North South University
Department of Electrical and Computer Engineering

Junior Design Project

Text Summarizer and Entity Checker

Rokeya Siddiqua	**ID#1813063042**
Nusrat Islam	**ID#1812037042**
Saima Alam Miduri	**ID#1812399042**

Faculty Advisor
Tanjila Farah
Department of Electrical and Computer Engineering

SPRING 2021

Chapter 1

ABSTRACT

Text summarization and Entity checker is a process of extracting or collecting important information from original text and presents that information in the form of summary. It has become a necessity in our daily life starting from academics to research purposes. Summarization helps to gain required information in less time to improve resources. Following this goal, we want to develop an NLP model for text summarization and create an end to end web application which can take a document or URL as input and generate a summary. Although there are so many text summarizers available currently but our target is to create a web application containing summarizers, entity checker, redactor and word details checker to provide a best version of Text Summarizer And Entity checker for users from any background. This paper interprets extractive text summarization methods with a less redundant summary and depth information.

Keywords: Text Summarization, Compare Summaries, Natural Language Processing (NLP), Machine Learning (ML), Count Vectorizer, Term Frequency–Inverse Document Frequency Vectorizer (TD-IDF), Application Programming Interface (API), Text Type, Natural Language Toolkit (NLTK), Parts of Speech (POS), Lemmatization (LEMMA), Generate Similar (GENSIM), JavaScript Object Notation (JSON), Entity Checker, Redactor.

Chapter 2

INTRODUCTION & BACKGROUND

Our goal is to develop a web based software called **Text Summarizer and Entity checker** which will generate the fluent and most accurate summary from a longer text document. The main idea behind this software is to find the summarized text with most essential information automatically and also check the entities with various information.

In our software, users can input into the summarizer tool in a three way- written text/copy paste text, importing file, inserting URL and get the summarized text as output. To show the output we will be using **Spacy**, **Gensim**, **Summy** libraries for summarized text in a three way like short, medium, longer etc. and also our software will provide an option where users can set the word limits. And this summarized text or output can be saved for the future uses and also possible to download the file and also it is possible to edit the summarized part with the user's own information.

Our software not only summarizes the text also has some other significant features and one of the features is **Entity checker** and other is **Redactor**. User can search by Name, Date, Place, Organizations to check the entities and our software will highlight those entity from the text and also user can redact those entity and the redactor will work as a way that it will remove the entity which user will want and will marked that particular entity as redacted. After redacting, our software will generate a text file which is possible to download. And there will be a download list where previously saved files will be stored as a list and from there the user can get the file and re-use or re-download the file.

Our software will contain some other features like user can search the information of key words or token and user can get the words info or type of word (lemma, shape, alpha, stopword) ,word meaning, sentiment (polarity, subjectivity), part of speech(tag, pos, dependency) etc. And also our software will contain a feature to show the **text type**. If a user wants to search for the text type our software will show that it is a private text or business text or any other type.
This table contains the information of the required libraries of our project:

Frontend	Backend	Libraries
FLASK: It's a framework. We will use it because it's fast.	**ML:** Our App will receive input from the end user and then process that input with our models. Finally, using Machine Learning classification we will predict the text type.	**Frontend:** Flask **Backend:** Scikit-learn, Pandas, Random Forest, Logistic regression, Support Vector Machine, Decision Tree
HTML: It will create our web app pages.	**NLP:** Tokenization, named entity recognition, pictorial form for the most common words and lemmatization in our text will be shown using NLP libraries.	**Backend:** Spacy, wordcloud
BOOTSTRAP: It will design our web app pages.	**API:** We will display our word details as a json format.	**Frontend:** Bootstrap **Backend:** Json, streamlit
JAVASCRIPT: It will check the form validation.	**DATABASE:** Feedback from the user will be stored here.	

2.3 Extra Libraries

Textblob, urllib, matplotlib, tifdfvectorize, countvectorizer,

NLTK, Gensim, SUMY, streamlit

2.4 Vectorizers

Since our models cannot work with text, we will need to vectorized them or convert them into numbers. Hence we will be using tifdfvectorizer and countvectorizer to vectorize our text into an array of numbers so that our ML model will be able to process them.

