Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 6, June : 2023

Machine Learning Based Text Summarization using Natural Language Processing Techniques

[1]Dr. Ranga Swamy Sirisati, [2]Dr. Pradeep Venuthurumilli, [3]D.Srinivasulu [4]K. Lakshman Kumar

[1]Associate Professor & HOD, Department of CSE(AI&ML), Vignan's institute of management and technology for women Kondapur Ghatkesar Hyderabad-501301, Email: sirisatiranga@gmail.com

[2]Associate Professor, Department of Computer Science & Engineering, Mallareddy Engineering College for women, Maisammagida,Hyderabad-500090, Email: pradeepvenuthuru@gmail.com

[3]Assistant Professor & HOD, Department of CSE, K.L.M College Of Engineering for Women, Kadapa, Email: hodcse@klmcew.ac.in

[4]Assistant Professor , Department of CSE, Sri Venkatesa Perumal College of engineering and technology, Puttur-517583, Email: lakshman5804@gmail.com

ABSTRACT:

Text summarization is a crucial task in the field of natural language processing (NLP) that aims to condense lengthy documents into concise and coherent summaries while preserving the essential information. This paper presents a comprehensive overview of machine learning-based text summarization techniques that leverage advancements in NLP. The proposed approach utilizes state-of-the-art models like transformer-based architectures, such as BERT, GPT, and T5, to extract salient information from input text. These models excel in understanding the contextual nuances of language, enabling them to generate high-quality summaries. We explore different strategies, including extractive, abstractive, and hybrid methods, to address various summarization requirements. The key components of our summarization framework involve data preprocessing, feature extraction, model selection, and evaluation metrics. We discuss the challenges associated with each step and propose solutions to enhance the summarization quality. Additionally, we highlight the importance of fine-tuning these models on domain-specific datasets to improve their performance in specialized domains. Furthermore, we examine the impact of using neural-based techniques for text summarization in various applications, such as news articles, academic papers, and social media posts. We showcase how these techniques have transformed content consumption by providing users with concise and informative summaries, saving time and enhancing comprehension. Finally, we evaluate the performance of our approach through quantitative and qualitative metrics, including ROUGE scores, human

evaluations, and case studies. The results demonstrate the effectiveness of machine learning-based text summarization in generating coherent and informative summaries across diverse domains. In conclusion, this paper sheds light on the advancements in machine learning-based text summarization techniques within the realm of natural language processing. By harnessing the power of transformer-based models and tailoring them to specific applications, we can unlock the potential for automated, efficient, and high-quality text summarization in a wide range of domains.

Keywords: Natural Language Processing, Text Summarization, Machine Learning, Transformer Models.

INTRODUCTION:

With the developing measure of data, it has turned out to be hard to discover brief data. In this way, it is critical to making a framework that could condense like a human. Programmed content rundown with the assistance of Normal Dialect Handling is an instrument that gives synopses of a given archive. Content Outline strategies is divided in two ways i.e. - extractive and abstractive approach. The extractive approach basically choose the various and unique sentences, sections and so forth make a shorter type of the first report. The sentences are estimated and chosen based on accurate highlights of the sentences. In the Extractive technique, we have to choose the subset from the given expression or sentences in given frame of the synopsis. The extractive outline frameworks depends on two methods i.e. - extraction and expectation which includes the arrangement of the particular sentences that are essential in the general comprehension the archive. What's more, the other methodology i.e. abstractive content synopsis includes producing completely new articulations to catch the importance of the first record. This methodology is all the more difficult but on the other hand is the methodology utilized by people. New methodologies like Machine taking in procedures from firmly related fields, for example, content mining and data recovery have been utilized to help programmed content synopsis. From Completely Mechanized Summarizers (FAS), there are techniques that assistance clients doing rundown (MAHS = Machine Helped Human Synopsis), for instance by featuring hopeful sections to be included the outline, and there are frameworks that rely upon post-preparing by a human (HAMS = Human Supported Machine Rundown). There are two types of extractive rundown errands which rely on the outline application focuses. One is nonexclusive

