TEXT SUMMARIZATION USING NATURAL LANGUAGE PROCESSING

Kota Prudhvi[1], A.Bharath Chowdary[1], P.Subba Rami Reddy[1], P. Lakshmi Prasanna[2]

[1] Student, Computer Science and Engineering, KL University, Vaddeswaram.

{prudhvi.kvnsn,bharathchowdary2909, padigapatiram}@gmail.com

[2] Associate Professor, KL University, Vaddeswaram.

pprasanna@kluniversity.in

Abstract. With the advancements in the technology most of the things in this world have become automated. The concept of text summarization came into limelight as summarization of text manually has become a tough and time-consuming task. So, the main purpose of text summarization is to overcome the difficulties faced during manual summarization of text documents or other information from various sources. Text Summarization is the process of extracting the main idea of the context or the text and briefly explaining about the context. This process is not only to extract key idea and phrases from the text sources but also generating meaningful summary in a concise and crisp way. The demand for text summarization is raising now a days because of the large amounts of data from multiple sources like internet, Twitter, Facebook, Instagram, research papers and other news articles. Text Summarization can be efficiently implemented using NLP as it has many packages and methods in Python or R. Text Summarization is also related to text mining as summary is generated based on classifying the given input text. There are different approaches for text summarization and some algorithms are identified to implement these approaches. In this paper, unsupervised learning approach is implemented and cosine similarity technique is used to find the similarity between sentences. To generate rank based on similarity, text rank algorithm is used and sentences with top rank are placed in summarized text.

Keywords: Text summarization, text mining, natural language processing, cosine similarity, text rank algorithm.

1 Introduction

Data and Information have become more important in our day to day life. This can be explained with the fact given by International Data Corporation (IDC) that the total amount of digital data travelling all around the world in the internet every year would increase from 33 zettabytes in 2018 to 175 zettabytes in 2025. With increase of online information and other resources, text summarization has become an important part in our fast-paced lives. Text Summarization is a part of text analytics which deals with analysing textual data in order to understand customer opinions, reviews, activities and their feedback. This text analytics has become an important

part for data scientists and engineers and they don't face much difficulties in dealing with structured or relational data. But they face difficulties in understanding and analysing online information like articles, blogs, social media data as these mediums generate unstructured data in form of audio, video, image.

Text Summarization can be said as a sub domain in domain of Natural Language Processing (NLP). Automatic Text Summarization can be said as a task which is used to generate brief and coherent version of larger text or documents. Researches these days show interest in summarizing the text due to presence of large volume of data in one hand and time constraint in other. Furthermore, automatic text summarization reduces the reading time of the users, speeds up the selection process of research among documents and automatic text summarization provides more effective and efficient summaries than human or manual summarization.

Automatic Text Summarization is a common problem in both the fields of Natural Language Processing (NLP) and Deep Learning (DL). But it can be better implemented in NLP as several methods and packages like nltk are available in Python and R. The main types of automatic text summarization are extractive based summarization and abstraction-based summarization. In extractive based summarization, algorithm identifies for key phrases and important sentences from the text or document and generates a summary using the same key phrases and sentences from the text without any change. In abstraction-based summary, algorithm generates a summary which consists of brief explanation of the text in few lines. In this paper we will be using extractive text summarization procedure to implement automatic text summarization.

2 Literature Review

An Overview of Extractive Text Summarization

Shohreh Rad Rahimi et al [1] put in the picture that NLP researches are showing more interest in summarizing textual information. In this paper, authors defined text summarization as process of automatically creating and reducing form of given document and retaining its information content source into shorter version with correct meaning. In this paper the authors also explained about the relationship between text mining and text summarization. The authors of this paper explained about different processes in text summarization like Case Folding, Bag of Words, Tokenization, N-grams, Stemming. This paper also discusses about different criteria in text summarization systems designing such as Summarization based on output summary, based on details, based on contents, based on limitation, based on number of input texts, based on language acceptance. Finally, this paper discusses about various approaches to text summarization such as Statistical approaches, Lexical

Chain based approaches, Graph based approaches, Cluster based approaches, Fuzzy logic-based approaches.