TfidfVectorizer: TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents. Which algorithm will follow this vectorizer is described in **figure 1** exhaustively.

$$tf\ idf\ (t, d, D) = tf\ (t, d)\ .\ idf\ (t, D)$$

Where:

$$tf\ (t, d) = log\ (1 + freq\ (t, d))$$

$$idf\ (t, D) = log\ (\ \frac{N}{count\ (d \in D : t \in d)}\ )$$

Fig 1: TFIDFVectorizer calculation

CountVectorizer: CountVectorizer creates a matrix in which each unique word is represented by a column of the matrix, and each text sample from the document is a row in the matrix. The value of each cell is nothing but the count of the word in that particular text sample. A sample matrix has shown in **figure 2** exhaustively.

	Jumps	The	brown	dog	fox	lazy	over	quick	the
Doc1	0	1	1	0	1	0	0	1	0
Doc2	1	0	0	1	0	1	1	0	1

Fig 2: CountVectorizer

Why it is important for users-

- Our software is really important as we can use it for studying purposes or writing any document like research and professional paper. It can give a new view and also can give a different or short idea of any particular topic.

- It is really productive for our daily life as it can save our valuable time by not losing its actual quality rather enhancing the documents and also we can use this time to explore our knowledge with any other sector or we can focus more on other important works. And also it can help users to do the last minute checks and can improve the writing.

- This is a more advanced type of summary as it paraphrased the parts of text which the user inputs into the summarizer tool. If we want to write the summary by our own words, it may not contain the exact quality which we were looking for. But this software can do this with a short amount of time as it is built with rich libraries and APIs.

- Our software not just summarizes the text, it has the ability to do other tasks like check entities, find token, sentiment, word information and check meaning, redact any words etc. so it's a platform where users can get all the features easily. So if any user wants to get any information of words or meaning or any other information they don't have to waste their time searching the internet as our software is providing all the features in a single platform.

Why it is important for us-

- Building this software is really important as we can get the knowledge about several new sectors. To develop this software, we have to get the knowledge about how to deal with NLP libraries and APIs. As NLP is part of AI knowledge, so we have to get ideas how AI works. Also it is a web based development so we can get ideas of building a website along with the knowledge of the framework.

Chapter 3

LITERATURE REVIEW

To summarize a text, a paragraph or even an essay, you can find a lot of tools online. The summary maker shows the reading time, which it saves for you, and other useful statistics. Here we'll list some of these, including those that allow choose the percent of similarity and define the length of the text you'll get. So, the first one is, **Summarize Bot** which is an easy-to-use and ad-free software for fast and accurate summary creation in our list. To summarize any text, you should only send the message in Facebook or add the bot to Slack. The app works with various file types: including PDF, mp3, DOC, TXT, jpg, etc., and supports almost every language. [1]. Then there is **SMMRY** which has everything you need for a perfect summary-easy to use design, lots of features, and advanced settings (URL usage). If you look for a web service that changes the wording, this one would never disappoint you. SMMRY allows you to summarize the text not only by copy-pasting but also with the file uploading or URL inserting. The last one is especially interesting. With this option, you don't have to edit an article in any way. Just put the URL into the field and get the result. [2] Thirdly, **Tools4Noobs summarize tool** is a comfortable article summarizer with a wide range of settings. You can use Threshold function to limit the number of sentences based on relevance or reduce the summary to a specific length. Here it is also possible to see the main keywords or highlight them in the text. The software works with texts you insert or you can give it URL you want to summarize.[3] Another tool is **Split Brain Summary Tool** is a helpful app to summarize texts and articles in a great variety of languages. You can choose one of thirty-nine languages to make a couple of sentences on your article. The difference in summaries also can be produced by the summarization ratio. There's also a possibility to insert an URL instead of the text. However, there is no option to import a file or export the result to PDF, DOC, or any other popular format. [4] Then there is the **TextSummarization tool** that allows you to put the text into the field or give a link to a source where your article is posted. Then, set the number of sentences you want to have in your text. This summary maker analyzes your nonfiction text and extracts the exact number of sentences you're aiming at. The website is free, however, it contains ads, so make sure you turned on the ad-blocker. Also, the tool doesn't let its users import files or export the result to TXT, PDF, or Doc for other popular formats. Another application available is the **Split-Brain Summary Tool** which is a helpful app to summarize texts and articles in a great variety of languages. You can choose one of thirty-nine languages to make a couple of sentences on your article. The difference in summaries also can be produced by the summarization ratio. You can change it from 5% to 80% controlling the density of paraphrasing. There's also a possibility to insert an URL instead of the text. However, there is no option to import a file or export the result to PDF, DOC, or any other popular format. The website is ad-free and contains lots of other useful tools for students. Then we have the **TextSummarization tool** that allows you to put the text into the field or give a link to a source where your article is posted. Then, set the number of sentences you want to have in your text. This summary maker analyzes your nonfiction text and extracts the exact number of sentences you're aiming at. The website is free, however, it contains ads, so make sure you turned on the ad-blocker. Also, the tool doesn't let its users import files or export the result to TXT, PDF, or Doc for other popular formats. The sixth tool available is **Text Compactor is a free summarizing** tool where you have to set the percentage of text to keep in summary. The website is ad-free and doesn't require registration. Its users can choose the output result within the range of 1-100%. If you are not satisfied with the result, change the percentage and try again. Although the tool is easy-to-use, it doesn't allow users to import files or URLs and save the result into the popular file types. [5] The seventh tool available is **Resoomer** which is another paraphrasing and summarizing tool that works with several languages. You're free to use the app in English, French, German, Italian, and Spanish. [6] Then there is the **Summarizer** which is another good way to summarize any article you read online. This simple Chrome extension will provide you with a summary within a couple of clicks. Install the add-on, open the article or select

