**AUTOMATION MULTI-MODAL SUMMARIZATION USING  
DEEP NEURAL NETWORKS**

(Major-Project (Phase -1))

A project report submitted to the Srinivas University as partial fulfilment for the award of the degree of

**Bachelor of Technology in Cloud Technology and Information Security**

Submitted By

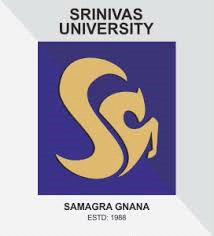
**Ajay Surya S**

**USN: 1SU19DS003**

Under the Guidance of

**Mrs. Renisha**

Assistant Professor



**Department of Cloud Technology & Data Science**

**College of Engineering and Technology**

**SRINIVAS UNIVERSITY**

**Mukka, Mangalore – 574146 December 2022**

**BONAFIDE CERTIFICATE**

This is to certify that this project report entitled” **AUTOMATION MULTI-MODAL SUMMARIZATION USING DEEP NEURAL NETWORKS** is submitted to Srinivas University College of Engineering and Technology, Mukka, is a bonafide record of work done by **Ajay Surya S** under my supervision from 1ST of December 2022 to 28th of December 2022

Mrs. Renisha

Assistant Professor

Prof. Daniel Selvaraj

Head of Department

Department of Cloud Technology and Data Science

Srinivas University, Mukka

Date

Place: Mukka

**Table of Contents**

[LIST OF FIGURES 4](#_Toc123547079)

[ABBREVIATIONS AND NOMENCLATURE 5](#_Toc123547080)

[ABSTRACT 6](#_Toc123547081)

[1 INTRODUCTION 7](#_Toc123547082)

[1.1 THE DOMAIN 7](#_Toc123547083)

[1.2 THE PROBLEM 8](#_Toc123547084)

[1.3 THE TECHNOLOGY 9](#_Toc123547085)

[2 SYSTEM ANALYSIS 11](#_Toc123547086)

[2.1 LITERATURE REVIEW 11](#_Toc123547087)

[2.2 EXISTING SYSTEM 15](#_Toc123547088)

[2.3 PROPOSED METHODOLOGY: 15](#_Toc123547089)

[3 REQUIREMENTS: 20](#_Toc123547090)

[4 REFERENCE 21](#_Toc123547091)

# LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| **FIGURE NUMBER** | **PAGE NUMBER** | **EXPLANATION** |
| FIGURE 1.1 | 8 | NEURAL NETWORKS. |
| FIGURE 1.2 | 9 | THE VOLUME OF DATA CREATED |
| FIGURE 1.3 | 10 | TECNOLOGY USED |
| FIGURE 1.4 | 11 | WEB SCRAPPING |
| FIGURE 1.5 | 16 | VIDEO SUMMMARY USING VIDEO FRAMES |
| FIGURE 1.6 | 16 | FEATURE EXTRACTION |
| FIGURE 1.7 | 18 | OVERVIEW OF VIDEO SUMMERISATION |
| FIGURE 1.8 | 18 | A TWO LAYER LSTM MODEL |
| FIGURE 1.9 | 19 | AUDIO SUMMARY GENERATION |
| FIGURE 2.0 | 20 | SCHEME OF AUTOMATIIC SPEECH SUMMERISATION |

# ABBREVIATIONS AND NOMENCLATURE

|  |  |  |
| --- | --- | --- |
| **ABBREVIATION** | **PAGE NUMBER** | **FULL FORM** |
| NLP | 10 | NATURAL LANGUAGE PROCESSING |
| DNN | 10 | DEEP NUERAL NETWORKS |
| HTML | 10 | HYPERTEXT MARKUP LANGUAGE |
| GAN | 10 | GENERATIVE ADVERSARIAL NETWORK |
| ASR | 10 | AUTOMATIC SPEECH RECOGNITION |
| LSTM | 10 | LONG SHORT-TERM MEMORY |

