Text Summarization Using Text Frequency Ranking Sentence Prediction

Meena S M
Department of Computer Science and Engineering
Thiagarajar College of Engineering
Madurai, India
meena59807@student.tce.edu

Ramkumar M P
Department of Computer Science and Engineering
Thiagarajar College ofEngineering
Madurai,
Indiaramkumar@tc.edu

Asmitha R E
Department of Computer Science and Engineering
Thiagarajar College of Engineering
Madurai, India
asmitha59613@student.tce.edu

Emil Selvan G SR
Department of Computer Science and Engineering
Thiagarajar College of Engineering
Madurai, India
emil@tce.edu

2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP) | 978-1-7281-6509-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCCSP49186.2020.9315203

Abstract— **In the era of information technology, data plays significant role. The data which prevails on the internet are unstructured and are not in a concise manner. To make the raw data into a structured, readable, coherent and concise and to extract the summary of data, the text summarization concept is introduced. The text summarization involves in providing asummary of the useful information from the raw data without dissolving the main theme of the data. Nowadays readers face the challenge of reading comments, reviews, news articles, blogs, etc., as they are too informal and noisy. Retrieving the correct gist of the text which is necessary for all the readers is a quite difficult task. In order to overcome the problems faced by the readers, TFRSP (Text Frequency Ranking Sentence Prediction) algorithmis proposedto generate a precise summary that uses supervised and unsupervised learning algorithms. The proposed approach uses the combination of TF-IDF-TR (Term Frequency – Inverse Document Frequency – Text Rank) as an unsupervised learning algorithm and Seq2Seq (Sequence to Sequence) model as a supervised learning algorithm to obtain the benefits of both extractive and abstractive summarization. The results of the proposed TFRSPapproach is compared with the existing methods of text summarization using the Recall Oriented Understudy of Gisting Evaluation (ROUGE) and attains a high ROUGE score, hence achieves high accuracy ofsummary.**

Keywords— Text Summarization, Natural Language Processing, Extractive Summarization, Unsupervised learning, Abstractive Summarization, Supervised learning, ROUGE

I. INTRODUCTION

The digitalized world makes rapid growth in the field of information retrieval. People are relying on a variety of resources to stay updated. Considering the time as a factor, people want the information to be in a short and precise manner. There exists a major issue while reading the news articles and online reviews or feedbacks which are hard to conclude unless they are completely read. This leads to the evolution of the text summarization concept for the betterment of information retrieval. Text summarization is the concept of extracting the main corpus of information as a summary from the original text in a brief, orderly and human interpreted manner [1]. Automatic text summarization uses the ideas of Natural Language Processing (NLP) to obtain the summary systematically. Automatic text summarization generates a human interpreted summary in the form of a system-generatedsummary.

Text summarization is classified as shown in fig.1, namely extractive and abstractivesummarization:

1. Extractive summarization: It extracts the most relevant and significant sentences which are actually the subclass of the sentences from the original text [2].Extractive summarization is similar to highlighting the prominent sentences from the original textdocument.

2. Abstractive Summarization: It extracts the main gist of the original text and generates the summary with its own words. Abstractive summarization is equivalent to recreating the original text with new phrases and produces the summary[2].

Both the extractive summarization and abstractive summarization has benefits and drawbacks of its own. Extractive summarization picks out the sentences that are vital and correct but has the problem of incoherency while abstractive summarization generates word-by-word summary that will be crisp and in readable format but it may lead to loss of key facts in case of huge documents.

The proposedTFRSP algorithm deals with the integration of both extractive summarization and abstractive summarization which uses supervised and unsupervised learning algorithms.The Term Frequency-Inverse Document Frequency (TF-IDF) is used in combination with the Text Rank (TR) algorithm in extractive summarization. It is an unsupervised learning technique that does not have the need to supervise the model, which is modified as TF-IDF-TR in the proposed TFRSP approach. Abstractive summarization involves the sequence to sequence (seq2seq) model which is a supervised learning algorithm that includes training and testing datasets. TheTFRSP method collects the dataset from Amazon Product reviews which is preprocessed and the TF-IDF-TR algorithm is used to generate the extractive summary in the first phase which is then fed into the second phase abstractive seq2seq model as an input to obtain a more precise summary, hence obtaining an effective summary when the performance is calculated using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE)score.

