International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT228421

Text Summarization Using Machine Learning Algorithm

Dr. Vidyagouri B H*[1], BibiSadiqa M D[2]

*[1]Assistant Professor, Dept of Computer Science and Engineering, SDM College of Engineering and Technology, Dharwad, Karnataka, India[1]

[2]Mtech Student Dept of Computer Science and Engineering, SDM College of Engineering and Technology, Dharwad, Karnataka, India

Article Info
Publication Issue :
Volume 8, Issue 4
July-August-2022

Page Number : 167-173

Article History
Accepted: 05 July 2022
Published: 22 July 2022

ABSTRACT

In the age of technology, data is critical. The data on the internet is formless and poorly organized. The concept of text summarization is introduced in order to convert data summaries. Text summarization is the process of extracting useful information from raw data without diluting the main theme of the data. Today's readers must contend with task of reading comments, reviews, news articles, blogs and other forms of informal and noisy communication. It is difficult to retrieve the correct gist of the gist, which is required by all readers. To achieve the benefits of both extractive and abstractive summarization, the proposed approach combines TF-TDF-TR(Term Frequency – Inverse Document Frequency – Text Rank) as an unsupervised learning algorithm and the seq2seq (Sequence to Sequence) model as a supervised learning algorithm. In terms of ROUGE score, the proposed TFRSP approach outperforms existing text summarization methods, resulting in high summary accuracy.

Keywords : Text Summarization, Natural Language Processing, Extractive Summarization, Abstractive Summarization, ROUGE.

I. INTRODUCTION

Data is extremely important in the age of information technology. The vast majority of internet data is disorganized and poorly presented. The text summarization concept is provided to extract the data summary and transform the raw data into something structured, understandable, cohesive, and brief. The text summarizing process entails extracting relevant information from raw data while retaining the data's main theme. Reading comments, reviews, news stories, blogs, and so on these days may be difficult due to their casual and boisterous nature. It might be challenging to find the precise substance of the text, which is essential for all readers. People like information that is brief and to the point, taking time into consideration. When reading news items and online evaluations or comments, there is a significant problem that makes it difficult to draw any conclusions before reading the entire piece. As a result, the idea of text summarization is developed for the benefit of information retrieval. In the modern day, where an enormous amount of knowledge is constantly being created online. Therefore, it is

essential to offer a better technique for quickly and most efficiently extracting the relevant information. Text summarizing is a technique for extracting key ideas from a paper or group of linked documents and condensing them into a shorter version while retaining their main points. It decreases the amount of time needed to read a lengthy article and solves the space issue caused by the need to store a lot of data. The idea behind text summarizing is to take the major body of information from the original text and condense, organize, and interpret it for human consumption [1]. To produce the summary methodically, natural language processing (NLP) principles are applied in automatic text summarization. Automatic data summarization generates a system-generated summary that humans can read and understand. The two types of text summarization are depicted in Figure 1.1: extractive and abstractive.

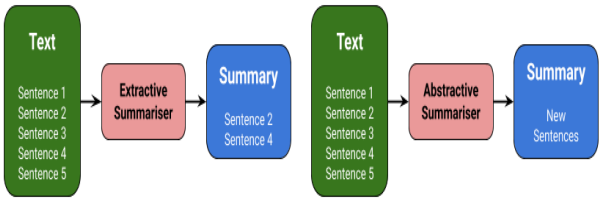

Fig 1. Text summarization by extraction and abstraction

1. Extractive Summarization: It takes the most relevant and significant from the original text and condenses them into a subset of these sentences [2]. Extractive summarization is comparable to underlining the key phrases within the source text. Extractive summarization is the process of extracting valuable information or paragraph from a text file or original document. The selection of valuable informative sentences is done using linguistic or statistical criteria in an extractive text summarizing technique.

2. Abstractive summarization: This technique takes the essential ideas from the original text and creates a summary in its own terms. The summary is created through abstractive summarization, which is the same as rewriting the original text using new words [2]. An abstractive text summary will attempt to comprehend the input or original file and will re-generate the output in a concise manner by determining the core idea of the input file.