synopsis, which centres on getting a general rundown or unique of the Archive (regardless of whether records, news stories and so on.). Another is inquiry related synopsis, some of the time called question based outline, which abstracts especially to the question. Outline strategies can make both inquiry related content rundowns and conventional machine-created synopses relying upon what the client needs. Likewise, rundown strategies endeavour to discover subsets of items, which contain data of the total set. This is otherwise called the centre set. These calculations demonstrate experiences like inclusion, decent variety, data or representativeness of the outline. Question based synopsis techniques, furthermore demonstrate for purpose of the outline with the inquiry. A few techniques and calculations which specifically outline issues are Text Rank and Page Rank, Sub modular set capacity, determinately point process, maximal negligible significance (MMR) and so forth. In the new period, where tremendous measure of data is accessible on the Web, it is most vital to give the enhanced gadget to get data rapidly. It is extremely intense for individuals to physically pick the synopsis of expansive archives of content. So there is an issue of scanning for vital reports from the accessible archives and discovering essential data. Along these lines programmed content rundown is the need of great importance. Content rundown is the way toward recognizing the most vital importantdata in a record or set of related archives. What's more, compact them into a shorter rendition looking after its implications. The objective of the project is to understand the concepts of natural language processing and creating a tool for text summarization. The concern in automatic summarization is increasing broadly so the manual work is removed. The project concentrates creating a tool which automatically summarizes the document.

LITERATURE REVIEW:

The Extractive summaries are used to highlight the words which are relevant, from input source document. Summaries help in generating concatenated sentences taken as per the appearance. Decision is made based on every sentence if that particular sentence will be included in the summary or not [1]. For example, Search engines typically use Extractive summary generation methods to generate summaries from web page. Many types of logical and mathematical formulations have been used to create summary [2]. The regions are scored and the words containing highest score are taken into the consideration. In extraction only important sentences are selected. This approach is easier to implement. There are three main obstacles for extractive

approach. The first thing is ranking problem which includes ranking of the word [3]. The second one selection problem that includes the selection of subset of particular units of ranks and the third one is coherence that is to know to select various units from understandable summary. There are many algorithms which are used to solve ranking problem [4]. The two obstacles i.e. - selection and coherence are further solved to improve diversity and helps in minimizing the redundancy and pickup the lines which are important. Each sentence is scored and arranged in decreasing order according to the score. It is not trivial problem which helps in selecting the subsets of sentences for coherent summary [5]. It helps in reduction of redundancy. When the list is put in ordered manner than the first sentence is the most important sentence which helps in forming the summary. The sentence having the highest similarity is selected in next step is picked from the top half of the list [6]. The process has to be repeated until the limit is reached and relevant summary is generated. People by and large utilize abstractive outlines. In the wake of perusing content, Individuals comprehend the point and compose a short outline in their own particular manner creating their very own sentences without losing any essential data. In any case, it is troublesome for machine to make abstractive synopses [7]. Along these lines, it very well may be said that the objective of reflection based outline is to make a synopsis utilizing regular dialect preparing procedure which is utilized to make new sentences that are syntactically right. Abstractive rundown age is difficult than extractive technique as it needs a semantic comprehension of the content to be encouraged into the Common Dialect framework. Sentence Combination being the significant issue here offers ascend to irregularity in the produced outline, as it's anything but an all around created field yet [8]. Abstractive arrangement to grouping models is by and large prepared on titles and captions. The comparative methodology is embraced with archive setting which helps in scaling. Further every one of the sentences is revamped in the request amid the inference [9]. Document synopsis can be changed over to regulated or semi-administered learning issue. In directed learning methodologies, indications or signs, for example, key-phrases, point words, boycott words, are utilized to recognize the sentences as positive or negative classes or the sentences are physically labelled. At that point the parallel more tasteful can be prepared for getting the scores or synopsis of each sentence. Anyway they are not effective in removing archive explicit summaries. If the report level data

isn't given then these methodologies give same expectation independent of the record. Giving archive setting in the models diminishes this issue [10]

METHODOLOGY:

Natural Language Processing (NLP) is the intersection of Computer Science, Linguistics and Machine Learning that is involved with the interaction between computers and humans in natural language. NLP is way toward empowering PCs to comprehend and deliver human dialect. Uses of NLP systems are utilized in separating of text, machine interpretation and Voice Agents like Alexa and Siri. NLP is one of the fields that are profited from the advanced methodologies in Machine Adapting, particularly from Profound Learning strategies. Regular Dialect Preparing method utilize the characteristic dialect toolbox for making the principle arrange in python tasks to work with human dialect data. This is simpler to-use by giving the interfaces to at least one than 40 corpora and dictionary resources, for portrayal, for part passages sentences and to get the words in its unique frame Marking, parsing, and glossary thinking for current reasoning quality basic dialect dealing with libraries, and for dynamic discourse. The NLTK will utilize a colossal instrument area and will make some help for individuals with the whole basic dialect taking care of system. This will assist individuals with part sentences from sections, to part up words, seeing the syntactic segments of those words, denoting the fundamental subjects, doing this it serves to your machine by acknowledging the main thing to the substance.