A Review Paper on Text Summarization

Deepali K. Gaikwad et al [2] explains that in text summarization the information that is important is extracted and collected from original text and presented in the form of summary. In this paper, the authors discussed about the two main approaches of text summarization which are extractive based summarization and abstractive based summarization. This paper explains about various features of text summarization like term frequency, location, cue method, sentence length, similarity, it also explains about various approaches present in abstractive based summarization and extractive based summarization. This paper is also an attempt to present the view of text summarizers for Indian languages and comparison of their performances.

An Overview of Text Summarization Techniques

Narendra Andhale et al [3], in this paper presented comprehensive survey of both the abstractive based and extractive based summarizations. The authors explain about the automatic text summarization becomes important in finding relevant information from long texts in short period of time with less efforts. The authors in this paper mainly focused on extractive based text summarization methods such as Term Frequency-Inverse Document Frequency method, cluster based method, text summarization using neural networks, graph based method, text summarization with fuzzy logic, latent semantic analysis method, machine learning approach, query based summarization. The authors also discussed about final observations of summarization methods.

Document Summarization Using Sentence Based Topic Modelling and Clustering

Augustine George et al [4] discuss about various approaches to identify significant portion of each document and determine that topic representation and modelling are intermediate representation of text that captures topics discussed in the input and aids the automatic summarization. This article attempts to provide comprehensive summary that includes sentence extraction, tokenization of extracted sentences. The authors discuss about Sentence based Structural Topic Modeling to determine important content for each domain in integrated document and sentences are grouped using k-means clustering. Finally, this paper discusses about lexical ranking score in summarized text.

A Novel Technique for Efficient Text Document Summarization as a Service

Anusha Bagalkotkar et al [5] discusses the need of automatic text summarizer and its effect in reducing human efforts. The authors also explain about text summarization as the process of generating summary of input document by

extracting important sentences from it. In this paper authors proposed a unique technique for generating summarization of domain specific text from single document from web by using statistical NLP techniques. The system proposed by authors have four distinct phases which are generation of list of frequent words, sentence generation, update details in database, setup web service, also proposed an algorithm for single document summarizer. Finally, authors discussed about experimental results, parameters, performance of proposed system.

Automatic Text Summarization using Neural Networks

Khosrow Kaikhah [6] proposed a different technique for summarizing news articles using neural networks. In this article, author discussed about a neural network that is trained to learn important features of sentences that should be included in summary of article, after that the neural network is modified to generalize and combine important features possible in summary sentences and the neural network that is modified is used as filter to summarize news articles. In this paper the author discussed about the three phases in text summarization process in neural network that are neural network training, feature fusion and sentence selection and analysed about their results.

Automatic Text Summarization Using Natural Language Processing

Pratibha Devihosur et al [7] proposed a system of automatic text summarization using unsupervised learning. The authors used a technique of simplified Lesk calculation to assess the significance and importance of sentence in information, an online semantic lexicon wordnet is utilized. This paper discussed about the advantages of proposed system and system architecture of proposed system. The system architecture consists of three stages that are data pre-processing, evaluation of weights, summarization and the overall representation of automatic text summarization using natural language processing consists of input document, pre-processing, Lesk algorithm, generation of summary and lesk algorithm is connected to WordNet. Finally, authors discussed about the outputs of the lesk algorithm and its calculation and future scope of the proposed system.

Automatic Text Summarization and it's Methods- A Review

Neelima Bhatia et al [8], in this survey paper investigated the popular and important work done in the field of single and multiple document summarizations. The authors discussed about the method based approaches for text summarization. These method based approaches include term based frequency method, graph based method, time based method, separation and merging based method, semantic dependency method, topic based approaches, discourse based approaches, Latent Semantic based approaches, approaches based on lexical chain, approaches based on fuzzy logic.

NLP Based Text Summarization Using Semantic Analysis

Hamza Shabbir Moiyadi et al [9] presented a unique technique for generating domain specific text by using semantic analysis for text summarization and it is the subset of Natural Language Processing. In this paper authors proposed a system that uses latent semantic analysis to summarize the documents from the user. According to this paper the product evaluation process is done in four distinct phases that are preparation, criteria establishment, characterization, testing and the stages in the proposed system are parsing, generation of terms by Singular Value Decomposition (SVD) matrix, analysis of SVD matrix, arranging of sentences from the final SVD matrix. The authors also suggested a code for implementation of Latent Semantic Analysis using python library.