the piece of text you want to summarize and click the button “Summarize”. [7] The **Simplifly** tool is one more way to summarize a scientific article for your research. The rules are the same – install the free Chrome extension, open the website, and get a summary. The software is ad-free, doesn’t require registration, and has no character limits. It works great with online articles and news websites; however, it doesn’t support PDF articles and scientific journals. You can’t summarize doc or any other file, vary or download the result. [8] Another well-known tool is the **Autosummarizer** which is a great tool for those in hurry. It has a minimum of functions and produces short summaries. Users can set up from 5 to 10 sentences of the output result. The tool is free and requires no registration. However, it doesn’t allow you to summarize files or web pages. Also, you should have ad-blocker to keep yourself away from ads on this website. [9] There is also **AappZaza** Article Summarizer that is mainly used for simple summary creator for your academic and professional needs. The software is free and doesn’t require registration. The app is simple and to reword a text, you need only type it or paste the article and click “Summarize Article” button. To get better results, try to summarize only well-structured documents. The software does not support import files or export the summary to any popular format. [10] And Lastly, we found the **Summary Generator** in our list which can be helpful for your experience in college or university. This is free open software everyone can use. The tool has only two buttons—one to summarize the document and the other to clear the field. With this software, you’ll get a brief summary based on your text. You don’t have to register there to get your document shortened. Speaking about drawbacks of the website, we would mention too many ads and no options to summarize a URL or document, set up the length of the result and export it to the popular file types. [11]

Chapter 4

PROPOSED RESEARCH

4.1 Problem and Investigation of that Problem

In the previous section, we discussed about SMMRY which allows to summarize by copy paste + URL but no download option. Hence, from areas suggested at the end of the previous section a particular problem that we propose to address is view Download List and Download one of those file again.

Do make the download work we will be using st.markdown() function by streamlit and set unsafe_allow_html to True which is a parameter of that function. Since for now there is no file download option for streamlit hence we will use a workaround made by @MarcSkovMadsen of awesomestreamlit.org. He utilized base64 and html anchor tags to make this work.

4.2 Difficulties may encounter

There will be some issues like-

1. Some given pdf by the user may be handwritten. Hence the summarizer may not understand few words and become confused during converting that into text. So this sector can be the next research sector for the experts so that it can be possible to make a text summarizer that can also understand this type of texts.

2. Some website like wikipedia have both Bangla and English mixed Paragraph. For example- "Language Movement Day (Bengali: ভাষা আন্দোলন দিবস Bhasha Andolôn Dibôs), also called State Language Day or Language Martyrs' Day (Bengali: শহীদ দিবস Shôhid Dibôs)" or ""Amar Sonar Bangla", also pronounced "Amar Shonar Bangla" (Bengali: আমার সোনার বাংলা, pronounced [amar ʃonar baŋla]". For this type of text, the summarizer will be confused, and unfortunately, won't be able to understand this Bangla words. So this sector can also be the next research sector for the experts so that it can be possible to make a text summarizer that can also understand this type of texts.

