Major Project Synopsis

on

Automatic Summarization of User Reviews

In partial fulfillment of requirements for the degree

of

BACHELOR OF ENGINEERING

IN

INFORMATION TECHNOLOGY

Submitted by:

Piyush Mali [19100BTIT06588]

Raghav Sood [19100BTIT06595]

Rishika Jain [19100BTIT06604]

Under the guidance of

Prof. Sujit K Badodia
Prof. Manorama Chouhan

DEPARTMENT OF INFORMATION TECHNOLOGY
SHRI VAISHNAV INSTITUTE OF INFORMATION TECHNOLOGY
SHRI VAISHNAV VIDYAPEETH VISHWAVIDYALAYA, INDORE

JULY-DEC 2022

Abstract

Text summarization is a technique for creating a precise, concise summary of long texts while concentrating on the passages that convey important information and maintaining the overall meaning. The goal of automatic text summarization is to reduce lengthy articles into shorter versions because doing it manually would be time-consuming and expensive. Machine learning algorithms may be used to examine papers and find the sections that contain pertinent facts and information before producing the necessary summary paragraphs.

The "Automatic Summarization of User Reviews" project serves as a template for creating a summary of all customer reviews. Because there is a wealth of material available online for any topic, condensing the pertinent data into a summary would benefit both producers and a large number of other users. Our model is required since there are more people shopping online and more options available to them. Before making a purchase, the user needs to know how reliable the product was when it was utilized by real customers.

Our programmed uses all of the product reviews as input to provide a succinct summary that closely follows the flow of all the evaluations. The reviews are first divided into sentences. It has been preprocessed to eliminate superfluous punctuation and leave just words with significance. Sentences are given relative weighting based on certain common characteristics, and evaluations are presented and categorized based on the salient scores assigned to each of the sentences.

INDEX

1. INTRODUCTION

In a world where the internet is booming with vast volumes of data every day, the ability to automatically summarize data is a critical issue. Summaries of lengthy papers, news articles, and even dialogues can improve and speed up how much material we consume. Natural Language Processing (NLP), which has attracted a lot of interest in recent years, has generated a lot of interest in Automatic Text Summarizations. Massive data transmission and data collection have suddenly entered our society. According to a survey by International Data Corporation (IDC), the amount of data generated online is predicted to rise from 4.4 ZB in 2013 to 180 ZB in 2025. That's a lot of information! IDC predicts that by 2025, there will be 180 zettabytes of digital data generated year worldwide, up from 4.4 zettabytes in 2013. That's a lot of details! IDC predicts that by 2025, there will be 180 zettabytes of digital data flying annually around the globe, up from 4.4 zettabytes in 2013. That's a lot of details! ML algorithms that can automatically summarize lengthy texts and provide accurate summaries that elegantly express the intended information are required because there is so much data generated and migrating online.

1.1 What is Natural Language Processing (NLP)?

A subset of AI is NLP. that manages interactions between computers and people using natural language. The ultimate goal of NLP is to effectively read, decode, comprehend, and comprehend human language. [1] To extract meaning from human language, the majority of NLP approaches rely on machine learning. The size of the global natural language processing market is anticipated to reach $29.5 billion by 2025, growing at a market growth rate of 20.5% CAGR throughout the projected period. Natural language processing employs a variety of ways for deciphering human language, from algorithmic and rules-based methods to statistical and machine learning techniques. Because text- and voice-based data, like actual applications, vary greatly, a wide variety of techniques are required. Tokenization, sorting, lemmatization/stemming, speech tagging, language detection, and semantic link identification are all fundamental NLP activities.

