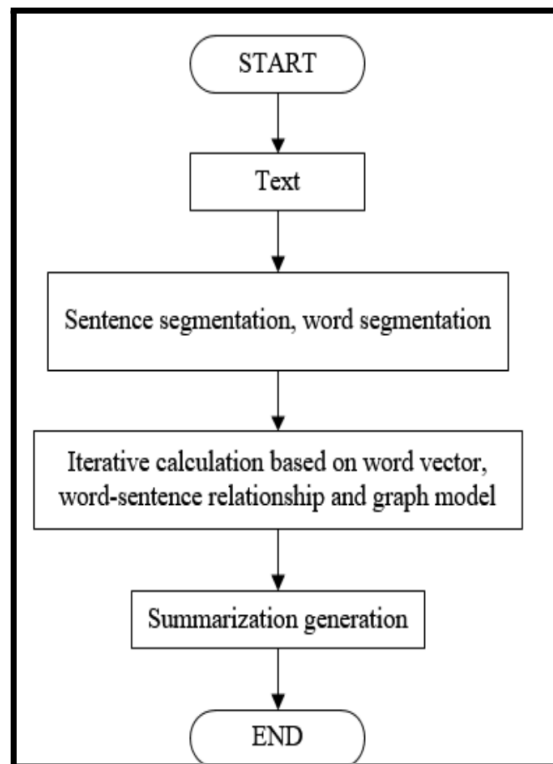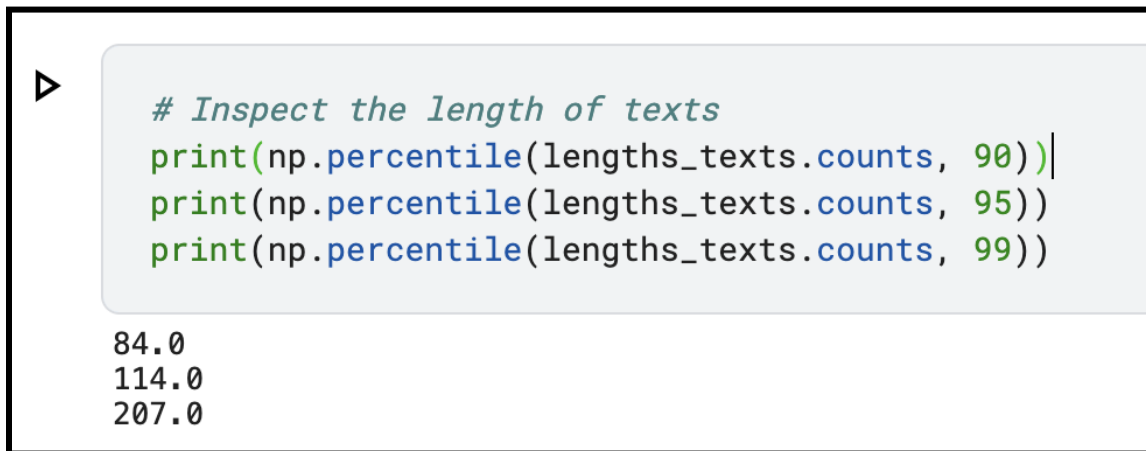
2.3 Algorithm and Process Design

Fig 1: Process Diagram

2.4. Details of Hardware & Software

- Python
- Google Colab
- nltk
- Pandas
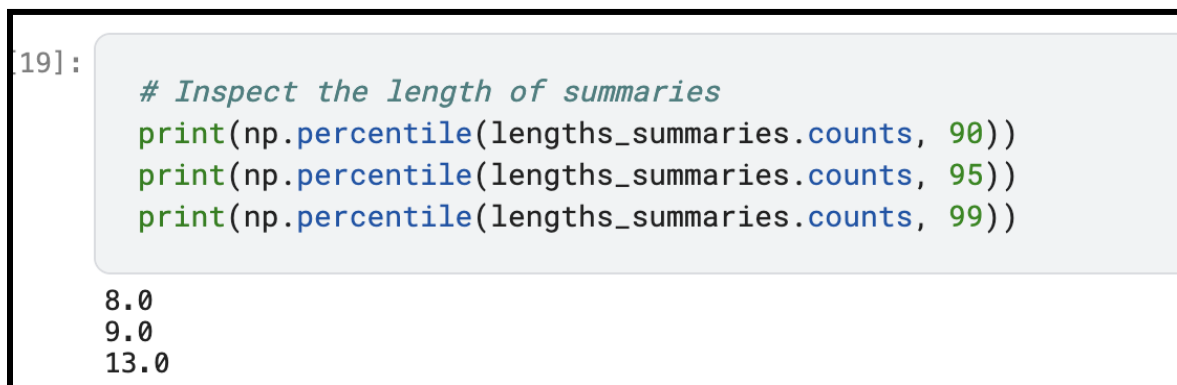- Numpy
- Matplotlib
- SeaBorn