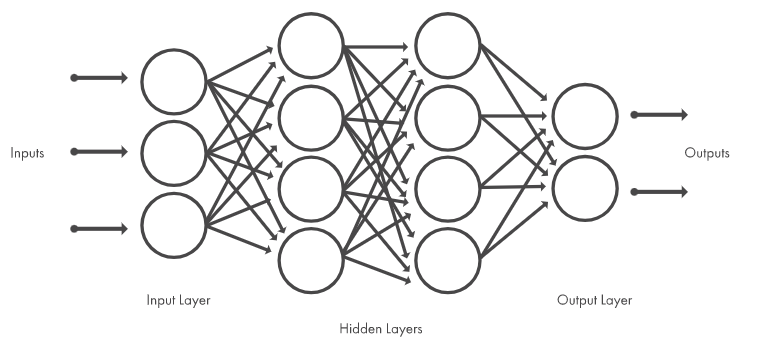
# ABSTRACT

In this project, we are creating a web application through which users can summarise their files. The input can be a link containing text, audio, and video files, or it can be normal text, audio, and video files. Automatic text summarization is basically a summary of the given paragraph using natural language processing and machine learning. We review the different processes for summarization and describe the effectiveness and shortcomings of the different methods. In addition to basic summarization, some attempt is made to address the issue of targeting the text at the user. There has been an explosion in the amount of data coming from a variety of sources. As the demand for video summarization techniques increases nowadays, many methods are proposed for how to extract the best-representing key frames of a video. While most of them rely on hand-crafted image features, we resort to the feature learning power of deep convolutional networks. Different features are video to short video summarization, audio to short audio summarization, video to text summarization, audio to text summarization, and text to text summarization. It will take original video-based features using a deep summarization network, then create small video segments from those selected frames, and then proceed as a short summed-up video. In video-to-text summarization, it will take video as input and then convert it to audio using Python libraries. Then the audio was divided into audio chunks for text extraction. The text ranking algorithm used for summarization work Extracted text showed up as output. In Audio to Short Audio Summarization, it will take audio as input. It will convert audio into text, then perform text processing and text summarization after the selected sentence with time stamps is used for short audio generation.

# 1 INTRODUCTION

## THE DOMAIN

Deep learning is a machine learning technique that teaches computers to do what comes naturally to humans: learn by example. Deep learning is a key technology behind driverless cars, enabling them to recognize a stop sign, or to distinguish a pedestrian from a lamppost. It is the key to voice control in consumer devices like phones, tablets, TVs, and hands-free speakers. Deep learning is getting lots of attention lately and for good reason. It’s achieving results that were not possible before. In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labelled data and neural network architectures that contain many layers. Deep learning models are trained by using large sets of labelled data and neural network architectures that learn features directly from the data without the need for manual feature extraction.

FIGURE 1.1

Neural networks, which are organized in layers consisting of a set of interconnected nodes. Networks can have tens or hundreds of hidden layers.

## THE PROBLEM

There is an enormous amount of data and it is only growing every single day. Think of the internet, comprised of web pages, news articles, status updates, blogs and so much more. The data is unstructured and the best that we can do to navigate it is to use search and skim the results. There is a great need to reduce much of this data to shorter, focused summaries that capture the salient details, both so we can navigate it more effectively as well as check whether the larger documents contain the information that we are looking for. Information in the form of digital documents quickly accumulates to huge amounts of data. Most of this large volume of documents is unstructured: it is unrestricted and has not been organized into traditional databases. Processing documents is therefore a perfunctory task, mostly due to the lack of standards.