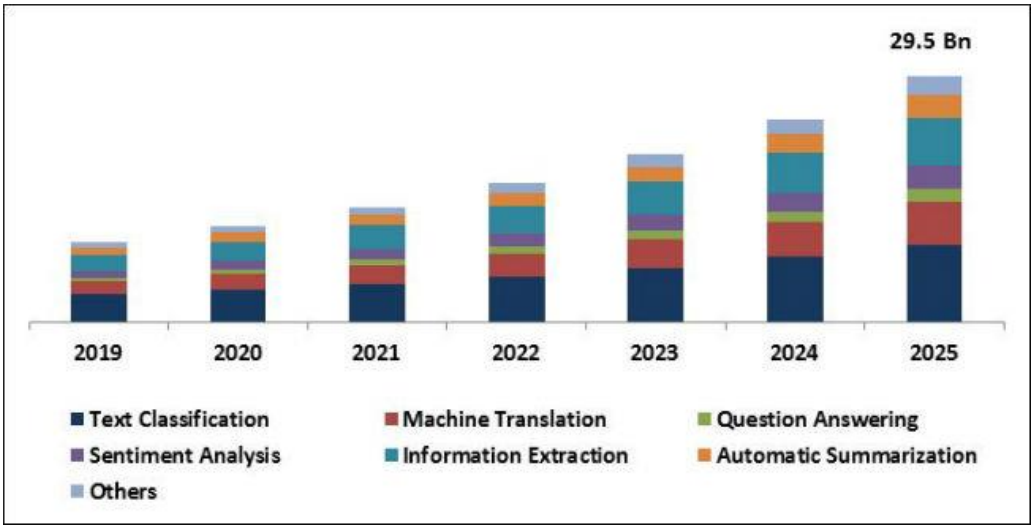


Fig 1.1 Global Natural Language Processing Market Size

1.2 Automatic Text Summarization:

Text summarization is the process of extracting the most critical or essential information from a source document (or multiple sources) to create a simplified version of a particular user (or multiple users) and a task (or multiple tasks).[3]Two approaches defined by researchers to automatic summarization extraction and abstraction.

1.2.1 Extractive Summarization:

Extract summaries are created by selecting some related statements from the original document.[3] The length of the summary depends on the compression ratio. This is a simple and robust way to summarize text. Here, some emphasis is assigned to the sentences in the document, then the sentences with the highest ratings are selected to create a summary.

Here is an example:

Source text: ***Durgesh and Harsh*** *rode on a donkey to* ***attend the*** *annual* ***event in Kapoorthala.*** *In the city,* ***Harsh*** *has* ***friend*** *named* ***Shubham****.*

Extractive summary: *Durgesh and Harsh attend the event in Kapoorthala. Harsh Friend Shubham.*

Bold phrases are extracted and displayed sometimes it can show grammatically strange summaries.

1.2.2 Abstractive Summarization:

It generates completely new sentences from the given document.[3] So we can say an abstract is a summary of ideas or concepts taken from the original document reinterpreted and presented in another format.

Here is an example:

Source text: ***King and Queen*** *rode on a donkey to* ***attend*** *the* ***annual*** *event in* ***Jerusalem****. In the city,* ***Queen*** *gave* ***birth*** *to a child named* ***Jesus****.*

Abstractive summary: *King and Queen came to Jerusalem where Jesus was born.*

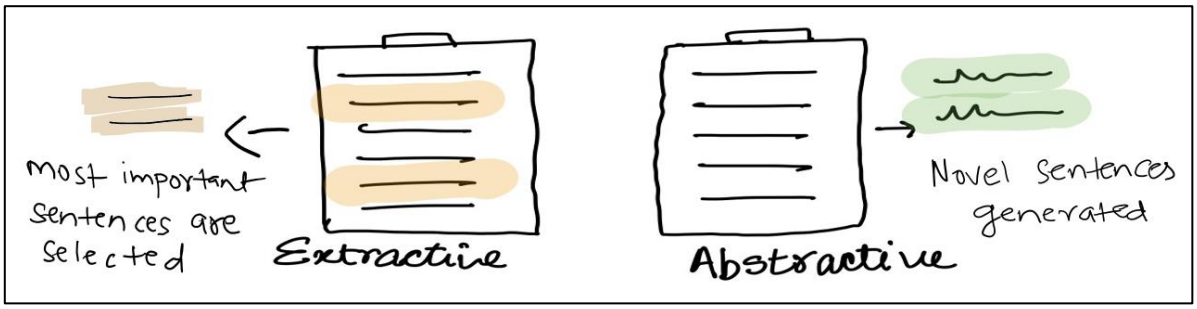


Fig 1.2Extractive & Abstractive Summarization

We'll be using the **Extractive method.**

2. PROBLEM DOMAIN

- Using NLP (Natural Language Processing) techniques, the challenge is to create and construct a model that may automatically summarize customer reviews for a product from an e-commerce website. The model must also be evaluated for correctness.

- NLP is challenging because of the complexity of processing human language. Natural language communication follows principles that are difficult for a machine to comprehend.

- Some of these laws may seem abstract and high-level, as when someone employs a sarcastic comment to convey information. However, some of these regulations may be of a low degree.

- Use the letter "s," for instance, to denote a number of objects. The use of machines to translate human language is fraught with numerous other issues.

- Due to the ambiguous nature of human language, there are no set, precise rules that can be taught to the computer. Instead, it is the ambiguity and imprecision of natural languages that make NLP challenging for machines to apply.

- Research and contrast the various text summarizing techniques. to find weaknesses in currently available text-summarization software. Create and put into use a model or tool for automatically summarizing user reviews.