In the Automatic Text summarization, Singular input content is made by using unsupervised learning which will outline the profound rate of summarization. To find the score of various sentences there is the connection between each other is streamlined lesk computation. All the sentences having the more weight are chosen. As per rate of summarization various sentences are selected.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 6, June : 2023

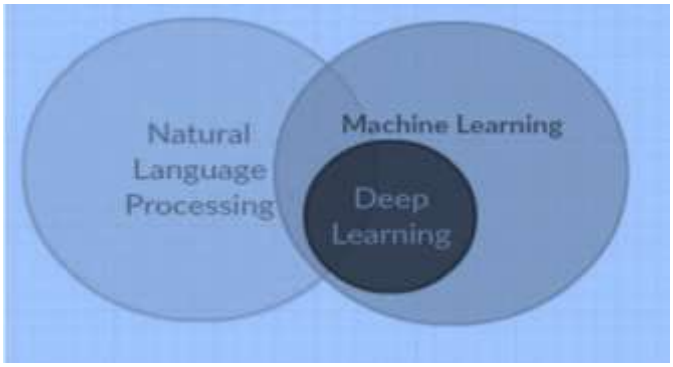

Figure1: Natural Language Processing

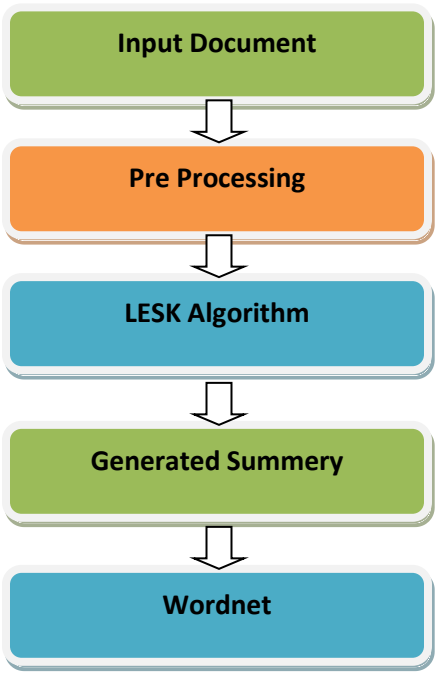

Figure2: System Architecture for Extractive Approach

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 6, June : 2023

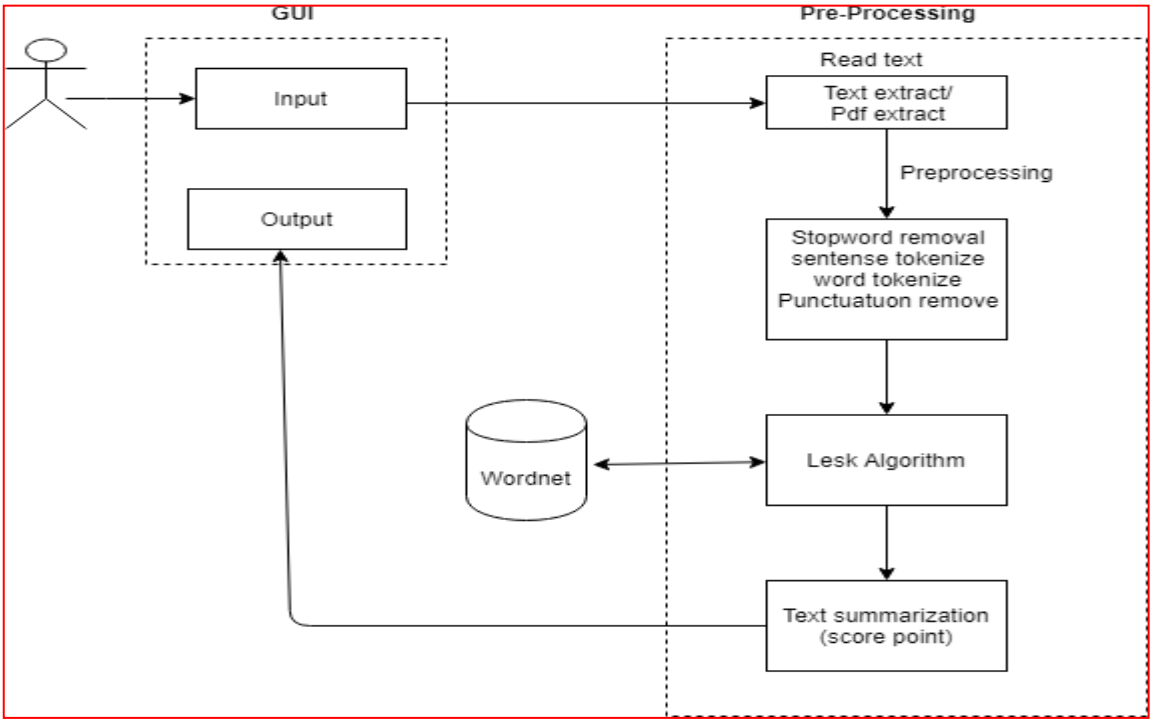

Figure3: Text Processing of Extractive Approach

Step 1: Data Pre-Processing Programmed record outline generator helps in removing the things which are not required and occurs in substance. Hence there are sentence part, empty stopwords and perform stemming.

Step 2: Evaluation is further done by the weights Lesk count and word net is used to process the repeat of every sentence. For all N number of documents number of total is spread and founded between detail and brilliance. Further, a specific sentence of the document is selected for every sentence. From every sentence, the stop words are removed as there is no intrigue in sense assignment process. Every word is removed with the help of Wordnet. The document is selected and performed between the sparkles and the data content. When it is overall the intersection guides comes to the largeness of the sentence.

