Automatic Text Summarization using Natural Language Processing

Project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Technology

In

Computer Science and Engineering/Information Technology

By

Sonali Behal 151433

Aayush Gupta 151220

Under the supervision of

Dr. Vivek Sehgal

To

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

CERTIFICATE

This is to certify that this project report entitled **Automatic Text Summarization Using Natural Language Processing** submitted to **Jaypee University of Information Technology**, is a bonafide record of work done by

151433 Sonali Behal

151220 Aayush Gupta

Under my supervision from **July 2018** to **May 2019** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science & Engineering**.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Dr. Vivek Sehgal Associate Professor Department
of Computer Science & Engineering and IT

ACKNOWLEDGMENT

We would like to express our special thanks of gratitude to our Project Supervisor Dr. Vivek Sehgal as well as our Project Coordinator Dr. Hemraj Saini who gave us the golden opportunity to do this wonderful project on the topic **Automatic Text Summarization Using Natural Language Processing**, which also helped us in doing a lot of Research and we came to know about so many new things we are really thankful to them. Secondly, we would also like to thank our parents and friends who helped us a lot in finalizing this project within the limited time frame.

Sonali Behal Aayush Gupta

ABSTRACT

Automatic text summarization is basically summarizing of the given paragraph using natural language processing and machine learning. There has been an explosion in the amount of text data from a variety of sources. This volume of text is an invaluable source of information and knowledge which needs to be effectively summarized to be useful. In this review, the main approaches to automatic text summarization are described. We review the different processes for summarization and describe the effectiveness and shortcomings of the different methods. Two types will be used i.e.-extractive approach and abstractive approach. The basic idea behind summarization is finding the subset of the data which contains the information of all the set. There is a great need to reduce unnecessary data. It is very difficult to summarize the document manually so there is the great need of automatic methods. Approaches have been proposed inspired by the application of deep learning methods for automatic machine translation, specifically by framing the problem of text summarization as a sequence-to-sequence learning problem.

TABLE OF CONTENT

Chapter-1

INTRODUCTION

1.1 Introduction

With the developing measure of data, it has turned out to be hard to discover brief data. In this way, it is critical to making a framework that could condense like a human. Programmed content rundown with the assistance of Normal Dialect Handling is an instrument that gives synopses of a given archive. Content Outline strategies is divided in two ways i.e. - extractive and abstractive approach. The extractive approach basically choose the various and unique sentences, sections and so forth make a shorter type of the first report. The sentences are estimated and chosen based on accurate highlights of the sentences. In the Extractive technique, we have to choose the subset from the given expression or sentences in given frame of the synopsis. The extractive outline frameworks depends on two methods i.e. - extraction and expectation which includes the arrangement of the particular sentences that are essential in the general comprehension the archive. What's more, the other methodology i.e. abstractive content synopsis includes producing completely new articulations to catch the importance of the first record. This methodology is all the more difficult but on the other hand is the methodology utilized by people.

New methodologies like Machine taking in procedures from firmly related fields, for example, content mining and data recovery have been utilized to help programmed content synopsis.

From Completely Mechanized Summarizers (FAS), there are techniques that assistance clients doing rundown (MAHS = Machine Helped Human Synopsis), for instance by featuring hopeful sections to be included the outline, and there are frameworks that rely upon post-preparing by a human (HAMS = Human Supported Machine Rundown).

There are two types of extractive rundown errands which rely on the outline application focuses. One is nonexclusive synopsis, which centres on getting a general rundown or

unique of the Archive (regardless of whether records, news stories and so on.). Another is inquiry related synopsis, some of the time called question based outline, which abstracts especially to the question. Outline strategies can make both inquiry related content rundowns and conventional machine-created synopses relying upon what the client needs.

Likewise, rundown strategies endeavour to discover subsets of items, which contain data of the total set. This is otherwise called the centre set. These calculations demonstrate experiences like inclusion, decent variety, data or representativeness of the outline. Question based synopsis techniques, furthermore demonstrate for purpose of the outline with the inquiry. A few techniques and calculations which specifically outline issues are Text Rank and Page Rank, Sub modular set capacity, determinately point process, maximal negligible significance (MMR) and so forth.

1.2 Problem Statement

In the new period, where tremendous measure of data is accessible on the Web, it is most vital to give the enhanced gadget to get data rapidly. It is extremely intense for individuals to physically pick the synopsis of expansive archives of content. So there is an issue of scanning for vital reports from the accessible archives and discovering essential data. Along these lines programmed content rundown is the need of great importance. Content rundown is the way toward recognizing the most vital important data in a record or set of related archives. What's more, compact them into a shorter rendition looking after its implications. .

1.3 Objectives

The objective of the project is to understand the concepts of natural language processing and creating a tool for text summarization. The concern in automatic summarization is increasing broadly so the manual work is removed. The project concentrates creating a tool which automatically summarizes the document.

1.4 Methodologies

For obtaining automatic text summarization, there are basically two major techniques i.e.-Abstraction based Text Summarization and Extraction based Text Summarization.

Extraction Based Extraction

The Extractive summaries are used to highlight the words which are relevant, from input source document. Summaries help in generating concatenated sentences taken as per the appearance. Decision is made based on every sentence if that particular sentence will be included in the summary or not. For example, Search engines typically use Extractive summary generation methods to generate summaries from web page. Many types of logical and mathematical formulations have been used to create summary. The regions are scored and the words containing highest score are taken into the consideration. In extraction only important sentences are selected. This approach is easier to implement. There are three main obstacles for extractive approach. The first thing is ranking problem which includes ranking of the word. The second one selection problem that includes the selection of subset of particular units of ranks and the third one is coherence that is to know to select various units from understandable summary. There are many algorithms which are used to solve ranking problem. The two obstacles i.e. - selection and coherence are further solved to improve diversity and helps in minimizing the redundancy and pickup the lines which are important. Each sentence is scored and arranged in decreasing order according to the score. It is not trivial problem which helps in selecting the subsets of sentences for coherent summary. It helps in reduction of redundancy. When the list is put in ordered manner than the first sentence is the most important sentence which helps in forming the summary. The sentence having the highest similarity is selected in next step is picked from the top half of the list. The process has to be repeated until the limit is reached and relevant summary is generated.

Abstraction Based summarization

People by and large utilize abstractive outlines. In the wake of perusing content, Individuals comprehend the point and compose a short outline in their own particular manner creating their very own sentences without losing any essential data. In any case, it is troublesome for machine to make abstractive synopses. Along these lines, it very well may be said that the objective of reflection based outline is to make a synopsis utilizing regular dialect preparing procedure which is utilized to make new sentences that are syntactically right. Abstractive rundown age is difficult than extractive technique as it needs a semantic comprehension of the content to be encouraged into the Common Dialect framework. Sentence Combination being the significant issue here offers ascend to irregularity in the produced outline, as it's anything but an all around created field yet.

Abstractive arrangement to grouping models is by and large prepared on titles and captions. The comparative methodology is embraced with archive setting which helps in scaling. Further every one of the sentences is revamped in the request amid the inference. Document synopsis can be changed over to regulated or semi-administered learning issue. In directed learning methodologies, indications or signs, for example, key-phrases, point words, boycott words, are utilized to recognize the sentences as positive or negative classes or the sentences are physically labelled. At that point the parallel more tasteful can be prepared for getting the scores or synopsis of each sentence. Anyway they are not effective in removing archive explicit summaries. If the report level data isn't given then these methodologies give same expectation independent of the record. Giving archive setting in the models diminishes this issue.

1.5 Organization

Chapter 1: Includes a brief introduction to the project. A basic idea of what we are doing and what we are trying to accomplish with this Project has also been provided and technologies we are using in this project have also been listed. .

Chapter 2: Includes literature survey. We have studied various papers and journal from reputed sources on machine learning and artificial neural network and have mentioned those in this chapter. .

Chapter 3: Includes details on system development. Explanation about the project design, models implemented and formulas applied have been mentioned.

Chapter 4: Includes result and result analysis. This section provides the results are implemented models are yielding and accuracy of those results have been scrutinized in this section. .

Chapter 5: Conclusion. Outcomes and the future scope of the project have been discussed briefly. .

Chapter-2

LITERATURE SURVEY

2.1 Summary of Papers

2.2.1

Title	**Automatic Text Summarization Approaches[1]**
Authors	Ahmad T. Al-Taani (Ph.D., MSc, BSc) Professor of Computer Science (Artificial Intelligence) Faculty of Information Technology and Computer Sciences Yarmouk University, Jordan.
Year of Publications	August 2017
Publishing Details	International Conference on Infocom Technologies and Unmanned Systems (ICTUS'2017)
Summary	Automated Text summarization systems are important in many aspects in a language like natural language processing. ATS creates the summary of given document which save time and resources. There are single and multi-document text summary. Only one document is extracted in case of single document summarization whereas group of documents is selected in multi document summarization. On other hand, mathematics techniques makes the extractive summarization language independent to theoretical ways. In this analysis, we tend to the thought of the utilization of extractive summarization methodology. There are two content-based summaries i.e. - generic and query-based summaries. In the generic summarization system if the user doesn't have knowledge about text then information measure equal level in information. Whereas in query-based summarization, before starting of the summarization technique, the topic is verified of the initial text. The system extracts

	that knowledge from the provision text and presents it defines. There are three main approaches to summarization i.e.- statistical, the graph-based, and machine learning approaches. The other approach is a clustering approach. In **statistical approaches**, researchers are based upon sentence ranking and the important sentences are selected from the given document, regarded as the important summary compression ratio. **Graph-based approaches** concentrate on the semantic analysis and relationship among sentences. The graph-based approach is used in the representation for text inside documents. **Machine learning approaches** helps in producing summary by applying machine learning algorithms. This approach deals with the summarization process as a classification problem. Based on the characteristics sentences are divided for summary.