3. SOLUTION DOMAIN

The main highlights of the Solution domain followed is:

1. **Dataset:** The dataset is obtained from the Kaggle repository.
2. **Review Extraction:** Only the reviews of specific types and features are extracted from the pool of reviews for summarization.
3. **Data Preprocessing and Sentence Tokenization:** The sentence tokens obtained are preprocessed to remove the stop words and unnecessary punctuations and the obtained reviews are further broken down into sentence tokens. This is done to extract the sentences for further analysis.
4. **Summarization:** The summarization method selected here is "Text Rank "to get the summary of user reviews.
5. **Summarized Data:** When summarization is done, we did the evaluation.

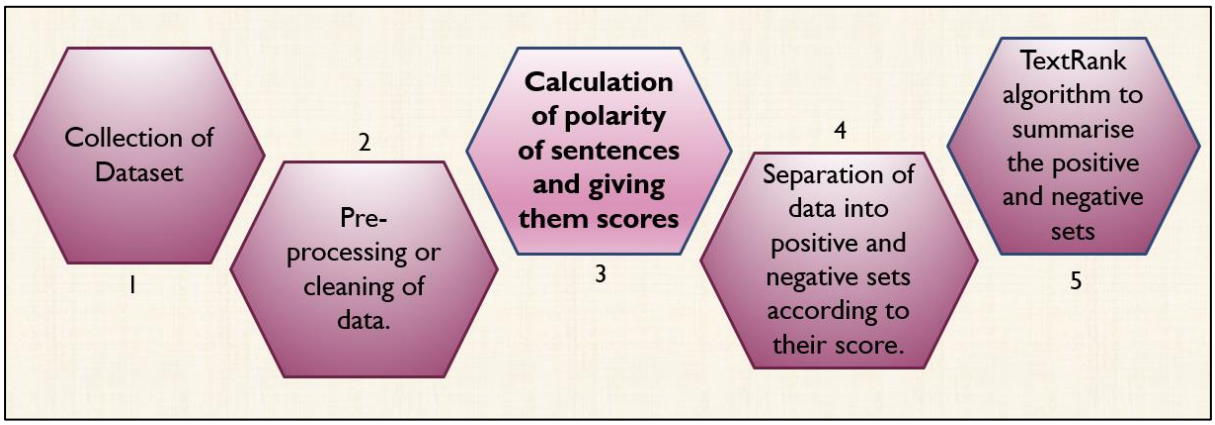


Fig 3.1Work Flow of Project

3.1.1 Source of Data

The dataset obtained from Kaggle in prompt cloud repository which contains more than 4 lakhs reviews of cellphones bought on amazon by the users we'll look into it to find useful insights with Reviews, Price, and their Relationship

Field of Feature sets: -

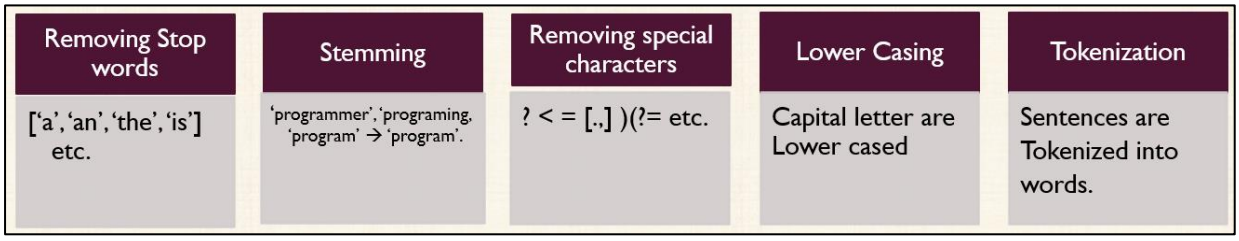
- Product Title
- Brand
- Price
- Rating
- Review Text
- Number of People who find those reviews helpful

Dataset can be found from the link below:

https://www.kaggle.com/PromptCloudHQ/amazon-reviews- unlocked-mobile-phones

3.1.2 Pre-processing or cleaning of data

Used Natural Language toolkit (NLTK) library to do the pre-processing of data.



3.1.3 Libraries Used in Sentiment Analysis

Textblob

The sentiment analyzer has two properties for the given sentence:

- **Polarity [-1,1],** -1 indicates negative sentiment and +1 indicates positive sentiments.
- **Subjectivity [0,1]** Subjectivity refers to opinion, emotion, or judgment.

```
Sentiment(polarity=1.0, subjectivity=0.75)
```

Vader

- It has a list of lexical features (e.g. word) which are marked as positive or negative as per semantic orientation.
- It returns the probability as 'positive', 'negative', and 'neutral' in a given sentence.

```
{'compound': 0.6588, 'neg': 0.0, 'neu': 0.406, 'pos': 0.594}
```

3.1.4 Separation of Data into Positive and Negative sets according to their sentiment score.

- Calculating Sentiment Score

We've calculated the sentiment score by multiplying the compound score and subjectivity score to identify these sets.

