VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

Department of Computer Engineering

Project Report on

Video Classification using Text Analysis

In complete fulfillment of the Fourth Year, Bachelor of Engineering (B.E.) Degree in Computer Engineering at the University of Mumbai Academic Year 2017-2018

Submitted by

Kajol Chawla (D17A - 16)

Revati Pathak (D17A - 56)

Aarti Raghani (D17A - 62)

Aditi Sawant (D17A - 67)

Project Mentor

Mrs. Lifna C. S.

(2017-18)

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

Department of Computer Engineering

Certificate

This is to certify that ***Kajol Chawla, Revati Pathak, Aarti Raghani, Aditi Sawant*** of Fourth Year Computer Engineering studying under the University of Mumbai have satisfactorily completed the project on "***VIDEO CLASSIFICATION USING TEXT ANALYSIS***" as a part of their coursework of PROJECT-II for Semester-VIII under the guidance of their mentor ***Mrs. Lifna C. S.*** in the year 2017-2018.

This thesis/dissertation/project report entitled ***Video Classification using Text Analysis*** by ***Kajol Chawla, Aditi Sawant, Aarti Raghani, Revati Pathak*** is approved for the degree of ***Bachelor of Engineering in Computer Science***.

Programme Outcomes	Grade
PO1,PO2,PO3,PO4, PO5,PO6,PO7, PO8, PO9, PO10, PO11, PO12, PSO1, PSO2	

Date:

Project Guide: Internal and External

Project Report Approval
For
B. E (Computer Engineering)

This thesis/dissertation/project report entitled ***Video Classification using Text Analysis*** by ***Kajol Chawla, Revati Pathak, Aarti Raghani, Aditi Sawant*** is approved for the degree of ***Bachelor of Engineering in Computer Science***.

Internal Examiner

External Examiner

Head of the Department

Principal

Date:
Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

-
(Signature)

-
(Name of student and Roll No.)

-
(Signature)

-
(Name of student and Roll No.)

-
(Signature)

-
(Name of student and Roll No.)

-
(Signature)

-
(Name of student and Roll No.)

Date:

ACKNOWLEDGEMENT

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to Assistant Professor **Mrs. Lifna C. S.** (Project Guide) for her kind help and valuable advice during the development of project synopsis and for her guidance and suggestions.

We are deeply indebted to Head of the Computer Department **Dr.(Mrs.) Nupur Giri** and our Principal **Dr. (Mrs.) J.M. Nair ,** for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement at several times.

Computer Engineering Department
COURSE OUTCOMES FOR B.E PROJECT

Learners will be to,

Course Outcome	Description of the Course Outcome
CO 1	Able to apply the relevant engineering concepts, knowledge and skills towards the project.
CO2	Able to identify, formulate and interpret the various relevant research papers and to determine the problem.
CO 3	Able to apply the engineering concepts towards designing solution for the problem.
CO 4	Able to interpret the data and datasets to be utilized.
CO 5	Able to create, select and apply appropriate technologies, techniques, resources and tools for the project.
CO 6	Able to apply ethical, professional policies and principles towards societal, environmental, safety and cultural benefit.
CO 7	Able to function effectively as an individual, and as a member of a team, allocating roles with clear lines of responsibility and accountability.
CO 8	Able to write effective reports, design documents and make effective presentations.
CO 9	Able to apply engineering and management principles to the project as a team member.
CO 10	Able to apply the project domain knowledge to sharpen one’s competency.
CO 11	Able to develop professional, presentational, balanced and structured approach towards project development.
CO 12	Able to adopt skills, languages, environment and platforms for creating innovative solutions for the project.

Abstract

With the advent of the Information Age, there has been an exponential growth in the amount of videos available for any given topic. The process of finding relevant videos requires the users to go through an enormous amount of information. This becomes a tedious and time-consuming task. Hence, internet users are relentlessly in search for foolproof video search algorithms. The objective of the paper is to rank videos using text summarization techniques. This application takes into account the users' requirements and this, clubbed with several other factors, is used to generate efficient textual summaries of the videos' content. Finally, this application will provide the users with the most relevant videos and their summaries, thereby saving time and increasing efficiency.

Table Of Contents:

CHAPTER - 1 : INTRODUCTION

This chapter aims at specifying the motivation and scope of the project. In this chapter we will discuss the problem definition of project along with the relevance and motivation for the project. This chapter also includes the methodology used for this project.

1.1 Motivation

In today's era of digital world, everyone uses internet for finding information. Query for a particular topic results in a heap of contents and it becomes difficult to manually discard irrelevant results. Also, there has been an enormous growth in the number of videos available for any given topic. A manual search for relevant videos takes much time and effort on the user’s part, making it a tedious process. Hence, there is a need to automatically summarize the videos based on their content. In such a scenario, a summary of videos will be helpful in getting a gist of their content, thereby enabling the users to make a decision without having to watch the video . Such a textual document will be helpful for naive users to narrow their search to relevant re-ranked videos. This re-ranking is obtained on the basis of user query. This list of re-ranked videos will be complemented by a summary of every video, obtained on the basis of its content, which will be helpful for users to decide whether to watch video or not.

1.2 Problem Definition

The main aim of this project is to provide the users an ease of search for videos. This is possible if users get a gist of the video’s context before having watched it, which is provided through the summary and re-ranking of the top videos. Many times, Youtube descriptions fail to give a clear idea about the content of the video.This is because providing descriptions is at the discretion of the video uploader. Hence, there is no standard for descriptions of videos. Moreover, often times descriptions are not provided at all. Lack of descriptions makes it difficult for the user to make a decision about whether to watch the video or not. This problem can be overcome by providing a summary of the video. Summary of the video is formed on the basis of its contents. Using this, the user can get a better idea about the video and decide whether to watch the video or not. Sometimes the videos which are more relevant are ranked lower in search results. In such a case, the user will waste time by first going through the higher-ranked videos before finding the desired one. Our project aims at making this searching process easier by providing the user with a re-ranking of the videos in the top search results. Along with this, a textual summary of the re-ranked videos is also provided, which will enable the user to get a gist of the video without the need for watching it.

