A Text Mining using Web Scraping for Meaningful Insights

Kishor Kumar Reddy C¹, Anisha P R², Nhu Gia Nguyen³ and Sreelatha G⁴

¹,²Post Doctoral Researcher, Duy Tan University, Vietnam
³Dean Graduate School, Duy Tan University, Vietnam
⁴Research Scholar, Lincoln University College, Malaysia
E-mail: Kishoar23@gmail.com

Abstract. This research involves the usage of Machine Learning technology and Natural Language Processing (NLP) along with the Natural Language Tool-Kit (NLTK). This helps develop a logical Text Summarization tool, which uses the Extractive approach to generate an accurate and a fluent summary. The aim of this tool is to efficiently extract a concise and a coherent version, having only the main needed outline points from the long text or the input document avoiding any type of repetitions of the same text or information that has already been mentioned earlier in the text. The text to be summarized can be inherited from the web using the process of web scraping or entering the textual data manually on the platform i.e., the tool. The summarization process can be quite beneficial for the users as these long texts, needs to be shortened to help them to refer to the input quickly and understand points that might be out of their scope to understand.

Keywords. Natural Language Processing, Text Summarization, Web Scraping

1. Introduction

With Technology nowadays is not just a tool; it is far more than what we think and what we believe in. Individuals in today’s generation are mainly looking forward to using things or any technology that might simplify their work to a large extent by saving their time as well as cost. It is also the application of scientific as well as analytical knowledge for the practical purposes used in the industry today. It is mainly an integration of the techniques, skills, processes, methodologies that can be used in the production industry that is required to carry out certain operations. Systems using technologies or applying it are taking in input from the users, modifying it according to the requirements and retaining an output that might satisfy the user. These systems are called technological systems or just technology systems [2].

The technology of Machine Learning (ML) basically deals with the study of the algorithms used by the computer or any machine that helps improve it automatically through the usage of data as well as experience. [28] It mainly helps in building a model using any sample data that might also be called as the ‘training data’, to help the machine make predictions and take decisions on its own without being explicitly programmed. The technology of Natural Language Processing (NLP) is a field in Artificial Intelligence (AI) which is related to interactions using the natural language carried out between the machines and humans [19]. The primary objective of this technology is for deciphering, reading the contents as well as understand the linguistics of human language to improve productivity and carry out certain tasks that require understanding and precision in human language [25]. The Natural Language Tool-kit is a package of python programs that comprises of a collection of different libraries and various other codes used in symbolic as well as statistical processing of natural language.

Data always need not be in the structured format, sometimes this data to be used is unstructured and needs to be pre-processed which involves a few other processes like data cleaning, visualization, integration etc., which can be carried out by the NLTK (Natural Language Tool-Kit).
Natural Language Toolkit is used as a major platform for structuring the programs using Python language that work on the natural language data i.e., human language. This tool-kit provides interfaces for the users and along with those also provides libraries for classification, parsing, semantic reasoning, tokenization etc. [22]

This era of Big Data, has experienced a huge outburst in the amounts of textual data inherited through a wide number of resources. Thus, this inestimable source of information needs to be utilized in a systematic way for providing an ease to the users who are involved with these huge amounts of data [26]. This process of condensing long texts into more meaningful summaries is quite challenging for the machines, as they lack the human expertise of understanding and analysing. Thus, these machines need to be trained in this context to develop the language capability and the sense of language understanding. This can be attained by the Machine Learning (ML) and the Natural Language Processing (NLP) tools, shown in figures 1, 2 and 3 respectively [23-24].

![How does Machine Learning Work?](image)

**Figure 1.** Working of Machine Learning

![Applications of NLP](image)

**Figure 2.** Applications of NLP

![Using NLTK to pre-process data.](image)

**Figure 3.** Using NLTK to pre-process data.

While referring to the documents, related to a specific topic on a given paper, one has to go through each and every topic in detail but some data might be out of their scope to understand. This creates difficulty for the users to create summaries on those topics. An Automatic Text Summarizer is an effective tool in such cases. A text summarizer tool effectively creates an abstract format of the long text given as the input data. Sometimes it is quite hectic for the users to copy the entire data from the website to the summarisation tool. In such cases, we use the process of Web scraping which extracts the content from the webpage mentioned and then copies it onto the summarizer for summarization purposes. The scraper tool can replicate the entire website data and can put it elsewhere [29-33]. The process of retrieving valuable knowledge and information data from the text is termed as Text Mining. Using text mining along with web scraping, one can reduce the bodies of text into a shorter and a more fluent version of the input but, restoring the original meaning or giving the great insight to the original text.
The main objectives of the study are:

i. Create a training model which will take the long length texts or documents as input data and use the Machine Learning and NLP technology to train itself.

