Major Project SRS

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

Table of Contents

1. Introduction

1.1 Purpose

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.

1.2 About this Document and its Readers

The system requirements specification document describes what the system is to do, and how the system will perform each function. The audiences for this document include the system developers and the users. The system developer uses this document as the authority on designing and building system capabilities. The users review the document to ensure the documentation completely and accurately describes the intended functionality.

This version – version 1.0 - provides general descriptions of the system. The system developer should review the document to ensure there is adequate information for defining an initial design of the system. The users should review the document to affirm the features described are needed, to clarify features, and to identify additional features needed within the system.

The next version – version 2.0 – will be the result of more detailed requirements analysis. When version 2.0 is written, the system developer and users will be asked to review this document. The document is structured to follow IEEE 830-1998 standards for recording system requirements.

1.3 Product Scope

Nowadays, automatic text summarization systems can successfully retrieve the summary sentences from the input documents. But it has many limitations such as inaccurate extraction to essential sentences, low coverage, poor coherence among the sentences, and redundancy. Our "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.

1.4 References

a. 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)

b. 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)

c. 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)

d. 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).

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

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

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

h. 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).

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

j. 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).

2. Overall Description

2.1 Product Perspective

This Product “Automatic Summarization of User Reviews” aims to provide an efficient and enhanced summarization tool for the users that can be used to perform automatic text summarization of all the users to determine the reliability of a product based on the past experiences of the customers who have used the product previously. Since there are many reviews present for a product and a user does not have the time to go through all of them, our model will aim to summarize all the reviews and pick only the Top 10 (both negative and positive out of those. This will help the user in making a faster decision.

2.2 Product Functions

In this project, we will tend to summarize the reviews posted by the users on the e-commerce giant Amazon for the mobile phones purchased by them through the website The dataset used will be Prompt cloud which contains approx. 4 Lakhs reviews. We'll be analyzing our results based on certain metrics like Model Efficiency and Time Efficiency. The accuracy of the summarizer will also be evaluated. The mixed Reviews present in the dataset will first be segregated into positive and negative and then a score will be given to each of those positive and negative reviews. The Top 10 reviews with the highest score will be presented to the user based on which a user will be able to conclude its analysis and make an informed decision.

2.3 User Classes and Characteristics

- **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.

2.4 Operating Environment

The rule for selecting hardware and software is that the components/application must be functionally efficient, capable of interfacing with other software, and easy to maintain.

a) OS: Windows 7 or later one

b) Processor: i3, i5 and equivalent Ryzen

c) Main Memory:8 GB RAM (Minimum)

d) Hard Disk:120 SSD or 1 TB HDD

e) Language: Python

f) Libraries: NLTK, Textblob, Vader, NumPy, Pandas, Networkx

g) Platform: Kaggle or Anaconda

h) Database: from Kaggle - Amazon Reviews: Unlocked Mobile Phones

2.5 Design and Implementation Constraint

- Some devices running version lower than, won’t able to use this model.
- This system shall be developed using open-source tools.
- Information or data flow can be controlled and more effective.
- This system shall be developed using Kaggle Platform.

2.6 User Documentation

The list below contains user documentation for the Data model mentioned in the SRS:

Python: https://www.python.org/doc/

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

Kaggle: https://www.kaggle.com/docs

NLTK: https://www.nltk.org/

TextBlob: https://textblob.readthedocs.io/en/dev/

Vader: https://pypi.org/project/vaderSentiment/

Glove Model: https://nlp.stanford.edu/projects/glove/

TextRank Algo: https://cran.r-project.org/web/packages/textrank/vignettes/textrank.html

Networkx: https://networkx.org/documentation/latest/

2.7 Assumption and Dependencies

No specific assumptions or dependencies are considered at this time.

3. External Interface Requirements

3.1 User Interface

While most participants completely agree the most for aspect-based summaries when we combined the ratings of completely agreed and agreed statistical summary where the most favored when the participants were asked to write a short summary of the rating based on provided criteria.