- Creating Sets

Sorted the stored products and the most negative and the most positive sentences are selected on the basis of sentiment score.

3.1.5 Proposed Architecture

We will be using Text Rank Algorithm(extractive and unsupervised), which uses a graphbased approach mentioned above, to develop our summarizer.[4] The approach is similar to Google's PageRank algorithm and it ranks sentences by importance.[8] Below is flow diagram of Text rank Algorithm.

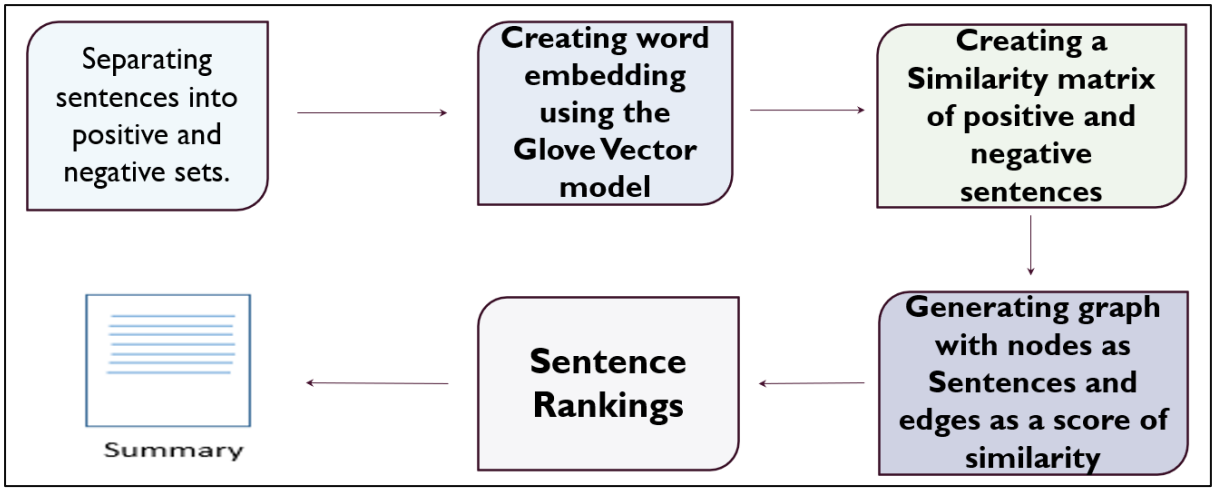


Fig 3.2Flow Diagram of Text rank algorithm

- The first step is to merge the individual articles into a single document.
- Then split that text into sentences.
- Vectorization of Each and Every Sentence
- Sentence vector similarities are calculated and stored in Matrix
- Similarity matrix then represented as a graph with sentences on the nodes and edges on the basis of similarity
- When the TextRank algorithm is applied to the graph and the sentences are sorted in descending order. The Highest-scoring 10 sentences have been displayed to the user.

3.1.6Calculation of similarity

Let's say S_i, S_j two sentences shown by a set of n words that in Si are shown as $S_i = w^i, w^i, \ldots, w^i$. The similarity function for this is shown as:

$$Sim(S_i, S_j) = \frac{|\{w_k \mid w_k \in S_i \& w_k \in S_j\}|}{\log(|S_i|) + \log(|S_j|)}$$

3.1.7 Evaluation

We also evaluated those results on based on two metrics:

1. **Model Efficiency (ME):** Percentage of the reviews that our model identified as positive and negative. Reviews, which could not be classified as positive or negative. We call those reviews neutral reviews and such reviews were left out by our model during classification.

$$ME = \frac{(Total\ positive\ reviews\ generated + Total\ negative\ reviews\ generated)}{Total\ reviews\ available}$$

2. Time efficiency (TE): Simply factor increase the efficiency of reading time. It's known that the average human reading rate is 225 words/ minute. We can then estimate the time taken to read all the reviews, given the average reading rate of humans. We will also know the execution time of our program and the number of reviews generated by our program during runtime Hence, we'll estimate the time took to generate the best reviews out of the review corpus. Finally, **we can calculate the ratio of time taken by a person to read all the reviews to the time taken by a person to read the best reviews generated**. This ratio will tell us the factor increase in the efficiency of reading time.