ii. Generate a short summary i.e., an abstract version of the given file to reduce reading time of the users.

iii. To reduce user’s time by providing a user-friendly User Interface (UI) that does not need any in depth understanding of how to use the platform or the tool.

iv. To reduce the dependency on the individuals who manually generate the summaries as it can be a quite lengthy and a time consuming, hectic job.

v. Highlights then key information from the voluminous data, generating a precise summary without losing any key information.[21]

vi. Helps to understand and read the text more easily.

vii. Changes the structure of the original voluminous text.

viii. Summarization tool improves the effectiveness of indexing.

2. Relevant Work

Julius Steen and Katja Markert presented a paper “How to Evaluate a Summarizer: Study Design and Statistical Analysis for Manual Linguistic Quality Evaluation (2021)”; reveals how manual assessment is important to be judged for the progress on the summarization process. This paper conducts two experiments on the two characteristics of summaries based on the quality of linguistics i.e., coherent and repetitiveness in order to compare the Like rt type and the annotations of ranking of sentence and words to show which might be the best method to pick for the evaluation purpose as it can vary from one criterion to another, regarding the choice of evaluation [11].

Ujjwal Rani and Karambir Bidhan presented a paper “Review Paper on Automatic Text Summarization (2020)”; reveals the approaches to text summarization. The abstractive summarization requires characteristic language preparing devices for indicating the records to create new sentences. The Extractive Text Summarizer requires factual, etymology and heuristic strategies in order to position the sentences as they use the same sentences to create an effective summary. It elaborates on the various techniques used in order to build effective summaries. Apart from the methodologies, this survey depicts the procedures utilized and the exhibition accomplished. Content Summarization has its own significance in both businesses, just as research and enhancing network. Abstractive
summarization needs more focus, as it advances and produces summaries by learning the data and creating new decrees for a well understood concise summary. It provides a better and a progressive summary as compared to the one generated through the Extractive methodology. [10]

Nouf Ibrahim Altmami and Mohamed El Bachir Menai in this paper “Automatic Summarization of scientific articles: A survey (2020)”; a scientific research process survey which reveals the ways of using abstract generation summarization and citation-based summarization and its methodologies utilised. [4]. Shai Erera, Michal Shmueli-Scheuer, Gey Feigenblat, Konopnicki presented a paper “A Summarization System for Scientific Documents (2019)”; presenting an original system building summaries for the various CS publications. User study depicts the most featured and important scenarios which are identified for learning, investigation and comprehension of the scientifically documents [13]

N. Moratanch and S. Chitrakalapresented the paper “A Survey on Extractive Text Summarization (2017)”; where they described the approach to get the concise and a fluent form of the dataset using the Extractive Text Summarization techniques [2]. Mehdi Allahyari, SaeidSafaei, SeyyedaminPouriyeh, KrysKochut, Mehdi Assefi, Juan B. Gutierrez, and Elizabeth D. Trippe presented a paper “Text Summarization Techniques: A Brief Survey (2017)”; describes how text summarisation can be quite a challenging task as machines lack the knowledge of understanding and language capability. This can now be achieved through Machine Learning by training the machines to understand from the inputs provided [5].

Deepali K. Gaikwad and C. Namrata Mahender presented a paper “A Review Paper on Text Summarization (2016)”; It describes the two major approaches i.e., Abstractive and Extractive Summarization. The textual summarizers find the key sentences from the input text data and concatenate them into a final summary that is a concise one [1]. Hans Christian, Mikhael Pramodana Agus, Derwin Suhartono presented a paper “Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF) (2016)”; There has been an outburst in the social media data in the recent years. This information still spreads all across the internet in the form of data in text. Generating a summary of the long texts can minimize the effort and time and extract a simplified version of a document which only showcases the relevant information needed. Natural Language Processing (NLP) is a field of computer linguistics which helps to extract full meaning from the text [12].

Jingqiang Chen and Hai Zhuge presented a paper “Summarizing Related Work through Citations (2016)”; This paper uses the citations to perform various comparisons to give rise to the related sections of work [8]. Maake Benard Magara, Sunday O. Ojo, Transo Zuva presented a paper “A Comparative Analysis of Text Similarity Measures and Algorithms in Research Paper Recommender Systems”; It is being quite a difficult task for the researchers and the online users to get relevant, appropriate and accurate information due to the outbursts on the web. Research on this article depicts the non-linear classification algorithms that use the text similarity metrics for summarization. These articles utilised Recursive Partitioning (rpart), RF and Boosting algorithms where the rpart algorithm performed comparatively better in contrast to the Random Forest and Boosting algorithms by achieving an average accuracy and a time efficiency of 80.73 and 2.354628 seconds respectively [14]

Marzieh Oghbaie and Morteza Mohammadi Zanjireh presented a paper “Pairwise document similarity measure based on present term set (2018)”; Many of the similarity measures judge the commonness between two documents based on their phrase weights as well as the information content that these documents share, but this measurement can be quite difficult if there are multiple documents with the same degree of commonness to a particular document [15].