The key factor of choosing the TFRSP algorithm is to combine the strengths of both abstractive and extractive summarization techniques. Extractive summarization works on huge documents and can produce the summary in the form of important corpus retrieval. But abstractive summarization works only on lesser documents but still can produce the summary with higher accuracy in the form of human generated summary. When these two different approachesare

978-1-7281-6509-7/20/$31.00 ©2020 IEEE

combined the resulting summary could be more accurate. The ROUGE score evaluates the summary produced by the system with the human-interpreted referencesummary.

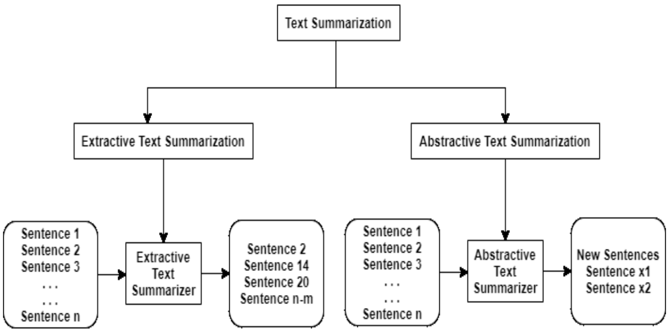


Fig. 1. Extractive and Abstractive Text Summarization

The key benefit of text summarization is to minimize the reading time of the end-users.Text summarization plays an important part throughout the fields of medicinal, news articles, financial and legal document analysis, blogs, literature, online reviews, etc. [3]

II. RELATEDWORK

The literature survey for the text summarization is briefly explained in the followingsegment. The main discussion is associated to the supervised and unsupervised algorithms and evaluation metric [4] which is used to evaluate the summary generated.

A. Term Frequency (TF) - Inverse Document Frequency (IDF)

The algorithm used in the extractive summarization is TF-IDF. Here, TF represents Term Frequency in which the frequency of the words is counted. The frequency acquired is used to find the importance of the word. Higher the frequency of the word higher is the importance of the word in the document [5] [6]. The simplest way of explaining TF is, it measures the frequent occurrence of the word in the document [7]. IDF represents the Inverse Document Frequency which allots a higher value to the rare words and lower value to the recurrent words. At times TF miscalculates the stop words as important as of their frequent occurrence. To rectify the issue faced by TF, IDF identifies the rare occurrence of words in the document. IDF denotes the inverse version ofTF; both when combined produce TF-IDF which is the multiplication of TF and IDF. The formula for TF-IDF is mentioned below [5][7]:

$$tf(i,j) = term(i)\ within\ the\ document\ j$$
$$(i,j) = \log \frac{N}{df_i} \quad (1)$$
$$TFIDF(i,j) = tf(i,j) \times idf(i,j)$$

In (1), (i, j)represents the frequency of the word $i$in the document j[7]. N denotes the number of documents in the dataset and df_idenotes the documents containing the word at least once. The higher is the value of $df_i$when the word is frequently used in multiple documents. TF–IDF is the value of the word $i$in the document j of the document N[5].

B. Text RankAlgorithm

The text rank is an unsupervised algorithm which is used for ranking the sentences with the help of weights as a value. Text rank algorithm has its base origin from Google's page rank algorithm which ranks the pages based on its hyperlinks and its importance[8]. A directed graph is constructed with the help of sentences as of its name graph-based ranking algorithm.The sentences are considered as nodes or vertices and the similarity between two nodes is connected with the help of edges[9].Text rank algorithm isa recommender based algorithm in which the importance of the sentences is recommended by the vertices connected by the edges within the graph.

C. Sequence to SequenceModel

The proposed TFRSP algorithm involves an abstractive summarization algorithm which is termed as sequence to sequence model which is useful in creating the new phrases by retaining the meaning of the source document. The Sequence to sequence model is first introduced by Google which powers applications like Google translate, image captioning, text summarization, online chat bots,etc. It is an encoder-decoder based model that maps the sequences of different lengths of input and output to each other [10].The encoder-decoder component has the subcomponent Long Short Term Memory(LSTM) that is useful in capturing long term dependencies. The encoder-decoder model consists of 2 phases - training and inference phases [11]. The encoder and decoder aremeantfor both the training and inference phase. In the training phase, the encoder reads the entire input word by word and processes the information present in the input sequence and stores it as a hidden state. The hidden state from the encoder is meant to initialize the decoder and is trained to predict the next word in the sequence from the previous hidden state word[12].In the inference phase, the sequence to sequence model is tested with new sequences for which the target summary sequence will not be known[13].