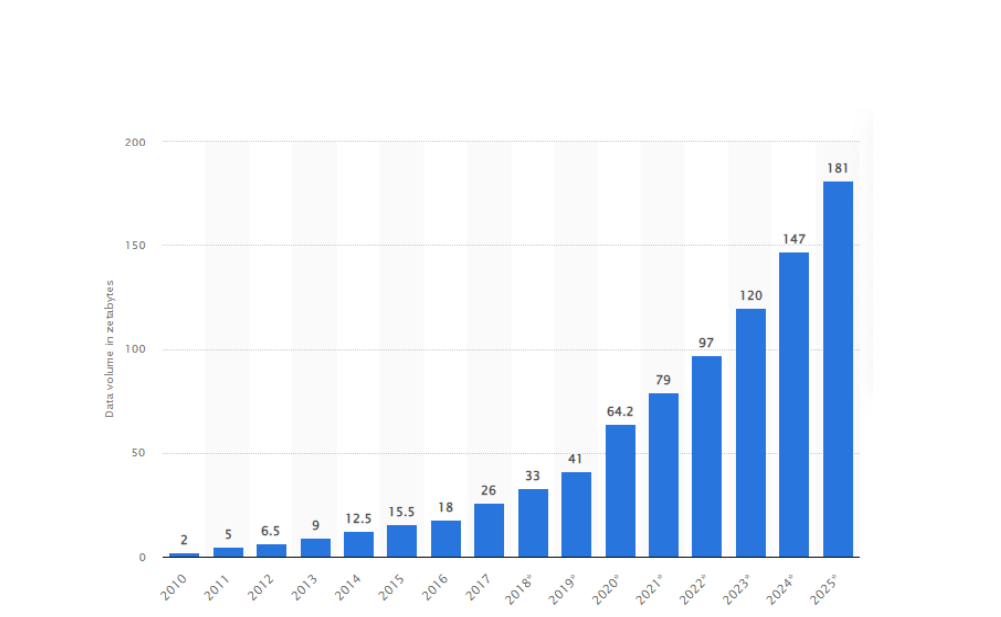


FIGURE 1.2

The volume of data created, captured, or copied and consumed worldwide from 2010 to 2025.

**BENEFITS:**

1. Summaries reduce reading time.
2. When researching documents, summaries make the selection process easier.
3. Automatic summarization improves the effectiveness of indexing.
4. Automatic summarization algorithms are less biased than human summarizers.
5. Personalized summaries are useful in question-answering systems as they provide personalized information.
6. Using automatic or semi-automatic summarization systems enables commercial abstract services to increase the number of texts they are able to process.

## THE TECHNOLOGY

FIGURE 1.3

* NLP text summarization is the process of breaking down lengthy text into digestible paragraphs or sentences. This method extracts vital information while also preserving the meaning of the text. This reduces the time required for grasping lengthy pieces such as articles without losing vital information.
* DNN is a type of machine learning that mimics the way the brain learns. It's been used for a variety of tasks; some that you might be familiar with, like language translation and image search tools
* Web scraping is the process of using bots to extract content and data from a website. Unlike screen scraping, which only copies pixels displayed onscreen, web scraping extracts underlying HTML code and, with it, data stored in a database.

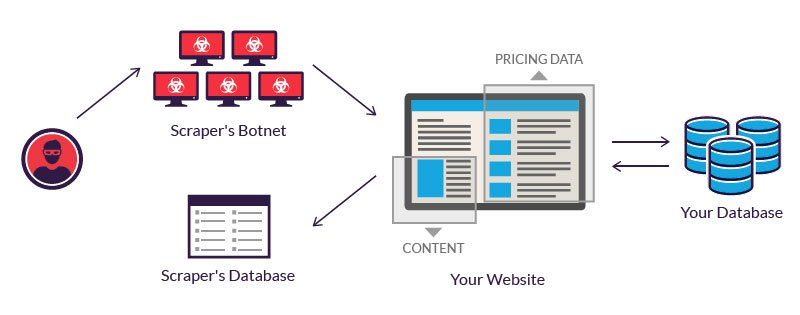


FIGURE 1.4 WEB SCRAPING

* An Encoder Long Short-Term Memory model (LSTM) reads the entire input sequence wherein, at each timestep, one word is fed into the encoder. It then processes the information at every timestep and captures the contextual information present in the input sequence
* Generative adversarial networks (GANs) are an exciting recent innovation in machine learning. GANs are generative models: they create new data instances that resemble your training data. For example, GANs can create images that look like photographs of human faces, even though the faces don't belong to any real person
* Flask is used for developing web applications using python, Advantages of using Flask framework are: There is a built-in development server and a fast debugger
* HTML is the standard markup language for creating Web pages.