Ranjeet Kumar and R.C. Tripathi presented a paper “Text mining and similarity search using extended tri-gram algorithm in the reference based local repository dataset (2015)”; This paper puts forward the methodologies for inspecting plagiarism out of the used references. The methodology used is the TRI-GRAM algorithm on the cited document using the citation-based approach. This also
describes the categories of the citations, Under Used citations, Usually Used and Over Drawn citations [6].

Said A. Salloum, Mostafa Al-Emran and Khaled Shaalan presented a paper “Using Text Mining Techniques for Extracting Information from Research Articles (2018)”; This paper includes the study that states that text mining has become one of the most trending fields that can be incorporated in various research fields like computational linguistics, Information Retrieval (IR) and data mining. Sometimes the search results are irrelevant according to the requirements, thus the key is to use text mining in order to find the appropriate information. The process of text mining includes the gathering of the documents through diverse resources [3].

G Ravi Kumar, S Rahamat Basha, Surya Bhupal Rao presented a paper “A Summarization on Text Mining Techniques for Information Extracting from Applications and Issues (2020)”; Growths in the field of web, digital libraries, technical documentation, medical data have made it easier for accessing large number of textual documents, coupled together to create useful resources, thus allowing text extraction or finding information from the databases, a challenging task to meet the standards of the natural language depths used by most of the documents nowadays. [9] [20]

Stephan Camilleri, Matthew R Agius, Joel Azzopardi presented a paper “Analysis of Online News Coverage on Earthquakes through Text Mining (2020)”; Many of the researchers have tried to find the factors determining the levels of news agencies that cover data about certain events. This study does not specify that the data was gathered automatically, clustered and processed for extraction of information in the real time and moreover the studies for this paper has been strictly focussed on the specific earthquakes and how the news agencies have been reporting these events. The objective of the algorithm used is to generate a dataset of the mined news articles mapped to the actual earthquakes that have occurred [16].

Erdinc Uzun presented the paper "A novel web scraping approach using additional information found on web pages (2016)"; Web pages contain unnecessary items such as menus, ads, banners, footer, site maps and other required content such as title, summary, keyword, price and description according to the information required by users. HTML parsers used for web scraping are based on the DOM tree structure. The most widely studied problem in web scraping is to determine the automatic rendering of patterns and these patterns are hand-made by developers / experts in the use of web filtering tools that are basically HTML filters [17-18]. Shreesha M, Srikara S B, Manjesh R presented the paper “A Novel Approach for News Extraction using Web Scrapping (2018)”; Smart parser is a useful program that provides the latest and latest news from top news sites. The program also has issues based on user location on these sites. This also provides a summary of cricket points for live reviews. [7].

3. Proposed System
The application being created provides the user with an accurate summary along with an option of retrieving the content from the website itself and summarizing it. There are various modules need to be focussed upon to create an accurate Automatic Text Summarization system.

3.1 Importing text onto the tool
The text to be summarized has to be in the form of sentences as needs to be copied onto the summarizer tool. This text is summarized into an abstract format and all the stop words such as articles (a, an, the) are removed from it. Along with these, punctuations, semicolons, commas, full stops, inverted commas, back slashes, forward slashes, exclamation marks and all the unnecessary symbols are removed from the text. This text is then converted into single sentences using the splitting operation. These sentences are considered as individual texts and vectorization is carried out on these sentences. While scraping or extracting any text from websites we just input a URL so that the text is automatically scraped in the form of text and then splitting is done on this text along with removing the symbols for individual sentences.
3.2 Vectorization

Word embedding is basically a form of word presentation that gives a mathematical meaning of words that have the same meaning. This process helps in capturing the natural meaning of the different words in the nearest reach to any other word representations vectors. After finding the vector representation of the words, then we consider the measurement / scale of the carriers to get the combined vector of the whole sentence. By finding the vector representation (embedding) of each sentence, the sentences in the text are represented by the vector shape of the X-Y-Z plane and the similarity between these vectors is determined by the metaphor of cosine similarity.
3.3 Finding similarity between two vectors