2.2.2

Title	**Automatic Text Summarization: Single and Multiple Summarizations[2]**
Authors	**Neelima Bhatia** Amity School of Engineering and Technology (ASET) Amity University Noida, India **Arunima Jaiswal** Amity School of Engineering and Technology (ASET) Amity University Noida, India
Year of Publications	May 2015
Publishing Details	International Journal of Computer Applications
Summary	There is large amount of information available in internet.. The actual

	results very long and hard to read . thus the demand of automatic summarization increased. Automatic summarization collect the important information from the given document and generate the summary which is important and save time. The review is based on single and multiple summarization methods. Automatic summarization takes out the data from the document and save the time. H.P. Luhn discovered automatic summarization of given text in the year 1958. `NLP invented the subfield of summarization. In automatic summarization important points are not lost. There are two approaches-Abstraction and extraction approach. Extraction is domain independent and provides summary. Abstraction is domain dependent and understands the whole document and produce the summary. Two types of summarization- Single document text summarization Multi-document text summarization The idea of single document summarization dropped and the focus was on multi-document which helps in size reduction , maintaining syntax and semantic relationship Abstractive seq to seq models is generally trained on titles and subtitles. The similar approach is adopted with document context which helps in scaling. Further all the sentences are rearranged in the order during the inference. Document summarization is changed to supervised or semi-supervised problems. Hints are given like topic words; blacklist words etc. in case of supervised learning approach which are used to identify the sentences as positive or negative classes or the sentences are manually tagged. Then the binary classifier helps in obtaining the scores of each sentence. However they are not successful in providing document specific summaries. The document helps in predicting if the document level information is not provided. **Abstraction** is domain dependent and understands the whole document and produce the summary. Two types of summarization-

	Single document text summarization **Multi-document text summarization** The idea of single document summarization dropped and the focus was on multi-document which helps in size reduction ,maintaining syntax and semantic relationship

2.2.3

Title	**Text Summarization Techniques: A Brief Survey[3]**
Authors	**Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe,** **Juan B. Gutierrez, Krys Kochut**
Year of Publications	November 2017
Publishing Details	(IJACSA) International Journal of Advanced Computer Science and Applications
Summary	This paper surveys the distinctive procedures for synopsis and portrays the adequacy and inadequacies of the diverse strategies. Content outline helps in shortening a content report into a packed form keeping all the imperative data. Programmed content synopsis is the undertaking of delivering a short rundown while protecting key data Substance and by and large meaning. There are numerous precedents like web crawlers create pieces as the sneak peeks of the archives, Different models helps in incorporating the new sites which helps in delivering packed portrayals of news points more often than not as features to encourage perusing. Programmed content synopsis is troublesome and non paltry errand. Luhn et al presented a technique to choose acclaimed sentences from the content utilizing highlights,

	for example, word and expression recurrence. The heaviness of the sentence was proposed. High recurrence words were given high inclination. The two methodologies for programmed rundown is extraction and reflection. Extractive outline strategies distinguish imperative segments of the content and creating them precisely. Abstractive synopsis techniques deliver Imperative component recently. Characteristic dialect handling is utilized to translate and inspect the content with the end goal to create another shorter content. Extractive rundown gives preferred outcomes over abstractive outline.

2.2.4

Title	**Sentiment Analysis and Text Summarization of Online Reviews: A Survey**
Authors	**Pankaj Gupta, Ritu Tiwari and Nirmal Robert**
Year of Publications	April 2016
Publishing Details	International Conzatiference on Communication and Signal Processing
Summary	Content rundown has incited the enthusiasm of numerous specialists in past few years, since the literary information is valuable for some genuine applications and issues. These surveys are expanding step by step and have turned out to be vital to condense the record. It is extremely troublesome for person to abridge the archive. In Content synopsis, significance of sentences is chosen dependent on correct component of sentences. This paper provides a far reaching diagram

	of content outline and gives phenomenal queries about research and methodologies for future appearances. Content synopsis acquires valuable data from this expansive information which can be utilized to make the summary. The fundamental point is to look at how changed techniques have been utilized to assemble rundown frameworks and perform surveys.

2.2.5

Title	**A Survey of Text Summarization Extractive Techniques**
Authors	**Vishal Gupta University Institute of Engineering & Technology, Computer Science & Engineering, Panjab University Chandigarh, India** **Gurpreet Singh Lehal Department of Computer Science, Punjabi University Patiala, Punjab, India**
Year of Publications	August 2010
Publishing Details	JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3,
Summary	Text Summarization flattens the document into summary which maintain its important information. It becomes very difficult for human beings to summarize the paragraph. There are basically two approaches i.e.-extractive and abstractive summarization. **Extractive approach**-The sentences which are important are selected from the provided document and converts into summary. Based on its statistical and semantic features the importance is

	decided of the particular sentence. **Abstractive Approach** - It consists of understanding the original text and converting in summary. It checks the test and interprets it. It describes by generating in shorter form which includes most important information from the given document. Therefore, a twofold problem is faced for important documents through number of documents available, and it absorbs the large quantity of important information. Automatic text summarization is use to short the source text into a shorter version protecting its information content and overall meaning. The advantage of the summary is that the reading time is reduced. The repetition is kept to be minimum. Summarization tools also search for headings to identify the key points of a document. Microsoft Word's AutoSummarize function is example of text summarization. Extractive text summarization process is divided into two steps 1) Pre Processing step 2)Processing step. Pre Processing is a structured description of the original text. It usually involves: a)Sentences boundary identification- it is identified with the appearance of dot at the end of decision. b) Stop-Word Elimination-Common words with no semantics. c) Stemming—The stemming is used to get the stem of every word, that highlight its semantics. In Processing step, the importance of given sentence is decided and the weight is assigned using weight learning method. The score is calculated using Feature-weight equation. The sentences containing highest ranking are converted for summary.

2.2.6

Title	**Abstractive document summarization with a Graph-Based attentional neural model**
Authors	**Jiwei Tan, Xiaojun Wan,Jianguo Xiao** **Institute of Computer Science and Technology,Peking University**
Year of Publications	August 2017
Summary	Document summarization helps in generating a fluent and short summary of provided document and helps in keeping important information as it is very difficult to read whole paragraph. People are facing lot of problem in today's era, document summarization has been investigated. Two approaches are used for text summarization i.e. - Extractive and Abstractive approach. In case of extractive approach summary is generated by extracting sentences or paragraph from provided document. It helps in preserving the meaning of original text along with the important sentences but it has a drawback of information redundancy and incoherence between sentences. Hence the other approach is used for the text summarization i.e.- abstractive method. It generates better summary with the help of arbitrary words and understanding the expression. Although it is very difficult approach as it involves various techniques including meaning representation, content organization, and surface realization. As the starting the abstractive approach cannot always guarantee grammatical abstracts. RNN enable an end-to-end framework for natural language generation. Although the approach is very successful in machine translation and captioning of images. Unfortunately, document summarization in case of abstractive approach is not straight forward. Encoding and decoding of multiple sentences, still does not provide relevant and satisfactory solutions. In this paper, the review of document is done through key factors of document summarization i.e.- the salience, fluency, coherence, and novelty requirements of the generated summary.

	.Distraction mechanism is used to avoid the redundancy. In this paper there is study of how neural summarization model generate or get to know salient information of a particular document. By studying and seeing graph-based extractive approach the novel based approach is discovered in the encoder-decoder framework. Seqtoseq model is also generated and discovered a new hierarchical decoding algorithm is determined with a reference mechanism which generate the abstractive summaries. The method helps to check constraints of saliency The proposed method, non-redundancy, information correctness, and fluency under various framework. Experiments have been conducted on two large-scale corpora with human generated summaries. Results are produced successfully that outperforms previous neural abstractive Summarisation models, and is also competitive with state-of-the-art extractive methods. Various methods, experiments have been provided in given paper.

2.2.7

Title	**Framework of automatic text summarization using Reinforcement learning**
Authors	**Seonggi Ryang, Graduate school of Information science and technology, University of Tokyo** **Takeshi Abekawa, National institute of informatics**
Year of Publications	August 2012

Summary	Well organized summary is generated of single and multiple documents. Multi-document summarization has become very important part of our daily lives as there is lot of information about one particular topic so it becomes very difficult to read. Summary of document helps to easily understand about the topic and important information is generated. The extractive approach is used which is popular for document summary. Summary is generated by selecting words and sentences from the provided document because it is difficult to guarantee the linguistic quality. Marginal relevance (MMR) is used which is used to score every textual unit and take out the highest score. Greedy MMR algorithm is also used but due to its greediness they don't take into account the whole quality. Global inference algorithm is also used for summary. However, these algorithms create lot of problem in formulation of integer linear programming for scoring and the time complexity is very hard. So there is great need of efficient algorithms. In this paper the new approach in generated called Automatic Summarization using Reinforcement Learning (ASRL), where the summary is generated within framework and scores the function of summary. The method is used and adapts to problem with automatic summarization in natural way. Sentence compression is also adapted as action of framework. ASRL is evaluated which is comparable with the state of ILP-style taking rouge score into consideration. Evaluation is done on basis of execution time. State space is searched efficiently for sub optimal solution underscore functions and the score function, and produce a summary whose score denotes the expectation of the score of the same features' states. The quality of summary only depends on score function.

