International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 03 | Mar -2020 ISSN: 2582-3930

News Article Text Summarization Using Lex Rank and LSA Algorithm

Student: Prerana Sugdare
Dept. of Information Technology
Pillai College of Engineering
New Panvel, India

Student: Tanmay Gurao
Dept. of Information Technology
Pillai College of Engineering
New Panvel, India

Student: Sanket Thale
Dept. of Information Technology
Pillai College of Engineering
New Panvel, India

Faculty: Gayatri Hegde
Dept. of Information Technology
Pillai College of Engineering
New Panvel, India

Abstract: **Text summarization is research field which helps to find out detailed but short information from documents which are often large in size, the documents can be from different fields such as finance, news , media, academics, politics, etc. Automatic Text summarization helps people to get more comprehensive information about the document. In other words, the process from which condensed form of document is created which tries to maintain information without losing or reducing the general meaning of the source document. The main goal is often to maintain the remarkable information. Automatic Text summarization is an important mean by which large information can be concluded into shorter text in less amount of time and minimal effort. Thus, making it an important field of active research. Approaches of Text summarization are classified into two categories: Extractive and Abstractive. Extractive summarization techniques produce summaries by using words that are present in the document itself. Abstractive model takes a lot of time for training the machine learning model, it involves deep neural network to train the model and requires large amount of corpus. Collecting such a large amount of corpus (i.e. around 400 to 500 gb minimum) and training time (i.e. about 1200 hours minimum) is hard and tedious. This paper focuses on extractive model like Lex Rank and LSA. Using Lex Rank and LSA the big news articles are summarized in order to generate a shorter news article, which is enough for reader to make sense and to get complete idea.**

Keywords - Extractive, Lex Rank, LSA (Latent Semantic Analysis), Stop words, Summarization.

I. INTRODUCTION

Text summarization plays a vital role day-to-day life. The continuing growth of content on world wide web and online text articles collections makes a large volume of information available to end users. The massive information either leads to wastage of significant time in browsing information or else useful information may be missed out. The text summarization technology is maturing and may provide a solution for the information overload problem. Text summarization involves the process which can automatically generate a compressed version which is a small paragraph of a given text that is useful information to users. Text summarization is a complicated task which ideally would involve deep natural language processing (NLP) capacities. In order to simplify issue, the method Extractive is going to use Lex Rank and LSA which are types of extractive algorithms. By implementing text summarization, it saves time to search or to get to conclusion of the article. Big companies like Google, Amazon uses text summarization for providing better relevance result to the user. For example, Google Assistant, Amazon Alexa. Extractive Summarization produces condense form to the original documents which helps in retrieval of necessary information from the huge volumes of text documents and identification of context of document. Relaying on summary, the user was provided with a facility to find the most desirable documents. Because of multiple irrelevant documents to a query there is a necessity of document summarization which summarizes the documents based on analysis of text and ranking them according to their values and generates new sentences where these sentences may not directly contain the actual keyword but conceptually related to the word that is used in search.

II. LITERATURE SURVEY

[1] M. Haque, et al., ”Literature Review of Automatic Multiple Docu- ments Text Summarization”, International Journal of Innovation and Applied Studies, vol. 3, pp. 121-129, 2013.

This paper is consist of the basics of multi-document summarization, then several approaches for extractive summarization and extractive methods. Extractive summarization provides the information according to the users input that describes the original document in a small but to the point sentences that may not be in order. This paper contains the comparison of various extractive methods that are used for the summarization. Many extractive methods have evolved but it is difficult to mention which method creates the more concise summary with the high performance.

[2] N. Moratanch, S. Chitrakala “A Survey on Extractive Text Summarization” IEEE International Conference on Computer, Communication, and Signal Processing (ICCCSP-2017)

This paper has shown assorted mechanism of extractive text summarization process. The implication of sentences is

determined based on linguistic and statistical features. In this paper,a comprehensive review of extractive text summarization process methods has been ascertained. In this paper, the various techniques, populous benchmarking datasets and challenges of extractive summarization have been reviewed.

[3] PavanKartheekRachabathuni, "A Survey on Abstractive Summarization Techinques". Department Of Computer Science and Engineering, IEEE Internation Conference. Part Number: CFP17L34-ART, ISBN:978-1-5384031-9.

