Perspectives in Communication, Embedded-Systems and Signal-Processing (PiCES) – An International Journal
ISSN: 2566-932X, Vol. 5, Issue 5, August 2021

# Abstractive Multi Document Text Summarization of User Reviews Using Graph Generation and TF-IDF

Soma Shrenika

Student, Computer Science and Engineering, Institute of Aeronautical Engineering, Hyderabad, Telangana, India, somashrenika@gmail.com

***Abstract: With the increase in number of e-commerce sites, one finds it difficult to choose and buy a product. It is not possible for a person to read hundreds of product reviews from various sources. This problem can be solved by using text summarization. The issue with most text summarizers is that, they summarize the text but they do not tend to preserve the underlying meaning. The aim of this paper is to overcome this problem by developing a abstractive multi document text summarizer which summarizes reviews that are obtained from multiple sources like amazon, trip advisor, etc. and are stored in the form of multiple documents. The implementation is carried out by first cleaning the data and using a graph based approach that considers tagging parts of speech and building edges with weights and then using TF-IDF to find the most important set of words. The final summary is obtained by using maximum weight graph traversal. Evaluation metric is ROUGE.***

***Keywords: Text Summarization; Abstractive Multi Document Text Summarize; Graph Based; TF-IDF; ROUGE.***

## I. INTRODUCTION

The data available on the internet is enormous. Most of which consists of unstructured data Extracting brief information from this data is becoming a tedious task. For this purpose, text summarization is used. Text summarization is a process of generating summary of given information without any change in the definition and connotation. The information can in be in various forms like news articles, product reviews, documents, social media posts, etc. Let us consider a scenario where a user is interested in obtaining specific information from various sources, it will take huge time in the removal of unwanted information. Summarizing it will thereby amplify the comprehensibility and saves time.

Text summarization is a part of NLP. Automation of this process will produce summaries without the requirement of human intervention. As computer does not understand human language and the sentiment associated with it, summarization be-comes a difficult work. To overcome this, many machine learning models are used. These models are trained in a way so as to comprise important information. Text summarization involves one document or on many documents. Single document text summarization contains single input document and produces single output summary. Multi documents text summarization consists of multiple input documents of the same main concept and it produces a single output summary.

Text summarization came into existence in the year 1958. The extractive and abstractive methods depend on domain. The ideology turned toward multiple documents in the early 2000's. The summarization task is performed by taking into account key features. These include, obtaining the frequency of occurrence in a sentence, presence of a word or sentence in a particular place in a sentence or paragraph, the sentiment of the sentence, vocabulary and parts of speech in which the priority is given to nous. In abstractive method, the summarization can be achieved by using structure and semantics. In this approach, the words from original document may or may not be present in the final summary.

## II. LITERATURE SURVEY

Ahmad T Al-Taani [1] et al. demonstrated the usage of extractive approach. In statistics based methods, emphasis is laid on mathematical values like ranking of sentences based on importance, occurrence, length and frequency. In graph methods, a graph is created using text data from small to huge structures. All the data present is interconnected to each other. Then summaries are generated from them. Using machine learning methods help us in tackling this process by using algorithms that are primarily focused on features. In clustering methods, the data is segregated into clusters and data for summaries are obtained from these clusters. These depend on selection of clusters.

Taner Uçkan [2] et al. proposed a method which uses independent sets. It takes into consideration only those sentences that are not a part of the independent sets. It was performed as three step processes. First, all the prepositions were discarded. Then a graph is constructed based on the mathematical values of dependence and

independence. Finally, Eigen vectors are used to generate the final summary by considering weights of nodes.

Shai Erera [3] et al. developed a system for summarizing scientific documents of computer science domain. The system accepts input either in the form of a query or by defined tasks. Evaluation was performed by comparing the summary with that written by human summaries. Most of the existing systems focus either on news or other data. But a system for scientific documents needs to focus on other important parameters like data implying figures and the underlying context. As research papers contain different sections, generation of a small summary of the entire paper is very difficult. So, Each and every section of the paper is individually summarized to reduce ambiguity and irrelevant information. After pre-processing the data, summarization task is carried out by using extractive method and follows the bag of words model.

Min Yang [4] et al. presented a novel abstractive summarization process which is closely related to how humans comprehend a particular information in proposed. Hierarchical deep neural networks are used to produce the summary. The implementation is divided into 3 parts which consists of different network modules. Hu-mans first skim the information, then give a detailed reading and then come to conclusion. The same implementation is applied here. Sequence to sequence model is used by using classification of the data. In the first step, the input information is converted to a knowledge based data. Then the meaning of the text is extracted. Then, a framework is used to enhance the summarizer. The obtained summary when compared to human written summaries. The evaluation parameters indicated that the model is performing the task with great accuracy.

V. Mohan Kalyan [5] et.al proposed, summarization of a document is performed by using the sentences or words that are already present in the document without the use of additional text. Here the occurrence of words is key feature. Weights are assigned to the occurrence. Higher the weight, higher the probability of the words being present in the final summary. Input can in the form of a text, a Uniform Re-source Locator or a file containing information. The data is then cleaned and processed by calculating term weights and the final step is to add them to the final summary. A user interface was provided to implement summarization.