2.2.8

Title	**Neural Abstractive text summarization with sequence-to - sequence models**
Authors	**Tian shi, Yaser Keneshloo, Naren ramakrishnan, Chandan K. Reddy, Senior member, IEEE**
Year of Publications	August 2018
Summary	It is very difficult and challenging to generate the summary of large number of textual data. Automatic text summarization helps in generating the summary. The basic task is to convert long paragraph into short summary. It conserver all the important information and meanings. In General two approaches are used. Extractive approach helps in producing correct summary. Abstractive summarization produces novel words using language generation models. It produces great quality of summary and includes external knowledge. Many evaluation measures have been used for better performance. In this paper concentration is on sequence-to-sequence models. SeqtoSeq have been used in natural language processing tasks for example machine translation, Generation of headlines and summary of given text. They first worked on a neural attention seq2seq model by concentrating on encoder and neural network language model (NNLM) decoder to the abstractive sentence. Various elements were used to RNN encoder-decoder architecture for difficult problems in case of abstractive approach that further includes encoder to capture keywords, switching generator-pointer to model out-of-vocabulary (OOV) word and the hierarchical attention to capture hierarchical document structures. There were various drawbacks as well for multi sentence documents like the salient information was not generated correctly and could not handle OOV words. It was very difficult to determine word and sentence repetitions. The pointer generator is used which helps in copying the words from text via pointer with the help of pointing

	important information is generated accurately and OOV words are also taken into consideration while summarizing a paragraph. Coverage mechanism, intra-decoder attention mechanisms and many more approaches have been used.

2.2.9

Title	**Neural Summarization by extracting sentences and words**
Authors	**Jianpeng Cheng,ILCC,school of informatics,University of Edinburgh** **Mirella Lapata,10 crichton street, Edinburgh**
Year of Publications	July 2016
Summary	It is very difficult to digest huge amount of data which has provided stimulus to generate automatic summarization where the main focus is on generating summary from given document. Focus is on sentence extraction. It helps to create the summary by identifying salient text units. The extractive approach involves the different words in title, nouns, word frequency, action nouns etc. There are various methods selected to generate summary like graph based algorithms, sentences ranging from binary classifiers, integer linear programming. The data driven approach has been used in this paper with the help of neural networks and sentence features. Machine translation has become very important part. Encoder Decoder is used by recurrent neural network. The encoder helps in reading the source sequence into the list whereas decoder helps in generating the target sequence. Framework is developed for single documents summarization which helps in extracting the sentences. The model includes neural network-based hierarchical document reader or encoder and an attention-based content extractor. Reader helps in telling the meaning of paragraph based on sentences. Model include different neural network to extract sentences. The model directly

	focuses on selecting sentences to generate output summary. Neural arch. Have also been used for geometry reasoning. With the help of number of transformation and scoring algorithms, highlights to document content are matched and two large training data sets are constructed. One is generated for sentence extraction and the other one is generated through word extraction. They viewed summarization as a problem analogous to statistical machine translation and generate headlines using statistical models for selecting and ordering the summary words. Model keeps on operating representations and helps in producing multi-sentence output and organizes summary words into sentences so that it is easy for reader to read. Meaning of sentences is also determined and employs neural network directly to generate actual summarization. Neural attention model is generated for abstractive sentence compression which is trained in on different pairs of headlines and different sentences in article. Rather than selecting whole vocabulary the decoder selects output symbols. Model helps in accommodating both generation and extraction. The evaluation is done by both the ways i.e. - automatically and by humans on both the datasets.

2.2.10

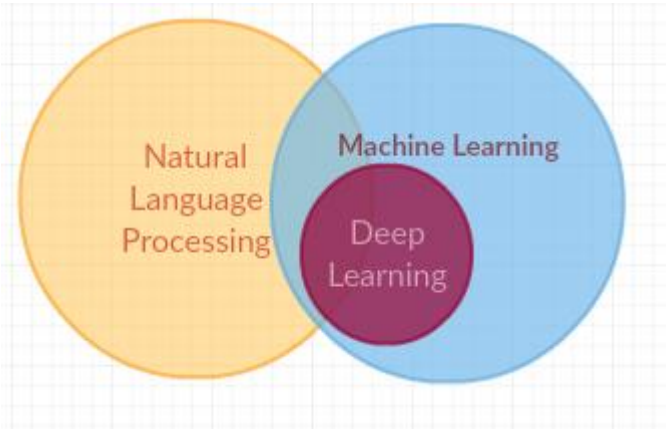
Title	**A Neural Attention Model for Abstractive Sentence summarization**
Authors	**Alexander M. rush, Facebook AI research/ Harvard SEAS** **Sumit Chopra, Facebook AI research** **Jason Weston, Facebook AI research**
Year of Publications	sep. 2015
Summary	It has become very important to summarize the document as it becomes very easy to get whole information in summary. The basic aim is to grasp the main and important meaning of the original document. Extractive summarization is very successful approach which crop out and combine whole portion to produce better version. Whereas, abstractive summarization produces bottom up summary. The focus is on task of sentence level summarization. From the previous work either linguistic-inspired constraints or with syntactic transformation of input text. Full data driven approach is used for generating abstractive summary. There is the combination of neural language model with a contextual input encoder. Both the models are trained on jointly on sentence summarization task. It also incorporates beam search decoder. It incorporates less linguistics structure than other approaches. Since the system does not make any assumption about the vocabulary of the summary so it is trained directly on any document summary pair.

Chapter-3

System Development

3.1 Natural Language Processing

Natural Language Processing (NLP) is the intersection of Computer Science, Linguistics and Machine Learning that is involved with the interaction between computers and humans in natural language.



NLP is way toward empowering PCs to comprehend and deliver human dialect. Uses of NLP systems are utilized in separating of text, machine interpretation and Voice Agents like Alexa and Siri. NLP is one of the fields that are profited from the advanced methodologies in Machine Adapting, particularly from Profound Learning strategies.

Regular Dialect Preparing method utilize the characteristic dialect toolbox for making the principle arrange in python tasks to work with human dialect data. This is simpler to-use by giving the interfaces to at least one than 40 corpora and dictionary resources, for portrayal, for part passages sentences and to get the words in its unique frame Marking, parsing, and glossary thinking for current reasoning quality basic dialect dealing with libraries, and for dynamic discourse. The NLTK will utilize a colossal instrument area and will make some help for individuals with the whole basic dialect taking care of system. This will assist individuals with part sentences from sections, to part up words, seeing the syntactic segments of those words, denoting the fundamental subjects, doing this it serves to your machine by acknowledging the main thing to the substance.

3.2 Lesk Algorithm

NLP is the way toward empowering PCs to comprehend and deliver human dialect. Uses of NLP systems are utilized in separating of text, machine interpretation and Voice Agents like Alexa and Siri. NLP is one of the fields that are profited from the advanced methodologies in Machine Adapting, particularly from Profound Learning strategies.

Regular Dialect Preparing method utilize the characteristic dialect toolbox for making the principle arrange in python tasks which work with human language data. This is simpler to-use by providing the interfaces to at least one than 40 corpora and dictionary resources, for portrayal, for part passages sentences and to get words in its unique frame. marking, parsing, and glossary thinking for current reasoning quality basic dialect dealing with libraries, and for dynamic discourse. The NLTK will utilize a colossal instrument area and will make some help for individuals with the whole basic dialect taking care of system. This will assist individuals with role sentences from sections, seeing the syntactic segments of those words, denoting the fundamental subjects, doing this it serves to the machine by acknowledging the main thing to the substance.

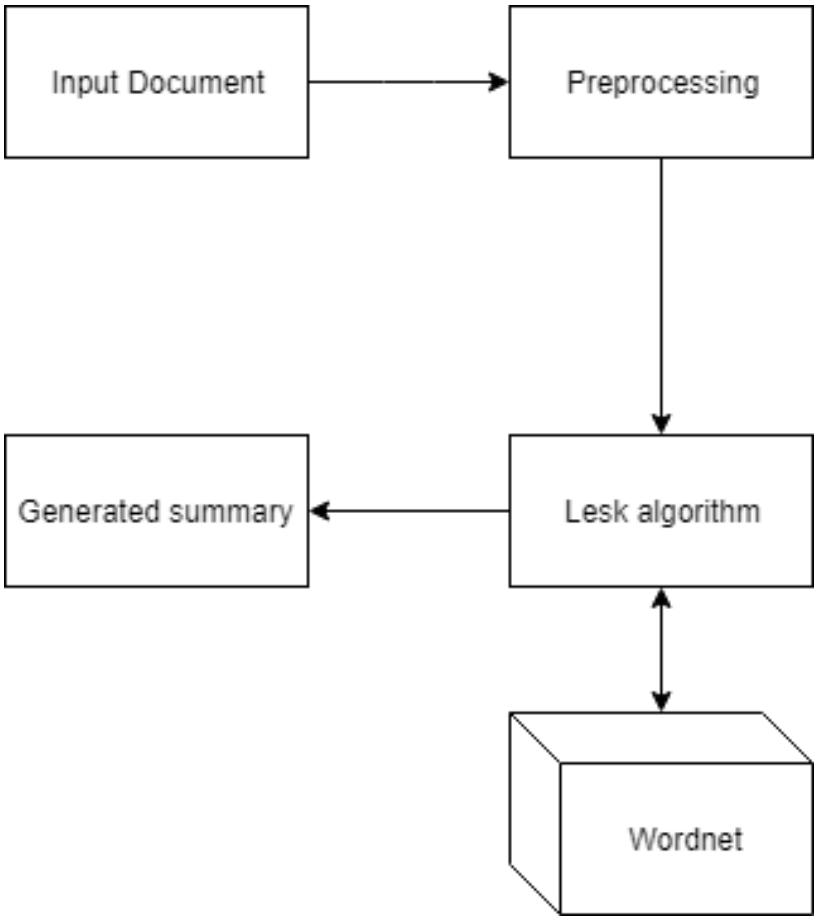
3.3 About WordNet

The wordnet is arranged semantically which creates the electronic database of verbs, adjectives etc. Similar words are grouped together to form the synonym sets. The algorithm is used which removes the words that belong to at least one synset and is known as wordnet words. Synsets are interrelated by semantics and lexical relation. Word net not only links the words but also the senses of the words. It group together all the English words and provides the short definitions. It is accessible to human users via a web browser and is used in automatic text review and artificial intelligence utilizations