D. ROUGEscore

ROUGE - Recall Oriented Understudy of Gisting Evaluation is a metric for text summarization. It is used to measure the n-gram matches of the system-summary retrieved from text summarization with the reference summary generated by the humans. ROUGE measure includesprecision, recall, and f-measure values which when combined will yield the ROUGE score. ROUGE-1 score evaluates the overlapping of unigrams from the system generated summary with the human-generated reference summary [4]. ROUGE -2 evaluates the bigrams overlapping in the similar way as

ROUGE -1. Among the various ROUGE -n scores ROUGE-1 is considered as having the highest accuracy in finding outthe overlapping words.

$$Recall = \frac{Number\ of\ wordsoverlapped}{Total\ words\ in\ human\ reference\ summary} \quad (2)$$

$$Precision = \frac{Number\ of\ words\ overlapped}{Total\ words\ in\ system\ summary} \quad (3)$$

$$F - measure = 2 * \frac{ision \times Recall}{Precision + Recall} \quad (4)$$

In (2), Recall refers to the extent to which the reference summary is related to system summary. In (3), Precision refers to the extent to which the summary generated by the system was relevant [1]. In (4), F-measure is the balanced score or harmonic mean of Precision (2) and Recall (1) [7].

E. LiteratureReview

The studies regarding the various techniques in text summarization are performed and it is explained in the section. The algorithms for text summarization which are proposed earlier are discussed in the forthcoming section.

In [5], Joo-Chang Kim and Kyungyong Chung proposed the associative feature extraction in health data. In the paper text preprocessing for the health big data is done.TF-IDF is implemented to find out the most relevant words from the document. Along with that TF-C-IDF is incorporated for associative feature extraction purpose from the retrieved corpus obtained in TF-IDF. The associative keywords are further analyzed with the help of the Apriori algorithm. Apriori algorithm is a data mining algorithm in which the associative rules are designed for large data relations in large data sets. The main advantage of the proposed method is to digitalize the health data and to extract the proper associative keywords.

In [14], ShahzadQaiser and Ramsha Ali discussed mainly on the TF-IDF's benefits, drawbacks and solutions for better and improvised algorithms.TF-IDF is a straight forward, easy algorithm to deploy and the most appropriate information from the data is retrieved. But at times the algorithm used in the paper cannot find out the most prompt word due to slight changes in the tense forms. In order to overcome the issues faced by TF-IDF a new approach of TF-IDF is proposed which involves the techniques of the combination of classification algorithms with TF-IDF.

In [15], Rajendra Kumar Roul and JajatiKeshariSahoo proposed the concepts of sentiment analysis with the extractive summarization. But the paper emphasized mainly on the extractive summarization techniques. The hierarchical summarization which involves the four main algorithms like Textrank, Lexrank, Latent Semantic Analysis (LSA) and sumbasic are introduced and found that the functionality of the Textrank algorithm is far better when compared with other algorithms.

In [16], Madhurima Dutta et al. have proposed an extractive summarization approach using the concepts of graph theory. Infomap clustering algorithm is used to cluster the sentences and the similarities among the sentences are found by constructing a graph.

In [1], ChanduParmaret al. have compared the algorithms related to the abstractive summarization. A comparison is done with sequence to sequence model and Long Short Term Memory (LSTM) bidirectional and according to the approach used in the paper, it is found that the summarization produced from the sequence to sequence model is more efficient when compared with LSTM bidirectional. The amazon reviews and CNN dataset are used for effective comparison and for better implementation of algorithms.

In [17], Ramesh Nallapati et al. have proposed a model on abstractive text summarization using Attentional Encoder-Decoder Recurrent Neural Network. The algorithm showed promising results with the multi-sentence document.

III. PROPOSEDSYSTEM

The steps that are involved in generating the summary are discussed in the section. Firstly, datasets are collected as the input raw data and secondly the preprocessing of the text is done to obtain the cleaned and structured text. Finally, the summarization of the cleaned text is performed using TFRSP algorithm.

