DocFix - One Website for All Your Document Needs

Varun Deokar[1], Mohit Jain[2], Aznan Khan[3], Vinita Bhandiwad[4]

[1]Student, Dept. of IT Engineering, Vidyalankar Institute of Technology, Maharashtra, India
[2]Student, Dept. of IT Engineering, Vidyalankar Institute of Technology, Maharashtra, India
[3]Student, Dept. of IT Engineering, Vidyalankar Institute of Technology, Maharashtra, India
[4]Professor, Dept. of IT Engineering, Vidyalankar Institute of Technology, Maharashtra, India

---

Abstract - *In today's life, everyone uses a computer, and information is exchanged in the form of text by students, teachers, working professionals, businessmen, even government agencies. When there are millions of documents being created daily, there needs to be an app that can take care of the numerous small issues that arise in text-based documents. For these reasons, we wanted to create a space that a user can navigate to handle all of his text related problems. This is where the idea of DocFix was born. DocFix is a website created for the quick and inclusive insight gathering of any text-based documents. It helps us understand articles, documents, summaries, and any other text-based information in a more time efficient and reliable way.*

***Key Words*: text-summarization, sentiment-analysis, content-downloader, transformers, NLP, Flask, document editing**

1.INTRODUCTION

DocFix is an all-inclusive website for any text related work. The purpose of DocFix is to get more analytical and esoteric insights about your document. Features such as Text-Summarization, Sentiment Analysis, Content Downloader and Keyword Finder can be used to not only understand the documents better but also to help save time and manpower to achieve the same results. This website helps automate the process of multiple tedious tasks and provides an easy-to-understand interface for the users.

The need for quick and reliable information gathering is paramount in the world today. This app is necessary in today's world as the computer has replaced pen and paper. People rarely take down notes in books any longer. They just open a text writing app and start typing. With so much typed data available. They need a website to analyze their data in a quick and efficient manner to find out what is important. People don't have the time to read through all their documents meticulously. This is where our website comes in handy.

The objective of the proposed system is to tackle all the challenges faced by people related to text documents. Editing and analyzing documents for further work. Our main focus is to build one website which can take care of all your document needs. Our goal is for "DOCFIX" to become a verb used for editing documents just as photoshop is used in sentences for 'editing pictures.' This application provides features such as summary generation, sentiment analysis, word count, and content downloader.

2. Literature Survey

In the paper titled "Sentiment analysis using product review data" by Xing Fang and Justin Zhan. [1]

Sentiment analysis or opinion mining is a field of study that analyses people's sentiments, attitudes, or emotions towards certain entities. This paper tackles a fundamental problem of sentiment analysis, sentiment polarity categorization. Online product reviews from Amazon.com are selected as data used for this study. A sentiment polarity categorization process has been proposed along with detailed descriptions of each step. Experiments for both sentence-level categorization and review-level categorization have been performed. This paper identifies and evaluates some algorithms which can be used to implement this feature.

In the paper titled "A survey on text summarization" by N Moratanch and Chitrakala Gopalan from Anna University, Chennai DOI – Jan 2017. [2]

The idea behind text summarization is to provide a concise summary of the document to the user such that it contains all the salient information.

Key learnings from the paper are that Text Summarization techniques are classified into abstractive and extractive summarization. We would be considering 2 categories of feature for selecting the vital sentences; those are Word level feature and Sentence level feature.Text summarization task is basically modelled as a classification problem wherein we would be categorizing each sentence as a summary sentence or a non-summary sentence.

3. Proposed System

As a student or a working professional, we all need several fixes for our document such as text summarization, sentiment analysis and many more. And for this we have to search around the web for each feature and still it is hard to find an optimal solution for our problem.

To overcome this tedious task, we came up with an idea to create an all-in-one platform for all the document needs. The application will provide the following functionalities: -

- Concise summary of the text or passage entered.
- Sentiment of the entered passage or a piece of text.
- Automated text download of provided website.

Such a system in place will help several users in their various walks of life.

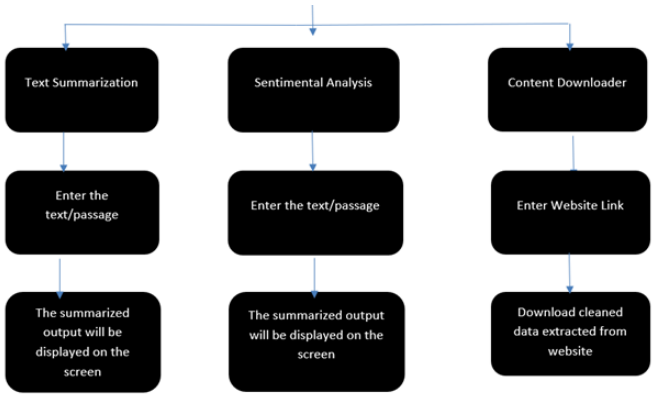

Fig -1: Flowchart for DocFix

A set of training documents along with its extractive summaries is fed as input to the training stage. [3] The sentences are restricted as a non-summary and summary sentence based on the feature possessed by the sentence. The probability of classification is learned from the training data. For this classification we would be either using Machine Learning approach Bayes rule or Neural Network based approach.