Both the extractive and abstractive summaries have advantages and disadvantages. While abstractive summarizing produces a word-for-word summary that is crisp and legible, it may result in the loss of important data when dealing with large texts. Extractive summarization selects important and correct phrases but suffers from incoherency. The primary advantage of text summarization is that it reduces end-user reading time. Text summarization is widely used in a variety of fields, including medicine, news stories, tax and legal content analysis, articles, writings, reviews online, and a variety of other fields. The current strategy combines TF-IDF-TR (Term Frequency – Inverse Document Frequency – Text Rank) as an unsupervised algorithm and the Sequence to sequence model as a supervised learning algorithm to reap the benefits of both extractive and abstractive summarization. Using the Recall Oriented Understudy of Gisting Evaluation (ROUGE), the proposed TFRSP strategy is tested against existing text summarization methods and achieves a high ROUGE score, resulting in good summary accuracy.

II. METHODS AND MATERIAL

A. Term Frequency Inverse Document Frequency

In feature extraction, the TF-IDF method is used. Term Frequency (TF) is a metric that calculates the frequency with which terms occur. The acquired frequency is being used to assess the importance of the word. More frequently term appears, the more important it is in the paper [5] [6]. The simplest explanation for TF is that it counts how many times a word appears in a document [7]. The inverse document frequency (IDF) gives uncommon words a

higher value and recurring terms a lower value. Because stop words occur frequently, TF sometimes overestimates their significance. To address TF's concern, IDF acknowledges the document's infrequent use of words. The Term Frequency calculates the frequency of each term. In inverse document frequency, rather than each word in the sentence, the frequency of sentences is calculated. The TF-IDF formula is as follows [5][7]:

Term Frequency = (How often a words appears in text) / (total number of words in text).

Inverse document frequency = log(total sentence count) / (number of times each word appear in sentences).

B. Text Rank Algorithm

The text rank is an unsupervised sentence-rating technique that applies to weights. The text rank algorithm is based on Google's page rank algorithm, which ranks sites according to their importance and number of hyperlinks[8]. As the name graph-based ranking algorithm implies, a directed graph is built using phrases. The pages are referred to as nodes or vertices, and the edges represent the similarity between two nodes[9]. The text rank algorithm is a recommender-based system in which the graph's vertices and edges advocate the significance of words. This algorithm is based on the assumption that the sentence that is similar to the majority of the other sentences in the paragraph is most likely the most important statement in the passage. Using this concept, one can build a network of sentences that are linked to all comparable sentences, and then run Google's PageRank algorithm on it to determine which sentences are the most important. The summary would then be formed by combining these words.

C. Sequence to Sequence Model

The proposed TFRSP algorithm employs the sequence to sequence model, which is an abstractive summarization algorithm capable of generating new phrases while retaining the meaning of the source document. It is a model of encoder-decoder with variable length input and output sequences [10]. Long Short Term Memory (LSTM) is a component of an encoder-decoder that aids in the capture of long-term dependencies. There are two phases to the encoder-decoder model: training and inference [11]. Both the encoder and decoder are designed to be used during the training and inference phases. The encoder reads the entire input word for word during the training phase, processing the information in the input sequence and storing it as a hidden state. The encoder hidden state is used to train the decoder to predict the next word in the sequence based on the previous hidden state word[12]. During the inference phase, the sequence to sequence model is tested with new sequences for which the input summary sequence is unknown [13].

D. ROUGE score

The acronym ROUGE stands for Recall Oriented Understudy of Gisting Evaluation, and it is a text summarization metric. It compares the n-gram matches in reference summaries generated by the system and those generated by humans. The ROUGE score is computed by adding the precision, recall, and f-measure values. The ROUGE-1 score evaluates the unigram overlap between the system-generated and human-generated reference summaries [4]. ROUGE - 2 evaluates overlapping bigrams similarly to ROUGE - 1. ROUGE-1 is thought to have the highest accuracy in detecting overlapping words among the various ROUGE -n scores.