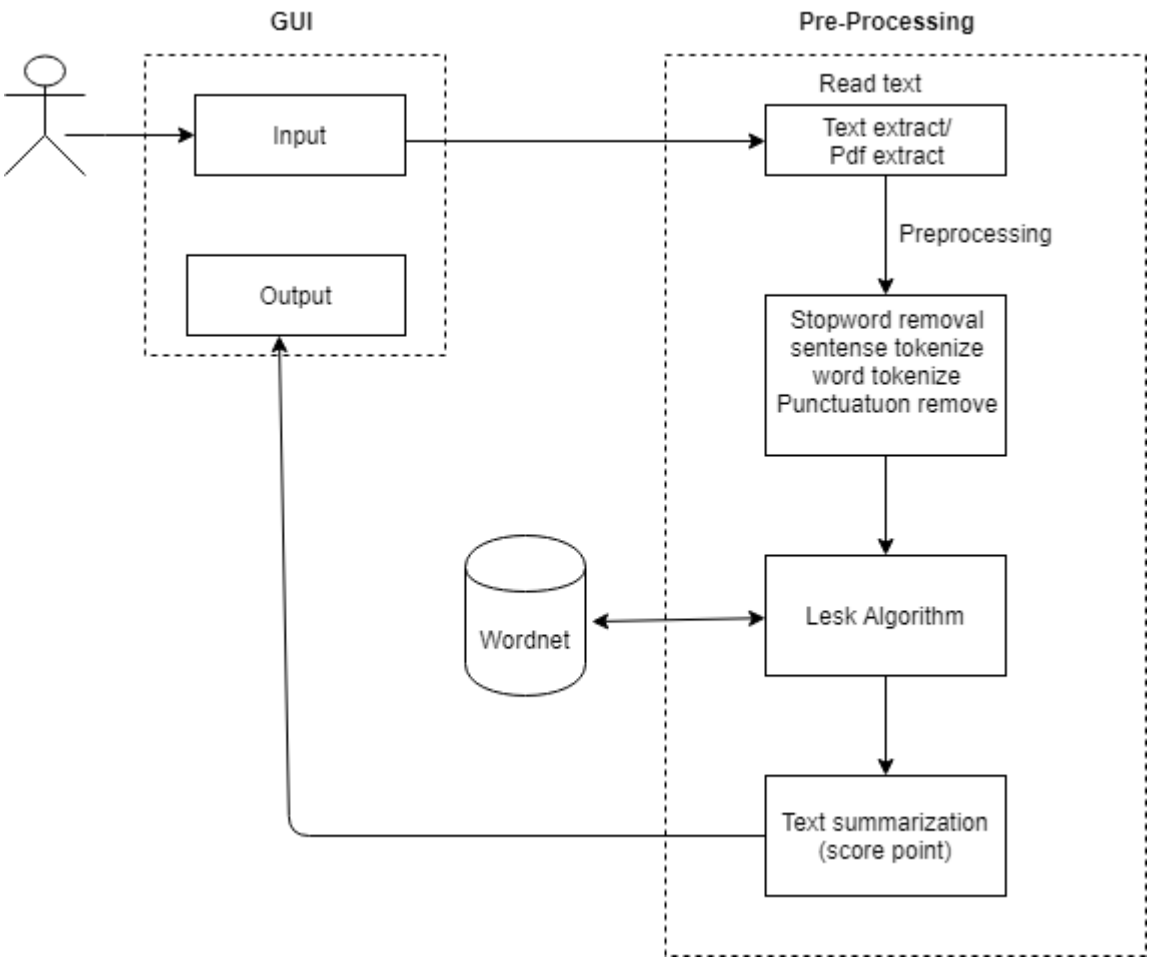
prepositions and other function words and includes nouns, verbs etc. All Synsets are connected by semantic relation.

3.4 Proposed System for Extractive approach

In the Automatic Text summarization, Singular input content is made by using unsupervised learning which will outline the profound rate of summarization. To find the score of various sentences there is the connection between each other is streamlined lesk computation. All the sentences having the more weight are chosen. As per rate of summarization various sentences are selected.



3.5 System Architecture for Extractive Approach



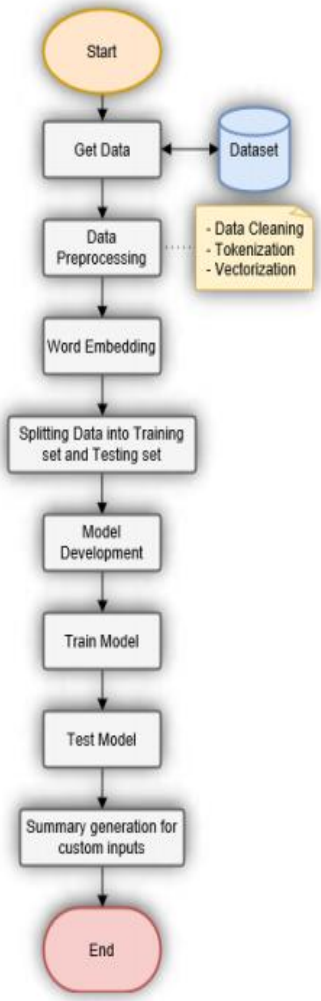
Step 1: Data Pre-Processing Programmed record outline generator helps in removing the things which are not required and occurs in substance. Hence there are sentence part, empty stopwords and perform stemming.

Step 2: Evaluation is further done by the weights Lesk count and word net is used to process the repeat of every sentence. For all N number of documents number of total is spread and founded between detail and brilliance. Further, a specific sentence of the document is selected for every sentence. From every sentence, the stop words are removed as there is no intrigue in sense assignment process. Every word is removed with the help of Wordnet. The document is selected and performed between the sparkles and the data content. When it is overall the intersection guides comes to the largeness of the sentence.

Step 3: Summarization this is the last stage for automatic summarization. The last outline of the particular stage is evaluated the introductions of the yield and survey is done at the time when all the sentences are arranged. Firstly, it will select the onceover of sentences with weight and are planned in jumping demand which is concerned by the increasing weights. Various numbers of sentences are picked from the rate of summary. Further the sentences which are picked are recomposed by the gathering in information. Further, the sentences which are selected are gathered without any dependence of any particular object rather than the denotative erudition lying in the sentence. Restrained matter once-over is without spoken language.

3.6 Proposed system for Abstractive Approach

The proposed approach uses machine and deep learning concepts. The flow chart for this approach is as follows:



3.8 Platform Used

3.8.1 Windows 10

Windows 10 is defined as the Microsoft which works with the particular framework for PCs, tablets, and inserted gadgets etc. Microsoft discharged Windows 10 is follow-up to Windows 8. It was said on July that the window 10 will be refreshed instead of discharging it and framework as a successor.

When the window 10 is selected or received can be updated by inheritance straightforwardly from window 7, 8 or window 10. Without performing meddling and the framework redesign methodology. For maintenance clients run the windows 10 which helps in exchanging the application on the past operating system and setting to window 10. Clients pickup and fill or refresh window 10. With the help of window refresh partner window 10 can be redesigned to physically start an overhaul for Windows.

Windows 10 is used to highlight work in capacities which through which IT offices enables to utilize mobile phones the board (MDM) programming to anchor and control gadgets helps in running working framework. For given boarding programming For example, Microsoft Framework Centre Arrangement Chief. Microsoft Windows 10 is used for multifaceted validation advances, for example, smart cards and tokens. Further, Windows Hi has the biometric verification to Windows 10, where clients can sign in with a unique finger impression, or facial acknowledgment.

The framework is used to incorporate virtualization-based security tools, for example, Secluded Client Mode, Windows Safeguard Gadget Watch and Windows Protector Qualification Monitor. The Windows 10 is used to keep the highlights of particular information, procedures and client certifications separated trying to resolve the problem from any strike. Windows 10 is newer version for Bit Locker encryption to confirm information between clients' gadgets, stockpiling equipment, messages and cloud administrations.

Windows 8 came up with the new idea and gave touch-empowered motion driven UI like those on cell phones and tablets, but there was not much interpretation of well to customary work area and workstation PCs, particularly in big business settings. In Windows 10, Microsoft venture to address this issue and different behaviour of Windows 8, for example, an absence of big business neighbourly highlights.

The declaration of Windows 10 in September 2014 from Microsoft was made and window insider was made that time. There was the discharge from Microsoft to Windows 10 by seeing the total population in July 2015. After that clients observed that Windows 10 is cordial than Windows 8 because it was more conventional interface, which echoes the work area engaging format of Windows 7.