1.3 Relevance of the Project

Video Classification and Summarization based on its content finds many day-to-day applications. An efficient summarization algorithm will help us in obtaining a more accurate summary, and the re-ranking algorithm will enable us to recognize the more relevant videos, based on the ranking of the video. Summarization of videos can be useful for:

- **Summarizing news articles on a daily-basis:** Everyday, there are many events taking place around the globe and various newspapers and blogs cover them, each with a different perception. It becomes difficult for a reader to read all available articles. This creates a need for an efficient summarization algorithm that can abbreviate the lengthy articles to a more concise format.
- **Summarizing NPTEL Videos :** Learners find it difficult to identify videos relevant to the query, as there are a number of videos available. More often than not, these videos are too lengthy to be watched in full and the learners are not sure if the video contain required part of information without fully watching it.Thus, there is a need of informative summary of videos ranked in the order of their relevance with the user query.

1.4 Methodology employed for development

The system provides two main functionalities - video content summarization and video re-ranking. Summarization is done by first taking a query as input from the user. Next, the top results from YouTube are extracted based on this query. The videos are converted into suitable textual format for summarization by first extracting the audio content from the video and then converting it to text. Next, this text document is summarized using extractive summarization technique that uses sentence scoring method. In the next phase, the top ranked videos are re-ranked on the basis of the scores awarded to each summary generated. This list of re-ranked videos, along with their respective summaries is provided to the user to expedite the searching process.

CHAPTER 2: LITERATURE SURVEY

A literature review is a text of a scholarly paper, which includes the current knowledge including substantive findings, as well as theoretical and methodological contributions to a particular topic.This chapter also includes patents regarding project.

2.1 Papers

1. Text Summarization using Sentence Scoring Method [1]

Inference drawn: In this paper, an extractive text summary has been developed by using Sentence Scoring Method which takes into consideration several factors such as sentence ranking, term frequency, etc. All of the factors listed in methodology result in a higher-quality summary as compared to systems using one or fewer number of factors. Query based summarization is not supported. It includes summarization based on title matching, which awards higher marks to a sentence containing the title but does not take into account the relevance of the title to the sentence.

Methodology: It includes four phases, namely PreProcessing, Sentence Scoring, Sentence Ranking and Summary generation. The Pre-Processing phase involves sentence segmentation, Tokenization, Stop word removal and Stemming. Sentence Scoring phase typically includes the following factors:

1. Frequency
2. Sentence Position
3. Cue words
4. Similarity with the Title.
5. Sentence length.
6. Proper noun.
7. Sentence reduction.

After each sentence is scored, the sentences are arranged in the descending order of the score awarded to them. After ranking the sentences based on their total score, the summary is generated selecting a certain number of top ranked sentences where the number of sentences required is provided by the user.

2. Using Lexical Chains for Text Summarization[2]

Inference drawn: In this paper, the proposed algorithm uses the concept of WordNet to find relatedness among words. Using WordNet, words having semantic relations like synonymy and hyponymy are linked together and included in the same category. This algorithm differs from other approaches as it creates chains at each sentence using relatedness criteria. Merging of different segments are done only if they satisfy strong criterion for relatedness, i.e. only if the two segments contain words having the same meaning or representing a similar concept. The chain representation approach completely avoids the problem of related terms in summary, because all these terms occur in the same chain, which reflects that they represent the same concept. This approach faces the problem of sentence granularity, i.e. it extracts sentences as a whole unit.

Methodology used: In the preprocessor step, all the words that appear as noun form are chosen as an input in WordNet. Later, it identifies important concepts in the domain and also it eliminates words that occur as modifier by using shallow parser. WordNet is used to identify semantic relations and generate lexical chains based on these relations. Once all the chains are generated, the final summary can be obtained based on different heuristics functions. They are as follows:

Heuristic 1 - For each chain m the summary representation, choose the sentence that contains the first appearance of a chain member in the text.

Heuristic 2 - For each chain in the summary representation, choose the sentence that contains the first appearance of a representative chain member in the text.

Heuristic 3 - Often, it happens that the same semantic concept appears in multiple places in the document. Hence, the lexical chain may be spread all over the document.

3. Ontology Based Text Document Summarization System using Concept Terms[3]

Inference drawn: In this paper, Ontology based Text Document Summarization is carried out using Concept terms. Hierarchical Representation is generated for concept terms. Later, a level for concept terms is set to generate summary. This approach also supports POS-tags i.e considering verbs and adjectives of term. This approach doesn’t give good results for single document summarization.

Methodology used: This approach has the following steps: Preprocessing, Concept Extraction, Clustering and Ontology Creation. In Preprocessing, tokenization is carried out by splitting the sentences into words by using white space as the separator. Then, Part-Of-Speech Tagger (POS Tagger) is applied to the tokenized words.. Next, a Stemming algorithm is used to find the root word. Stemming process is continued by removing Stopword. For the Concept extraction Phase, the TF-IDF algorithm is used. It is followed by Clustering. This is done by using the k-means algorithm which is one of the simplest unsupervised learning algorithms that solves the well known clustering problem. Finally, Ontology can be built in two ways. One way is developing tools that are used by knowledge engineers or domain experts to build the ontology like Protégé and Jena. They are called the ontology modelling tools. Another way is semi-automatic or automatic building of ontologies by learning it from different information sources.

4. Text Rank: A Novel Concept for Extraction Based Text Summarization[4]

Inference drawn: The proposed system includes an indexing structure in which index is built on the basis of context of the document rather than on the terms basis. In this approach, a graph of a document is generated. After using wordnet to link documents, Textrank algorithm is used find the weight of each term. Finally, similarity between sentences is calculated by indexing approach.

Methodology used: Initially, WordNet is used to understand the links between different parts of the document; subsequently the Lexical associations between two document terms are which are most relevant are extracted. Later, Textrank Algorithm is used to find a context sensitive indexing weight of each term by generating graph of a document. Here, each edge gives the lexical association between the terms corresponding to the vertices. Next step is to find similarity between sentences using context based indexing weights which find similarity between words. Then, this similarity between words is used to find similarity between sentences. So for each sentence in the document, the sentence vector is built. Later, sentences are extracted based on higher sentence score. So, this approach can be used to retrieve results within short span.