Automatic Amharic Text Summarization using NLP Parser

Getahun Tadesse Mekuria et al [10] proposed a system that investigates the problem of building the domain based single and multiple document Amharic text summarization. In this paper authors suggested that multi-document summarization targets to compress the most important information from a set of documents to produce a short summary and also suggested that text summarization can be performed based on input, purpose and output. In this paper the authors proposed a system that solves the existing problem by developing the combinations of extractive and abstractive based approach on single and also multiple document input from the user. The authors discussed about the importance of page rank algorithm as it is used in finding out sentence score and weights of sentence in document. In this paper, authors also concluded that the proposed model summarizes only text documents but in future a system would be proposed to develop text summarization for all types of documents including images, graphs, pictures, videos.

Text Summarization Using Latent Semantic Analysis Model in Mobile Android Platform

Oi-Mean Foong et al [11] in this paper, presented about Latent Semantic Analysis in Automatic Text Summarization on single English document in android platform of mobiles. The authors objective in this paper is to investigate the LSA model by examining the semantic relationship between terms and sentences in a document for text summarization. In this paper authors proposed a model for text summarization that consists of three steps that are Pre-processing which includes Stop words removal, Word Stemming, POS Tagging; Latent semantic analysis which includes input matrix, Singular Value Decomposition (SVD); Generation of Summary.

Text summarization using Latent Semantic Analysis

Makbule Gulcin Ozsoy et al [12] explains about text summarization as it solves the problems of presenting the information needed by a user in compact form. The authors of the paper explained different Latent Semantic Analysis based

summarization algorithms and two algorithms were proposed by the authors. In this paper authors used ROGUE scores to evaluate the performance of the algorithms. The authors mainly discussed about Latent Semantic Analysis approach and Sentence selection approach for text summarization. One of the algorithms proposed by authors produced best scores and both the algorithms equally performed well on summarizing English and Turkish document texts.

Text Summarization within the Latent Semantic Analysis Framework: Comparative Study

Rasha Mohammed Badry et al [13] explained that text summarization is used to provide shorter version of original text by retaining the overall meaning. In this papers authors discussed about various approaches of text summarization using Latent Semantic Analysis that are Gong and Liu's approach, Steinberger and Jezek's approach, Murray, Renals and Carletta's approach, Ozsoy's approach which consists of Cross method and Topic method. Finally, in this paper author presented a table that explains about the comparative summary among the LSA based summarization algorithms.

A Survey on Extractive Text Summarization

N. Moratanch et al [14] explained about the technique of text summarization is that extracted information is obtained as summary report and presented as a short summary to the user. In this paper authors discussed about word level features, sentence level features and various extractive text summarization methods. Word level features consists of content word feature, title word feature, cue phrase feature, biased word feature, upper case word feature. Sentence level features consists of sentence location feature, sentence length feature, paragraph length feature, sentence-to-sentence cohesion. Various summarization methods are categorized based on supervised and unsupervised learning methods. In this paper, unsupervised learning methods discussed are graph based approach, fuzzy logic-based approach, concept-based approach and supervised learning methods are machine learning approach based on Bayes rule, neural network-based approach, conditional random fields. The authors in this paper suggested some of the evaluation metrics such as human evaluation, Rogue score, Recall, Precision, F-measure, Compression ratio.

Literature Review on Extractive Text Summarization Approaches

Saiyed Saziyabegum et al [15] discussed about the basis on which types of summarization are classified. The main types of summarization are classified based on pre-processing techniques that include abstractive, extractive; based

on purpose of pre-processing that include indicative summaries, informative summaries, critical or evaluative summaries, update summaries; based on audience that include generic summaries, query-based summaries, user-specific or topic-specific summaries; based on number of documents that include single document summarization, multi document summarization; based on language that include single, multi-language summarization.The authors in this paper discussed about problems of text summarization, discussed about features of extractive text summarization. This paper also explained about approaches for summarization of single document that include term frequency-inverse document frequency based approach, cluster based approach, naïve-bayes based approaches, rich feature and decision trees based approaches, hidden markov model based approach, log-linear model approach, neural network based approach, graph theoretic approach, latent semantic based approach, concept obtained approach, fuzzy logic based approach, genetic algorithm and mathematical regression model based approach, query-biased and structure-preserving based approaches, lexical chain based approach, ranking-based sentence clustering approach, hybrid approaches.