3.2 Hardware Interface

System runs on any regular PC/Laptop If the user intends on training the model on a different dataset, then a machine with 4 GB or more RAM and processor of 2.7 GHz or more is required. Commercial servers, workstations, and other high-end PCs may have more than one physical processor. Windows 7 Professional, Enterprise, and Ultimate allow for two physical processors, providing the best performance on these computers. Windows 7 Starter, Home Basic, and Home Premium will recognize only one physical processor.

3.3 Software Interface

- Text Rank Algorithm

TextRank is a graph-based ranking algorithm like Google's PageRank algorithm which has been successfully implemented in citation analysis. We use text rank often for keyword extraction, automated text summarization and phrase ranking. Basically, in the text rank algorithm, we measure the relationship between two or more words.

- GloVe Model

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

- Natural Language Processing

NLP is a field of artificial intelligence that allows the computers to read, understand, and determine meaning from a human dialect in a keen and valuable way. It acts as a middle point of computer science, artificial intelligence, and computational etymology. Owing to the complicated nature of natural human language, processing of the data using this technique can prove to be difficult. Despite this complication, NLP has found a wide range of applications in different domains. By utilizing NLP, designers can organize and structure information to perform errands such as programmed summarization, interpretation, named substance acknowledgment, relationship extraction, sentiment analysis, speech recognition, and topic segmentation.

- Fuzzy Logic

Fuzzy logic is used to estimate the degree of significance and relationship and also to highlight the vital phrases to form summarization. Fuzzy Logic is mainly based on natural language. It may be a numerical device utilized for dealing with uncertainty, imprecision, ambiguity, and unclearness. Fuzzy logic could be a form of many-valued logic that bargains with inexact reasoning instead of settled and exact reasoning. Fuzzy logic is utilized to handle the concept of halfway truth where its truth esteem ranges between fully genuine and fully wrong.

3.4 Communication Interface

The system will use the communications resources provided by the Kaggle. The basic assumption of this work is that various representation components have differences in contributing to the structure of a text and the formation of the structure significantly influences the differences. Human summarization is about knowledge, understanding and language use while automatic summarization concerns the modeling of text. There are two streams of models on texts: one stream (including the vector space model and topic model) assumes that words are independent of each other, and the other stream (including semantic link network) assumes that words are inter-related to render themes (this work distinguishes theme from topic according to the semantic link point of view). Human reading involves different strategies in different cases. Integrating the two streams of models is a way to establish a powerful model for summarization.

4. System Features

4.1 Since there are a large number of reviews present for a product and a user does not have the time to go through all of them, our model will aim to summarize all the reviews and pick only the Top 10 (both negative and positive out of those.

4.2 This will help the user in making a faster decision. In this project, we'll tend to summarize the reviews posted by the users on the e-commerce giant Amazon for the mobile phones purchased by them through the website The dataset used will be Prompt cloud which contains approx. 4 Lakhs reviews.

4.3 We'll be analyzing our results based on certain metrics like Model Efficiency and Time Efficiency. The accuracy of the summarizer will also be evaluated.

4.4 The mixed Reviews present in the dataset will first be segregated into positive and negative and then a score will be given to each of those positive and negative reviews. The Top 10 reviews with the highest score will be presented to the user based on which a user will be able to conclude its analysis and make an informed decision.

5. Other Nonfunctional Requirements

5.1 Performance Requirement

- Application must be simple and easy to use.
- Application must be intuitive and simple in the way it displays all relevant data and relationships.

5.2 Safety Requirement

The application must inform the user in situations of a crash or error should be up for working as and when required.

5.3 Security Requirements

System runs on any regular PC/Laptop If the user intends on training the model on a different dataset, then a machine with 4 GB or more RAM and processor of 2.7 GHz or more is required.

5.4 Software Quality Attribute

The quality of the automatic summaries be performed and intrinsic online evaluation which is one of the standard evaluation approaches used in the field of text summarization.

This evaluation involves the active participation of human judges who rate each of the automatic summaries based on the own perception of its internal quality.