5. A Survey on Extractive Text Summarization[5]

Inference drawn: Various approaches for extractive text summarization are discussed. The extractive text summarization is a process of selecting important sentences from the document and including those sentences as they are in the final summary of the document, and the selection procedure of sentences is done on the basis of statistical and linguistic features of the sentences. Sentence scoring efficiency depends on the various factors considered for scoring. In this, query based summarization approach and domain based approach are not explained.

Methodology used: There are several features which are used to score the sentences. They are Cue-Phrase Feature, Title Word Similarity Feature, Sentence Length Feature, Proper Noun Feature, Font Feature for Sentences, Uppercase Feature, Sentence Position Feature, Numeric Data Feature, Average TF-ISF (Term Frequency Inverse Sentence Frequency) Feature, Sentence Centrality Feature, Pronoun Feature, Non-Essential Information Feature. A brief introduction to extractive text summarization methods such as Term Frequency-Inverse Document Frequency (TF-IDF) Method, Cluster Based Method, Graph Theoretic Approach, Machine Learning Method, LSA Method, Text Summarization with Neural Networks, Text Summarization Based on Fuzzy Logic, Query Based Extractive Text Summarization, etc. is given.

6.LexRank:Graph-based Lexical Centrality as Salience in Text Summarization[6]

Inference drawn: In this paper, a graph-based unsupervised algorithm for extractive summarization is proposed. It uses the concept of centrality to identify the most relevant sentences in a cluster. It uses TF-IDF (Term Frequency- Inverse Document Frequency) to calculate how important a word is in multiple documents. Sentences subsuming the information of other sentences get higher scores than individual sentences; thus a compact summary is generated and redundancy is reduced. This approach prevents unnatural boosting of sentences containing an irrelevant topic. LexRank provides an improvement over TextRank as it uses a cosine similarity function, instead of the linear function used by TextRank. It performs well even on noisy data. Variations of LexRank such as Continuous LexRank are available which improve its performance. LexRank primarily works on multiple documents and uses TF-IDF for calculating the importance of a word. Since we are aiming for a single document summarization, TF-IDF will have to be substituted by and TF-ISF.

Methodology used: A vector is defined for every sentence where the entry in the vector for every word is the value of its frequency in the sentence multiplied by its idf. The dimensionality of this vector is equal to the number of all words in the targeted language. An idf-modified-cosine function is defined to compute the similarity between two sentences. A cosine similarity matrix is formed with entries for the similarity calculated above. A threshold value is chosen to choose only the significantly similar sentences and discard the rest. LexRank provides an improvement over TextRank as it uses a cosine similarity function, instead of the linear function used by TextRank. It performs well even on noisy data. Variations of LexRank such as Continuous LexRank are available which improve its performance. LexRank primarily works on multiple documents and uses TF-IDF for calculating the importance of a word. Since we are aiming for a single document summarization, TF-IDF will have to be substituted by and TF-ISF.

7. A Recursive TF-ISF Based Sentence Retrieval Method with Local Context[7]

Inference drawn: In this paper, a recursive TF-ISF based method that takes into account the local context of a sentence is proposed. By context, what is meant is the previous and next sentence of current sentence. The new method, on comparing to the TF-ISF baseline and to an earlier unsuccessful method that also incorporates a similar context into TF-ISF gave statistically significant improvements of the results in comparison to both of the methods. TF-ISF method achieve explicit modeling of sentences.

Methodology used: Summary is generated by considering relevance between previous and next sentence. It is achieved by defining recursive ranking function.It does not take into account Linear Combination of word frequency in document, Sentence positional value, Cue Words, Similarity with the title of the document, Sentence length and Proper noun etc factors is considered while summarizing the documents.

8. Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency(TF-IDF)[8]

Inference drawn: An extractive text summarization with TF-IDF method is used to build the summary. TF-IDF is a numerical statistic which reflects on how important a word is to a document in the collection or corpus, but it is offset by the frequency of the word in the corpus, which helps to control the fact that some words are more common than others. The frequency term means the raw frequency of a term in a document. Moreover, the term regarding inverse document frequency is a measure of whether the term is common or rare across all documents in which can be obtained by dividing the total number of documents by the number of documents containing the term. The result of this research produces 67% of accuracy with three data samples which are higher compared to the other online summarizers. TF-IDF is proven to be a powerful method to generate the value which determines how important a word inside the document is. Better result can be produced by making summary biased on the title. So that high TF-IDF value can be given to that words which appears in the title. It should involve more respondents to evaluate the system by determining the number of correct, wrong, or missed sentences within the summary.

Methodology used: Preprocessing involves operation needed to enhance feature extraction such as tokenization, part of speech tagging, removing stop words, and word stemming. Second step is feature extraction. It is used to extract the features of the document by obtaining the sentence in text document based on its importance and given the value between zero and one. Third, sentence selection and assembly is when the sentences are stored in descending order of the rank, and the highest rank is considered as the summary. Lastly, summary generation is the sentences that are put into the summary in the order of the position in the original document. After Preprocessing each sentence's TF-IDF value is calculated.

9. Multi-document text summarization - A survey[9]

Inference drawn: Here, an approach for summarizing multiple documents is discussed. This application will allow the user to automatically summarize relevant information from various sources. This application will not only save time but also render higher scope of efficiency. Multilingual summarization is not possible. Recordings or discussions cannot be summarized.

Methodology used: This paper uses Lexrank, Relevance graph generation to find the summary. Initially, it generates relevance, later it converts assign score to each document; after assigning a score irrelevant documents are discarded by setting up threshold values. Final summary is generated using Lexrank algorithm and results displayed are summary of documents in visual form. Different summarization approaches such as cluster based approach, Topic based approach, Lexical chains approach are discussed. First, summaries of individual documents are extracted. Then algorithm generates relevance graph where documents are nodes and assigns weights based on similarity. This will be followed by pruning of documents which are poorly linked with others.

10. A Comparison of Document, Sentence, and Term Event Spaces[10]

Inference drawn: In this paper, a comparison on various information retrieval techniques is discussed. TF-IDF gives the relation between the documents. It is used for multiple documents. However, in case of single document these results can get better with TF-ISF algorithm which uses multiple sentences as an input. For a single document TF-ISF gives better results than TF-IDF.The TF-ISF along with local content gives better results since the context of content of a line cannot be understood from a single line.