## III. DATA DESCRIPTION

For the purpose of multiple document text summarization, we have used “Opiniosis” dataset. The dataset consists of user review data. There are almost 100 reviews on each topic and there are 51 topics. These reviews were obtained from various ecommerce websites. It also has 4 handwritten summaries for each topic. These are used for the process of evaluation.

### *A. Natural Language Processing (NLP)*

NLP deals with diminishing the barriers between human beings and computers. It aids machines to process information in a way similar to that of humans. It is a part of a huge field named artificial intelligence. This processing has gained much importance in recent years. This field came into existence in 1950’s. It is considered as one of the difficult tasks in machine learning. There are many problems that are solved by using NLP. It is employed in almost all application we use in our daily life. Some of them include virtual personal assistants like Alexa, Siri, Google Assistant, applications like Grammarly which predict the text quality and sentiment of the sentence. There are two major analysis methods with deal with syntax and semantics of information. As the name suggests, in syntax analysis importance is given to formation, position and vocabulary. Whereas, in semantic analysis the idea mainly depends on the meaning of the information.

## IV. METHODOLOGY

Below is the block diagram.

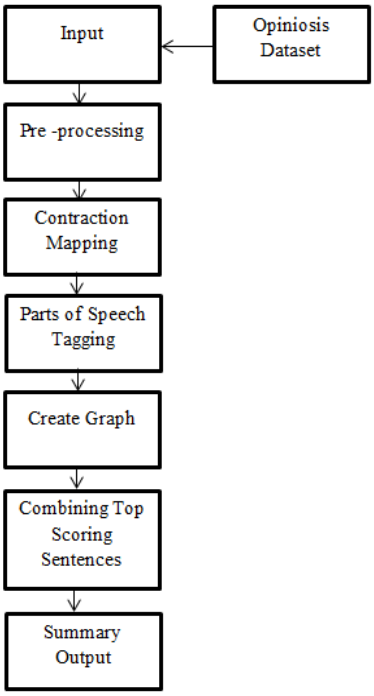


Fig 1. Block Diagram for Proposed Methodology

### *A. Input*

Input to the system are text documents obtained from the opiniosis dataset. It consists of 100 reviews given by different people. All of them vary with regard to sentiment and opinions. End of each review is marked by using a period.

### *B. Pre-Processing*

It is nothing but removing unwanted data by cleaning it. First the entire data is converted into lower case. Newline characters, tab spaces, comma, carriage return character, etc. are removed. All the stop words defined in nltk library are imported are removed and each sentence is converted into a list of tokens. A sample of tokenizing and converting to lower case from amazon kindle reviews is given below:

Input: By the way, Kindle battery lasts forever indeed ! Output: ['by', 'the', 'way', 'kindle', 'battery', 'lasts', 'forever', 'indeed']

### *C. Contraction Mapping*

It is used to convert the contractions into their extended form. A dictionary is de-fined which contains the key and value pairs for the above. Some examples include, {"ain't": "is not", "aren't": "are not","can't": "cannot", "'cause": "because", "could've": "could have", "couldn't": "could not}

### *D. Parts of Speech Tagging*

Assigning each word with its parts of speech is termed as tagging parts of speech. We implement this extract nouns from the text. These are represented as starting nodes in the graph. These are used to understand the semantics of the data and help in knowing the frame of reference.

Input: ['by', 'the', 'way', 'kindle', 'battery', 'lasts', 'forever', 'indeed']

Output: [('by', 'IN'), ('the', 'DT'), ('way', 'NN'), ('kindle', 'JJ'), ('battery', 'NN'), ('lasts', 'VBZ'), ('forever', 'RB'), ('indeed', 'RB')]

### *E. Create Graph*

A directed graph is constructed for the vocabulary in reviews. All the linking verbs are hubs. The occurrence of words is found and sorted in ascending order. An edge is added for every two words in a sentence. If those two words appear again, then the weight of edge is incremented by one. Add weights to each edge based on occurrence relationship between two words. We find the word number and its aver-age position

### *F. Combine Top Scoring Sentences*

TF-IDF is used to find the most important set of words . We apply sub linear term frequency by replacing term frequency by 1+log(term frequency). Then we fit each and every review and then transform it. Then we combine the sentences by obtaining starting point from the graph and traversing the neighbours by using tf-idf values until all the nodes are processed.

### *G. Output*

The output is an abstractive summary.

**Recall-Oriented Understudy for Gisting Evaluation (ROUGE)**

Used for evaluating text summary. Here, the comparison is done between the generated summary and human summary. For this paper implementation, the following rouge metrics are used:

- ROUGE-1

It produces the overlap of unigram among both the summaries.

- ROUGE-2

It produces the overlap of bigrams among both the summaries.

- ROUGE-L

It finds the longest common subsequence.

## V. RESULTS

This section is used to show the results of above implementation.