A. DatasetCollection

There are two ways to obtain datasets. It can be a well-defined format as .csvfile which is obtained from Kaggle or other dataset collection websites. Another way of dataset collection is done through web-scraping which extracts raw data from the websites. Here the Amazon Product Reviews dataset is collected from Kaggle [18]. The dataset is preprocessed and then the summary will be generated through combined unsupervised and supervised techniques.

B. Preprocessing

The raw input data collected from different sources are preprocessed through various libraries in Natural Language Processing (NLP) [19]that is used for string or text processing. Manipulation of large amounts of natural data is done using NLP. Text summarization has a set of activities which involvesNLP.

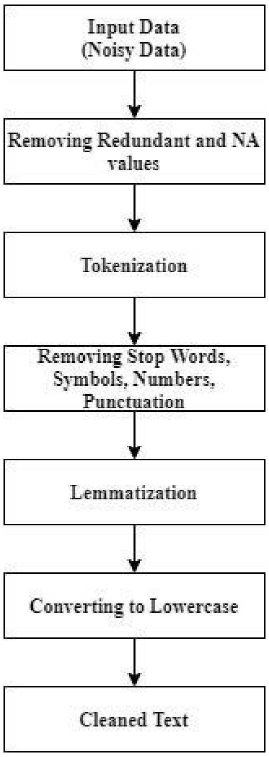

Input Data
(Noisy Data)
Removing Redundant and NA values
Tokenization
Removing Stop Words, Symbols, Numbers, Punctuation
Lemmatization
Converting to Lowercase
Cleaned Text

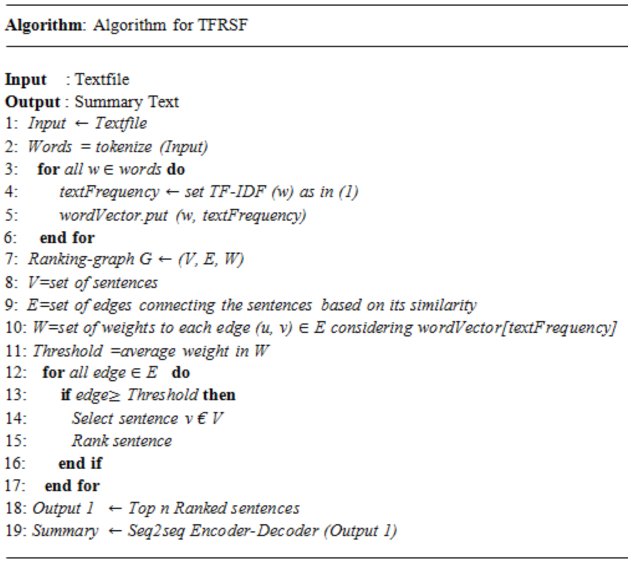
Fig. 2. Text Preprocessing

The raw input dataset undergoes various preprocessing stages. Firstly the input noisy data is checked for redundancy and null values. The null values are removed and the data is sent for the process of tokenization in which the whole text document is tokenized into sentences and then to words. From the collection of words obtained, stop word removal is performed [16] [20]. It is followed by lemmatization which is a technique to obtain the root word from theextracted keywords. Lemmatization is a process of obtaining a dictionary word which lags in stemming process. The words are converted into lowercase and cleaned text is obtained. The preprocessing steps are shown in the flow in the fig.2.

The TFRSP algorithm used to obtain a concise and meaningful summary is shown in the fig. 3. The TFRSP algorithm consists of two phases namely unsupervised phase and supervised phase as represented in the fig. 4.

C. UnsupervisedPhase

The data after preprocessing is fed into the stream of extractive summarization where the TF-IDF-TR algorithm is applied over the preprocessed text document. Here the concept of TF-IDF is used where the most frequent words are extracted and the ranking of sentences is done using the text rank algorithm. Using the concept of TF-IDF-TR the coherency of the summary is maintained as the sentences are ranked based on the similarity of the sentences. The weighted graph is constructed after the TF-IDF algorithm and the wordsareconsidered as the vertices in the graph. Text rank algorithm uses the cosine similarity matrix for finding the most important words or sentences with the help of similarity and relevancy for the sentences. But in TFRSP algorithm Text rank algorithm is combined with TF-IDF to produce the most import sentences based on ranking algorithm along with term frequencycalculation.