3 Existing System

The above research papers in literature survey suggests that many researches have done their research in text summarization using Latent Semantic Analysis and Fuzzy logic. The proposed system or model in many research papers consists of Latent Semantic Analysis which mostly included Singular Value Decomposition (SVD) and some researchers proposed neural networks to implement text summarization. Most of the researches used python libraries for implementation of text summarization using LSA. Fuzzy logic-based approaches were also used by some researches. The evaluation criteria was mostly done by using parameter called ROGUE score to evaluate performance of various algorithms.

4 Proposed Method

In this paper, a model was proposed by combining two techniques that are used for implementing automatic text summarization. The techniques used to summarize the text are cosine similarity and textrank method. Cosine similarity can be defined as measurement of similarity between documents that consists of group of sentences or vectors of single sentence (non-zero vector) and this cosine similarity approach measures the cosine angle between them. In this paper, we represented each sentence in a vector so we can measure the similarity between the sentences using cosine similarity. If the angle measured using cosine similarity technique is equal to zero then the

sentences present in the vectors are similar. Another technique used in this paper is text rank algorithm which is commonly used algorithm for text summarization in NLP. Textrank algorithm is similar to PageRank algorithm but in pagerank algorithm online web pages are ranked in online search results using pagerank score. But in text rank algorithm sentences are used in place of webpages and rank is generated. Textrank is unsupervised learning method and used for extractive text summarization. Generally, text rank algorithm alone is used for generating text summaries. But in this paper, text rank algorithm is combined with cosine similarity to extract summaries.

5 Flowchart

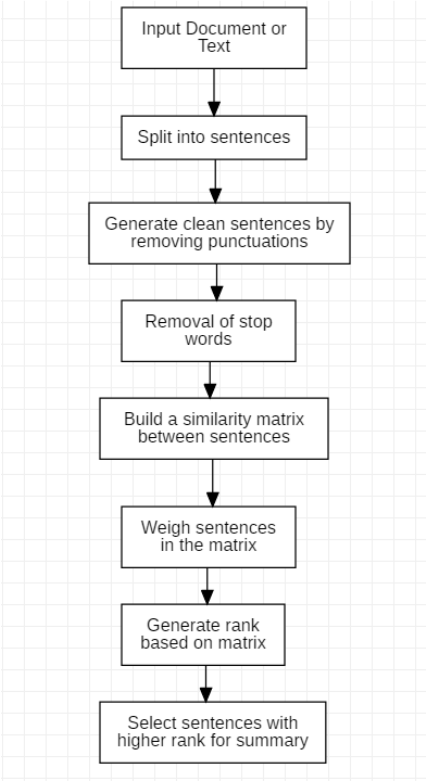

Fig. 1. Flowchart for the implementation of text summarization

In this architecture, initially a document which consists of the user's text is passed as an input then the group of sentences in the paragraph are divided into vectors of sentences where each vector holds each sentence. In the next step pre-processing techniques like removal of punctuations and removal of stopwords is applied. Then a technique called cosine similarity is used to build a similarity matrix between the vectors that consists of sentences. Then weigh the sentences using an algorithm called textrank algorithm to generate ranks in the matrix. Then the proposed summarizer generates summary that includes on top ranked sentences.

6 Procedure

- Import required packages like nltk (as it includes stopwords, cosine_distance), numpy (for using text rank algorithm)
- Define a function to read the document as input and divide the document into sentences
- Define a function to measure similarity between the sentences using cosine similarity that is consider two sentences at each time and convert these sentences into vectors, then return the value of cosine similarity
- Define a function to build a matrix basing on the cosine similarity value which is measured in the above function.
- Rank the sentences in the similarity matrix using text rank algorithm imported from numpy, networkx packages and sort the sentences using the rank and place top ranked sentences in the summary