4. SYSTEM DOMAIN

Hardware Requirement –

The following describes the hardware needed in order to execute and develop the Automatic summarization of user review:

OS: Windows 7 or later one

Processor: i3, i5 and equivalent Ryzen

Main Memory:8 GB RAM (Minimum)

Hard Disk:120 SSD or 1 TB HDD

Software Requirement –

The following describes the software needed in order to execute and develop the Automatic summarization of user review:

- **Tech Stack:**

  - Language: Python

  - Libraries: NLTK, Textblob, Vader, NumPy, Pandas, Networkx

  - Platform: Kaggle

- **Algorithm used:**

  - TextRank: To Rank Sentences according to its importance in text, based on Google Page Rank Algorithm.

  - Glove Model: To Create Word Embeddings. We've used Stanford's GloVe 100d word embeddings for our project.

5. APPLICATION DOMAIN

- **Media vigilance:** The issue of information overload and "content shock" has received a lot of attention. The option to break down the never-ending stream of information into manageable chunks is provided by automatic summarization.

- **Newsletters:** Summarization would enable businesses to add more value to newsletters by adding a stream of summaries (instead of a list of links), which can be a particularly useful format on mobile.

- **Financial research:** Systems designed to summarise financial documents, such as earnings reports and financial news, can speed up the process by which analysts extract market signals from content.

- **Social media marketing:** Organizations that create lengthy content, such as whitepapers, e-books, and blogs, may be able to use summary to simplify their work and make it easier to share on social media platforms like Twitter or Facebook. This would enable businesses to use current content again.

- **E-learning and class assignments:** Many lecturers frame their lectures with case studies and current events. By creating a summary report, summarization can assist teachers in updating their content more rapidly.

- **Legal contract analysis:** This point is connected to point 4 (internal document workflow), where it is possible to create more specialised summary systems to analyse legal documents. In this situation, a summarizer could be useful by distilling a contract down to its main provisions or by assisting you in comparing contracts.

6. EXPECTED OUTCOME

Since the Internet has increased user interaction, the number of consumer reviews written online has grown significantly. However, it can be challenging for marketers and business analysts to comprehend client concerns due to the sheer volume of customer evaluations that are placed on websites like Amazon.com. In this presentation, we outline a method for automatically summarising customer reviews and discuss the initial findings of our study on Amazon.com product reviews. We also evaluated those results on based on two metrics: Model Efficiency and Time efficiency. Our research, we hope, will advance the methods and comprehension of customer review summaries and will be advantageous to web marketers, business intelligence, and company owners alike. study in the fields of e-commerce and text mining.

7. REFERENCES

1. M. F. Mridha, A. A. Lima, K. Nur, S.. Das, M. Hasan and M. M. Kabir, "A Survey of Automatic Text Summarization: Progress, Process and Challenges," in IEEE CAccess, volume 9, pp. 156043-156070, 2021, doi:10.1109/ACCESS.2021.3129786. (2021)

2. Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, Hoda K. Mohamed,Automatic text summarization: A comprehensive survey, Expert Systems with Applications,Volume 165,113679 ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2020.113679 (2021)

3. Kumar, Y., Kaur, K. & Kaur, S. Study of automatic text summarization approaches in different languages. ArtifIntell Rev 54, 5897–5929 https://doi.org/10.1007/s10462-021-09964-4 (2021)

4. Atif, Naomie and Yogan. P ,Genetic semantic Graph Approach for multi document abstractive Summarization. Fifth International Conference on Digital Information Processing and Communications (ICDIPC)(2015).

5. GloVe: Global Vectors for Word Representation: Jeffrey Pennington, Richard Socher, Christopher D. Manning https://nlp.stanford.edu/projects/glove/ (2014)

6. Ramesh, Bowen, Cicero. Abstractive Text Summarization using Sequence-to sequence RNNs and Beyond , https://arxiv.org/abs/1602.06023 (2016).

7. Hongyan, Sentence Reduction for Automatic Text Summarization, P. (2008).

8. Atif, Naomie and Yogan. P ,Genetic semantic Graph Approach for multi document abstractive Summarisation. Fifth International Conference on Digital Information Processing and Communications (ICDIPC) (2015).

9. Dipanjan, Andr´e . A Survey on Automatic Text Summarization. Language Technologies Institute Carnegie Mellon University, {dipanjan, afm} @cs.cmu.edu (2000)

10. de Chalendar, G., Ferret, O., et al. . Taking into account inter-sentence similarity for update summarization. In Proceedings of the Eighth International Joint Conference on NLP (Volume 2: Short Papers), volume 2, pages 204–209 (2017).