Recall = The phrases that overlap in the human reference summary/the total word count.

Precision = total phrases in the system summary/number of overlapping phrases.

III. RESULTS AND DISCUSSIONS

The steps involved in creating the summary are described in this section. To begin, raw data input is collected, followed by text pre processing to produce washed and organised words. At last, the cleaned text

is summarised using TFRSP and the machine learning algorithm.

A. Dataset Collection

There are two ways to obtain sources of data. It could be a standard format, such as a.csv file from Kaggle or another set of data gathering website. Another method of gathering datasets is content, which extracts raw data from websites. Kaggle [18] is used to obtain the Amazon product review dataset, which is then compared to a shoes review dataset. The dataset is pre processed, and the summary is produced by combining unsupervised and supervised techniques.

B. Pre-processing steps

The raw input dataset goes through several pre-processing stages. First, the noisy input is examined for duplication and void values. The records is sent to the tokenization process after the null values have been removed. Tokenization is process of breaking down a large text document into sentences, which are then broken down into words. Stop words, numbers, punctuation, and special symbols are removed from the word collection [16] [20]. The extracted keywords are then lemmatized, a technique for determining the root word. Because lowercase letters have higher ASCII values than uppercase letters, all uppercase letters are converted to lowercase letters. The text is cleaned up and the words are converted to lowercase.

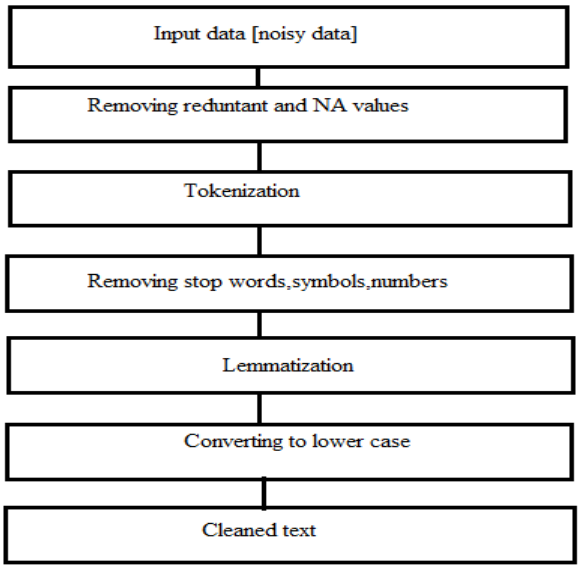

Fig 2. Text pre-processing

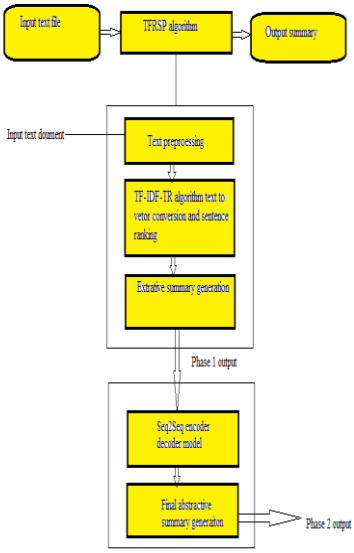

Fig 3. Unsupervised and Supervised phase

The ROUGE score is used to compare the performance of the TFRSP method to that of other existing summarization methods. Rouge is the package that is used to compute the ROUGE score. ROUGE is an evaluation metric for summary accuracy that compares human-generated reference summaries to system-generated summaries. As shown in Table 5.1, the ROUGE-1 score is used to analyze the performance of Amazon product review summarization, which includes accuracy, as well as recollect.