Methodology used: In this paper, different information retrieval methods are discussed-document, sub-document.It researches whether tf-idf retrieval method can be replaced by tf-isf or tf-itf. It is found that tf-idf gives same result as tf-isf or tf-idf with some added weight i.e. they are highly correlated. However transformation from a document retrieval to sub document retrieval doesn't give good results. In this paper, only stemmed words are considered for experiments.More research is required for language usage to understand the factors affecting them.

2.2 PATENTS

1. Patent N0.:US 8,402,369 B2- MULTIPLE-DOCUMENT SUMMARIZATION USING DOCUMENT CLUSTERING

Systems and methods are disclosed for summarizing multiple documents by generating a model of the documents as a mixture of document clusters, each document in turn having a mixture of sentences, Wherein the model simultaneously representing summarization information and document cluster structure; and determining a loss function for evaluating the model and optimizing the model.

2. Patent N0.: US 20140324883 A1- GENERATING A SUMMARY BASED ON READABILITY

A technique to generate a summary of a set of sentences. Each sentence in the set can be evaluated based on a criterion, such as informativeness of the sentence. The sentences may also be evaluated for readability based on a readability measure. Sentences can be selected for inclusion in the summary based on the evaluations.

CHAPTER - 3 : REQUIREMENTS

In product development and process optimization, a requirement is a singular documented physical and functional need that a particular design, product or process must be able to perform. It is most commonly used in a formal sense in systems engineering, software engineering, or enterprise engineering. It is a statement that identifies a necessary attribute, capability, characteristic, or quality of a system for it to have value and utility to a customer, organization, internal user, or other stakeholder. This chapter gives detailed information regarding the various functional, non-functional and software requirements.

3.1 Functional Requirements

In Software engineering and systems engineering, a functional requirement defines a function of a system or its component. The Functional Requirements Specification documents the operations and activities that a system must be able to perform.Following are as listed below:

- Users has an ability to input query for whom videos are required.
- Users achieve summary of the accessible videos related to input query.
- User are able to see re-ranked videos based on query.
- Users are able to access database.
- Database contain summary of videos based on video content and links of videos along with keyword based on which re-ranking happen.
- Users can also give speech as input to obtain a summary.

3.2 Non-Functional Requirements

In systems engineering and requirements engineering, a nonfunctional requirement (NFR) is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behaviors.Following are as listed below:

- Accuracy- The system aims at achieving more accuracy in video content summarization as compared to the existing video summarization system.
- Response time- Ability to respond in minimum time span.
- Performance- Ability to re-rank videos properly based on query.
- Portable - The system is platform independent.
- Availability - The system should be available to all the users all round the year.
- Recoverable - If something wrong happens system should be able to recover with minimum time span.

3.3 Constraints

- Multilingual- Restricted to only English language.
- Abstract Summary- As a lot of Cpu size is required.
- Music Video not Summarized- Videos containing content are only preprocessed.

3.4.Requirements

3.4.1 Software Requirements

- Python 3.5 - Python is a widely used high-level, general-purpose, interpreted, dynamic programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code
- NLTK POS-Tagging Library - The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the Python programming language. NLTK includes graphical demonstrations and sample data.
- Youtube-dl - An open source command line youtube video downloader and extractor to get speech out of it.
- Speech Recognition - Google Cloud Speech API enables developers to convert audio to text

by applying powerful neural network models in an easy to use API.

- Scipy – SciPy (pronounced “Sigh Pie”) is an open source Python library used by scientists, analysts, and engineers doing scientific computing and technical computing. SciPy contains modules for optimization, linear algebra, integration, interpolation, special functions, etc.
- Numpy – It is an extension to the Python programming language, adding support for large, multidimensional arrays and matrices, along with a large library of high-level mathematical functions to operate on these arrays.
- Textblob – TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.
- PyLexicon - Python library to find synonyms,hyponyms of word from the thesaurus dictionary.
- Wordnet - It is a lexical database for the English language. It groups English words into sets of synonyms called synsets, Hyponyms,Hypernyms. It also provides similarity function between synsets of words

3.4.2 Hardware Requirements

- Working CPU.
- Good Access to Internet.

3.5 System Block Diagram

The diagram shown in Fig. 1.describes the system. The user inputs the search query over the internet, query is searched for in the database and if the corresponding query is accessible in the database, the results, that is, the relevant re-ranked videos, along with the videos links and the summary of those videos is displayed to user. But if query is not available, it is redirected to YouTube and the listed videos are pre-processed by initially converting them to text document, summarizing their content and finally re-ranking them based on the video content and the input query.