2.5. Experiment and Results :

```python
# Inspect the length of texts
print(np.percentile(lengths_texts.counts, 90))
print(np.percentile(lengths_texts.counts, 95))
print(np.percentile(lengths_texts.counts, 99))
```
```
84.0
114.0
207.0
```

fig 1: Length of texts before summarization

```python
# Inspect the length of summaries
print(np.percentile(lengths_summaries.counts, 90))
print(np.percentile(lengths_summaries.counts, 95))
print(np.percentile(lengths_summaries.counts, 99))
```
```
8.0
9.0
13.0
```

fig 2 : Length of text after summarization

Some results of this Summarization are:

```
Review # 1
bought several vitality canned dog food products found good quality product looks like stew processed meat s
mells better labrador finicky appreciates product better
Summary # good quality dog food

Review # 2
product arrived labeled jumbo salted peanuts peanuts actually small sized unsalted sure error vendor intende
d represent product jumbo
Summary # not as advertised

Review # 3
confection around centuries light pillowy citrus gelatin nuts case filberts cut tiny squares liberally coate
d powdered sugar tiny mouthful heaven chewy flavorful highly recommend yummy treat familiar story c lewis li
on witch wardrobe treat seduces edmund selling brother sisters witch
Summary #  delight  says it all

Review # 4
looking secret ingredient robitussin believe found got addition root beer extract ordered good made cherry s
oda flavor medicinal
Summary # cough medicine

Review # 5
great taffy great price wide assortment yummy taffy delivery quick taffy lover deal
Summary # great taffy

Review # 6
got wild hair taffy ordered five pound bag taffy enjoyable many flavors watermelon root beer melon peppermin
t grape etc complaint bit much red black licorice flavored pieces particular favorites kids husband lasted t
wo weeks would recommend brand taffy delightful treat
Summary # nice taffy

Review # 7
saltwater taffy great flavors soft chewy candy individually wrapped well none candies stuck together happen
expensive version fralinger would highly recommend candy served beach themed party everyone loved
Summary # great  just as good as the expensive brands
```

fig 3 : Summarization of 1st 7 reviews

```
Review # 8
taffy good soft chewy flavors amazing would definitely recommend buying satisfying
Summary # wonderful  tasty taffy

Review # 9
right mostly sprouting cats eat grass love rotate around wheatgrass rye
Summary # yay barley

Review # 10
healthy dog food good digestion also good small puppies dog eats required amount every feeding
Summary # healthy dog food

Review # 11
know cactus tequila unique combination ingredients flavour hot sauce makes one kind picked bottle trip brough
t back home us totally blown away realized simply could find anywhere city bummed <br ><br >now magic interne
t case sauce ecstatic <br ><br >if love hot sauce mean really love hot sauce want sauce tastelessly burns thr
oat grab bottle tequila picante gourmet de inclan realize taste never want use sauce <br ><br >thank personal
incredible service
Summary # the best hot sauce in the world

Review # 12
one boys needed lose weight put food floor chubby guy protein rich product food higher skinny boy jump higher
food sits going stale really go food chubby boy losing ounce week
Summary # my cats love this  diet  food better than their regular food

Review # 13
cats happily eating felidae platinum two years got new bag shape food different tried new food first put bowl
s bowls sit full kitties touch food noticed similar reviews related formula changes past unfortunately need f
ind new food cats eat
Summary # my cats are not fans of the new food

Review # 14
good flavor came securely packed fresh delicious love twizzlers
Summary # fresh and greasy
```

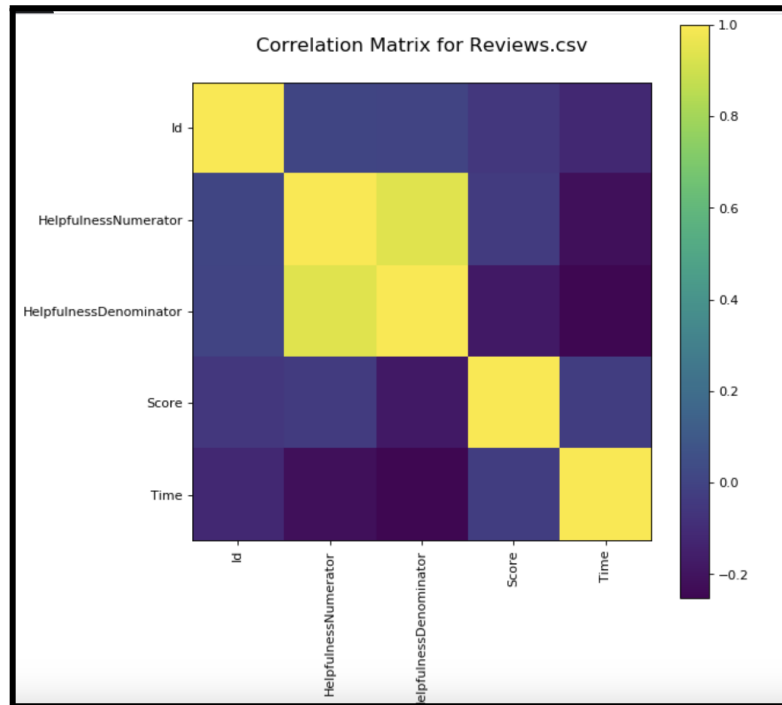**fig 4 : Summarization for next 7 reviews**

Fig 4: Correlation matrix

2.6 Conclusion

When tested on a test set, the model summarizes the text quite efficiently and concisely. The average size of review text was 90 words and the average size of summary generated was 13, demonstrating its capacity for accurate summarization . Additionally, the total number of UNK words was dropped down to just 1.42%.This was achieved by using full forms instead of their contractions as well as removing stop words. These visual aids improve our comprehension of the behavior of the system and direct future improvements for more effective product selection by consumer.

References:

1.   Sridevi Gadde,” summarization using Machine Learning and Deep Learning Techniques”2021 7th International Conference on Advanced Computing & E-commerce Systems (ICAECS).

2. Suparna Das Gupta, “Concise reviews for online shopping ” 2021 *J. Phys.: Conf. Ser.* 1797 012017.

3.   M.Rubin Julis, S.Alagesan , “Natural Language processing techniques on Amazon reviews” INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 02, FEBRUARY 2020 ISSN 2277-8616

4.   Sifat Ahmed, Faisal Muhammad,” Using Boosting Approaches to summarize texts” 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)

5.   Shaghayegh Hosseinpour, Hadi Shakibian, “ NLTK for summarizing paragraphs” 2023 9th International Conference on Web Research (ICWR)