The Windows 10 consecrate Refresh, which turned out in August 2016, made some modifications to the assignment bar and Begin Menu. It additionally presented program augmentations in Edge and gave client's access to Cortana on the bolt screen. In April 2017, Microsoft discharged the Windows 10 Makers Refresh, which made Windows Hi's facial acknowledgment innovation quicker and enabled clients to spare tabs in Microsoft Edge to see later.

The Windows 10 Fall Makers Refresh appeared in October 2017, adding Windows Safeguard Adventure Monitor to secure against zero-day assaults. The refresh likewise enabled clients and IT to put applications running out of sight into vitality productive mode to safeguard battery life and enhance execution.

3.8.2 Ubuntu 18.04

Ubuntu is a free and open-source working framework and Linux conveyance which dependents on Debi an. Ubuntu has three authority versions: Ubuntu Work area for PCs,

Ubuntu Server for servers and the cloud, and Ubuntu Centre for the Web of things gadgets and robots. Ubuntu which are new happen at regular intervals, while long haul bolster (LTS) discharges happen like clockwork, and the latest one is, 18.04 LTS (Bionic Beaver), is upheld for a long time.

Ubuntu is formed by Standard and the designer network, under a meritocratic administration display. Ubuntu is named after the Southern African rationality of ubuntu, which Accepted proposes can be inexactly made an interpretation of as "mankind to other people" or "I am what I am a direct result of who we as a whole are".

Ubuntu is the working framework for the cloud and is the reference working framework for Open Stack. With our committed server designs, you don't need to share the assets. You are qualified to utilize 100% of the gave server to deal with the activity to your site and deal with the business. On the off chance that your requirements develop with the time you can redesign the business to a greater and quicker server. We ensure that we develop as your business develops. We give the most secure, solid and adaptable Ubuntu based committed server facilitating administrations. Hostingraja gives concentrated and altered a devoted server on Ubuntu working framework.

It is the second most utilized Linux enhance for committed facilitating and VPS facilitating in the facilitating Enterprises. Being the mainstream decision of the holsters for running Linux virtual machines or devoted server, the second most well known is Ubuntu. Ubuntu is for the most part utilized in cloud stage or any cloud facilitating arrangements.

Some bundles are -

“ATutor, b2evolution, CMS Made Straightforward, CMSimple, CMSimple_XH, Coderity, Composr CMS, concrete5, Contao, DokuWiki, Dotclear, Drupal, Type CMS, eZ Stage, eZ Distribute, Geeklog, GetSimple CMS,Grav, Habari, ImpressCMS, ImpressPages, Jamroom, Joomla!, Kajona, Known, Magento, Mambo, MediaWiki,

MiaCMS, Microweber, Midgard CMS, MODX, Novius OS, Core CMS, OctoberCMS, Omeka, OpenCart, papaya CMS, pH7CMS, Phire CMS, PHP-Nuke, phpWebLog, phpWiki, Pimcore, PivotX, Pixie (CMS), PmWiki, Prestashop, ProcessWire, Luck, SilverStripe, SMW+, SPIP, Textpattern, Tiki Wiki CMS Groupware, TYPO3, WordPress, Xaraya, XOOPS"

This is the main PHP related stages there are numerous stages lick JAVA, Java bundles/package, Microsoft the Authority Microsoft ASP.NET Site, Perl, Python, etc the rundown simply continue endlessly HostingRaja assists with all the innovation which is utilized by Ubuntu working framework.

3.9 Python 2.7

Python is termed as interpreted, object-oriented, high-level programming language with dynamic semantics. The high-level is made in data structures, in combination with dynamic typing and binding, which helps in making it very attractive for Rapid Application Development. It helps in scripting languages where the components are together. Python is termed as very simple, easy to learn and has simple syntax. It helps in reducing the cost. Python supports packages which encourages program code reuse. The extensive standard library is available in source and can be freely distributed.

Often, programmers like coding in Python because it provides productivity. Editing testing and debugging cycle is very fast. Debugging is very easy in python. Whenever the interpreter finds an error it generates the exception. If this does not happen then interpreter prints a stack trace. A source level debugger helps in inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The fastest way to debug a program is by adding few print statements to the source.

Chapter-4

Performance Analysis

4.1 Approaches to Sentence Extraction

The basic idea of extractive-summarization is to distinguish and extricate vital record sentences and set up them collectively as a rundown; for example produced rundown is an accumulation of unique sentences. There are many ways to deal with sentence extraction. The accompanying subsections will depict three methodologies, in particular, recurrence based methodology, include based methodology and machine learning based methodology.

4.1.1 Frequency based approach

In each work based on content summarization, which was spearheaded, it was expected that vital words in the record are rehashed ordinarily contrasted with the different words in the record. Along these lines, demonstrate of the significance of sentences in the record by utilizing word recurrence. From that point forward, a considerable lot of the summarization frameworks utilize recurrence of the approaches in the extraction of the sentences. Two procedures that utilization recurrence as an essential frame measures in the content summarization is: word likelihood what's more, term recurrence reverse record recurrence.

4.1.1A Word Probability

It was expected that one of the least complex methods for utilizing recurrence is done by including the crude recurrence of the word i.e., by essentially including every word event the archive. Nonetheless, the actions are enormously affected by the report length. One approach is to get the modification for the report length is by processing the word likelihood. The equation 1 shows the probability of the particular word:

$$f(w) = \frac{n(w)}{N} \qquad (1)$$

Where:
$n(w)$ = The frequency count of the word w in the document
N = The total number of words in the document

The discoveries from the examination conveyed based on human-composed outlines demonstrate that individuals will, in general, utilize word recurrence to decide the key subjects of an archive. In case of summarization framework that misuses word likelihood to make outlines. The Sum Basic framework initially processes the word likelihood from the information archive. Each sentence Sj, it processes the sentence weight as a capacity of word likelihood. Best scoring is picked up on the basis of sentence weight.

$$weight(S_j) = \frac{\sum_{w \in S_j} f(w)}{\left|\left\{w \middle| w \in S_j\right\}\right|} \qquad (2)$$

4.1.1B Term Frequency–Inverse Document Frequency

Term frequency-inverse document frequency is customarily utilized in data recovery to manage visit happening terms in a corpus comprising the similar documents. The motivation is to inscribe the accompanying inquiry: Are on the whole substance words that as often as possible show up in documents are similarly vital? For example, an accumulation of news articles investigating seismic tremor fiasco will clearly contain the word 'quake' in all documents. In this manner the possibility of tf-idf is to diminish the weight age of visit happening words by looking at its corresponding frequency in the document gathering. This property has made the tf-idf to be one of the generally utilized terminologies in extractive synopsis. The frequency is defined as:-

$$tf_{i,j} = \frac{n_{i,j}}{\sum n_j} \quad (3)$$

Where:
$n_{i,j}$ represents the frequency count of the word i in document j.

Every word is partitioned and standardized from aggregate number of the various words in document j. The term is used to weight the calculation is like the word likelihood calculation given in Condition 1. The inverse document frequency of a word I is processed:

$$idf_i = log \frac{|D|}{|\{d | t_i \in d\}|} \quad (4)$$

Where, the total number is further divided in various number of the document in the corpus which has different words. Based on Condition 3 and 4, the of word I in document j is calculated:

4.1.2 Feature-based approach

The characteristic method to decide the importance of the particular sentence to recognize the highlights that mirror the importance of that sentence. The three highlights are characterized considered characteristic to sentence significance i.e., sentence position, the nearness of title word and prompt words. For instance, the starting sentences in a archive, as a rule, depicts the primary data concerning the archive. Along these lines, choosing sentences in view of the position can be the sensible methodology. The following highlights are typically determining sentence pertinence.

Sentence Position

The starting sentences in a record as a rule depict the primary data concerning the archive.

Title/Headline Word

Title words showing up in a sentence could recommend that the sentence contains critical data.

Term Weight

Words with high event inside the record are utilized to decide the significance of the document.

Sentence Length

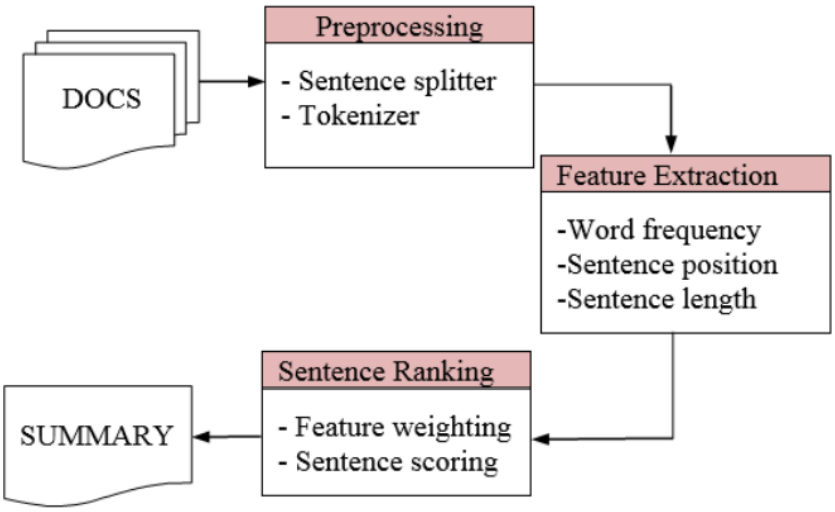
Sentences which are very short contain very less data and the sentences which are long are not proper to speak to outline.