Cosine parallelism is a measure of the similarity between two vectors in the internal product space. The cosine of an angle is usually measured between two vectors and then it determines whether the two vectors point in approximately the same direction or not. Cosine Parallels are used to measure how similar texts are regardless of their size. Statistically, it measures the cosine of an angle between two vectors displayed in a space wide. The similarity of the cosine is advantageous because even though two identical texts / vectors are far apart in the Euclidean (due to the size of the document), they are more likely to be closely related. The angle is smaller; it will resemble a cosine similar. The most common method of matching similar texts / sentences is based on calculating the maximum number of common words between the sentences in this text. But this approach has a natural error as the size of the document grows; the number of common words often increases, even though the texts refer to different topics. The similarity of cosine helps to overcome this fundamental error in the 'normal-word-words' method or the Euclidean distance method.

3.4 Storing the similarities in a matrix

These cosine similarities are then stored in a matrix of rows and columns where for each element (i , j) in the similarity matrix , each of the row and column represents the commonness between the i\textsuperscript{th} row sentence and j\textsuperscript{th} column sentence forming the Similarity Matrix.
This matrix of similarity with sentences is then transformed into a graphical representation to better visualize the scenario. The nodal elements of the graph represent the phrases or sentences and the edges or links represents the similarities. Each sentence is given a rank according to the similarity which is termed as Sentence Ranking. These sentences with the largest amount of similarity are picked for the final output abstract which is shorter in length and has a lesser reading time than the prior document given as the input.

4 Results & Discussion

The Home page of the Summarization Tool displays a slider that shows the different libraries that are used to transform the long texts into an abstracted one in the form a summary. The various libraries used are SpaCy, Gensim, Sumy and the NLTK. The ‘Refresh’ button provided clears all the text from the text boxes, if needed to change and insert new text or remove it for later use. The Title Page that slides after the home page displays title that is., the function of the summarizer that performs the Text Mining along with Web Scraping. It also shows what domain is the application based on. The ‘Learn More’ button provided redirects us to the about page which gives a description about the application. We are provided with two textboxes where one was for the text to copy onto for text summarization using Text Mining Approach. The other allows us to insert a URL that may be a link to a website or any link address, which extracts the text from that web page and copies it to out output screen using Web Scraping. The ‘Clear’ and ‘Summarize’ buttons are if clear, the text from the text boxes and summarizes the text respectively.

The Output Screen displays the text that was copied by the user or extracted from the web page. The Summarised text block displays the final abstract of the long text. This application also displays the Reading time for each text whether it is the actual text or the summarised one. Time Elapsed depicts the actual time taken for the summarization to happen.
Text about Mercedes was copied onto the Summarization tool for Text mining and on clicking the ‘Summarize’ button, the summarizer tool converts the entire text that was copied into a short and a coherent format.

The actual reading time of the original text was assumed 4mins whereas for the Summarized text, the reading time was reduced to 1.1mins. Thus, reading the summarized text can be time saving. The below figure displays the **result of Text mining** approach.
Figure 17. Result of Text Mining

Web Scraping, here allows us to copy or insert a URL for any web page where the data from that web page can be extracted through a web scraper can is copied onto the Summarizer for segmentation. This might take a few seconds to extract all the data from the website and to summarize it successfully. To clear the text box, we can use the ‘Clear’ button and to summarize we use ‘Summarize’ button.

Figure 18. Input of URL for Web Scraping

The result of the Web Scraping is displayed in the figure below along with the reading time of the original long text and the summarized text. The reading time for the original web scraped data is 81mins but for the Summarized text, it was reduced to 1.09mins. This is quite a time saving process.
5 Conclusion

After performing the various empirical tests on Machine Learning models and on the user interface, the result in the above section depicts that the most optimal and accurate results whether it may be related for text summarization. The important keys taken away from this research could be, to have a thorough understanding of different domains and data sets, importance of data related to how to extract it, how to manage large amounts of data, how to generate it in an optimized manner as well as drives the process in order to give a better solution for all the issues that are being faced. This research shows how automation makes our life easier as it is quite easy for any layman to utilize this platform so as to get benefitted. We have also seen how the process of extracting data from different web pages happens and the data is summarized in order to save the reading time of the users.