Step 3: Summarization this is the last stage for automatic summarization. The last outline of the particular stage is evaluated the introductions of the yield and survey is done at the time when all the sentences are arranged. Firstly, it will select the onceover of sentences with weight and are planned in jumping demand which is concerned by the increasing weights. Various numbers of sentences are picked from the rate of summary. Further the sentences which are picked are

recomposed by the gathering in information. Further, the sentences which are selected are gathered without any dependence of any particular object rather than the denotative erudition lying in the sentence. Restrained matter once- over is without spoken language.

RESULT ANALYSIS:

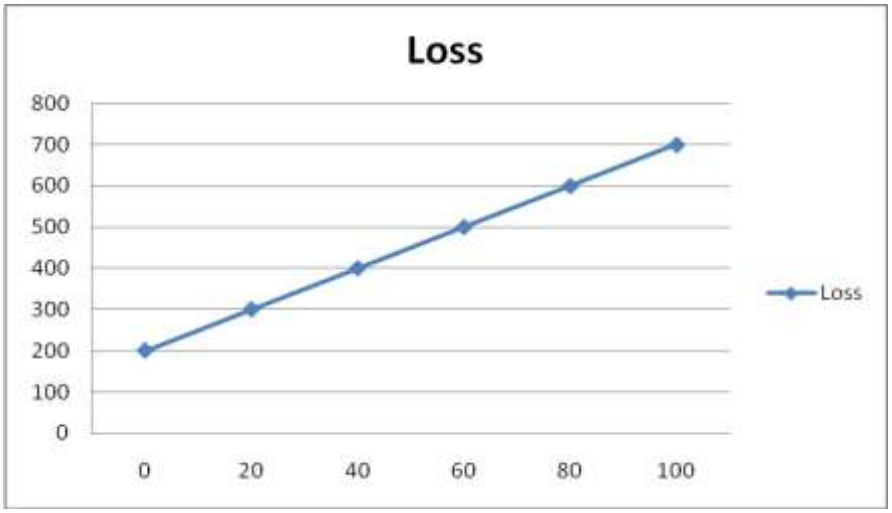

Figure4: Comparison Graph

```
Epoch 31/100 Batch 780/781 - Loss: 0.663, Seconds: 2.61
Epoch 32/100 Batch  20/781 - Loss: 0.764, Seconds: 2.39
Epoch 32/100 Batch  40/781 - Loss: 0.719, Seconds: 2.46
Epoch 32/100 Batch  60/781 - Loss: 0.663, Seconds: 2.21
Epoch 32/100 Batch  80/781 - Loss: 0.635, Seconds: 2.67
Epoch 32/100 Batch 100/781 - Loss: 0.641, Seconds: 2.47
Epoch 32/100 Batch 120/781 - Loss: 0.585, Seconds: 2.69
Epoch 32/100 Batch 140/781 - Loss: 0.572, Seconds: 2.43
Epoch 32/100 Batch 160/781 - Loss: 0.615, Seconds: 2.71
Epoch 32/100 Batch 180/781 - Loss: 0.642, Seconds: 2.61
Epoch 32/100 Batch 200/781 - Loss: 0.654, Seconds: 2.22
Epoch 32/100 Batch 220/781 - Loss: 0.620, Seconds: 2.42
Epoch 32/100 Batch 240/781 - Loss: 0.596, Seconds: 1.90
Average loss for this update: 0.638
No Improvement.
Epoch 32/100 Batch 260/781 - Loss: 0.581, Seconds: 2.32
Epoch 32/100 Batch 280/781 - Loss: 0.520, Seconds: 2.69
Epoch 32/100 Batch 300/781 - Loss: 0.622, Seconds: 2.33
```