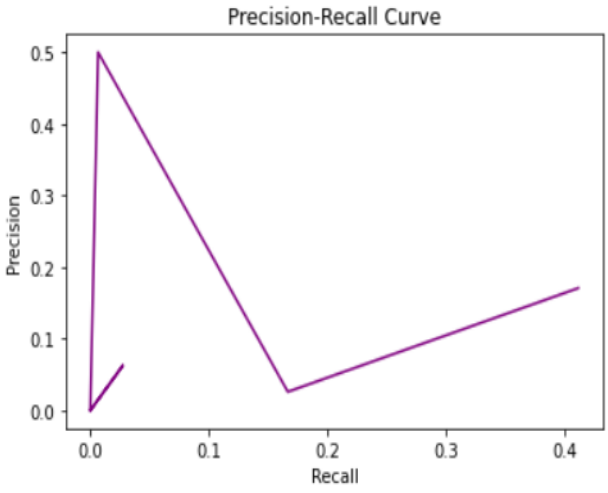

Fig 4. Precision Recall graph

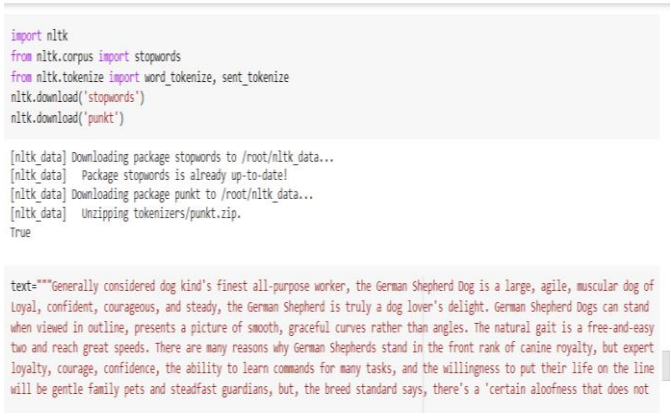

Fig 5. Summarization of the text

The necessary libraries and packages are imported and downloaded in the above screenshot, and we provide the input text to be summarized. The stop words are then removed, and a frequency table is created. A frequency table counts the number of times a specific word appears in our data. A frequency table displays a set of values as well as how frequently they occur. They assist us in determining which data values are common and which are uncommon.

Fig 6. Summarized text

The method sent tokenize is used in the above screenshot to tokenize the sentences, which are broken down into words or tokens. After all of this, the required summary is generated without changing the text's actual meaning.

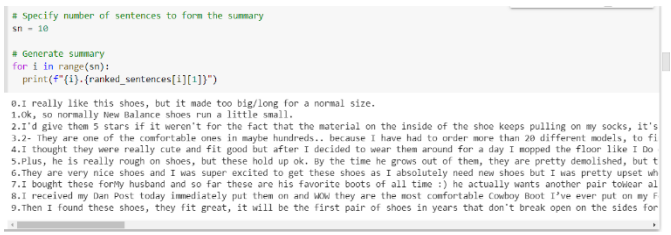

Fig 7. Summary of top 9 sentences

A summary of the top nine sentences is generated in the above screen shot after applying the text rank algorithm.

Fig 8. Result of Extractive summarization

The above screen shot shows an extractive summary of the dataset, which is then subjected to abstractive summarization.

Fig 9. Result of abstractive summarization

We import the necessary packages and libraries for abstractive summarization and use the LSTM approach, which employs two methods. The encoder encodes the data and keeps it hidden, while the decoder decodes and provides the precise summary.

IV. CONCLUSION

The combined overseen and unsupervised summary results in a 38.42 percent increase in the ROUGE score of the existing methods. The proposed approach could be improved further by combining classification techniques like Naïve Bayes, Decision tree, and so on with the TF-IDF. The proposed method could be conducted on various datasets, and the accuracy of the generated summary can be improved the epochs.

V. REFERENCES

[1]. Meena S M, Ramkumar M P, Asmitha R E. " Text Summarization Using Text Frequency Ranking Sentence Prediction." 2020 4th

International Conference on Computer, Communication and Signal Processing (ICCCSP).

[2]. Parmar, Chandu, RanjanChaubey, and Kirtan Bhatt. "Abstractive Text Summarization Using Artificial Intelligence." Available at SSRN 3370795 (2019).

[3]. Kim, Joo-Chang, and Kyungyong Chung. "Associative feature information extraction using text mining from health big data." Wireless Personal Communications 105, no. 2 (2019):691-707.