References

[1] Deepali K. Gaikwad, C. Namrata Mahender, A Review Paper on Text Summarization, Internal Journal of Advanced Research in Computer and Communication Engineering, Vol 5, Issue 3, March 2016.
[2] N. Moratach, S. Chitrakala, A survey on Extractive Text Summarization, IEEE International Conference on Computer, Communication and Signal Processing (ICCCSP-2017).
[3] Said A. Salloum, Mostafa Al-Emran and Khaled Shaalan, Using Text Mining Techniques for Extracting Information from Research Articles, Studies in Computational Intelligence, January 2018.
[4] Nouf Ibrahim Altamimi, Mohamed El Bachir Menai, Automatic Summarization of scientific articles: A survey, Journal of King Saud University- Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.020.
[5] Mehdi Allahyari, SaeidSafariSayedaminPouriyeh, KryksKochut, Mehdi Assefi, Juan B. Gutierrez, and Elizabeth D. Tripppe, Text Summarization Techniques: A Brief Survey.
[6] Kanjeeet Kumar, R.C. Tripathi, Text mining and similarity search using extended tri-gram algorithm in the reference based local repository dataset,International Conference on Communication,Management and Information Technology(ICCMIT-2015).
[7] Shreesha M, Srikara S B, Manjesh R, A Novel Approach for News Extraction using Web Scraping, International Journal of Engineering Research & Technology (IJERT), ISSN:2278-0181, Vol 6: Issue 15,2018.
[8] Ingyang Chen, Hai Zhuge, Summarizing Related Work through Citations, IEEE 12th International Conference on Semantics, Knowledge and Grids, 2016.
[9] Ramzan Talib, Muhammad Kashif Hanif, Shaheela Ayesha and Fakeeha Fatima, Text Mining: Techniques, Applications and Issues, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7 No. 11, 2016.
[10] Ujjwal Rani, Karambir Bidhan, Review Paper on Automatic Text Summarization,International Research Journal of Engineering and Technology (IRJET), Vol 07, Issue 04, April 2020.
[11] Julius Steen, Katja Markert, How to Evaluate a Summarizer: Study Design and Statistical Analysis for Manual
Linguistic Quality Evaluation, arXiv:2101.11298v1 [cs.CL], 27th January, 2021.

[12] Hans Christian, Mikhail Pramodana Agus, Derwin Suhartono, Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF), ComTech Computer Mathematics and Engineering Applications, December 2016.

[13] Shai Erera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora Peled Nakash, Odellia Boni, A Summarization System for Scientific Documents, Proceedings of the 2019 EMNLP and the 9th IJCNLP (System Demonstrations), Hong Kong, China, November 2019.

[14] Maake Benard Magara, Sunday O. Ojo, Tranos, A Comparative Analysis of Text Similarity Measures and Algorithms in Research Paper Recommender Systems, IEEE Conference of Information Communications Technology and Society (ICTAS), 2018.

[15] Marzieh Oghbaie, Morteza Mohammad Zanjireh, Pairwise document similarity measure based on present term set, Journal of Big Data-2018.

[16] Bo Zhao, Web Scraping, Encyclopaedia of Big Data, Chapter-May 2017.

[17] Hossam El-Din Hassanien, Web Scraping Scientific Repositories for Augmented Relevant Literature Search Using CRISP-DM, Applied System Innovation, 2019.

[18] Erdinc Uzun, A novel web scraping approach using the additional information obtained from web pages, Preparation of Papers for IEEE Transactions and Journals, Vol 4, 2016.

[19] Stephan Camilleri, Matthew R Agius, Joel Azzopardi, Analysis of Online News Coverage on Earthquakes through Text Mining, Frontiers in Earth Science, Volume 8, Article 141, May 2020.

[20] G Ravi Kumar, S Rahamat Basha, Surya Bhupal Rao, A Summarization on Text Mining Techniques for Information Extracting from Applications and Issues, National Conference on Recent Trends & Challenges in Engineering, Special Issue, No.-5, January-2020.

[21] https://blog.floydhub.com/gentle-introduction-to-text-summarization-in-machine-learning/

[22] https://www.nltk.org/

[23] https://www.imperva.com/learn/application-security/web-scrapin

[24] https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/

[25] https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32

[26] https://medium.com/luisfredgs/automatic-text-summarization-with-machine-learning-an-overview-68ded5717a25

[27] https://spacy.io/usage

[28] https://en.m.wikipedia.org/wiki/Machine_learning#:~:text=Machine%20learning%20(ML)%20is%20a%20part%20of%20artificial%20intelligence

[29] https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/

[30] https://medium.com/analytics-vidhya/what-you-need-to-know-about-web-scraping-fa626e447343

[31] https://pypi.org/project/gensim/

[32] https://www.geeksforgeeks.org/ml-text-summarization-of-links-based-on-user-query/

[33] https://heartbeat.fritz.ai/extractive-text-summarization-using-neural-networks-5845804c7701