Algorithm: Algorithm for TFRSF

```
Input  : Textfile
Output : Summary Text
Input ← Textfile
Words = tokenize (Input)
  for all w ∈ words do
      textFrequency ← set TF-IDF (w) as in (1)
      wordVector.put (w, textFrequency)
  end for
Ranking-graph G ← (V, E, W)
V=set of sentences
E=set of edges connecting the sentences based on its similarity
W=set of weights to each edge (u, v) ∈ E considering wordVector[textFrequency]
Threshold =average weight in W
  for all edge ∈ E  do
      if edge≥ Threshold then
        Select sentence v € V
        Rank sentence
      end if
  end for
Output 1 ← Top n Ranked sentences
Summary ← Seq2seq Encoder-Decoder (Output 1)
```

Fig. 3. Algorithm for TFRSP

D. SupervisedPhase

The summary extracted from the extractive summarization in the unsupervised phase is fed as an input to the abstractive summarization phase in which the unsupervised algorithm sequence to sequence model is used. As the size of the original document is reduced after extractive summarization it is a benefit for abstractive summarization. The sequence to sequence model is applied over the summary. The model is initially trained with the reference summary datasets. The previous output of the extractive summarization is considered to be the hidden state in the encoder-decoder model. The final summary is predicted with the help of the hidden state and the trained reference model. The summary obtained from the abstractive summarization is more precise and the loss of datais minimized as the initial process of extractive summarization is done to consider the most important facts.

The complete working of the TFRSP approach for summary generation is shown in the figure 4. The input text document is preprocessed and extractive summary is obtained from the unsupervised phase. Then the output of phase 1 is fed into the supervised phase which generates the final summary as of shown in figure4.

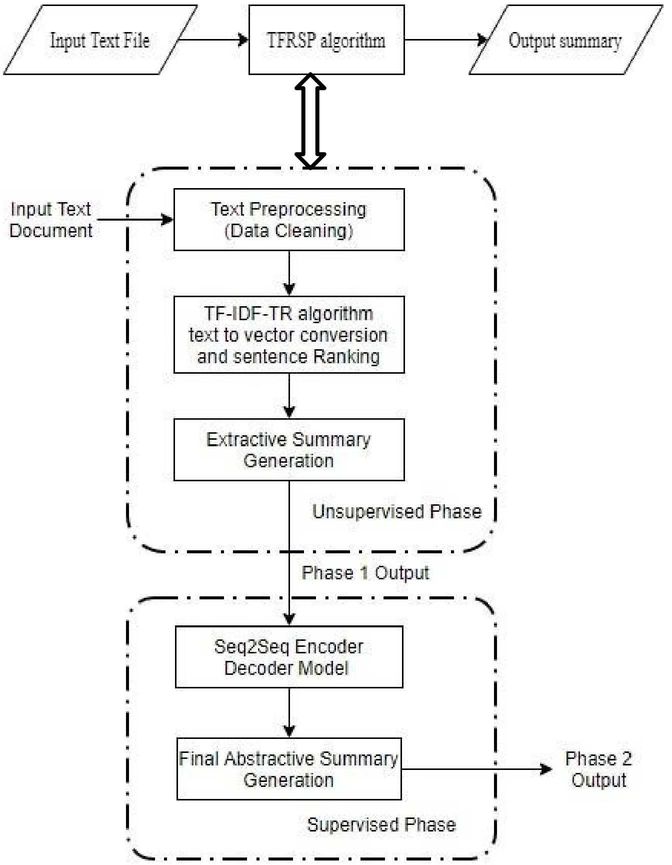

Input Text File
TFRSP algorithm
Output summary
Input Text Document
Text Preprocessing (Data Cleaning)
TF-IDF-TR algorithm text to vector conversion and sentence Ranking
Extractive Summary Generation
Unsupervised Phase
Phase 1 Output
Seq2Seq Encoder Decoder Model
Final Abstractive Summary Generation
Phase 2 Output
Supervised Phase

Fig. 4. Unsupervised and supervised phases of TFRSP

IV. RESULTS ANDDISCUSSION

The integrated approach for summary generation is implemented in the Python 3Jupyter Notebook [21]. The original review text given as input to the first phase of the proposed TFRSP algorithm to produce the phase 1 summary as an intermediate result and it is then served as input to the next phase of TFRSP algorithm to generate the final output summary as shown in fig.5.