This paper solves climacteric problems in furnishing information to the necessities of user. This makes user impractical to read entire documents and select the desirables. To this problem summarization is a novel approach which surrogates the original document by not deviating from the theme helps the user to find documents easily.

[4] Haroran Li, Junnan Zhu, Cong Ma, Jiajun Zhang and ChengqingZong, "Read, Watch, Listen and summarize: Multi-model Summarization for Asynchronus Text, Image, Audio and Video", IEEE Transaction On Knowledge and data engineering, Vol. X, No. Y, Month year.

In this paper, authors propose an approach to a generate textual summary from a set of asynchronous documents, images, audios and videos on the same topic. They formulate the MMS task as an optimization problem with a budgeted maximization of submodular functions. They investigate various approaches to identify the relevance between the image and texts, and find that the image match model performs best.

III. EXISTING SYSTEM

Existing system of text summarizer uses no logical approach. It includes following steps:

1) Download the contents/article to be extracted.
2) Extract the article from the html.
3) Figure out the 3 or 5 most important sentences from the article.

A. Algorithm of existing system

1) Download the Article from URL.
2) Get rid of html tags and everything else other than the article (use beautiful soup).
3) Split the article into words. (Use NLTK function like word-tokenize and sent-tokenize).
4) Eliminate the stop words. (Is, this, the, a)
5) Find how often each remaining word is repeated. (Frequency of particular word in the article)
6) The more common the word appears, the more important it is. So, for each sentence, find a score of how important the words in the sentence are.
7) Rank the sentence by that score. (Select top 3-5 sentences)

B. Pros of existing system

They are quite simple since they don't make changes in the document. They just try arranging them in the form of highest priority. They use existing natural language phrases which are the input of the model.

C. Cons of existing system

They miss flexibility since there is no use of grammar, figure of speech and they also lack use of novel words or connectors. It's impossible for them to explain or summarize like people do.

IV. PROPOSED SYSTEM.

Here, the news articles from the news website's such as Inshorts, NewsHunt, etc. Then we are going to scrape the website to get the new article. What this tool does is that it extracts data from website which is HTML file and save's it into database. The database also known as corpus. For scrapping the article from different websites, we will be using python and it's libraries like urllib, beautiful soup, etc. The python script will be deployed on Visual Studio Code in order to get real time articles from the website and the extracted output will be saved in database. With the help of Visual Studio Code we will host our flask application. Flask is python-based web framework. Extractive Summarization produces condense form to the original documents which helps in retrieval of necessary information from the huge volumes of text documents and identification of context of document. Relaying on summary, the user was provided with a facility to find the most desirable documents.

Steps:

1. Train the model using Visual Studio Code.
2. Create a flask environment that will have an API endpoint which would encapsulate our trained model and enable it to receive inputs (features) through GET requests over HTTP/HTTPS and then return the output after de-serializing the earlier serialized model.
4. Upload the flask script along with the trained model on Visual Studio Code.
5. Make requests to the hosted flask script through a website or capable of sending HTTP/HTTPS requests.