# 2 SYSTEM ANALYSIS

## 2.1 LITERATURE REVIEW

In this paper author propose a novel recurrent neural network for the problem of abstractive sentence summarization. Model consists of a conditional recurrent neural network, which acts as a decoder to generate the summary of an input sentence, much like a standard recurrent language model. In addition, at every time step the decoder also takes a conditioning input which is the output of an encoder module. Depending on the current state of the RNN, the encoder computes scores over the words in the input sentence. These scores can be interpreted as a soft alignment over the input text, informing the decoder which part of the input sentence it should focus on to generate the next word. Both the decoder and encoder are jointly trained on a data set consisting of sentence-summary pairs. In this encoder is more sophisticated, in that it explicitly encodes the position information of the input words. Lastly, the encoder uses a convolutional network to encode input words. These extensions result in improved performance.

Chun-I Tsai: In this paper, author wish to use the least human effort to produce reliable results on both extractive and abstractive summarizations. Firstly, author experiment how deep learning approaches encode representations merely based on words yet still outperform empirical features on extractive text summarization. By taking advantages of convolutional neural network, which has shown remarkable ability on learning a sentence representation in a hyperspace, author further extract sentence-document pair features during training without exhausted feature engineering. To bridge the gap between text document and spoken document, we introduce Long Short-Term Memory (LSTM) to capture sematic flow on the sentence level so as to prevent our model from misleading of ASR errors. Secondly, we analyse influences of different word representation techniques. Finally, we investigate the latest method on abstractive summarization, which is constructed by LSTM encoder and decoder with the attention mechanism. This framework has been proven can automatically learn probability distribution of translating a word to another language even without any 6 predefined rules. Author also demonstrates, that the neural summarization is able to compress the meaning of multiple words and generate a word not featured in current article for abstractive summarization. The results show our neural summarization methods outperform classic methods with traditional features for extractive spoken document. On the other hand, different from previous assumptions, the abstractive summarization method we used effectively produces grammatical summary by an end-to-end module which is not a cascaded system

Chin-Yew Lin In: this paper author introduced Recall Oriented Understudy for Gisting Evaluation ROUGE. That is an automatic evaluation package for text summarization. The paper also introduced four different measures of ROUGE: - ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-S. It measures the quality of summary by comparing the generated summary with other ideal summaries that are created by humans. These methods are efficient for automatic evaluation of single document summary as well as multi-document summaries.

Akshil Kumar In this paper author has analysed and compared the performance of three different algorithms. Firstly, the different text summarization techniques explained. Extraction based techniques are used to extract important keywords to be included in the summary. For comparison three comparison three keyword extraction algorithms namely Text Rank, Lex Rank, Latent Semantic Analysis (LSA) were used. Three algorithms are explained and implemented in python language. The ROUGE 1 is used to evaluate the effectiveness of the extracted keywords. The results of the algorithms compared with the handwritten summaries and evaluate the performance. In the end, the Text Rank Algorithm gives a better result than other two algorithms.

Pankaj Gupta In this paper author has reviewed different techniques of Sentiment analysis and different techniques of text summarization. Sentiment analysis isa machine learning approaching which machine learns and analyse the sentiments, emotions present in the text. The machine learning methods like Naive Bayes Classifier and Support Machine Vectors (SVM) are used. These methods are used to determine the emotions and sentiments in the text data like reviews about movies or products. In Text summarization, uses the natural language processing (NPL) and linguistic features of sentences are used for checking the importance of the words and sentences that can be included in the final summary. In this paper, a survey has been done of previous research work related to text summarization and Sentiment analysis, so that new research area can be explored by considering the merits and demerits of the current techniques and strategies.

Kavita Ganesan. In this research paper the author proposed graph-based text summarization framework Opinosis. It generates abstractive summaries. Opinosis works on redundant data like human reviews on movies or products and provides abstractive summaries. Firstly, it creates the direct Opinosis-Graph of the text. Where nodes represent the word units of the text. Three unique graph properties: Redundancy capture, Collapsible structures and Gapped subsequence capture is used to explore and explore different sub-paths that help in the creation of abstractive summaries of the text. The valid path is selected and marked with high redundancy score, collapsed path and summary generation. Then all paths ranked in descending order according to scores. The duplicate paths are removed using Jaccard measure the results are compared with human summaries. Results show Opinosis summaries has better agreement with human summaries. For future work author proposed to use a similar idea to overlay parse trees.

