Journal of Engineering Sciences

Vol 13, Issue 06, June/2022
ISSN NO:0377-9254

NLP based Machine Learning Approaches for Text Summarization

Ms.G. Sri Lakshmi, Assistant Professor, MTech, Department of IT, sree.gpk@gmail.com

J.Vishnu, BTech, Department of IT, vishnujadhavvj666@gmail.com

K.Yamini Devi, BTech, Department of IT, yaminikatta122@gmail.com

T.Sravani, BTech, Department of IT, sravanitellakula2399@gmail.com

P.Bhadra Reddy,BTech, Department of IT, bhadrareddypodapala@gmail.com

P.Siva Rahul, BTech, Department of IT, psrk3838@gmail.com

ABSTRACT: Due to the plethora of data available today, text summarization has become very essential to gain just the right amount of information from huge texts. We see long articles in news websites, blogs, customers' review websites, and so on. This review paper presents various approaches to generate summary of huge texts. Various papers have been studied for different methods that have been used so far for text summarization. Mostly, the methods described in this paper produce Abstractive (ABS) or Extractive (EXT) summaries of text documents. Querybased summarization techniques are also discussed. The paper mostly discusses about the structured based and semantic based approaches for summarization of the text documents. Various datasets were used to test the summaries produced by these models, such as the CNN corpus, DUC2000, single and multiple text documents etc. We have studied these methods and also the tendencies, achievements, past work and future scope of them in text summarization as well as other fields.

Keywords: *Abstractive (ABS) or Extractive (EXT)*

1. INTRODUCTION

Today's world is centralized on computers and data. Data are our intangible thoughts and imagination. We are producers and consumers of data at the same time. Every little thing in our mundane lives are either a source or receiver of data. For, example when we drive there's data involved, the speed of the car, mileage, distance traveled, etc. Since 20th century, data has been a significant part of our lives, but these days we infer more from data. We store and access them through electronic and wireless systems. Since the advent of the internet, there's an enormous amount of data available today. The Internet is a storehouse of data. Information on news, movies, education, medicine, health, nations, weather, geology, etc. is available on the internet. This could be statistical, numerical, mathematical or text data. Text data is more difficult to interpret due to larger amount of characters. Due to this gigantic amount of information, there must be a system in order to get only the essential parts of the information we access. Text summarization is a way of doing this. Text summarization has been a topic of research and study since decades. Various models have been proposed

and tested on different datasets to generate concise summaries. They are compared with different comparison scores. Text summarization can be EXT or ABS, single document or multidocument, and query-based or generic. EXT text summarization is a way of generating summaries by using the same sentences as in the document. ABS is more general and focuses on key concepts of the document. Similarly, single document summarization techniques give summaries of the text of a single document, and multidocument generates summaries of multiple documents.

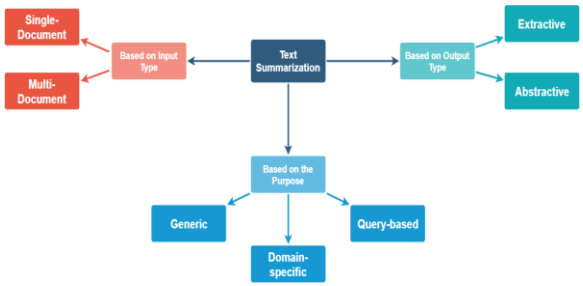

Fig.1 Text summarization

Moreover, these days, there's a need for summarizing text based on queries. Query-based summarization models give summaries of the text based on a specific area as described by the query given by the user, whereas generic summaries are mostly ABS that focus on the general area of the text input. Text summarization has been extensively used in various fields like science, medicine, law, engineering, etc. Researchers have focused on generating summaries of doctor's prescriptions, and that has been proved very useful to patients. Similarly, long news articles have been summarized and this way readers can gain a lot of information on several topics within a short span of time[1]. In this paper, we have discussed the various methods used in text summarization for past five years. The most common methods were found to be Machine Learning(ML), NNs, reinforcement learning, sequence to sequence modeling and fuzzy logic. Similarly, various optimization methods have been used to optimize the proposed objective function for the purpose of text summarization. We can see that various methods were tested on the same dataset, and their accuracy scores were found to be different. Moreover, we can also see that some researchers have combined the different methods and found out the summaries are more accurate than when a single method is used. When NLP processing has been used as a technique to summarize text documents, we see that python libraries such as scikit learn, nltk, spacy, and fastai has been used.