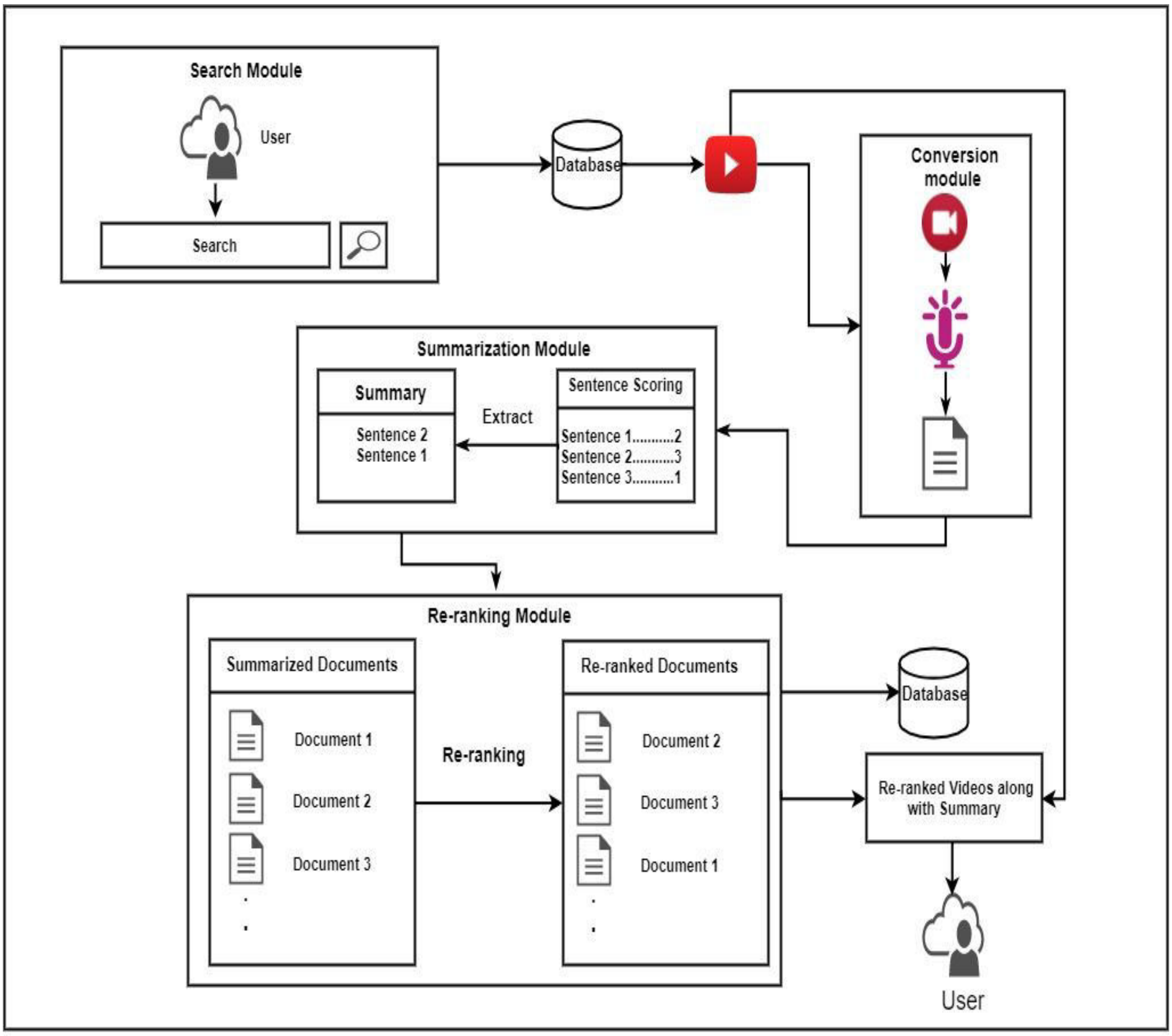


Fig 1. System Block Diagram

The system shown in Fig(1). has the following modules.(1) Search Module; (2) Audio Extraction Module; (3) Speech-to-Text Conversion Module; (4) Text Summarization Module and (5) Re-ranking Module.

- **Search Module :** The query input by user is searched for in the database. If found, relevant video links and generated summaries are displayed to the user. Else, the query is redirected to YouTube.
- **Audio Extraction Module :** The video for which summary is to be generated is the input and the audio is extracted from it.
- **Speech-to-Text Conversion Module :** The audio extracted in the preceding step is converted to textual format. This text document forms the input to the Summarization Module.
- **Text Summarization Module :** The text document obtained is summarized using extractive summarization technique. The factors considered for sentence scoring are: (1) Position of sentence; (2) Length of Sentence; (3) Similarity with the Title; (4) Proper Nouns; (5)Sentence similarity; (6) Similarity with query; (7) Unnecessary Sentences. These scores are combined and the best scored sentences are ranked better.
- **Re-ranking Module :** After summaries are generated for the videos, they are re-ranked on the basis of their summaries’ relevance to the query.

CHAPTER 4: PROPOSED DESIGN

Product design as a verb is to create a new product to be sold by a business to its customers. It is a set of strategic and tactical activities, from idea generation to commercialization, used to create a product design. This gives a detailed of our system design.

4.1 System Design/Conceptual Design(Architectural)

Conceptual Design is an early phase of the design process, in which the broad outlines of function and form of something are articulated. It includes the design of interactions, experiences, processes and strategies. It involves an understanding of people's needs - and how to meet them with products, services, & processes.