Dharmendra Hanh In this paper the author uses the extractive text summarization. The author gives the Wikipedia Articles as input to the system and identifies text scoring. Firstly, the sentences are Tokenized through pattern matching using regular expressions. Then we get data in form of set of words then stop words are removed from the set of words. The words are then stemmed. Then traditional methods are used for scoring of the sentences. Scoring helps in classifying the sentences if they included in summary or not. It is found that scoring sentences based on citation give better results.

|  |  |  |  |
| --- | --- | --- | --- |
| AUTHOR | YEAR | TECHNOLOGY USED | OUTCOME |
| An Improved Algorithm for Video Summarization – A Rank Based Approach | 2016 | summarization approach | summarization approach is used to summarize Wikipedia Articles. |
| Summit Chopra | 2018 | Attentive Recurrent neural networks | Recurrent neural network for abstractive summarization |
| Chun-I Tsai | 2019 | Neural Network Based Methods for Summarization | A Study on Neural Network Modelling Techniques for Automatic Document Summarization |
| Chin-Yew Lin | 2004 | Graph based approach | Abstractive Summary generation of redundant data |
| Akshil Kumar | 2017. | Graph based approach, Semantic based approach | Performance of Three different algorithms compared Text Rank, Lex Rank and LSA. Text Rank outperforms other two. |
| Pankaj Gupta | 2016 | Sentiment Analysis, Text Summarization Techniques | A Survey is performed on current research in sentiment analysis and Text summarization |
| Kavita Ganesan | 2010 | Graph based | New framework Opinions generates abstractive summary of redundant data |
| Dharmendra Hinhu | 2015 | Extractive Text | Extractive text |

## 2.2 EXISTING SYSTEM

There are some existing systems like which I have shown above in which they are implementing the summarization using python libraries and deep learning techniques. In video summarization they will convert into audio using some python libraries and deep learning concepts and then they will extract then audio and convert into text summary. Text rank algorithm is used for summarization, Text to text summarization they use NLP techniques.

## 2.3 PROPOSED METHODOLOGY:

* It will be a web application
* The input for this can be in two formats that is file formats and through link
* It has five modules those are:
* Video to short video summarization
* Video to text summarization
* Audio to audio summarization
* Audio to text summarization
* Text to text summarization
* Deploying all modules in cloud.

**1.Video to video summarization:**

* The ability to record videos with powerful sensors and share the contents on online social platforms like YouTube, Instagram, Facebook, and Twitter has caused a tremendous increase in the amount of data available in the modern world.
* Video summarization aims to provide an automated way of generating short and informative versions of the original video by identifying the most important and relevant content.
* The frames containing important information are selected as key frames from a given video sequence.
* The proposed model includes Resnet 18 feature extraction technique for features extraction. The Non-Maximum Suppression algorithm and Kernel Temporal Segmentation algorithm are used for Redundancy removal and generating the video summary.

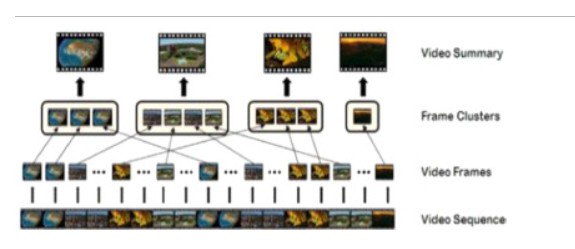


FIGURE 1.5

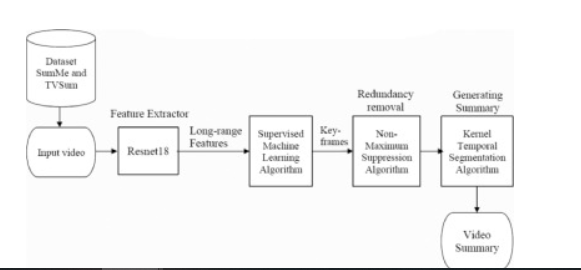


FIGURE 1.6

**2. Video to text summarization:**

Our proposed approach consists of four main components:

1. Identiﬁcation of interesting segments from the full video;

2. Key frame extraction from these interesting segments;

3. Annotations for these key frames are generated using a deep video-captioning network; and

4. The annotations are summarized to generate a para-graph description of the sequence of events in the video

Key frame selection done by open cv

* Video clip captioning is achieved by modifying S2VT with new frame features and introduction of key frame selection. Each key frame is passed through the 152-layer ResNet CNN model
* These key frame feature vectors are passed sequentially into a Long Short-Term Memory (LSTM) network, a recurrent neural network approach used in the speech recognition, language translation, as well as visual annotation.
* The S2VT framework ﬁrst encodes frames, one frame at a time to the ﬁrst layer of a two-layer LSTM.
* Then decoded into a natural language sentence one word at a time, feeding the output of one-time step into the second layer of the LSTM in the subsequent time step
* At the end for text summarization NLTK library is used.