7 Results

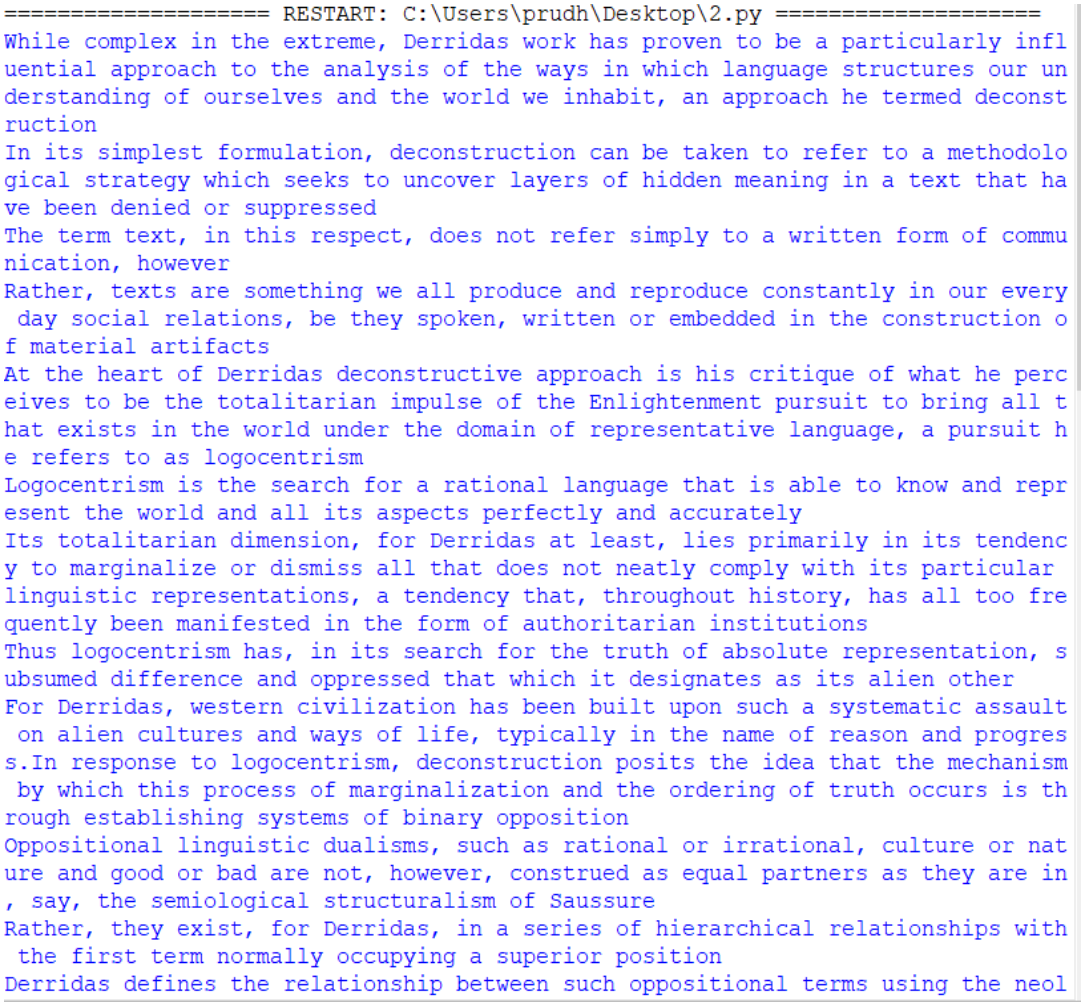

Fig. 2. Screenshot 1 of the text present in the file

Derridas defines the relationship between such oppositional terms using the neol
ogism difference
This refers to the realization that in any statement, oppositional terms differ
from each other for instance, the difference between rationality and irrationali
ty is constructed through oppositional usage, and at the same time, a hierarchic
al relationship is maintained by the deference of one term to the other in the p
ositing of rationality over irrationality, for instance
It is this latter point which is perhaps the key to understanding Derridas appro
ach to deconstruction.For the fact at any given time one term must defer to its
oppositional other, means that the two terms are constantly in a state of interd
ependence
The presence of one is dependent upon the absence or absentpresence of the other
, such as in the case of good and evil, whereby to understand the nature of one,
we must constantly relate it to the absent term in order to grasp its meaning
That is, to do good, we must understand that our act is not evil, for without th
at comparison the term becomes meaningless
Put simply, deconstruction represents an attempt to demonstrate the absent-prese
nce of this oppositional other, to show that what we say or write is in itself n
ot expressive simply of what is present, but also of what is absent
Thus, deconstruction seeks to reveal the interdependence of apparently dichotomo
us terms and their meanings relative to their textual context that is, within th
e linguistic power relations which structure dichotomous terms hierarchically
In Derridas own words, a deconstructive reading must always aim at a certain rel
ationship, unperceived by the writer, between what he commands and what he does
not command of the patterns of a language that he uses.It attempts to make the n
ot-seen accessible to sight.Meaning, then, is never fixed or stable, whatever th
e intention of the author of a text
For Derridas, language is a system of relations that are dynamic, in that all me
anings we ascribe to the world are dependent not only on what we believe to be p
resent but also on what is absent
Thus, any act of interpretation must refer not only to what the author of a text