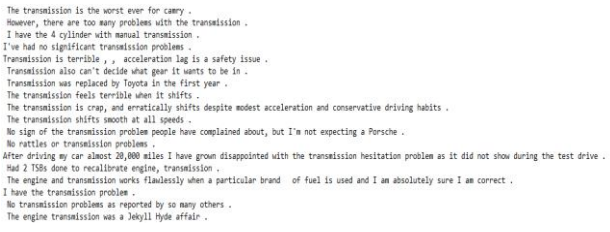


Fig 2. Sample review data on transmission_toyota_camry

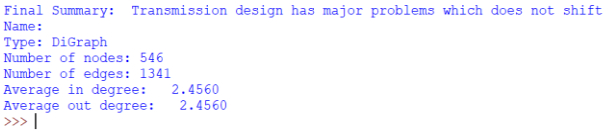


Fig 3. Output showing the final summary for transmission_toyota_camry

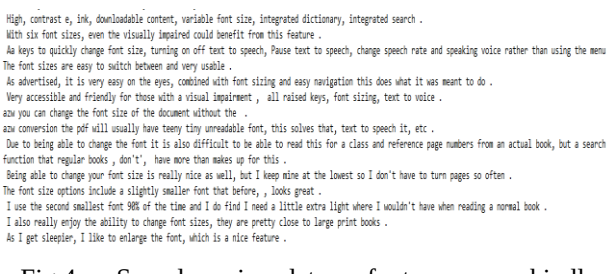


Fig 4. Sample review data on fonts_amazon_kindle

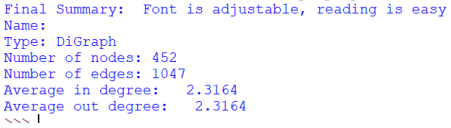


Fig 5. Output showing the final summary of fonts_amazon_kindle

| Document Name | ROUGE 1 | | | ROUGE 2 | | | ROUGR L | | |
|---|---|---|---|---|---|---|---|---|---|
| transmission_toyota_camry | F | P | R | F | P | R | F | P | R |
| | 0.57 | 0.66 | 0.66 | 0.42 | 0.50 | 0.36 | 0.49 | 0.60 | 0.50 |
| fonts_amazon_kindle | F | P | R | F | P | R | F | P | R |
| | 0.54 | 0.50 | 0.60 | 0.29 | 0.35 | 0.30 | 0.39 | 0.40 | 0.40 |

Table 1. Table showing the ROUGE metrics for above two topic data

## VI. CONCLUSION AND FUTURE SCOPE

This paper aims at helping a user to get an accurate overview of a product by using abstractive multi document text summarization. First , obtain the dataset and pre process it so that it does not contain any unwanted data. Then, resolve the contractions and assign every token with its parts of speech and construct a directed graph with nodes and edges. Finally, use term frequency to find the important words and obtain top ranked sentences are generate the final summary. In future, a user interface for this process can be developed and a system for results for custom data can be developed.

## REFERENCES

[1] Ahmad T Al-Taani, Automatic Text Summarization Approaches ,International Conference on Infocom Technologies and Unmanned Systems, 2017.

[2] Taner Uçkan, Ali Karcı, Extractive Multi-Document Text Summarization Based On Graph Independent Sets, Egyptian Informatics Journal, 2020

[3] Shai Erera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora Peled Nakash, Odellia Boni, Haggai Roitman, Doron Cohen, Bar Weiner, Yosi Mass, Or Rivlin, Guy Lev, Achiya Jer-bi, Jonathan Herzig, Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, David Konopnicki, A Summarization System for Scientific Documents ,Proceedings of the EMNLP and the 9th IJCNLP, 2019

[4] Min Yang, Chengming Li, Ying Shen, Qingyao Wu, Zhou Zhao, Xiaojun Chen, Hierar-chical Human-Like Deep Neural Networks for Abstractive Text Summarization, IEEE Transactions on Neural Networks and Learning Systems, 2020

[5] V. Mohan Kalyan, Chukka Santhaiah, M. Naga Sri Nikhil, J. Jithendra, Y. Deepthi, and N. V. Krishna Rao, Extractive Summarization Using Frequency Driven Approach , Machine Learning Technologies and Applications: Proceedings of ICACECS, 2020.

[6] Akash Ajampura Natesh , Somaiah Thimmaiah Balekuttira , Annapurna P Patil, Graph Based Approach for Automatic Text Summarization, International Journal of Advanced Research in Computer and Communication Engineering, 2016.

[7] Yang Gao, Wei Zhao, Steffen Eger, SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization, ACL 2020.

[8] Samer Abdulateef, Naseer Ahmed Khan, Bolin Chen and Xuequn Shang, Multidocument Arabic Text Summarization Based on Clustering and Word2Vec to Reduce Redundancy, Natural Language Generation and Machine Learning, 2020.

[9] Mohammad Bidoki, Mohammad R.Moosavi, MostafaFakhrahmad,A semantic approach to extractive multi-document summarization: Applying sentence expansion for tuning of con-ceptual densities, 2020.

[10] Dominik Ramsauer and Udo Kruschwitz, Exploring the Incorporation of Opinion Polarity for Abstractive Multi-Document Summarisation, 2021.