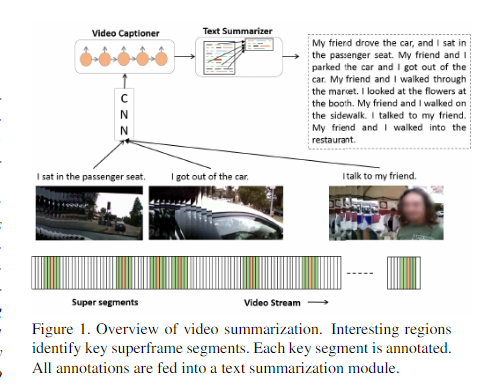


FIGURE 1.7

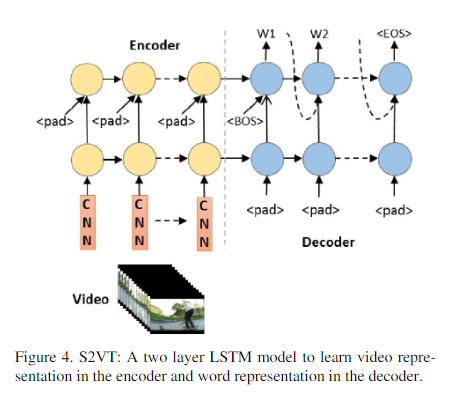


FIGURE 1.8

**3. Audio to audio summarization:**

* It generates a transcript of the audio using an ASR module and parses the text transcript into individual sentences. It then uses a text summarization model to select relevant sentences, along with their time offsets in the audio, and generates the final audio summary associated with the text summary
* Automatic Speech Recognition: ASR methods perform the task of automatic speech-to-text transcription. we choose to use a well known and publicly-available solution for this task, namely AWS Transcribe2
* The transcripts obtained from the ASR module contain the text for the individual words and punctuation marks, their start and end times in the audio, and their confidence scores regarding the prediction. We choose to use an open-source library for NLP, namely spaCy
* Text Summarization: We generate text summaries by selecting relevant sentences from the transcripts, using automatic extractive summarization. We used the recently-proposed PreSumm4 model , which builds upon BERT .
* Audio Generation: We use the time offsets of the selected sentences in the text summary to identify the corresponding audio sections in the podcast and stitch them together to form the final audio summary.

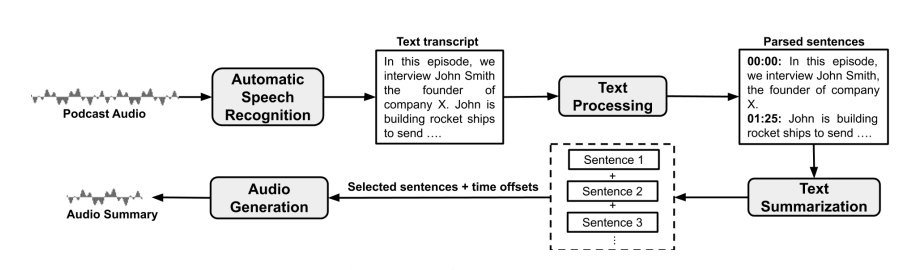


FIGURE 1.9

**4.Audio to text summarization**

* In audio to text summarization, it will take audio as input and using Automatic summarization recognition it will convert audio to text after performing text processing summarization models used to generate text summary
* Automatic Speech Recognition: ASR methods perform the task of automatic speech-to-text transcription
* For text processing and summarization, we are choosing NLP approaches