2. LITERATURE REVIEW

2.1 Scalable machine learning computing a data summarization matrix with a parallel array DBMS:

Big data analytics requires scalable (beyond RAM limits) and highly parallel (exploiting many CPU cores) processing of machine learning models, which in general involve heavy matrix manipulation. Array DBMSs represent a promising system to manipulate large matrices. With that motivation in mind, we present a high performance system exploiting a parallel array DBMS to evaluate a general, but compact, matrix summarization that benefits many machine learning models. We focus on two representative models: linear regression (supervised) and PCA (unsupervised). Our approach combines data summarization inside the parallel DBMS with further model computation in a mathematical language (e.g. R). We introduce a 2-phase algorithm which first computes a general data summary in parallel and then evaluates matrix equations with

reduced intermediate matrices in main memory on one node. We present theory results characterizing speedup and time/space complexity. From a parallel data system perspective, we consider scale-up and scale-out in a shared-nothing architecture. In contrast to most big data analytic systems, our system is based on array operators programmed in C++, working directly on the Unix file system instead of Java or Scala running on HDFS mounted of top of Unix, resulting in much faster processing. Experiments compare our system with Spark (parallel) and R (single machine), showing orders of magnitude time improvement. We present parallel benchmarks varying number of threads and processing nodes. Our 2-phase approach should motivate analysts to exploit a parallel array DBMS for matrix summarization.

2.2 A freeWeb API for single and multi-document summarization.

In this work we present a free Web API for single and multi-text summarization. The summarization algorithm follows an extractive approach, thus selecting the most relevant sentences from a single document or a document set. It integrates in a novel pipeline different text analysis techniques - ranging from keyword and entity extraction, to topic modelling and sentence clustering - and gives SoA competitive results. The application, written in Python, supports as input both plain texts and Web URLs. The API is publicly accessible for free using the specific conference token1 as described in the reference page2. The browser-based demo version, for summarization of single documents only, is publicly accessible at http://yonderlabs.com/demo.

2.3 Text Summarization using Neural Networks and Rhetorical Structure Theory

A new technique for summarization is presented here for summarizing articles known as text summarization using neural network and rhetorical structure theory. A neural network is trained to learn the relevant characteristics of sentences by using back propagation technique to train the neural network which will be used in the summary of the article. After training neural network is then modified to feature fusion and pruning the relevant characteristics apparent in summary sentences. Finally, the modified neural network is used to summarize articles and combining it with the rhetorical structure theory to form final summary of an article.

2.4 Automatic text document summarization based on machine learning

The need for automatic generation of summaries gained importance with the unprecedented volume of information available in the Internet. Automatic systems based on extractive summarization techniques select the most significant sentences of one or more texts to generate a summary. This article makes use of Machine Learning techniques to assess the quality of the twenty most referenced strategies used in extractive summarization, integrating them in a tool. Quantitative and qualitative aspects were considered in such assessment demonstrating the validity of the proposed scheme. The experiments were performed on the CNN-corpus, possibly the largest and most suitable test corpus today for benchmarking extractive summarization strategies.

2.5 K nearest neighbor for text summarization using feature similarity

In this research, we propose a particular version of KNN (K Nearest Neighbor) where the similarity between feature vectors is computed considering the similarity among attributes or features as well as one among values. The task of text summarization is

viewed into the binary classification task where each paragraph or sentence is classified into the essence or non-essence, and in previous works, improved results are obtained by the proposed version in the text classification and clustering. In this research, we define the similarity which considers both attributes and attribute values, modify the KNN into the version based on the similarity, and use the modified version as the approach to the text summarization task. As the benefits from this research, we may expect the more compact representation of data items and the better performance. Therefore, the goal of this research is to implement the text summarization algorithm which represents data items more compactly and provides the more reliability.

3. IMPLEMENTATION

Massimo Mauro, et al. have used a sentence extraction method to generate EXT summaries. In this method, the sentences are checked for their relevance and scored accordingly. Similar sentences were then clustered together to discover the most informative sentences and they were selected on the basis of sentence scores.

Drawbacks:

- We have seen various researchers presenting ML methods for text summarization. Various supervised ML models such as Naïve Bayes, Random Forest, SVD have been used to generate EXT text summaries.

We have also seen many researchers use deep learning techniques for EXT as well as ABS text summarization. Deep learning is an area of ML. Various NN techniques have been used. Similarly, reinforcement learning, Convolutional NN(CNN), RNN have also been applied to generate text summaries[10]. There's also a study of sequence-to-sequence models for text summarization these days. These methods are extension of ML. We describe some of the papers that use the above mentioned methods to generate summaries of text.