Original Review Text: *I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The product looks more like a stew than a processed meat and it smells better. My Labrador is finicky and she appreciates this product better thanmost.*

Phase 1 Summary:*The product looks more like a stew than a processed meat and itsmells better.*

Final summary: *great product.*

Fig. 5. Example of sample review text and its summary

The performance of the TFRSP method is compared with the other existing summarization methods [22] using the ROUGE score. The package installed in calculating the

ROUGE score is rouge-0.3.2[23]. ROUGE measure is a summary accuracy evaluation metric in which it compares the human-generated reference summary with system generated summary. The performance for the Amazon product review summarization is analyzed using the ROUGE-1 score which comprises F-measure, precision, and recall as mentioned in the Table1.

TABLE 1. Performance Comparison of existing algorithms with proposed TFRSP algorithm using ROUGE score

Algorithms	**ROUGE – 1 Score**		
	F-measure	**Precision**	**Recall**
Extractive (Text Rank)	0.051387	0.027464	0.077614
Abstractive (Seq2Seq)	0.121599	0.149289	0.106128
TFRSP Method	**0.248323**	**0.287440**	**0.204339**

From the experimental results obtained in the Table 1, it is found that the ROUGE score for the integrated approach of the text summarization using TFRSP algorithm is higher when compared with the existing extractive and abstractive methodologies separately as shown in the fig. 5. The proposed approach using TFRSP algorithm generates a precise summary which is similar to a human interpreted summary.

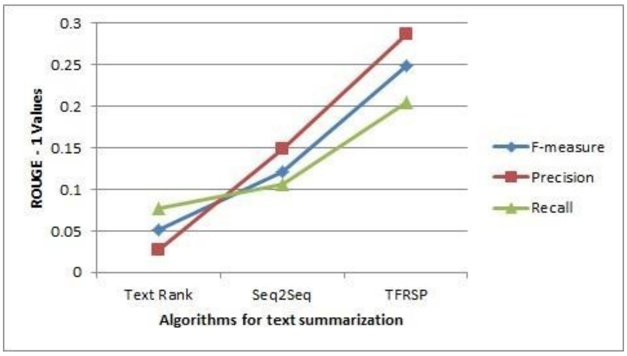


Fig. 5. Performance analysis of the experimental results

V. CONCLUSION

The TFRSP algorithm generates a summary for the Amazon product reviews combining the techniques of unsupervised (extractive summarization) and supervised (abstractive summarization) algorithms, producing an integrated approach with the increased accuracy of 87.58 % when compared with the traditional methods of the text summarization. The summary generated from the combined supervised and unsupervised learning results in 38.42 %increase in the ROUGEscore ofthe existing methods. The proposed method could be further improved by combining the classification techniques such as Naive Bayes, Decision tree, etc. along with the TF-IDF. The proposed method could also be tested for various datasets and the accuracy of the summary generated can be increased by increasing theepochs.

REFERENCES

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

[2] 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.

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Mallick, Chirantana, Ajit Kumar Das, MadhurimaDutta, Asit Kumar Das, and ApurbaSarkar. "Graph-based text summarization using modifiedTextRank."InSoftComputinginDataAnalytics,pp.137-146. Springer, Singapore, 2019.

[10] 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.

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

[12] Understanding Encoder - Decoder Sequence to sequence model : https://towardsdatascience.com/understanding-encoder-decoder- sequence-to-sequence-model-679e04af4346

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

[14] 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.

[15] 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.

[16] 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.

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

[18] Kaggle Dataset :https://www.kaggle.com/skathirmani/amazon-reviews

[19] "Natural Language Processing and ChineseComputing", Springer Science and Business MediaLLC,2018

[20] Bhavadharani, M., M. P. Ramkumar, and Emil Selvan GSR. "Information Retrieval in Search Engines Using Pseudo Relevance Feedback Mechanism." In 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), pp. 1-5. IEEE,2019.

[21] Python 3 Jupyter Notebook : https://jupyter.org/

[22] "Computational Intelligence in Data Mining", Springer Science and Business Media LLC,2020

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