intends, but also to what is absent from his or her intention
This insight leads, once again, to Derridas further rejection of the idea of the
definitive authority of the intentional agent or subject
The subject is decentred, it is conceived as the outcome of relations of differe
nce
As author of its own biography, the subject thus becomes the ideological fiction
of modernity and its logocentric philosophy, one that depends upon the formatio
n of hierarchical dualisms, which repress and deny the presence of the absent ot
her

Fig. 3. Screenshot 2 of the text present in the file

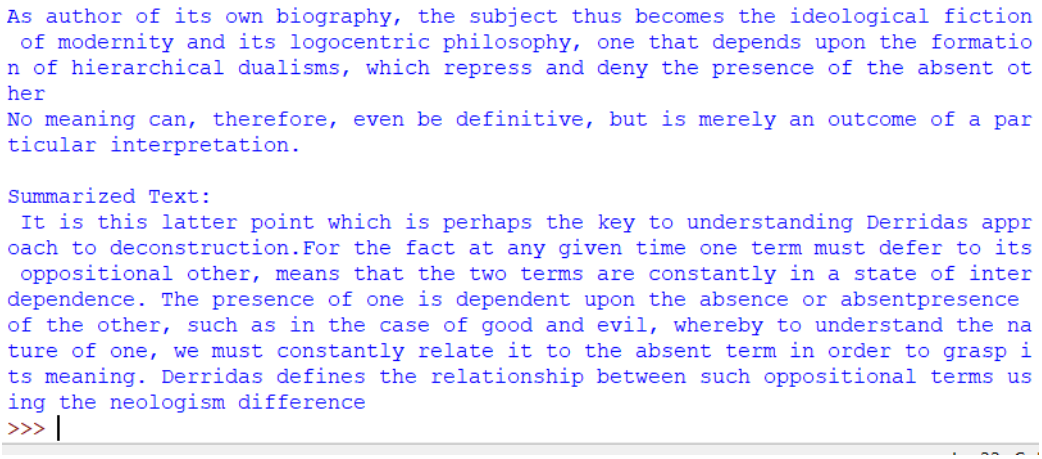

Fig. 4. Screenshot of the output that consists of the summarized text

8 Conclusion

Text Summarization which is one of the important problems in Natural Language Processing has become an interesting research area for many data scientists and researchers. There are various methods and several approaches in implementing automatic text summarization and still more new techniques and approaches are being identified, experimented and implemented to derive automatic text summarization. In this paper initially literature survey is done on various research papers and existing model for text summarization is explained. Later, a model is proposed with slight extensions to improve the performance of text summarization. The methods that are introduced in this paper are slight extension to previous identified methods but these extensions improve the performance of generating summaries.

References

[1] Shohreh Rad Rahimi, Ali Toofanzadeh Mozhdehi, Mohamad Abdolahi, "An Overview on Extractive Text Summarization", 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI) Dec. 22nd, 2017(Iran University of Science and Technology) – Tehran, Iran

[2] Deepali K. Gaikwad, C. Namrata Mahender, "A Review Paper on Text Summarization", International Journal of Advanced Research in Computer and Communication Engineering Vol.5, Issue 3, March 2016

[3] Narendra Andhale, L.A. Bewoor, "An Overview of Text Summarization Techniques"

[4] Augustine George, Dr. Hanumanthappa, "Document Summarization Using Sentence Based Topic Modelling and Clustering", International Journal of Advanced Research (IJAR), Received: 04 March 2018, Final Accepted: 06 April 2018, Published: May 2018