Using the above approaches, we will train our ML model and will categorize each sentence of the document as summary and non-summary sentences.

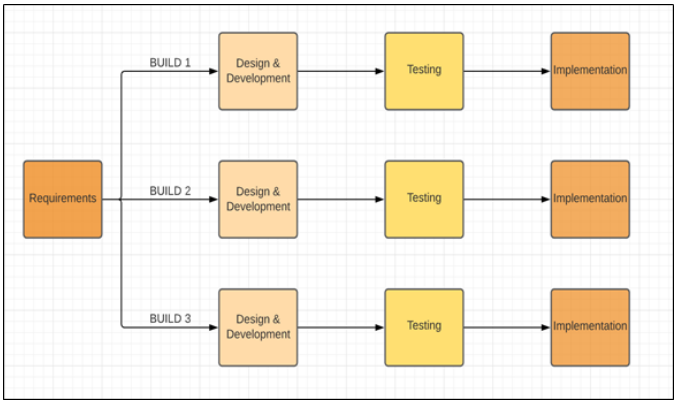

Fig -2: Iterative Process Model

The iterative development model develops a system through building small portions of all the features. This helps to meet initial scope quickly and release it for feedback.

In the iterative model, you start off by implementing a small set of the software requirements. These are then enhanced iteratively in the evolving versions until the system is completed. This process model starts with part of the software, which is then implemented and reviewed to identify further requirements.

Like the incremental model, the iterative model allows you to see the results at the early stages of development. This makes it easy to identify and fix any functional or design flaws. It also makes it easier to manage risk and change requirements.

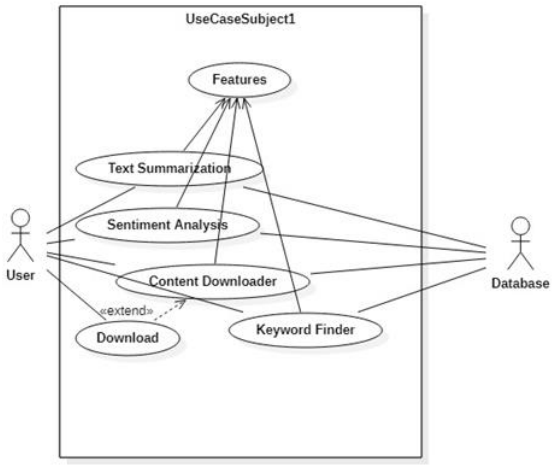

Fig -3: Use Case Diagram

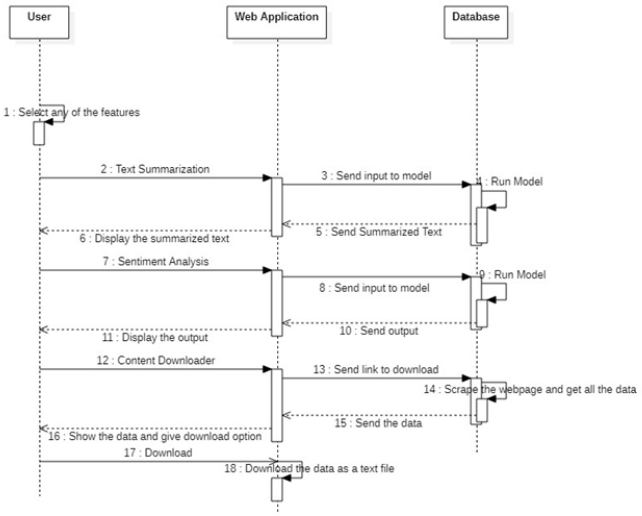

Fig -4: Sequence Diagram

4. Methodology

The domains utilized in this application are-

1. Machine Learning
2. Deep Learning
3. Natural Language Processing

The key features of our application are

- Text Summarization
- Sentiment Analysis
- Content Downloader

Text Summarization is the process of obtaining salient information from an authentic text document. In this technique, the extracted information is achieved as a summarized report and conferred as a concise summary to the user.