Based on Figure it delineates the basic model of an element based on the summarizer. The scores are registered for each and every element what's more, joined for sentence scoring. Before scoring of the sentence, the highlights are offered weights to decide its dimension of significance. For this situation, highlight weighting will be connected to decide the weights related to each highlight and the sentence score is then registered utilizing the direct mix of each component score duplicated by its comparing weight:

$$Score = \sum_{i=1}^{n} w_i \times f_i \qquad (6)$$

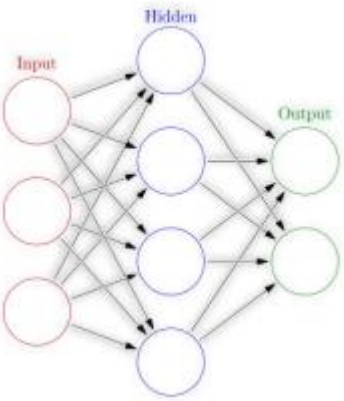
Where:
w_i – The weight of feature i
f_i = The score of feature i

Binwahlan et al. (2009) proposed a content synopsis display dependent on Molecule Swarm Enhancement to decide component weights. The scientists utilized hereditary calculation of rough the best weight blend for the multiple record summary. Different development calculation is additionally been utilized to scale the pertinence of highlight weights. Examination on the impact of various element mix was conveyed by Hariharan where it was discovered that better outcomes were gotten to consolidating term recurrence weight with position and hub weight.



4.2 Approach Using Deep learning

In this project we are going to use the concept of Deep Learning for abstractive summarizer based on food review dataset. So before developing the model, let's understand the concept of deep learning. The basic structures of neural network with its hidden layer are shown in the following figure.



Neural Networks (NN) are also used for Natural Language Processing (NLP), including Summarizers. Neural networks are effective in solving almost any machine learning classification problem. Important parameters required in defining the architecture of neural network (NN) are total amount of hidden layers used, number of hidden units to be present in each layer, activation function for each node, error threshold for the data, the type of interconnections, etc. neural networks can capture very complex characteristics of data without any significant involvement of manual labour as opposed to the machine learning systems. Deep learning uses deep neural networks to learn good representations of the input data, which can then be used to perform specific tasks.

4.2.1 Recurrent Neural Network (RNN)

Recurrent Neural Networks was formed in the year 1980's but is very popular helps in increasing the power which is computational from GPU. They are useful in terms of sequential data because neuron can use its internal memory. It helps in maintaining the information about the past input. This is great because in cases of language, "I had washed my house" is very different than "I had my house washed". The network helps in gaining deep understanding of the given statement. A RNN contain loops in them where the information is taken across neurons while reading the given input.

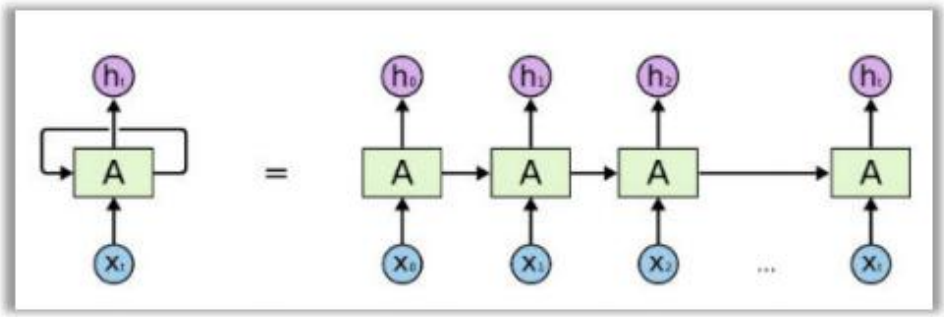


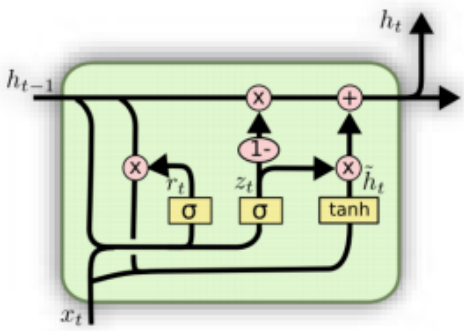
Figure · Unrolled RNN Unit

Here xt is the i/p, A is termed as a part of the RNN and ht is the o/p. The words are feed from given sentences. Or even various characters from a string as xt and will come upwards with ht. The ht is used as o/p and the comparison is done to given test data. Hence, the error rate will be determined. After the comparison with o/p from test data the back propagation technique is used. BPTT checks again with the help of network and

check and adjusts the weight depending on error rate.RNN is used to handle context from the starting of the sentence where the prediction is correct.

4.2.2 Long Short Term Memory (LSTM) Units

The LSTM is RNN architecture which can remember past contextual values. These stored values do not change over time while training the model. There are four components in LSTM which are LSTM Units, LSTM Blocks, LSTM Gates and LSTM Recurrent Components. LSTM Unit store values for long time or for short time. LSTM has no activation functions for their recurrent components. Since there are no activation function the values of units does not change for some period until the context is changed. A LSTM Block contains such many units. LSTM's are considered as deep neural networks. These LSTM's are implemented in parallel systems. LSTM blocks have four gates to control the information flow. Logistic functions are used to implement these gates, to compute a value between 0 and 1. To allow or deny information flow into or out of the memory, multiplication of values with these logistic functions is done. To control the flow of new values into memory, input gate plays key role. The extent to which a value remains in memory is controlled by forget gate. Output gate helps in controlling the extent where the value in given memory helps in computing o/p. activation of the block. In some cases, the input and forget gates are merged together into a single gate, hence we can see even 3 gate representations of LSTM. When new value which is worth remembering is available then we can forget the old value. This represents the combining effect of input and forget gate of LSTM.



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Figure ¯ Architecture of LSTM Unit

4.2.3 Encoders and Decoders

For predicting sequence to sequence problems which is effective is known as Encoder-Decoder LSTM. It contains two models: “one for reading the input sequence and encoding it into a fixed-length vector, and a second for decoding the fixed-length vector and outputting the predicted sequence”. Encoder-Decoder LSTM is designed specifically for sequence to sequence problems. It was developed for NLP problems where it gave state-of-the-art performance, majorly in translation of text called statistical machine translation. The method for this thing is sequence embedding. During the task translation, when the i/p sequence was reversed then model was more effective and it was effective on long i/p sequences. This approach has also been used with image inputs.

The approach involves implementation of Bi-directional Encoders. The following figure shows the structure of Bi-directional LSTM’s.

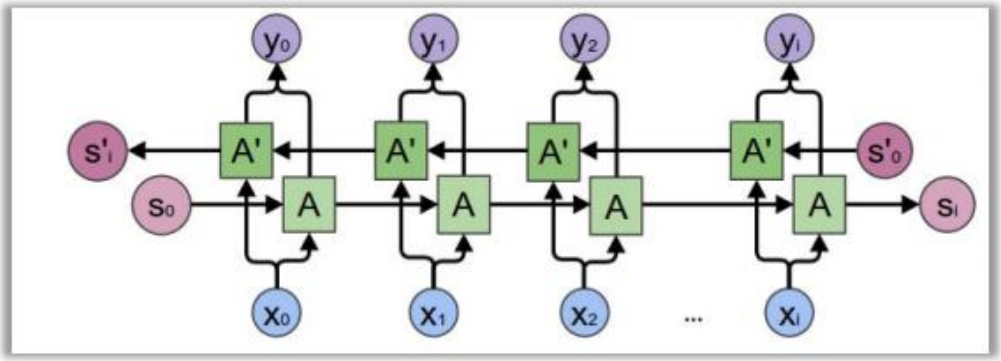


Figure : Bi-directional LSTM's

4.2.4 Attentive Recurrent Architecture

While designing Decoder we will implement Bahdanau Attention. Let, x be the i/p sentence which consists of a sequence of “M words where x = [x1, . . . ,xM], where each word xi is part of vocabulary V, of size |V| = V . Our task is to generate a target sequence y = [y1, . . . , yN], of N words, where N < M, such that the meaning of x is preserved: y =

argmaxy P(y|x)", y is termed as a random variable which denotes a sequence of N words. Conditional probability is modelled by a parametric function with parameters θ: P(y|x) = P(y|x; θ). We need to find θ that helps in maximizing the conditional probability of sentence-summary pairs in the training corpus. If models generates the next word then conditions can be factorized into a product of individual conditional probabilities:

This Conditional probability is implemented using RNN Encoder-Decoder. This model is also called as Recurrent Attentive Summarizer.

$$P(y|x; \theta) = \Pi^{N}_{t=1} p(y_t \mid \{y_1, \ldots, y_{t-1}\}, x; \theta)$$

4.3 Training Dataset

This dataset includes the fine foods review from Amazon. The data includes review from more than 10 years and has around 500,000 reviews up to October 2012. All review include product information, rating and text review. The snippet of dataset is as follows.

Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	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 than most.
2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanuts...the peanuts were actually small sized unsalted. Not sure if this was an error or if the vendor intended to represent the product as "Jumbo".
3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a few centuries. It is a light, pillowy citrus gelatin with nuts - in this case Filberts. And it is cut into tiny squares and then liberally coated with powdered sugar. And it is a tiny mouthful of heaven. Not too chewy, and very flavorful. I highly recommend this yummy treat. If you are familiar with the story of C.S. Lewis' "The Lion, The Witch, and The Wardrobe" - this is the treat that seduces Edmund into selling out his Brother and Sisters to the Witch.
4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient in Robitussin I believe I have found it. I got this in addition to the Root Beer Extract I ordered (which was good) and made some cherry soda. The flavor is very medicinal.
5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wide assortment of yummy taffy. Delivery was very quick. If your a taffy lover, this is a deal.
6	B006K2ZZ7K	ADT0SRK1MGOEU	Twoapennything	0	0	4	1342051200	Nice Taffy	I got a wild hair for taffy and ordered this five pound bag. The taffy was all very enjoyable with many flavors: watermelon, root beer, melon, peppermint, grape, etc. My only complaint is there was a bit too much red/black licorice-flavored pieces (just not my particular favorites). Between me, my kids, and my husband, this lasted only two weeks! I would recommend this brand of taffy -- it was a delightful treat.
7	B006K2ZZ7K	A1SP2KVKFXXRU1	David C. Sullivan	0	0	5	1340150400	Great! Just as good as the expensive brands!	This saltwater taffy had great flavors and was very soft and chewy. Each candy was individually wrapped well. None of the candies were stuck together, which did happen in the expensive version, Fralinger's. Would highly recommend this candy! I served it at a beach-themed party and everyone loved it!
8	B006K2ZZ7K	A3JRGQVEQN31IQ	Pamela G. Williams	0	0	5	1336003200	Wonderful, tasty taffy	This taffy is so good. It is very soft and chewy. The flavors are amazing. I would definitely recommend you buying it. Very satisfying!!
9	B000E7L2R4	A1MZYO9TZK0BBI	R. James	1	1	5	1322006400	Yay Barley	Right now I'm mostly just sprouting this so my cats can eat the grass. They love it. I rotate it around with Wheatgrass and Rye too
10	B001I1APVA	A21BT40VZCCYT4	Carol A. Reed	0	0	5	1351209600	Healthy Dog Food	This is a very healthy dog food. Good for their digestion. Also good for small puppies. My dog eats her required amount at every feeding.

4.4 Training Snippet

```
Epoch  31/100 Batch  780/781 - Loss:  0.663, Seconds: 2.61
Epoch  32/100 Batch   20/781 - Loss:  0.764, Seconds: 2.39
Epoch  32/100 Batch   40/781 - Loss:  0.719, Seconds: 2.46
Epoch  32/100 Batch   60/781 - Loss:  0.663, Seconds: 2.21
Epoch  32/100 Batch   80/781 - Loss:  0.635, Seconds: 2.67
Epoch  32/100 Batch  100/781 - Loss:  0.641, Seconds: 2.47
Epoch  32/100 Batch  120/781 - Loss:  0.585, Seconds: 2.69
Epoch  32/100 Batch  140/781 - Loss:  0.572, Seconds: 2.43
Epoch  32/100 Batch  160/781 - Loss:  0.615, Seconds: 2.71
Epoch  32/100 Batch  180/781 - Loss:  0.642, Seconds: 2.61
Epoch  32/100 Batch  200/781 - Loss:  0.654, Seconds: 2.22
Epoch  32/100 Batch  220/781 - Loss:  0.620, Seconds: 2.42
Epoch  32/100 Batch  240/781 - Loss:  0.596, Seconds: 1.90
Average loss for this update: 0.638
No Improvement.
Epoch  32/100 Batch  260/781 - Loss:  0.581, Seconds: 2.32
Epoch  32/100 Batch  280/781 - Loss:  0.520, Seconds: 2.69
Epoch  32/100 Batch  300/781 - Loss:  0.622, Seconds: 2.33
```

Epoch 32/100 Batch 500/781 - Loss: 0.599, Seconds: 2.70

Average loss for this update: 0.607

New Record!

Epoch 32/100 Batch 520/781 - Loss: 0.645, Seconds: 2.78

Epoch 32/100 Batch 540/781 - Loss: 0.575, Seconds: 2.20

Epoch 32/100 Batch 560/781 - Loss: 0.576, Seconds: 2.29

Epoch 32/100 Batch 580/781 - Loss: 0.601, Seconds: 2.28

Epoch 32/100 Batch 600/781 - Loss: 0.592, Seconds: 2.23

Epoch 32/100 Batch 620/781 - Loss: 0.666, Seconds: 2.60

Epoch 32/100 Batch 640/781 - Loss: 0.591, Seconds: 2.74

Epoch 32/100 Batch 660/781 - Loss: 0.653, Seconds: 2.47

Epoch 32/100 Batch 680/781 - Loss: 0.563, Seconds: 2.32

Epoch 32/100 Batch 700/781 - Loss: 0.587, Seconds: 2.67

Epoch 32/100 Batch 720/781 - Loss: 0.631, Seconds: 2.60

Epoch 32/100 Batch 740/781 - Loss: 0.763, Seconds: 2.17

Epoch 32/100 Batch 760/781 - Loss: 0.756, Seconds: 2.26

Average loss for this update: 0.631

No Improvement.

Epoch 32/100 Batch 780/781 - Loss: 0.652, Seconds: 2.68

Epoch 33/100 Batch 20/781 - Loss: 0.749, Seconds: 2.43

Epoch 33/100 Batch 40/781 - Loss: 0.702, Seconds: 2.40

Epoch 33/100 Batch 60/781 - Loss: 0.657, Seconds: 2.16

Epoch 33/100 Batch 80/781 - Loss: 0.628, Seconds: 2.59

Epoch 33/100 Batch 100/781 - Loss: 0.638, Seconds: 2.33

Epoch 33/100 Batch 120/781 - Loss: 0.581, Seconds: 2.36

Epoch 33/100 Batch 140/781 - Loss: 0.572, Seconds: 2.34

Epoch 33/100 Batch 160/781 - Loss: 0.596, Seconds: 2.57

Epoch 33/100 Batch 180/781 - Loss: 0.635, Seconds: 2.46

Epoch 33/100 Batch 200/781 - Loss: 0.650, Seconds: 2.21

Epoch 33/100 Batch 220/781 - Loss: 0.635, Seconds: 2.32

Epoch 33/100 Batch 240/781 - Loss: 0.596, Seconds: 1.94

Average loss for this update: 0.633

No Improvement.

Epoch 33/100 Batch 260/781 - Loss: 0.590, Seconds: 2.33

Epoch 33/100 Batch 280/781 - Loss: 0.520, Seconds: 2.64

Epoch 33/100 Batch 300/781 - Loss: 0.610, Seconds: 2.39

Epoch 33/100 Batch 320/781 - Loss: 0.596, Seconds: 2.38

Epoch 33/100 Batch 340/781 - Loss: 0.676, Seconds: 2.12

Epoch 33/100 Batch 360/781 - Loss: 0.622, Seconds: 2.65

Epoch 33/100 Batch 380/781 - Loss: 0.593, Seconds: 2.65

Epoch 33/100 Batch 400/781 - Loss: 0.604, Seconds: 2.65

Epoch 33/100 Batch 420/781 - Loss: 0.560, Seconds: 2.38

Epoch 33/100 Batch 440/781 - Loss: 0.585, Seconds: 2.37

Epoch 33/100 Batch 460/781 - Loss: 0.602, Seconds: 2.41

Epoch 33/100 Batch 480/781 - Loss: 0.637, Seconds: 2.44

```
Epoch  33/100 Batch  480/781 - Loss:  0.637, Seconds: 2.44
Epoch  33/100 Batch  500/781 - Loss:  0.578, Seconds: 2.70
Average loss for this update: 0.6
New Record!
Epoch  33/100 Batch  520/781 - Loss:  0.637, Seconds: 3.31
Epoch  33/100 Batch  540/781 - Loss:  0.565, Seconds: 2.12
Epoch  33/100 Batch  560/781 - Loss:  0.565, Seconds: 2.30
Epoch  33/100 Batch  580/781 - Loss:  0.587, Seconds: 2.37
Epoch  33/100 Batch  600/781 - Loss:  0.606, Seconds: 2.51
Epoch  33/100 Batch  620/781 - Loss:  0.667, Seconds: 2.73
Epoch  33/100 Batch  640/781 - Loss:  0.599, Seconds: 2.86

Epoch  33/100 Batch  660/781 - Loss:  0.647, Seconds: 2.45
Epoch  33/100 Batch  680/781 - Loss:  0.549, Seconds: 2.27
Epoch  33/100 Batch  700/781 - Loss:  0.578, Seconds: 2.68
Epoch  33/100 Batch  720/781 - Loss:  0.630, Seconds: 2.45
Epoch  33/100 Batch  740/781 - Loss:  0.763, Seconds: 2.09
Epoch  33/100 Batch  760/781 - Loss:  0.740, Seconds: 2.22
Average loss for this update: 0.627
No Improvement.
Epoch  33/100 Batch  780/781 - Loss:  0.657, Seconds: 2.45
Epoch  34/100 Batch   20/781 - Loss:  0.742, Seconds: 2.41
Epoch  34/100 Batch   40/781 - Loss:  0.697, Seconds: 2.42
Epoch  34/100 Batch   60/781 - Loss:  0.649, Seconds: 2.19
Epoch  34/100 Batch   80/781 - Loss:  0.631, Seconds: 2.62
Epoch  34/100 Batch  100/781 - Loss:  0.630, Seconds: 2.26
Epoch  34/100 Batch  120/781 - Loss:  0.583, Seconds: 2.35
Epoch  34/100 Batch  140/781 - Loss:  0.569, Seconds: 2.33
Epoch  34/100 Batch  160/781 - Loss:  0.601, Seconds: 2.57
Epoch  34/100 Batch  180/781 - Loss:  0.631, Seconds: 2.52
```

Epoch 34/100 Batch 180/781 - Loss: 0.631, Seconds: 2.52

Epoch 34/100 Batch 200/781 - Loss: 0.636, Seconds: 2.18

Epoch 34/100 Batch 220/781 - Loss: 0.619, Seconds: 2.40

Epoch 34/100 Batch 240/781 - Loss: 0.593, Seconds: 1.91

Average loss for this update: 0.628

No Improvement.