[5] Anusha Bagalkotkar, Ashesh Khandelwal, Shivam Pandey, Sowmya Kamath S, "A Novel Technique for Efficient Text Document Summarization as a Service", 2013 Third International Conference on Advances in Computing and Communications

[6] Khosrow Kaikhah, "Automatic Text Summarization with Neural Networks", Second IEEE International Conference on Intelligent Systems, June 2004

[7] Pratibha Devihosur, Naseer R, "Automatic Text Summarization Using Natural Language Processing" International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 08, Aug-2017

[8] Neelima Bhatia, Arunima Jaiswal, "Automatic Text Summarization and its Methods-A Review", 2016 6th InternationalConference - Cloud System and Big Data Engineering (Confluence)

[9] Hamza Shabbir Moiyadi, Harsh Desai, Dhairya Pawar, Geet Agrawal, Nilesh M.Patil, "NLP Based Text Summarization Using Semantic Analysis", International Journal of Advanced Engineering, Management and Science (IJAEMS) Infogain Publication (Infogainpublication.com), Vol-2, Issue-10, Oct- 2016

[10] Getahun Tadesse Mekuria, Aniket S. Jagtap, "Automatic Amharic Text Summarization using NLP Parser", International Journal of Engineering Trends and Technology (IJETT) – Volume 53 Number 1 November 2017

[11] Oi-Mean Foong, Suet-Peng Yong, Farha-Am Jaid, "Text Summarization Using Latent Semantic Analysis Model in Mobile Android Platform", 2015 9th Asia Modelling Symposium

[12] Makbule Gulcin Ozsoy and Ferda Nur Alpaslan, Ilyas Cicekli, "Text summarization using Latent Semantic Analysis", Article in Journal of Information Science, August 2011

[13] Rasha Mohammed Badry, Ahmed Sharaf Eldin, Doaa Saad Elzanfally, "Text Summarization within the Latent Semantic Analysis Framework: Comparative Study", International Journal of Computer Applications (0975 – 8887), Volume 81 – No.11, November 2013

[14] N. Moratanch, Chitrakala Gopalan, "A survey on extractive text summarization", IEEE International Conference on Computer, Communication, Signal Processing (ICCCSP-2017)

[15] Saiyed Saziyabegum, Priti S. Sajja, PhD, "Literature Review on Extractive Text Summarization Approaches", International Journal of Computer Applications (0975 – 8887), Volume 156 – No 12, December 2016

[16] Aysa Siddika Asa, Sumya Akter, Md. Palash Uddin, Md. Delowar Hossain, Shikhor Kumer Roy, Masud Ibn Afjal, "A Comprehensive Survey on Extractive Text Summarization Techniques", American Journal of Engineering Research (AJER) 2017

[17] Arpita Sahoo, Dr. Ajit Kumar Nayak, "Review Paper on Extractive Text Summarization", International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 5, Issue 4, April 2018

[18] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut, "Text Summarization Techniques: A Brief Survey", arXiv, July 2017, USA

[19] Josef Steinberger, Karel Ježek, "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation", Article January 2004

[20] Chirantana Mallick, Ajit Kumar Das, Madhurima Dutta, Asit Kumar Das and Apurba Sarkar, "Graph-Based Text Summarization Using Modified TextRank" , Conference Paper · August 2018

[21] Tanwi, Satanik Ghosh, Viplav Kumar, Yashika S Jain, Mr. Avinash B, "Automatic Text Summarization using Text Rank", International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue: 04 | Apr 2019

[22] Rizki Darmawan, Romi Satria Wahono, "Hybrid Keyword Extraction Algorithm and Cosine Similarity for Improving Sentences Cohesion in Text Summarization", Journal of Intelligent Systems, Vol. 1, No. 2, December 2015

[23] Mohd. Saif Wajid, Shivam Maurya, Mr. Ramesh Vaishya, "Sentence Similarity based Text Summarization using Clusters", International Journal of Scientific & Engineering Research, Volume 4, Issue 5, May-2013

[24] Mukesh Kumar Jadon, Ayush Pareek, "A method for Automatic Text Summarization using Consensus of Multiple Similarity Measures and Ranking Techniques", D S Sharma, R Sangal and A K Singh. Proc. of the 13th Intl. Conference on Natural Language Processing, pages 137–143, Varanasi, India. December 2016. §c 2016 NLP Association of India (NLPAI)