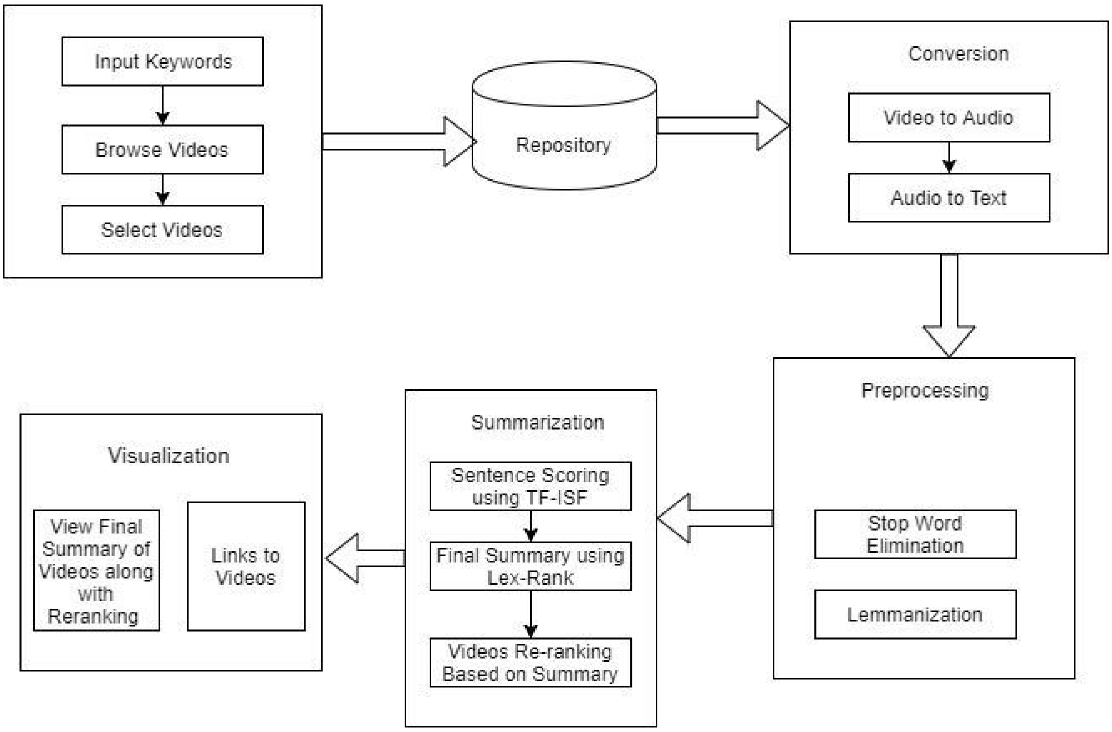


Fig 2. Conceptual System Design

Fig 2. depicts conceptual design of the system. The system delivers summary and link to re-ranked videos using the conversion and preprocessing modules. All the different modules includes various processes and strategies such as audio to text conversion, lemmanization to give output, which are input to the another module.

4.2 Detailed Design

4.2.1 Flowchart

A flow chart shows the flow of events in chronological order. It also incorporates conditionals and evaluates all cases. A flow chart starts with an input or a computation, depicts all possible scenarios in the process and then shows the stopping point.

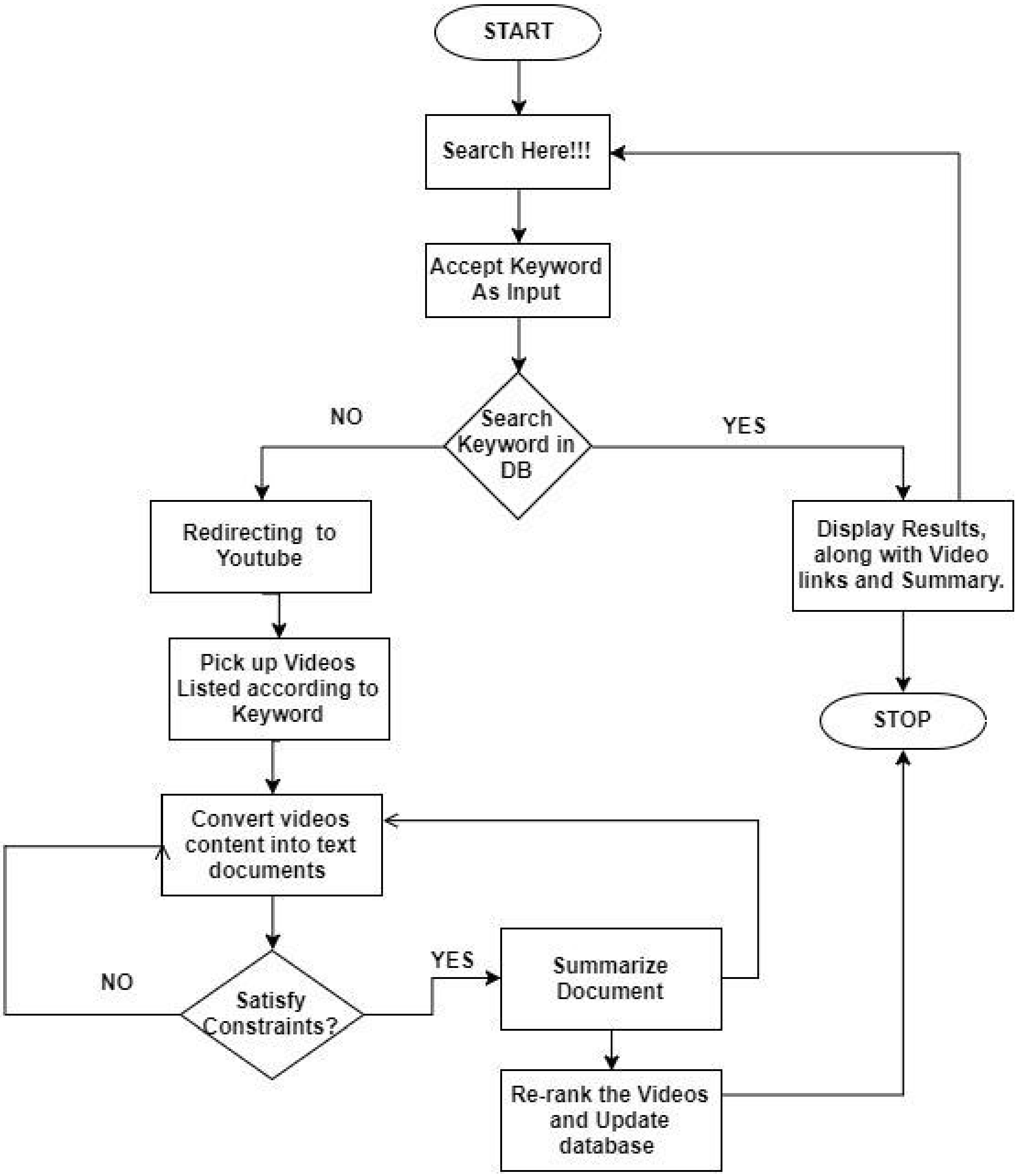

START
Search Here!!!
Accept Keyword As Input
Search Keyword in DB
NO
YES
Redirecting to Youtube
Display Results, along with Video links and Summary.
Pick up Videos Listed according to Keyword
STOP
Convert videos content into text documents
NO
Satisfy Constraints?
YES
Summarize Document
Re-rank the Videos and Update database

Fig 3. Flow Chart

4.2.2 DFD

A data flow diagram (DFD) is a graphical representation of the "flow" of data through an information system, modelling its process aspects.Level 0 represents the input and output of the system.Level 1 represents the tasks performed to give the primary functional requirements.Level 2 depicts the tasks in depth, performed to give all the functional requirements to the user.

LEVEL 0:

Level 0 represents the input and output of the system.Here,User enters input a keyword and the output generated is summary of videos and videos arranged in re-ranked order.