3. Converting a summary from text containing 500 words into 10 words might create a problem in our App. For example, if the user give a text containing more than 500 words and try to set the summarized text size to 5, this might not be possible for our App to summarize the text within 10 words. For this type of text, the user might not be get what he wants. Therefore, this sector can also be the next research sector for the experts so that it can be possible to make a text summarizer that can also understand this type of texts.

4.3 Significant features

There are a lot of text summarizing Apps or websites. Some of them find the text summary or entity checking facilities only, some of them give the facilities to find summary from a doc/pdf file, some of them find summary from the url, etc. Hence, the user need to find the summary from a text summarizer App, then copy that text and check the entity in another App. Moreover, to find other information like key words, sentiment, lemma, etc the user need to search another App.

None of the text summarized Apps or websites give the text summary from text/url/doc/pdf with summarized text limitation, entity/redactor checking, words information checking, comparing summaries, text type checking, etc facilities within a single Application.

However, our application will provide all these facilities along with other facilities within a single Application to reduce user's time. I believe this will be the significance of our proposed research.

4.4 Timetable

WEEKS	APPROXIMATE DATE	FEATURES
WEEK 1	07/03/2021	Project Proposal Report writing and software installation
WEEK 2	14/03/2021	UI Design and Text Summarize start
WEEK 3	21/03/2021	Text Summarize & Compare, Token, Words Info, Named Entities, POS, Sentiment, Meaning
WEEK 4	28/03/2021	Entity Checker & Redactor, Save File & Download List
WEEK 5	04/04/2021	Text Type
WEEK 6	11/04/2021	Check remaining works

4.5 Application design

There is a great need to reduce much of this text data to shorter, focused summaries that capture the whole design of our Application. Here, **figure 3** describes the user application part in short.

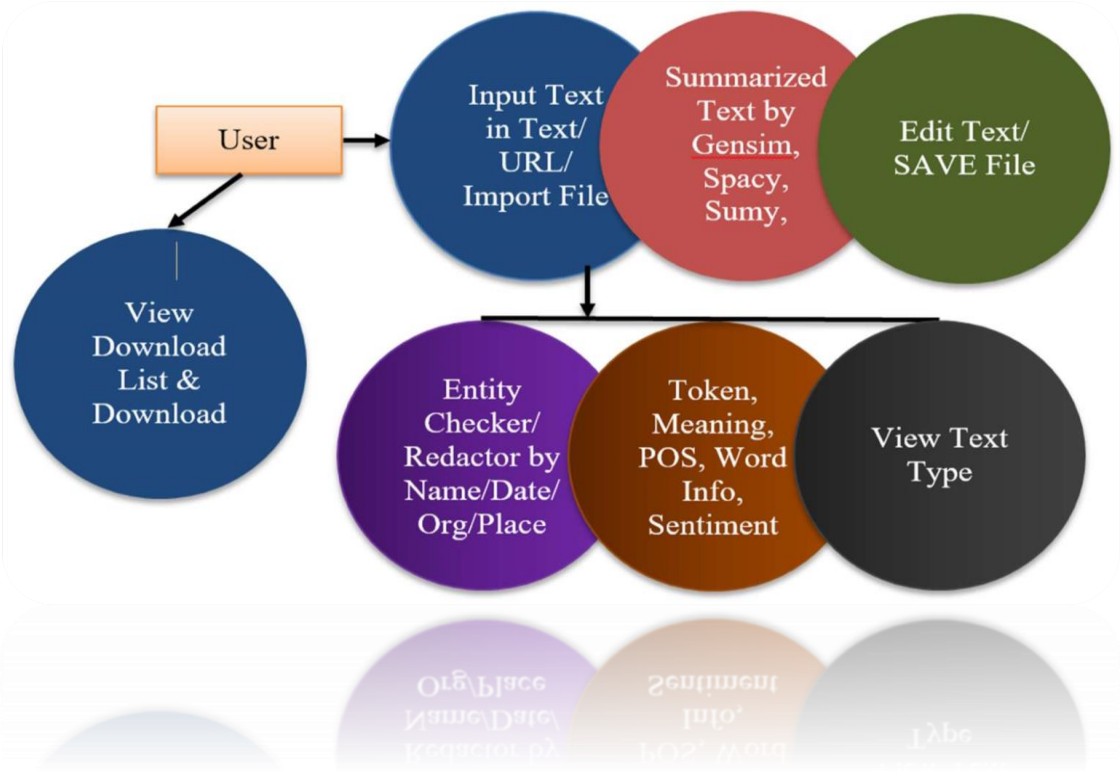

Fig 3: Workflow of the system

From the **figure 3** we can see, in our software, users can input into the summarizer tool in a three way- written text/copy paste text, importing file, inserting URL and get the summarized text by **Spacy**, **Gensim**, **Summy**, **NLTK** libraries as output. And this summarized text will be editable and can be saved and download for the future uses.