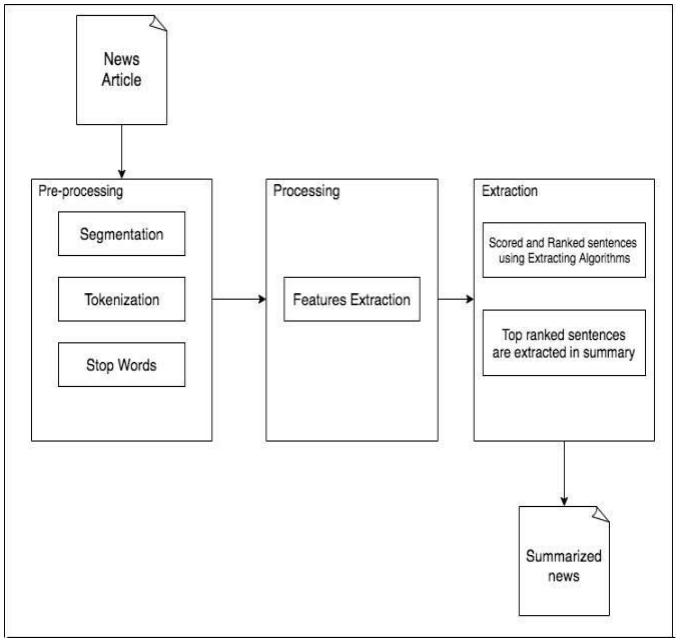

Fig. 1. Proposed System

1. *LexRank* is an unsupervised graph-based approach. IDF-modified used by Lex Rank Cosine as the similarity measure between two sentences. Lex Rank uses a technique which makes sure if all the sentences with high priority are not similar to each other. The problem of extracting a sentence that represents the contents of a given document or a collection of documents is known as extractive summarization problem. In extractive summarization problem, we want to extract one representative sentence that capture as broad as possible the content of the corpus, whether it is one document (single document summarization) or several documents (multi-document summarization). The new method, named Lex Rank, is identified by PageRank method. This method works by generating a graph, every sentence represents one node, and the edges are similarity relationship between sentences in the corpus. In this research, they measure similarity between sentences by considering every sentence as bag-of-words model. Frequency contributes to the similarity strength as the number of word occurrences is higher. This is then used as a measurement for similarity between sentences. Basically, calculating the 'distance' between two sentences x and y. More the similar two sentences, more the 'closer' they are to each other. To extract the most important sentences, from the resulting similarity matrix we apply a thresholding mechanism. The result is a subset of the similarity graph, from where we can pick one node that has the highest number of degrees.

2.*Latent semantic analysis* is an unsupervised summarization method which along with finding the frequency of important terms it decomposes to find the singular value for better and efficient summarization. LSA works by projecting the data into a lower dimensional space without any heavy loss of information. Latent semantic analysis uses spatial decomposition. In spatial decomposition the singular vectors of the words which recurring in the corpus. The higher the magnitude of the singular value the higher is the importance of the word in the document. Latent Semantic Analysis (LSA), also known as Latent Semantic Indexing (LSI) literally analyse the documents to find underlying meaning or concepts of those documents. If each word only meant one concept, and each concept was only described by one word, then LSA would be easy since there is a simple mapping from words to concepts.

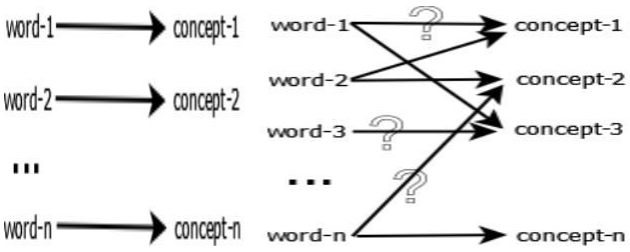

Fig. 2. LSA Similarity Graph

Unfortunately, English has different words and ways to represent the sentences which have similar meanings. There are many words with same meaning(synonyms), words with multiple meanings, and all the other type of ambiguities and redundancy in meaning which have hard time understanding.

V. RESULTS

Here, we used two extractive algorithm Lex Rank and LSA. The output of both algorithm was too good then the existing system. To evaluate the text summarization quality, we used ROGUE-N metric and BLEU metric. Rouge-N is a word N- gram that measures how much efficiency between the model and the summarized output. It finds the ratio of the no of counts of phrases which occur in both model and summary known as N-gram. BLEU metric is a modified form of precision, extensively used in machine translation evaluation. Precision is the ratio of the number of words that co-occur in both gold and model translation/summary to the number of words in the model summary. Unlike ROUGE, BLEU directly accounts for variable length phrases – unigrams, bigrams, trigrams etc., by taking a weighted average.

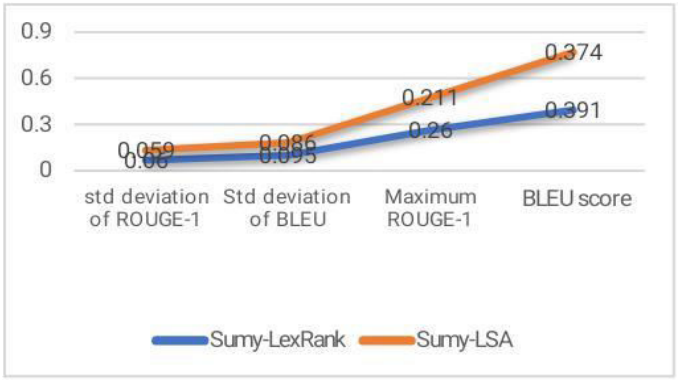

Fig. 3. Algorithm Performance Graph

Our graph tells us that Lex Rank outperforms LSA. A good practice would be to run both the algorithms and use the one which gives more satisfactory summaries.