Advantages:

- Accuracy Increased.

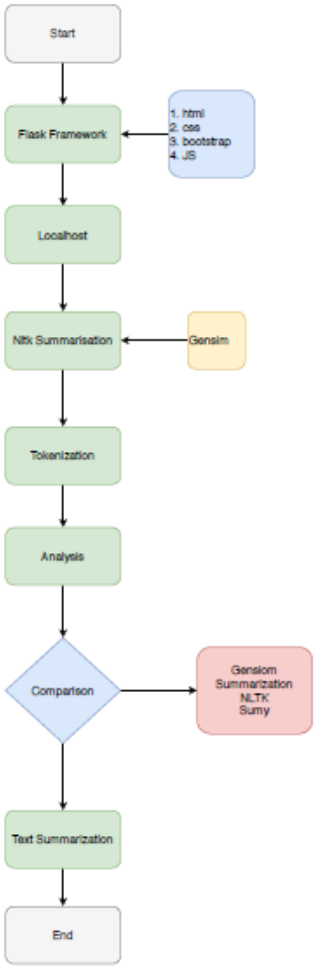

Fig.2: System architecture

Since the advent of the internet, there's an enormous amount of data available today. The Internet is a

storehouse of data. Information on news, movies, education, medicine, health, nations, weather, geology, etc. is available on the internet. This could be statistical, numerical, mathematical or text data. Text data is more difficult to interpret due to larger amount of characters. Due to this gigantic amount of information, there must be a system in order to get only the essential parts of the information we access. Text summarization is a way of doing this. The objective of this project is text Summarization using Machine Learning Approaches for We have seen the use of various algorithms and methods for this purpose. These methods, in individual and together give different types of summaries. Their accuracy score can be compared to find the better and more concise summaries. For this purpose, ROGUE score has been used more frequently. Similarly, in some cases TF_IDF scores have been used too.

4. ALGORITHMS

NLP:

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken and written -- referred to as natural language. It is a component of artificial intelligence (AI). NLP has existed for more than 50 years and has roots in the field of linguistics. It has a variety of real-world applications in a number of fields, including medical research, search engines and business intelligence.

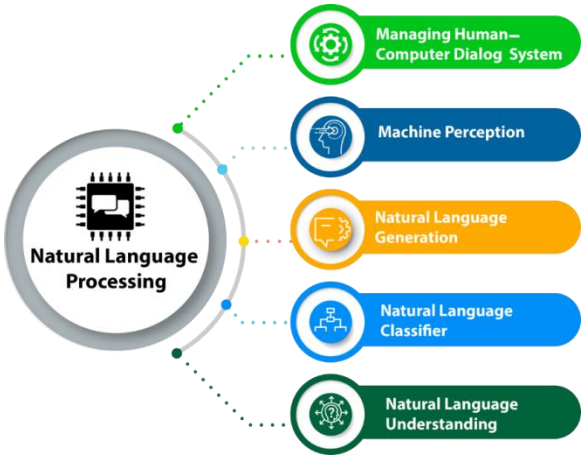

Fig.3: NLP model

TOKENIZATION:

Tokenization is breaking the raw text into small chunks. Tokenization breaks the raw text into words, sentences called tokens. These tokens help in understanding the context or developing the model for the NLP. The tokenization helps in interpreting the meaning of the text by analyzing the sequence of the words.

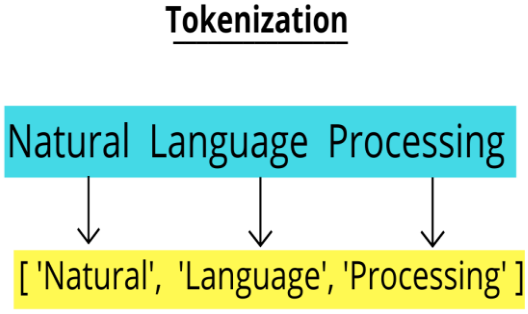

Fig.4: Tokenization

CNN:

CNN is a type of artificial neural network that is widely utilised. It's a type of deep, feed-forward artificial neural network that's often used for image analysis. CNN was first created by LeCun in the early

1990s. It's also a popular method for extracting features and classifying time-series data. A CNN is a multilayer perceptron that is similar to a multilayer perceptron (MLP). The architecture of the model permits CNN to exhibit translational and rotational invariance thanks to this particular structure. In general, a CNN consists of one or more convolutional layers, as well as associated weights and pooling layers, and a fully connected layer at the top.