In addition, users can **check entities** by name, date, organization, place, etc. and our software will highlight those entity from the text. Moreover, users can **redact** name, place, date, organization from the input text by removing the searching information and replacing that as redacted information [e.g. redacted name]. The users can also edit and save the redacted text as a text file from our system.

Furthermore, there will be a **download list** where previously saved files will be stored as a list and from there the user can get the file and re-use or re-download the file.

Additionally, our software will contain some other features like users can view the key words as token and get those **word's information** (e.g. lemma, shape, alpha, stopword), meaning, sentiment (e.g. polarity, subjectivity), part of speech (e.g. tag, pos, dependency) etc.

Our software will contain a feature to show the **text type**. The users will give the text and can view what is the text about [e.g. private/business text].

For further understanding an **USE CASE DIAGRAM** is given in **figure 4** where the user's and admin's activities have been shown in details. For instance, users can Find Summary in 4 different formats e.g. **Spacy**, **Gensim**, **Summy**, **NLTK**, Compare among those 4 libraries, Check Entity by name, date, place, organization, Get Redacted Text by redacting any user required name, place, date, organization, View Words Information e.g. token, sentiment, meaning, Check Text Type e.g. privet/business text, Give Feedback, Save redacted file and view Download list facilities. However, admin can view user's feedback and update software according to user requirements.