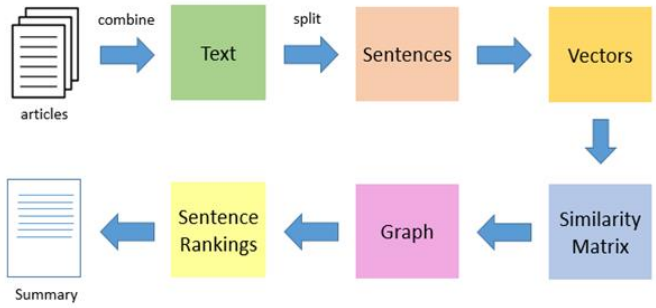

Fig -5: Working of Text Summarization

So, we will be using Text summarization technique for our website in which we will be selecting vital sentences, paragraphs, etc. [4] from the original manuscript and concatenating them into a shorter form. The significance of sentences is strongly based on statistical and linguistic features of sentences. Hence the concise summary of the document will be generated. [5]

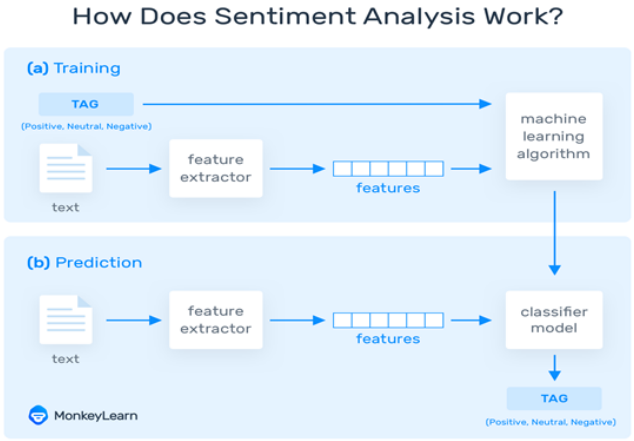

Fig -6: Working of Sentiment Analysis

Sentiment Analysis is the most common text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative our neutral. You can input a sentence of your choice and gauge the underlying sentiment.

- The input from the user which could either be a small piece of text or a passage is stored.
- Data preprocessing is done on the entered input.
- The preprocessed data is then fed to the trained machine learning model. [6]
- The model then outputs the sentiment of the preprocessed data which is displayed on the website.

To eliminate the redundant task of copy pasting articles online to store it locally, we created the content downloader feature.

This feature takes in a link to any website and extracts all the relevant textual data from it while skipping over ads and other unrelated data. [7]

After extracting this data, you have an option to download it on your local machine. The file name would begin with the prefix of 'docfix' followed by the time you asked for it to be downloaded.

This feature can also be run as a preprocess to text summarization or sentiment analysis if you want to run those features on data from a website.