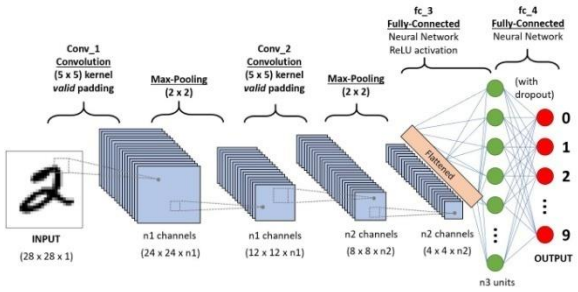

Fig.5: CNN model

LSTM:

A time recurrent neural network is known as an LSTM (RNN). A cell, an input gate, an output gate, and a forget gate make up a typical LSTM unit. The three gates govern the flow of information into and out of the unit, and the unit remembers values throughout any time interval. A simple one-layer neural network controls the forget gate in the memory block structure..

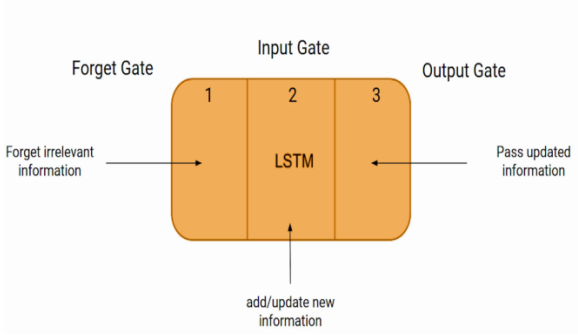

Fig.6: LSTM model

5. EXPERIMENTAL RESULTS

```
data.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 88421 entries, 0 to 99999
Data columns (total 10 columns):
Id                      88421 non-null int64
ProductId               88421 non-null object
UserId                  88421 non-null object
ProfileName             88421 non-null object
HelpfulnessNumerator    88421 non-null int64
HelpfulnessDenominator  88421 non-null int64
Score                   88421 non-null int64
Time                    88421 non-null int64
Summary                 88421 non-null object
Text                    88421 non-null object
dtypes: int64(5), object(5)
memory usage: 7.4+ MB
```