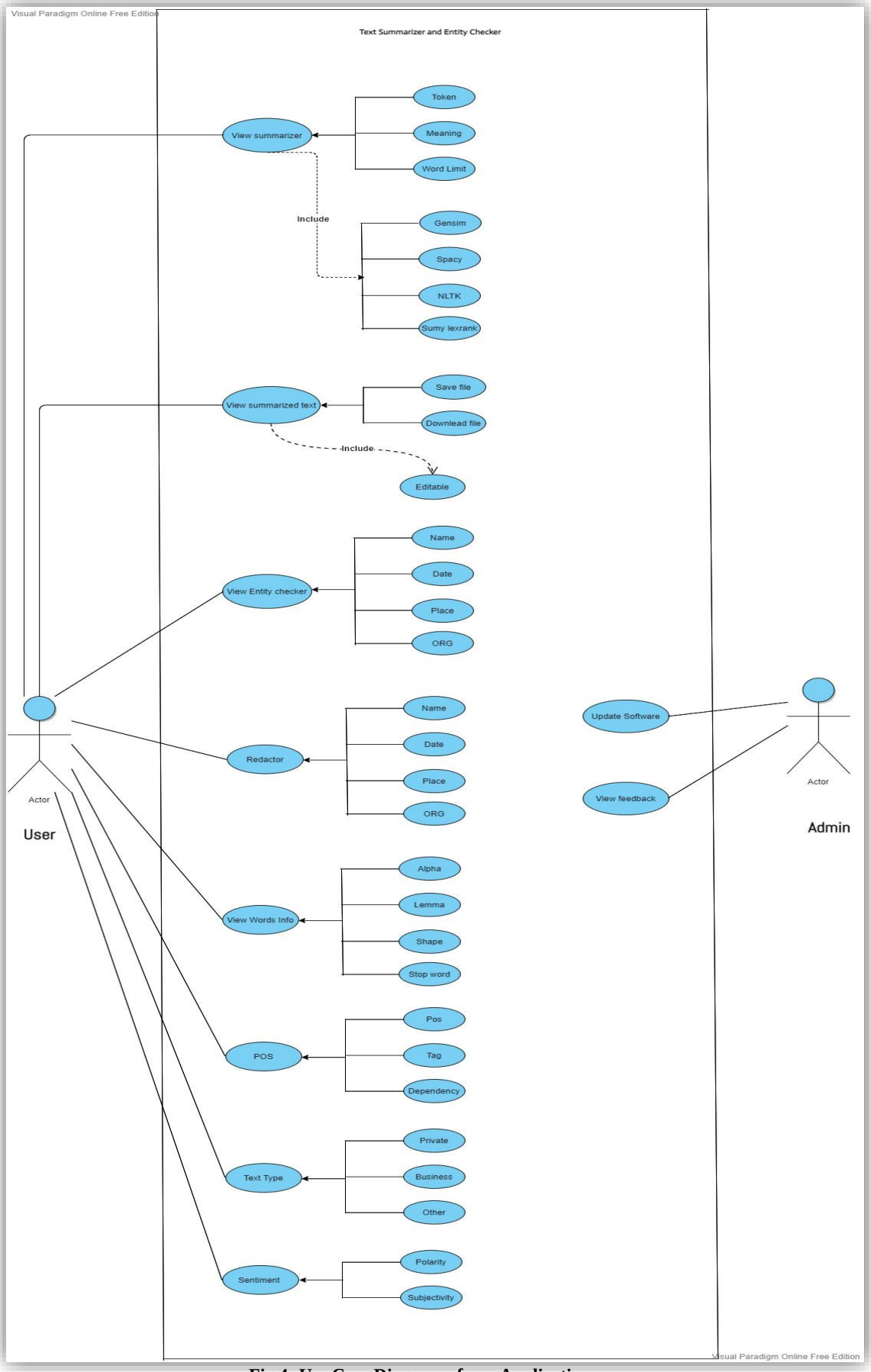

Fig 4: Use Case Diagram of our Application

Chapter 5

CONCLUSION

Text summarization is growing as sub – branch of NLP as the demand for compressive, meaningful, abstract of topic due to large amount of information available on net. Text summarization has its importance in both commercial as well as research community. As abstractive summarization requires more learning and reasoning, it is bit complex then extractive approach but, abstractive summarization provides more meaningful and appropriate summary compare to extractive. Through the study it is also observed that although there are so many text summarizers but students mainly focus on using specific tools like, URL facility, import and export of files and the grammar check capability.

Our goal is to create an advance Text summarizer that will have majority of the important tools that we found in our research and also have an entity checker. So that, this text summarizer can be used by people from different profession with all the useful tools to ease their life.

Chapter 6

REFERENCE

References

[1] "SummerizeBot," 2019. [Online]. Available: https://www.summarizebot.com/.[Accessed 3 March 2021].
[2] "Smmry," 2021. [Online]. Available: https://smmry.com/. [Accessed 3 March 2021].
[3] "tools4noobs," tools4noobs, 2007-2020.[Online]. Available:https://www.tools4noobs.com/summarize/.[Accessed 3 March 2021].
[4] 2001. [Online]. Available: https://www.splitbrain.org/services/ots. [Accessed 4 March 2021].
[5] "Textcompactor," Knowledge by Design Inc., 2010-2016. [Online]. Available: https://www.textcompactor.com/. [Accessed 4 March 2021].
[6] "Resoomer," 2021. [Online]. Available: https://resoomer.com/en/. [Accessed 4 March 2021].
[7] "Summarizer," 26 July 2013.[Online].Available:https://chrome.google.com/webstore/detail/summarizer/nehmaijilccnklmhkjlmmcnbjobdodhp?hl=en. [Accessed 28 February 2021].
[8] "Simplifly," 5 January 2018.[Online].Available:https://chrome.google.com/webstore/detail/simplifly/jioeflccmbeaeelfklklmbdjcbdeccnh?hl =en.[Accessed 28 February 2021].
[9] "Autosummarizer," 2013. [Online]. Available: https://autosummarizer.com/. [Accessed 6 March 2021].
[10]"AappZaza Article Summarizer," 2013. [Online]. Available: https://appzaza.com/articlesummarizer/. [Accessed 6 March 2021].
[11]"Summarygenerator," unknown. [Online]. Available: https://summarygenerator.com/. [Accessed 6 March 2021].
[10] S. N.Moratanch*, "A Survey on Extractive Text Summarization," 2017. [Accessed 6 March 2021].
[12] D. K. G. a. C. N. Mahender2, "IJARCCE," vol. 5, no. 3, p. 2, March 2016. [Accessed 6 March 2021].
[13]"Smmry," 2021. [Online]. Available: https://smmry.com/. [Accessed 6 March 2021].