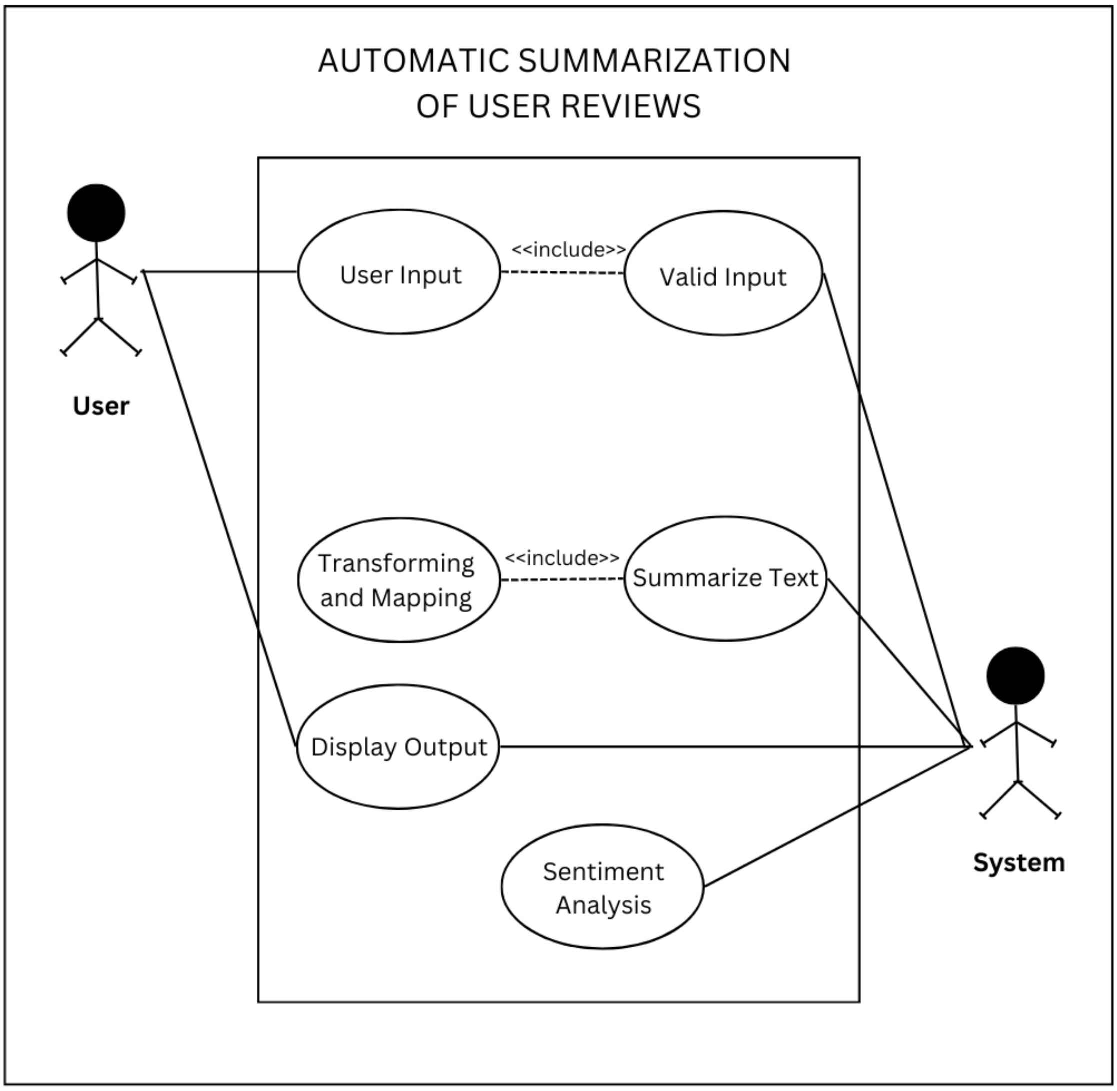
5.5 Business Rules

As the different techniques came out, each had its advantages and its disadvantages. Some performed better on particular domains, for example automobile articles, while others had a better overall performance. Keeping this in mind it was decided to study the effectiveness of different machine learning and deep learning models along with finding out which metrics are more influential in forming a summary. By finding what produces the most coherent summary through the research, it is aimed to aid future endeavors in the process of summarization and, subsequently, building a more accurate model. To train the model, a manually compiled dataset from different sources and domains was used. The decision to use of varied domains was an attempt to avoid bias towards particular topics and give the model equal exposure to different domains. The dataset comprises of individual sentences from each article and whether or not it was included in the summary.