[4]. Qaiser, Shahzad, and Ramsha Ali. "Text mining: use of TF-IDF to examine the relevance of words to documents." International Journal of Computer Applications 181, no. 1 (2018):25-29.

[5]. Roul, Rajendra Kumar, and JajatiKeshariSahoo. "Sentiment Analysis and Extractive Summarization Based Recommendation System." In Computational Intelligence in Data Mining, pp. 473-487. Springer, Singapore, 2020.

[6]. Dutta, Madhurima, Ajit Kumar Das, ChirantanaMallick, ApurbaSarkar, and Asit K. Das. "A Graph Based Approach on Extractive Summarization." In Emerging Technologies in Data Mining and Information Security, pp. 179-187. Springer, Singapore,2019.

[7]. Dutta, Madhurima, ChirantanaMallick, ApurbaSarkar, and Asit K. Das. "A Graph Based Approach on Extractive Summarization." In Emerging Technologies in Data Mining and Information Security, pp. 179- 187. Springer, Singapore,2019.

[8]. Nallapati, Ramesh, Bowen Zhou, CaglarGulcehre, and Bing Xiang. "Abstractive text summarization using sequence-to-sequence rnns and beyond." arXiv preprint arXiv:1602.06023(2016).

[9]. Jasmeet singh, Prbjot singh, Prateek chikkara. “An Ensemble Approach for extractive text summarization.” 2020 International conference on ETITE.

[10]. Gupta, Vanyaa, NehaBansal, and Arun Sharma. "Text summarization for big data: A comprehensive survey." In International Conference on Innovative Computing and Communications, pp. 503-516. Springer, Singapore, 2019.

[11]. Applications of automatic summarization : https://blog.frase.io/20- applications-of-automatic-summarization-in-the-enterprise/.

[12]. ShanmugasundaramHariharan. "Studies on intrinsicsummary evaluation", International Journal of ArtificialIntelligenceand Soft Computing, 2010.

[13]. Bhavadharani, M., M. P. Ramkumar, and Selvan GSR Emil. "Performance Analysis of Ranking Models in Information Retrieval." In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), pp. 1207-1211. IEEE,2019.

[14]. Pan, Suhan, Zhiqiang Li, and Juan Dai. "An improved TextRank keywords extraction algorithm." In Proceedings of the ACM Turing Celebration Conference-China, pp. 1-7.2019.

[15]. Mihalcea, Rada. "Graph-based ranking algorithms for sentence extraction, applied to text summarization." In Proceedings of the ACL Interactive Poster and Demonstration Sessions, pp. 170-173.2004.

[16]. Roul, Rajendra Kumar, and JajatiKeshariSahoo. "Sentiment Analysis and Extractive Summarization Based Recommendation System." In Computational Intelligence in Data Mining, pp. 473-487. Springer, Singapore, 2020.

[17]. Song, Shengli, Haitao Huang, and TongxiaoRuan. "Abstractive text summarization using LSTM-CNN based deep learning." Multimedia Tools and Applications 78, no. 1 (2019):857-875.

[18]. "Advances in Computational Intelligence", SpringerScience and Business Media LLC,2019,

[19]. Understanding Encoder - Decoder Sequence to sequence model :

https://towardsdatascience.com/understanding-encoder-decoder- sequence-tosequence-model-679e04af4346 .

[20]. Text Summarization using Sequence to sequence encoder decoder model: https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-textsummarization-using-deep- learning-python/ .

[21]. Kaggle Dataset :https://www.kaggle.com/skathirmani/amazon-reviews [19] "Natural Language Processing and ChineseComputing", Springer Science and Business MediaLLC,2018. [22] Python 3 Jupyter Notebook : https://jupyter.org/.

[22]. ROUGE :https://pypi.org/project/rouge.

Cite this article as :

Dr. Vidyagouri B H, BibiSadiqa M D, "Text Summarization Using Machine Learning Algorithm", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 4, pp. 167-173, July-August 2022. Available at doi : https://doi.org/10.32628/CSEIT228421
Journal URL : https://ijsrcseit.com/CSEIT228421