Fig.7: Dataset

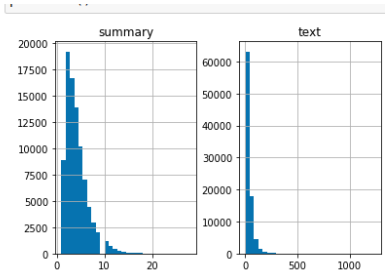

Fig.8: Visualization of text and summary words length

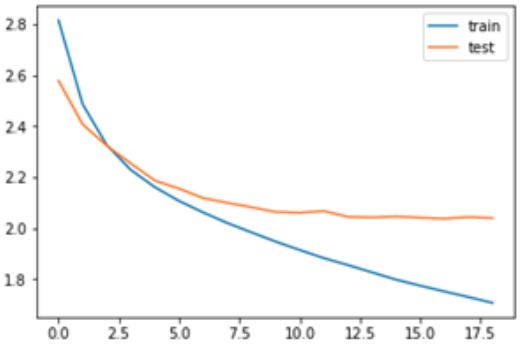

Fig.9: Graphical representation

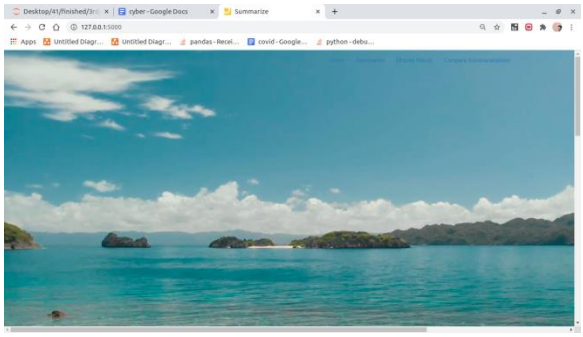

Fig.10: Home screen

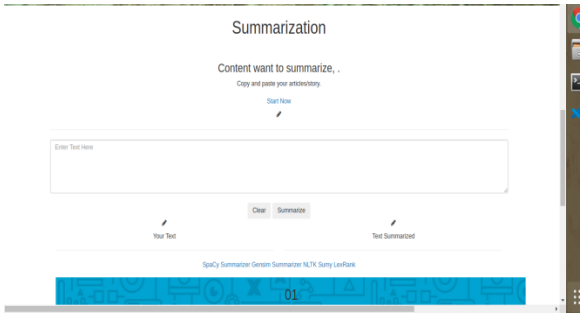

Fig.11: Content giving screen

6. CONCLUSION

We have seen that due to abundant availability of data, text summarization has a very vital role in saving user's time, as well as resources. Text summarization is indeed an important tool for today. We have seen the use of various algorithms and methods for this purpose. These methods, in individual and together give different types of summaries. Their accuracy score can be compared to find the better and more concise summaries. For this purpose, ROGUE score has been used more frequently. Similarly, in some cases TF_IDF scores have been used too. The summaries generated using these methods are not always up to the mark. Sometimes, it's also irrelevant to the original document. Therefore, this topic is ongoing and people have done various works on this. There isn't any specific model that generates best summaries.

7. FUTURE SCOPE

For future, the models discussed can be modified for more accurate summaries. For e.g., we could use GAN's and transfer learning. For future, this can give a way to develop and enhance further ideas for text summarization.

REFERENCES

[1] C. Ordonez, Y. Zhang, and S. L. Johnsson, "Scalable machine learning computing a data summarization matrix with a parallel array DBMS," Distrib. Parallel Databases, vol. 37, no. 3, pp. 329–350, 2019, doi: 10.1007/s10619-018-7229-1.

[2] M. Mauro, L. Canini, S. Benini, N. Adami, A. Signoroni, and R. Leonardi, "A freeWeb API for single and multi-document summarization," ACM Int. Conf. Proceeding Ser., vol. Part F1301, 2017, doi: 10.1145/3095713.3095738.

[3] A. T. Sarda and M. Kulkarni, "Text Summarization using Neural Networks and Rhetorical Structure Theory," Int. J. Adv. Res. Comput. Commun. Eng., vol. 4, no. 6, pp. 49–52, 2015, doi: 10.17148/IJARCCE.2015.4612.

[4] G. Silva, R. Ferreira, S. J. Simske, L. Rafael Lins, M. Riss, and H. O. Cabral, "Automatic text document summarization based on machine learning," DocEng 2015 - Proc. 2015 ACM Symp. Doc. Eng., pp. 191–194, 2015, doi: 10.1145/2682571.2797099.

[5] T. Jo, "K nearest neighbor for text summarization using feature similarity," Proc. - 2017 Int. Conf. Commun. Control. Comput. Electron. Eng. ICCCCEE 2017, pp. 1–5, 2017, doi: 10.1109/ICCCCEE.2017.7866705.

[6] B. Mutlu, E. A. Sezer, and M. A. Akcayol, "Multi-document extractive text summarization: A comparative assessment on features," Knowledge-Based Syst., vol. 183, p. 104848, 2019, doi: 10.1016/j.knosys.2019.07.019.

[7] M. Afsharizadeh, H. Ebrahimpour-Komleh, and A. Bagheri, "Query-oriented text summarization using sentence extraction technique," 2018 4th Int. Conf. Web Res. ICWR 2018, pp. 128–132, 2018, doi: 10.1109/ICWR.2018.8387248.

[8] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 Int. Conf. Data Sci. Commun. IconDSC 2019, pp. 1–3, 2019, doi: 10.1109/IconDSC.2019.8817040.

[9] K. Kandasamy and P. Koroth, "An integrated approach to spam classification on Twitter using URL analysis, natural language processing and machine learning techniques," 2014 IEEE Students' Conf. Electr. Electron. Comput. Sci. SCEECS 2014, pp. 1–5, 2014, doi: 10.1109/SCEECS.2014.6804508.

[10] M. A. Fattah, "A hybrid machine learning model for multidocument summarization," Appl. Intell., vol. 40, no. 4, pp. 592–600, 2014, doi: 10.1007/s10489-013-0490-0.

[11] N. Giamblanco and P. Siddavaatam, "Keyword and Keyphrase Extraction using Newton's Law of Universal Gravitation," Can. Conf. Electr. Comput. Eng., pp. 1–4, 2017, doi: 10.1109/CCECE.2017.7946724.

[12] R. Alguliyev, R. Aliguliyev, and N. Isazade, "A sentence selection model and HLO algorithm for extractive text summarization," Appl. Inf. Commun. Technol. AICT 2016 - Conf. Proc., pp. 1–4, 2017, doi: 10.1109/ICAICT.2016.7991686.

[13] M. Moradi, G. Dorffner, and M. Samwald, "Deep contextualized embeddings for quantifying the informative content in biomedical text summarization," Comput. Methods Programs Biomed., vol. 184, p. 105117, 2020, doi: 10.1016/j.cmpb.2019.105117.