5. Results

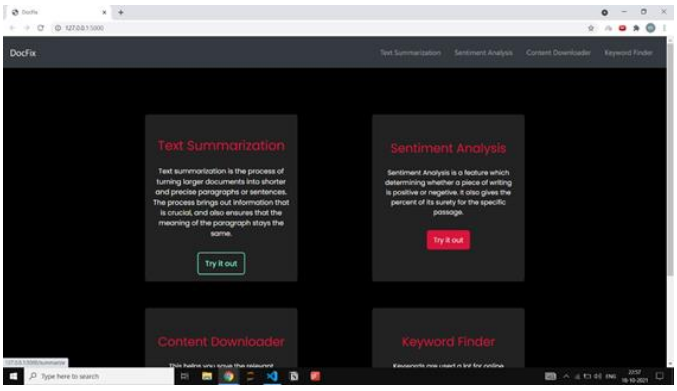

Fig -7: Homepage of DocFix website

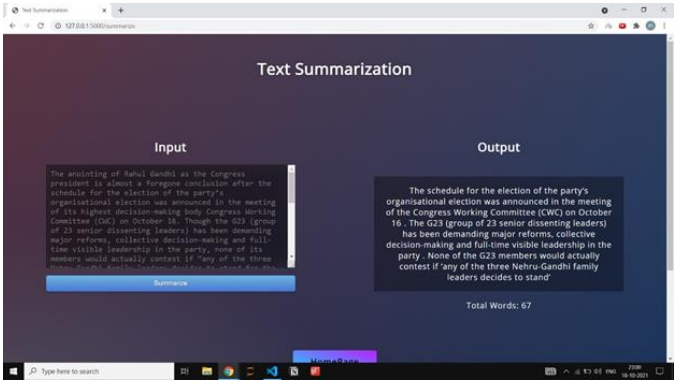

Fig -8: Output view of Text Summarization

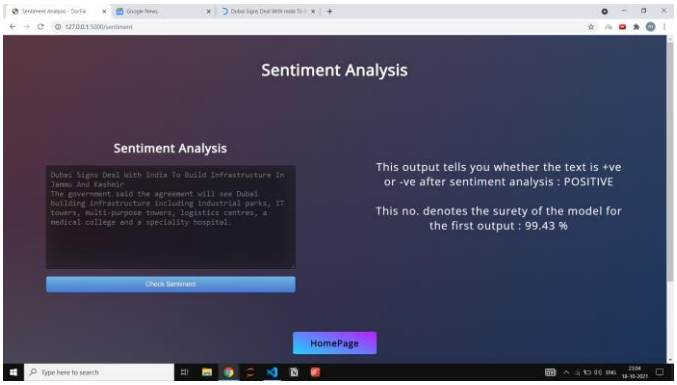

Fig -9: Output view of Sentiment Analysis

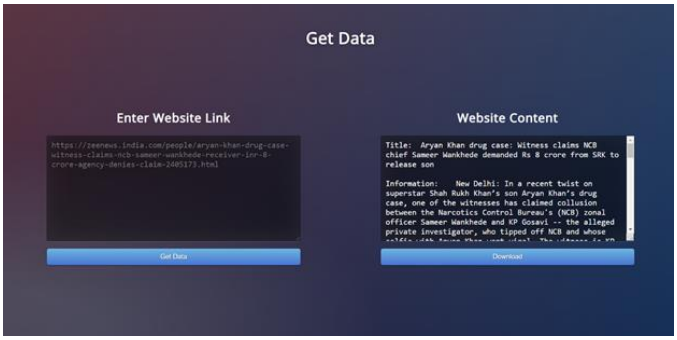

Fig -10: Output view of Content Downloader

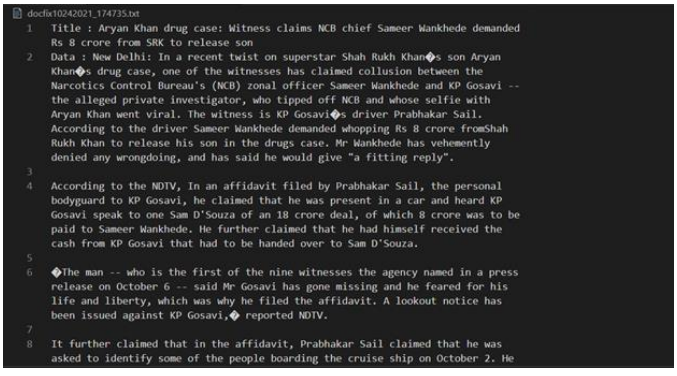

Fig -11: Text file downloaded through content downloader page

6. CONCLUSIONS

DocFix is a website built using Flask with technologies of natural language processing and machine learning incorporated within it. It is used to analyze and gain insights into text and text-based documents using features such as text summarization, sentiment analysis and content downloader. Our project could be a revolutionary product in the field of online document analysis. If implemented correctly our project can help everyone who deals with text-based data, from students to teachers, there will always be more features to be implemented and better methods of implementation of previous features with advancements in technology. The only thing that limits the scope of our project is our imagination.

REFERENCES

[1] Fang, X., Zhan, J. Sentiment analysis using product review data. Journal of Big Data 2, 5 (2015). https://doi.org/10.1186/s40537-015-0015-2

[2] "A survey on text summarization" by N Moratanch and Chitrakala Gopalan from Anna University, Chennai DOI – Jan 2017.

[3] Mike Lewis and Yinhan Liu and Naman Goyal and Marjan Ghazvininejad and Abdelrahman Mohamed and Omer Levy and Veselin Stoyanov and Luke Zettlemoyer (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. CoRR, abs/1910.13461.

[4] I. Awasthi, K. Gupta, P. S. Bhogal, S. S. Anand and P. K. Soni, "Natural Language Processing (NLP) based Text Summarization - A Survey," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1310-1317, doi: 10.1109/ICICT50816.2021.9358703.

[5] R. Boorugu and G. Ramesh, "A Survey on NLP based Text Summarization for Summarizing Product Reviews," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 352-356, doi: 10.1109/ICIRCA48905.2020.9183355.

[6] Thomas Wolf and Lysandre Debut and Victor Sanh and Julien Chaumond and Clement Delangue and Anthony Moi and Pierric Cistac and Tim Rault and Rémi Louf and Morgan Funtowicz and Jamie Brew (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. CoRR, abs/1910.03771.

[7] Richardson, L. (2007). Beautiful soup documentation. April.