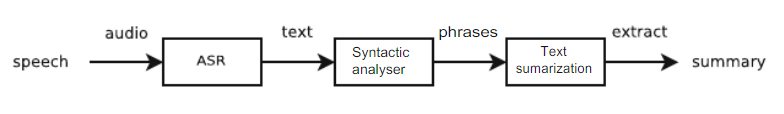




FIGURE 2.0

**5.Text to text summarization**

* There are three main steps for summarizing documents. These are topic identification, interpretation and summary generation.
* Topic Identification: The most prominent information in the text is identified. There are different techniques for topic identification are used which are Position, Cue Phrases, word frequency. Methods which are based on the position of phrases are the most useful methods for topic identification.
* Interpretation: Abstract summaries need to go through interpretation step. In This step, different subjects are fused in order to form a general content.
* Summary Generation: In this step, the system uses text generation method

# 3 REQUIREMENTS:

**Hardware requirements**

Computers

Internet connection

**Software requirements:**

Front end:

FLASK, HTML/CSS, STREAMLIT

HEROKU CLOUD

Python

# 4 REFERENCE

(1) [Rafiq M et al, “Deep Video Events (DVE): A Deep Learning Approach for Sports Video Summarization”, Proceedings of 2018 IEMEK Symposium on Embedded Technology (ISET 2018)

(2) A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in International Conference on Neural Information Processing Systems, pp. 1097–1105, 2012.

(3) M. S. Lew, et-al “Content-Based Multimedia Information Retrieval: State of the Art and Challenges”, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 2, February 2006.

(4) E. Mendi, et-al “Sports video summarization based on motion analysis”, Computers and Electrical Engineering, vol. 39, pp. 790–796, April 2013. 9 Video Summarization: A Review on Different Approaches

(5) Y. Tabii and R. O. H. Thami, “A new method for soccer video summarization based on shot detection, classification and finite state machine”, SETIT 2009 5th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, 2009.

(6) S. Zhu, Z. Liang, and Y. Liu, “Automatic Video Abstraction via the Progress of Story”, Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, vol. 6297, pp. 308–318.

(7) Marise Ramsaran, “Automated Highlight Generation from Cricket Broadcast Video”, 2016 8th International Conference on Computational Intelligence and Communication Networks.

(8)Ali Javed et al, “A Hybrid Approach for Summarization of Cricket Videos”, 2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia).

(9) Punith Kumar M B et al, “The Extraction of Events and Replay in Cricket Video”, International Conference on Emerging Research in Electronics, Computer Science and Technology – 2015

(10) Ehsan Anjum M et al, “Video Summarization Sports Highlights Generation”, 978-1-4799-3043-2/13/$31.00 ©2013 IEEE

(11)Video Summarization Using Deep Neural Networks: A Survey by Evlampios Apostolidis , Eleni Adamantidou , Alexandros I. Metsai , Vasileios Mezaris , Senior Member, IEEE, and Ioannis Patras , Senior Member, IEEE

(12)Video Transcription and Summarization using NLP By KhushiPorwal, Harshit Srivastava, Ritik Gupta , ShiveshPratap Mall, Nidhi Gupta

(13) Text Summarization Techniques: A Brief Survey Mehdi Allahyari ,Seyedamin Pouriyeh , Saeid Safaei , Mehdi Assef, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut

(14) Automatic Speech Summarisation: A Scoping Review By Dana Rezazadegan1,2,\*, Shlomo Berkovsky2 , Juan C. Quiroz3,2, A. Baki Kocaballi4,2, Ying Wang2 , Liliana Laranjo5,2, Enrico Coiera2

(15) Text Summarization:An Overview by [Samrat Babar](https://www.researchgate.net/profile/Samrat-Babar)

(16) PodSumm: Podcast Audio Summarization ANEESH VARTAKAVI∗ and AMANMEET GARG∗ , Gracenote Inc

(17) Robust video summarization algorithm using supervised machine learning

(18) Author links open overlay panelSunil SHarakannanavaraShaik RoshanSameeraVikashKumaraSunil KumarBeheraaAdithyaV AmberkaraVeena IPuranikmathb

(19) Robust video summarization algorithm using supervised machine learning

(20) Author links open overlay panelSunil SHarakannanavaraShaik RoshanSameeraVikashKumaraSunil KumarBeheraaAdithyaV AmberkaraVeena I.Puranikmathb