Epoch 34/100 Batch 260/781 - Loss: 0.581, Seconds: 2.36

Epoch 34/100 Batch 280/781 - Loss: 0.524, Seconds: 2.63

Epoch 34/100 Batch 300/781 - Loss: 0.604, Seconds: 2.36

Epoch 34/100 Batch 320/781 - Loss: 0.598, Seconds: 2.42

Epoch 34/100 Batch 340/781 - Loss: 0.671, Seconds: 2.13

Epoch 34/100 Batch 360/781 - Loss: 0.642, Seconds: 2.72

Epoch 34/100 Batch 380/781 - Loss: 0.617, Seconds: 2.61

Epoch 34/100 Batch 400/781 - Loss: 0.605, Seconds: 2.66

Epoch 34/100 Batch 420/781 - Loss: 0.561, Seconds: 2.24

Epoch 34/100 Batch 440/781 - Loss: 0.583, Seconds: 2.41

Epoch 34/100 Batch 460/781 - Loss: 0.597, Seconds: 2.41

Epoch 34/100 Batch 480/781 - Loss: 0.634, Seconds: 2.44

Epoch 34/100 Batch 500/781 - Loss: 0.587, Seconds: 2.80

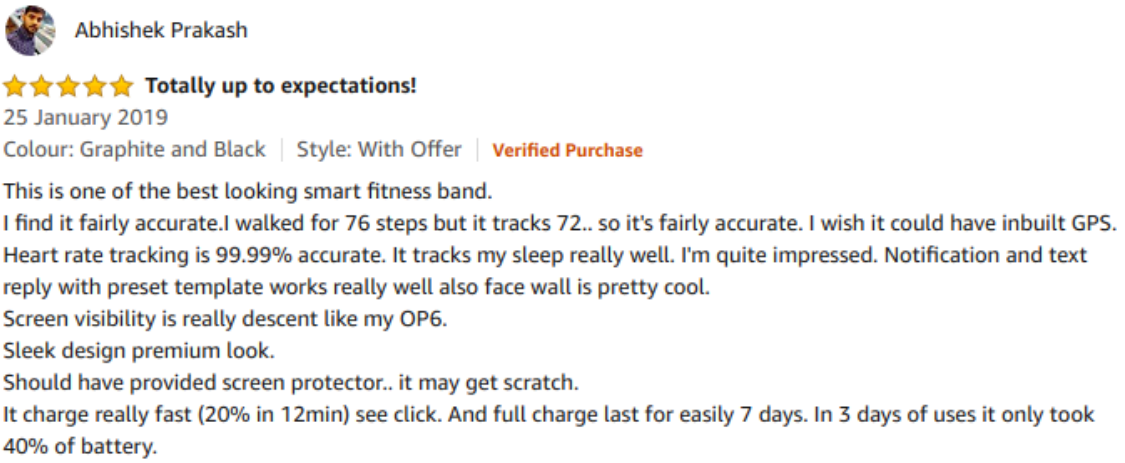
Average loss for this update: 0.605

No Improvement.

Stopping Training.

4.5 Custom Input and Output

4.5.1 Input



Abhishek Prakash

Totally up to expectations!

25 January 2019

Colour: Graphite and Black | Style: With Offer | **Verified Purchase**

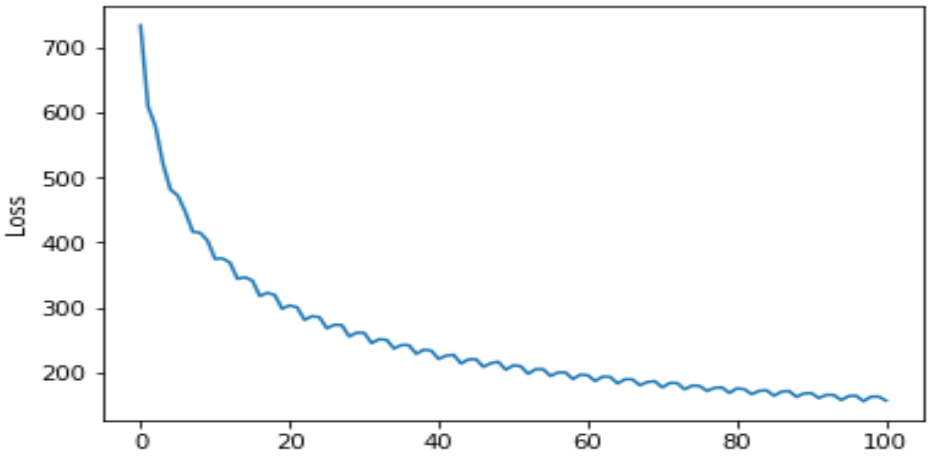
This is one of the best looking smart fitness band.
I find it fairly accurate.I walked for 76 steps but it tracks 72.. so it's fairly accurate. I wish it could have inbuilt GPS.
Heart rate tracking is 99.99% accurate. It tracks my sleep really well. I'm quite impressed. Notification and text reply with preset template works really well also face wall is pretty cool.
Screen visibility is really descent like my OP6.
Sleek design premium look.
Should have provided screen protector.. it may get scratch.
It charge really fast (20% in 12min) see click. And full charge last for easily 7 days. In 3 days of uses it only took 40% of battery.

4.5.2 Output

```
INFO:tensorflow:Restoring parameters from ./best_model.ckpt
Original Text: This is one of the best looking smart fitness band.I find it fairly accurate.I walked for 76

Summary
  Response Words: great for a smart product
```

4.6 Loss Graph



4.7 Rogue Score

Rogue Score after 43 Test Results:

```
2019-05-04 12:26:20
Precision is :0.9230769230769231
Recall is :0.48
F Score is :0.6315799023545738
Sum of ROUGE Score:  13.61002001826115
Average ROUGE Score =  0.3165120934479337
Count:  43
```

Chapter-5

Conclusion

As with time internet is growing at a very fast rate and with it data and information is also increasing. it will going to be difficult for human to summarize large amount of data. Thus there is a need of automatic text summarization because of this huge amount of data. Until now, we have read multiple papers regarding text summarization, natural language processing and lesk algorithms. There are multiple automatic text summarizers with great capabilities and giving good results. We have learned all the basics of Extractive and Abstractive Method of automatic text summarization and tried to implement extractive one. We have made a basic automatic text summarizer using nltk library using python and it is working on small documents. We have used extractive approach to do text summarization.

We have successfully implemented state-of-the-art model for abstractive sentence summarization to recurrent neural network architecture. The model is a simplified version of the encoder-decoder framework for machine translation. The model is trained on the Amazon-fine-food-review corpus to generate summaries of review based on the first line of each review. There are few limitations of the model which can be improved in further work. First limitation is that it sometimes generates repeated words in the summary, the other problem is it takes too much time to generate a summary if the input text size is large enough, the other issue is that for large text input it sometimes miss interpret the context and generates exactly opposite context summary.

5.1 Future Scope

We have implemented Automatic text summarization using abstractive method. Further, after using RNN and LSTM the accuracy is still very low for summarizer. Furthermore, we will be using machine learning for semantic text summarization for more accurate summaries and will try to make a grader which will grade the document according to English grammar. There are many text summarizers available but all does not give appropriate result. Thus we will be using machine learning algorithm to increase the effectiveness of the automatic summarizer.

References

[1] Ahmad T. Al-Taani. “,Automatic Text Summarization Approaches” International Conference on Infocom Technologies and Unmanned Systems (ICTUS'2017)

[2] Neelima Bhatia, Arunima Jaiswal, “Automatic Text Summarization: Single and Multiple Summarizations ”, International Journal of Computer Applications

[3] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut, “ Text Summarization Techniques: A Brief Survey”, (IJACSA) International Journal of Advanced Computer Science and Applications

[4]Pankaj Gupta, Ritu Tiwari and Nirmal Robert,”Sentiment Analysis and Text Summarization of Online Reviews: A Survey”International Conzatiference on Communication and Signal Processing,August 2013

[5]Vishal gupta,Gurpreet Singh Lehal,”A Survey of Text Summarization Extractive Techniques.”JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3, AUGUST 2010

[6]Jiwei Tan, Xiaojun Wan,Jianguo Xiao Institute of Computer Science and Technology,Peking University “Abstractive document summarization with a Graph-Based attentional neural model. ”

[7]Seonggi Ryang, Graduate school of Information science and technology, University of Tokyo Takeshi Abekawa, National institute of informatics “Framework of automatic text summarization using Reinforcement learning”

[8]Tian shi, Yaser Keneshloo, Naren ramakrishnan, Chandan K. Reddy, Senior member, IEEE " Neural Abstractive text summarization with sequence-to -sequence models"

[9]Jianpeng Cheng,ILCC,school of informatics,University of Edinburgh Mirella Lapata,10 crichton street, Edinburgh "Neural Summarization by extracting sentences and words"

[10]Alexander M. rush, Facebook AI research/ Harvard SEAS Sumit Chopra, Facebook AI research Jason Weston, Facebook AI research " A Neural Attention Model for Abstractive Sentence summarization"