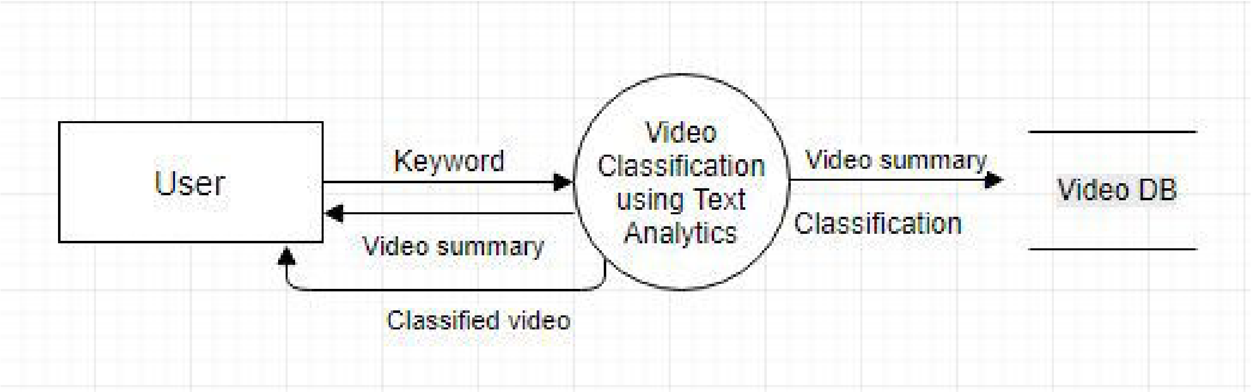


Fig 4. DFD Level 0

LEVEL 1:

Level 1 represents the tasks performed to give the primary functional requirements i.e query entered by user will be search in Video DB if found relevant videos summary in arranged order is displayed to user if not found the query will access videos and processing will take place further.

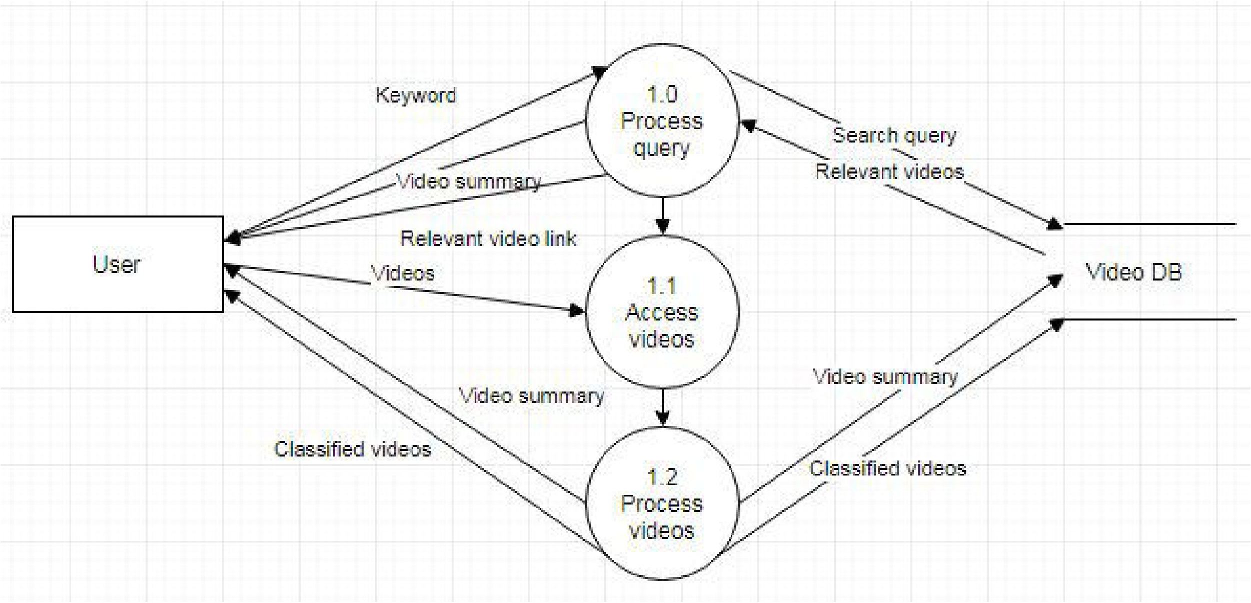


Fig 5. DFD Level 1

LEVEL 2:

Level 2 depicts the tasks in depth, performed to give all the functional requirements to the user. Query entered by user will be search in Video DB if found relevant videos summary in arranged order is displayed to user if not found the query will access videos and processing i.e preprocessing steps ,summarization and re-ranking of videos will take place and finally results are displayed to user.