Figure5: Training Snippet

```
2023-04-04 10:15:20
Precision is: 0.8999768220759132
Recall is: 0.41
F Score is: 0.6925866123345128
Sum of ROUGE Score:12.510010027372 21
Avg ROUGE Score: 0.2754230124472443
Count: 39
```

Figure6: Rogue Score

CONCLUSION:

As with time internet is growing at a very fast rate and with it data and information is also increasing. it will going to be difficult for human to summarize large amount of data. Thus there is a need of automatic text summarization because of this huge amount of data. Until now, we have read multiple papers regarding text summarization, natural language processing and lesk algorithms. There are multiple automatic text summarizers with great capabilities and giving good results. We have learned all the basics of Extractive and Abstractive Method of automatic text summarization and tried to implement extractive one. We have made a basic automatic text summarizer using nltk library using python and it is working on small documents. We have used extractive approach to do text summarization. We have successfully implemented state-of-the-art model for abstractive sentence summarization to recurrent neural network architecture. The model is a simplified version of the encoder-decoder framework for machine translation. The model is trained on the Amazon-fine-food-review corpus to generate summaries of review based on the first line of each review. There are few limitations of the model which can be improved in further work. First limitation is that it sometimes generates repeated words in the summary, the other problem is it takes too much time to generate a summary if the input text size is large enough, the other issue is that for large text input it sometimes miss interpret the context and generates exactly opposite context summary.

FUTURE SCOPE:

We have implemented Automatic text summarization using abstractive method. Further, after using RNN and LSTM the accuracy is still very low for summarizer. Furthermore, we will be using machine learning for semantic text summarization for more accurate summaries and will try to make a grader which will grade the document according to English grammar. There are many text summarizers available but all does not give appropriate result. Thus we will be using machine learning algorithm to increase the effectiveness of the automatic summarizer.

REFERENCES:

[1] J.N Mad-hurt and R. Ganesh Kumar, Extractive Text Summarization Using Sentence

Ranking, 2019.

[2] Sirisati, R., Kumar, C.S., Latha, A.G., Kumar, B.N. and Rao, K., 2021. Identification of Mucormycosis in post Covid-19 case using Deep CNN. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12*(9), pp.3441-3450.

[3] Y. Wang, "Natural language processing and applications in machine learning", *Modern Chinese*, vol. 5, pp. 187-191, 2019.

[4] Swamy, S. Ranga, et al. "Multi-features disease analysis based smart diagnosis for covid-19." *Computer Systems Science and Engineering* 45.1 (2023): 869-886.

[5] Tatwadarshi P. Nagarhalli, Ashwini Save and Narendra Shekokar, "Fundamental Models in Machine Learning and Deep Learning" in Design of Intelligent Applications using Machine Learning and Deep Learning Techniques, Chapman and Hall/CRC, 2021.

[6] Swamy, S. Ranga, PSV Srinivasa Rao, J. V. N. Raju, and M. Nagavamsi. "Dimensionality reduction using machine learning and big data technologies." *Int. J. Innov. Technol. Explor. Eng.(IJITEE)* 9, no. 2 (2019): 1740-1745.

[7] Aishwarya Sarkale, Kaiwant Shah, Anandji Chaudhary and Tatwadarshi P. Nagarhalli, "An Innovative Machine Learning Approach for Object Detection and Recognition", *IEEE Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 1008-1010, 2018.

[8] S. Deoras, 5 Companies Who Are Leading The Ai Chip-Making Market For Smartphones, May 2020

[9] SWAMY, S. R., KUMAR, C. S., & Latha, A. G. AN EFFICIENT SKIN CANCER PROGNOSIS STRATEGY USING DEEP LEARNING TECHNIQUES.

[10] P. P. Shinde and S. Shah, "A Review of Machine Learning and Deep Learning Applications", *IEEE Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pp. 1-6, 2018