VI. PROJECT INPUT AND OUTPUT AND SCREENSHOT

Home UI:

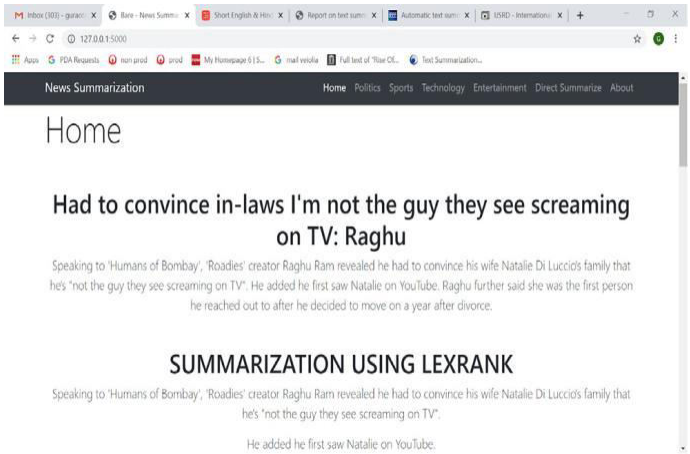

Fig. 4.1. Home UI

Fig. 4.1 depicts the front end of application. The front end consist of 7 tabs namely Home, Politics, Sports, Technology, Entertainment, Direct Summarize and About.

Entertainment UI:

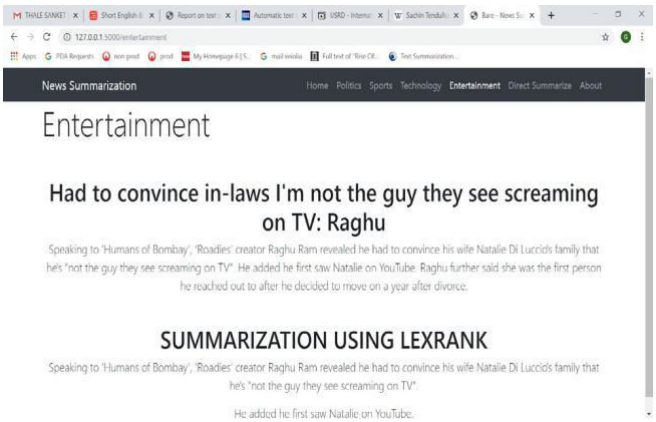

Fig. 4.2. Entertainment UI

When this web service is activated by the user, the Home UI appears first which consist of summarization activity of news articles. When the user opens Entertainment tab, then that user gets summarized information about articles which depends upon Entertainment stuff shown in Fig. 4.2

Technology UI:

When this web service is activated by the user, the Home UI appears first which consist of summarization activity of news articles. When the user opens Technology tab, then that user gets summarized information about articles which depends upon Technologies shown in Fig. 4.3.

Fig. 4.3. Technology UI

Sports UI:

When the user opens Technology tab, then that user gets summarized information about articles which depends upon Sports news shown in Fig. 4.4.

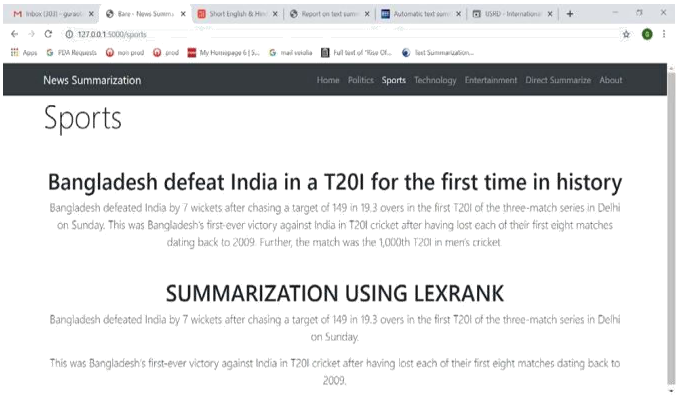

Fig. 4.4. Sports UI

Politics UI:

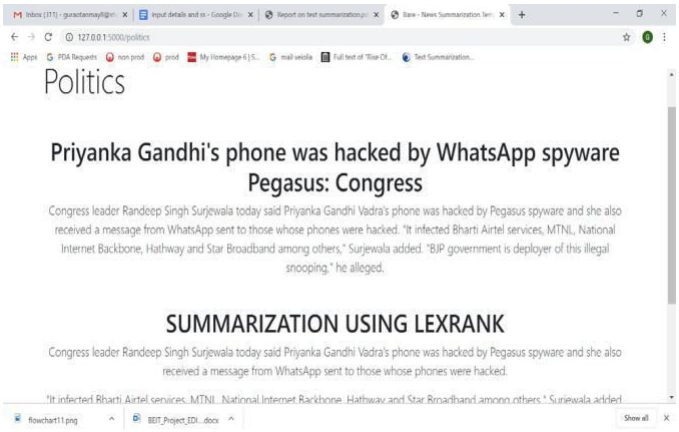

Fig. 4.5. Politics UI

When the user opens Politics tab, then that user gets summarized information about articles which depends upon Politics news shown in Fig. 4.5

VII. CONCLUSION AND FUTURE SCOPE

We are able to summarize news articles by analyzing the content of the news. In this process we analyze that each and every website uses different pattern to display their news so it is difficult to scrap data from various sites. Hence now we are able to scrap data from inshort news website which allow us to scrap article and body from same page. After we used two extractive algorithms i.e. Lex Rank and LSA because the results was too good then existing system. We intend to create a summarization model which can create a summary of news articles which are generated every day. This will help the users to get whole idea of the news without reading the entire news. In future we are going to use news API which allows us to get news, headlines, articles from over 30,000 news sources which allows us to import various type of news in our website and moreover we are going to build mobile application for both Android and IOS platform which becomes handy to people to read news anywhere anytime. And if we get access to enough resources then we can switch to abstractive method which can allows news to summarize in more contextual form making the summary more precise and correct. We may also build a complete automatic process

pipeline for fetching news, scraping the news and then summarizing and displaying it.

VIII. ACKNOWLEDGMENT

A project is as golden opportunity for learning and self-development. We consider ourselves very lucky and honoured to have so many wonderful people lead us through in completion of this project. We would like to thank Mrs. Gayatri Hegde for the opportunity and help which was provided by her. Our grateful thanks to Dr. Satishkumar L. Verma (Head of Department of Information Engineering) who in spite of extraordinary busy with his duties, took time out to hear, guide and keep us on the correct path. We choose this moment to acknowledge his contribution gratefully. We also indebted to Dr. Satishkumar L. Verma, for extending the help to academic literature. We express our gratitude to Dr. Sandeep M. Joshi (Principal) for their constant encouragement, Co-operation and support.

IX. REFERENCES

[1] M. Haque, et al.," Literature Review of Automatic Multiple Documents Text Summarization", International Journal of Innovation and Applied Studies, vol. 3, pp. 121-129, 2013.

[2] N.Moratanch ,S.Chitrakala "A Survey on Extractive Text Summarization" IEEE International Conference on Computer, Communication, and Signal Processing (ICCCSP-2017)

[3] Gune ¨ ¸s Erkan, Dragomir R. Radev, "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization".Department of EECS University of Michigan, Ann Arbor, MI 48109 USA Journal of Artificial Intelligence Research 22 (2004) 457-479.

[4] Haroran Li, Junnan Zhu, Cong Ma, Jiajun Zhang and Chengqing Zong, "Read, Watch, Listen and summarize: Multi-model Summarization for Asynchronous Text, Image, Audio and Video", IEEE Transaction on Knowledge and data engineering, Vol. X, No. Y, Month year.

[5] Prakhar Sethi, Sameer Sonawane, Saumitra Khanwalker, R.B. Keskar "Automatic Text Summarization of News Articles" 2017 International Conference on Big Data, IoT and Data Science.

[6] Luciano Cabral, Rinaldo Lima, Rafel Lins".2015 Fourteenth Mexican International Conference on Artificial Intelligence, "Automatic Summarization of News Articles for Mobile Devices" 978-1-5090-0323-5/15 2015 IEEE.