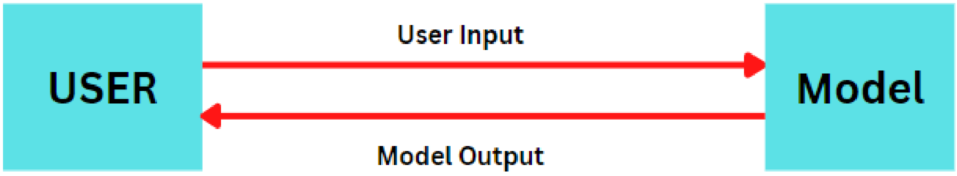
6. Other Requirements

6.1 Analysis Model

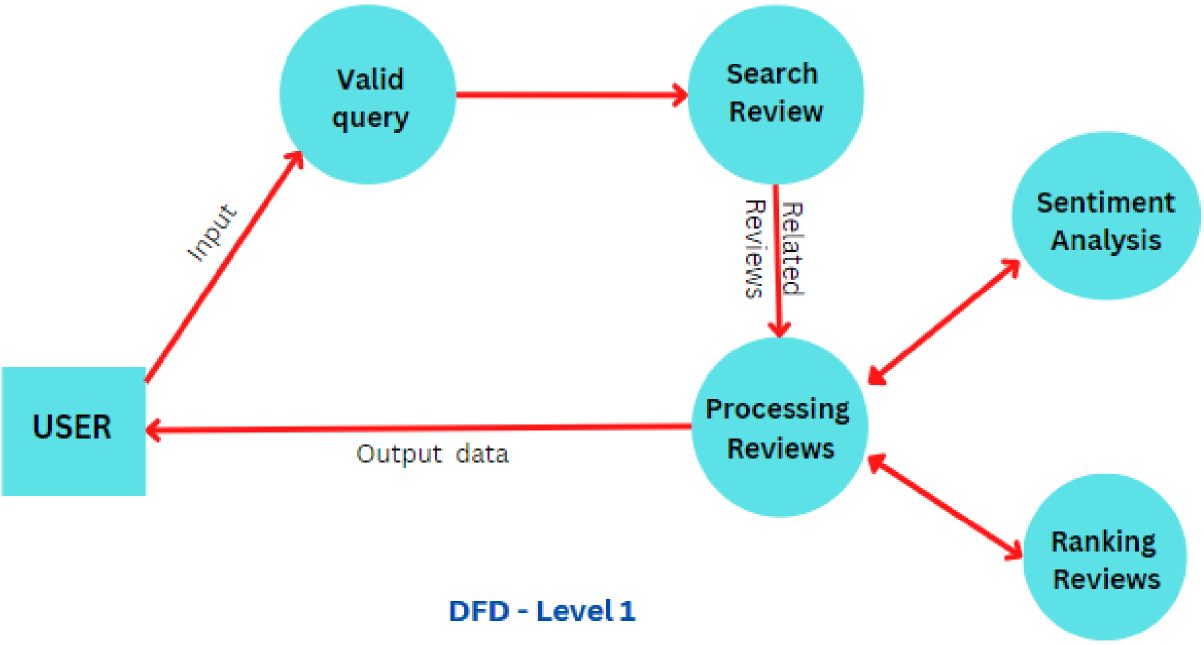
- **Use Case Diagram**



- **Data Flow Diagram**

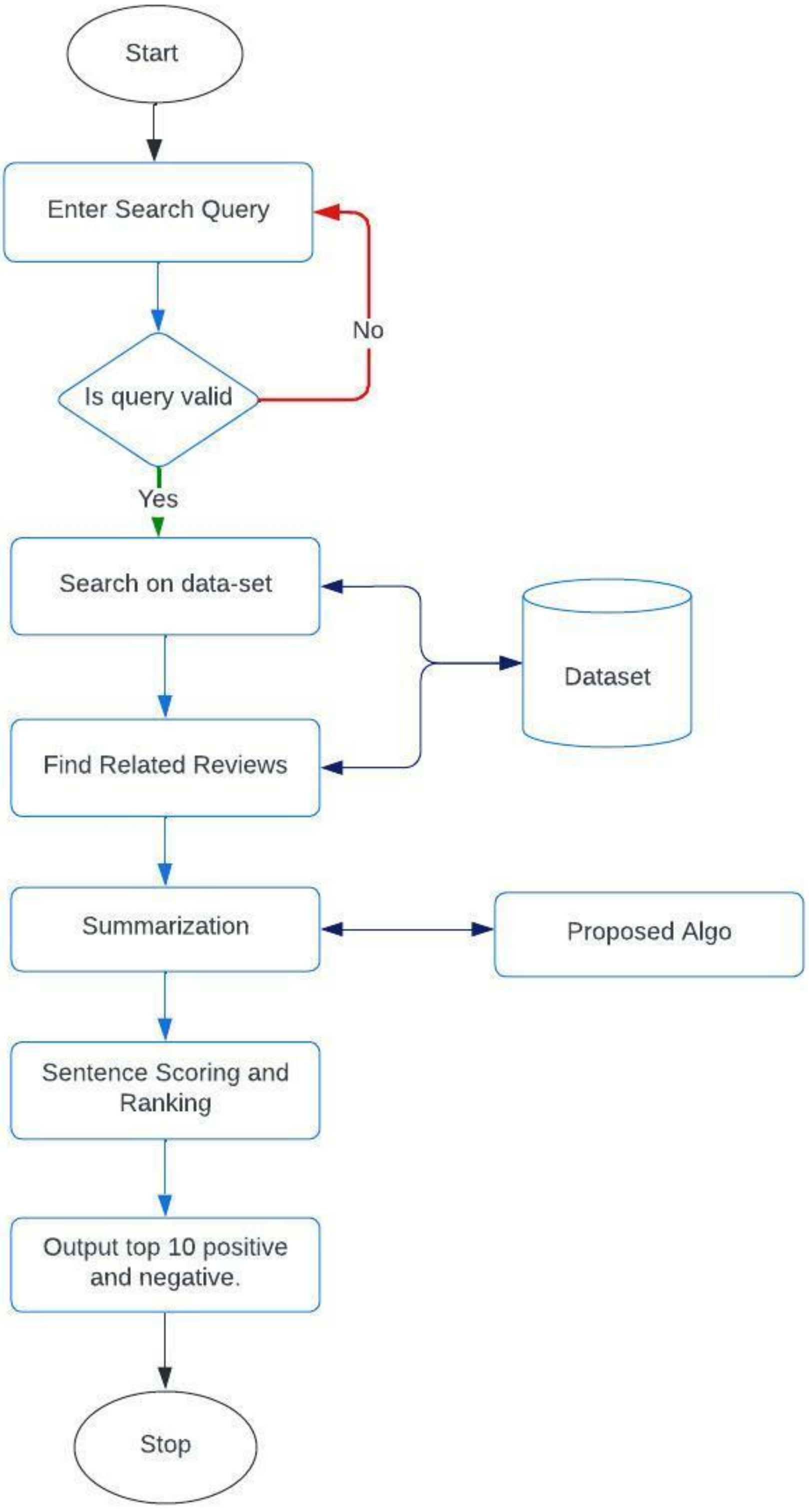


DFD - Level 0

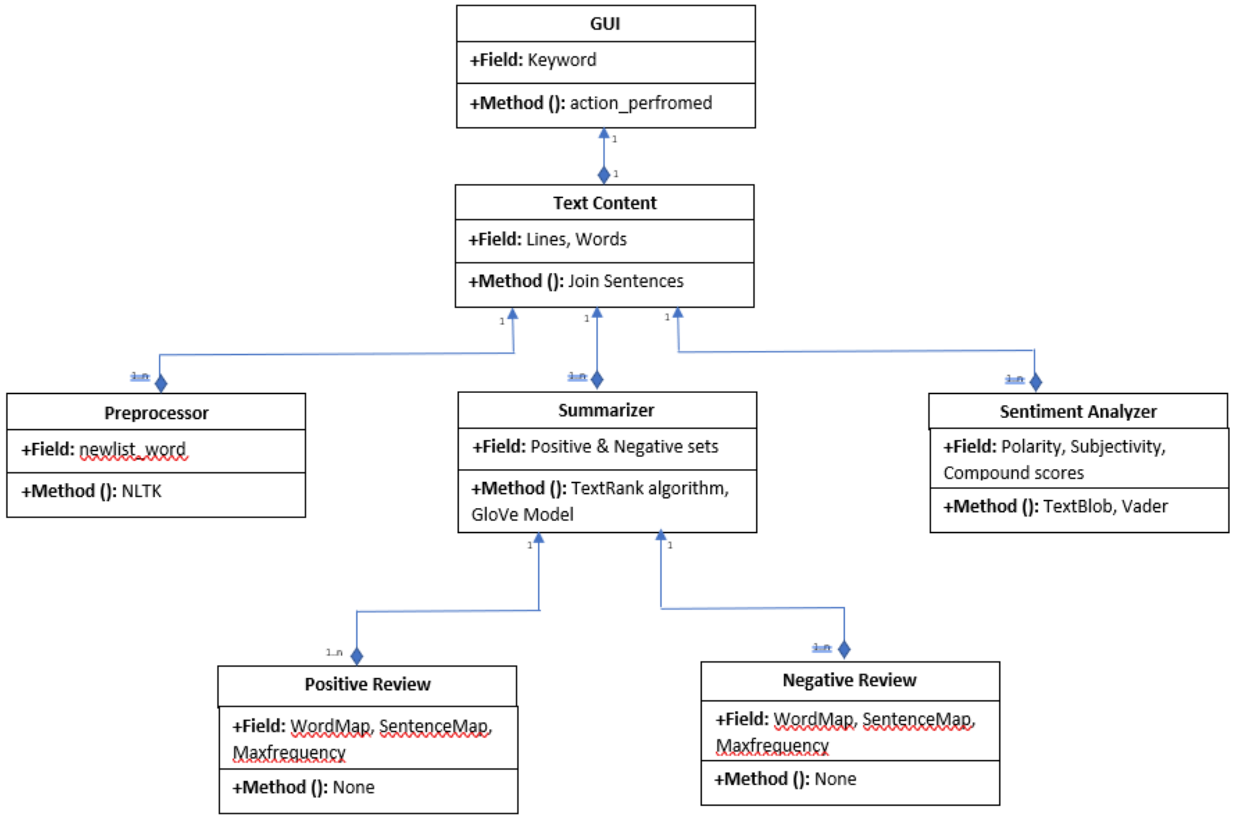


DFD - Level 1

- **Activity Diagram**



- ## Class Diagram


GUI
+Field: Keyword
+Method (): action_perfromed
1
1
Text Content
+Field: Lines, Words
+Method (): Join Sentences
1
1
1
1..n
1..n
1..n
Preprocessor
+Field: newlist_word
+Method (): NLTK
Summarizer
+Field: Positive & Negative sets
+Method (): TextRank algorithm, GloVe Model
Sentiment Analyzer
+Field: Polarity, Subjectivity, Compound scores
+Method (): TextBlob, Vader
1
1
1..n
1..n
Positive Review
+Field: WordMap, SentenceMap, Maxfrequency
+Method (): None
Negative Review
+Field: WordMap, SentenceMap, Maxfrequency
+Method (): None