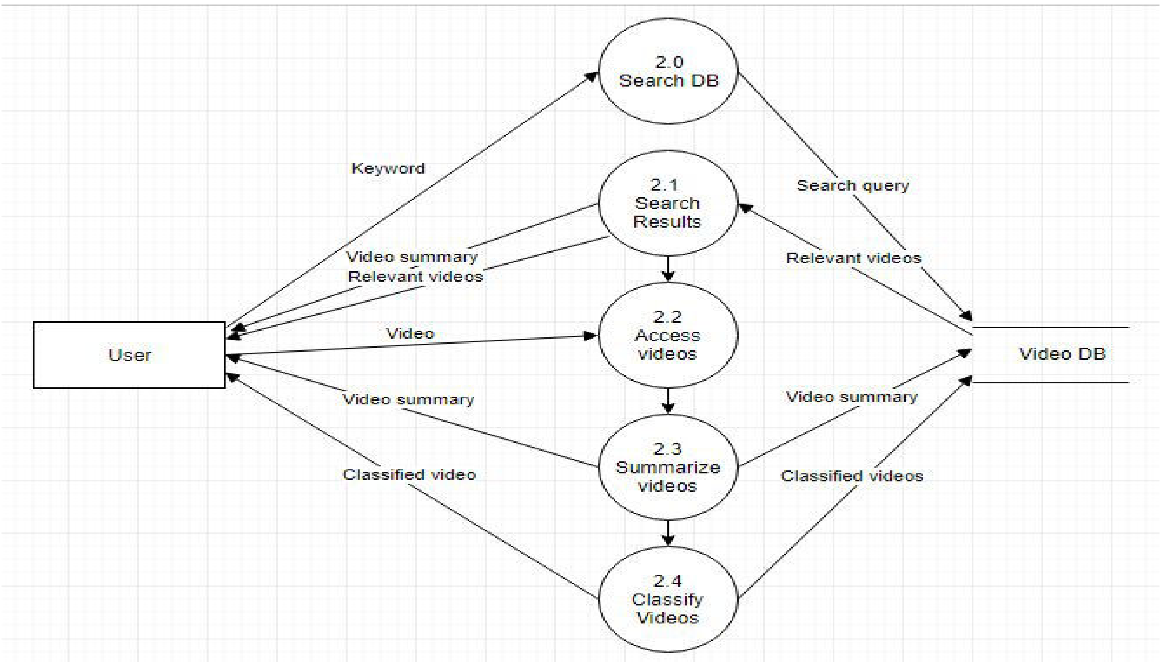


Fig 6. DFD Level 2

4.2.3 Activity Diagram

An activity diagram is much like a flow chart, except it depicts the flow of control instead of the flow of data in the system. The above diagram shows how the control passes to different processes in the system.

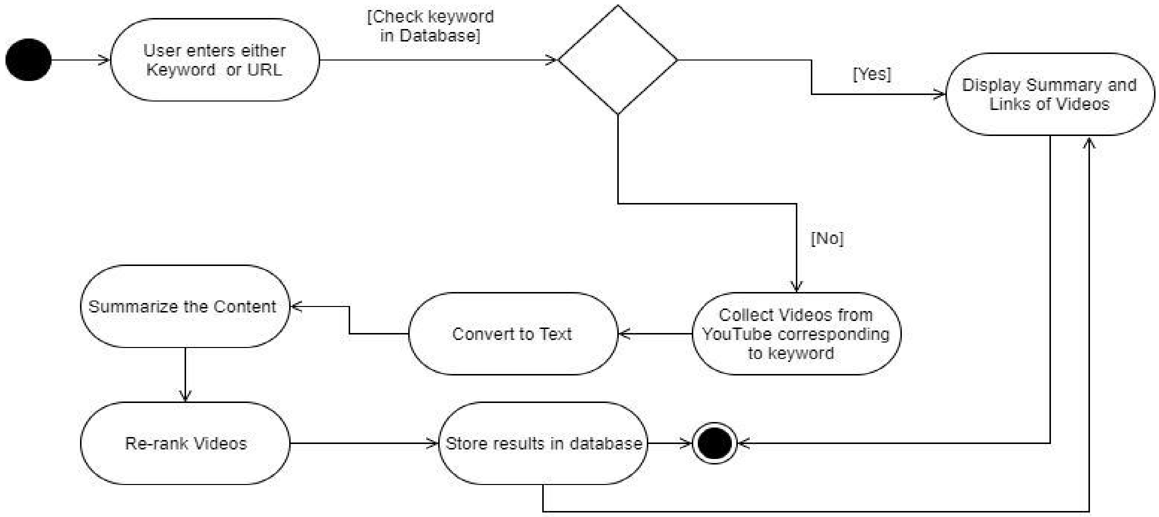
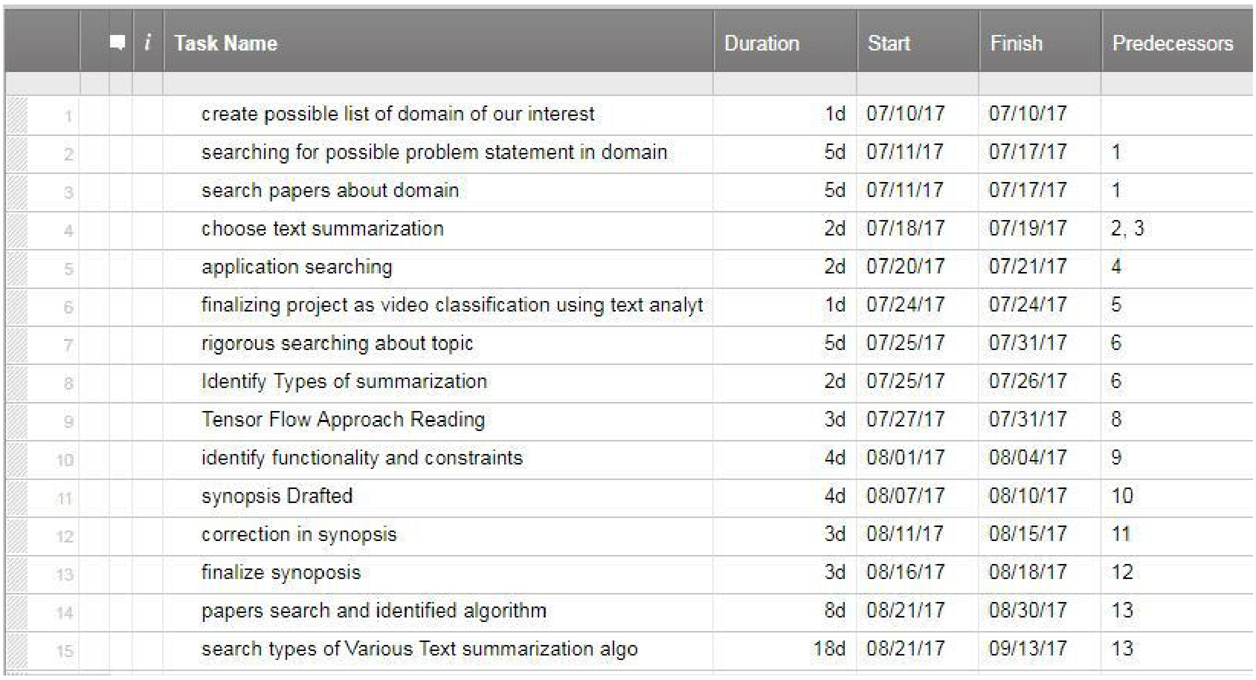

User enters either Keyword or URL
[Check keyword in Database]
[Yes]
Display Summary and Links of Videos
[No]
Collect Videos from YouTube corresponding to keyword
Convert to Text
Summarize the Content
Re-rank Videos
Store results in database

Fig 7. Activity Diagram

4.4 Project Scheduling & Tracking using Time line / Gantt Chart

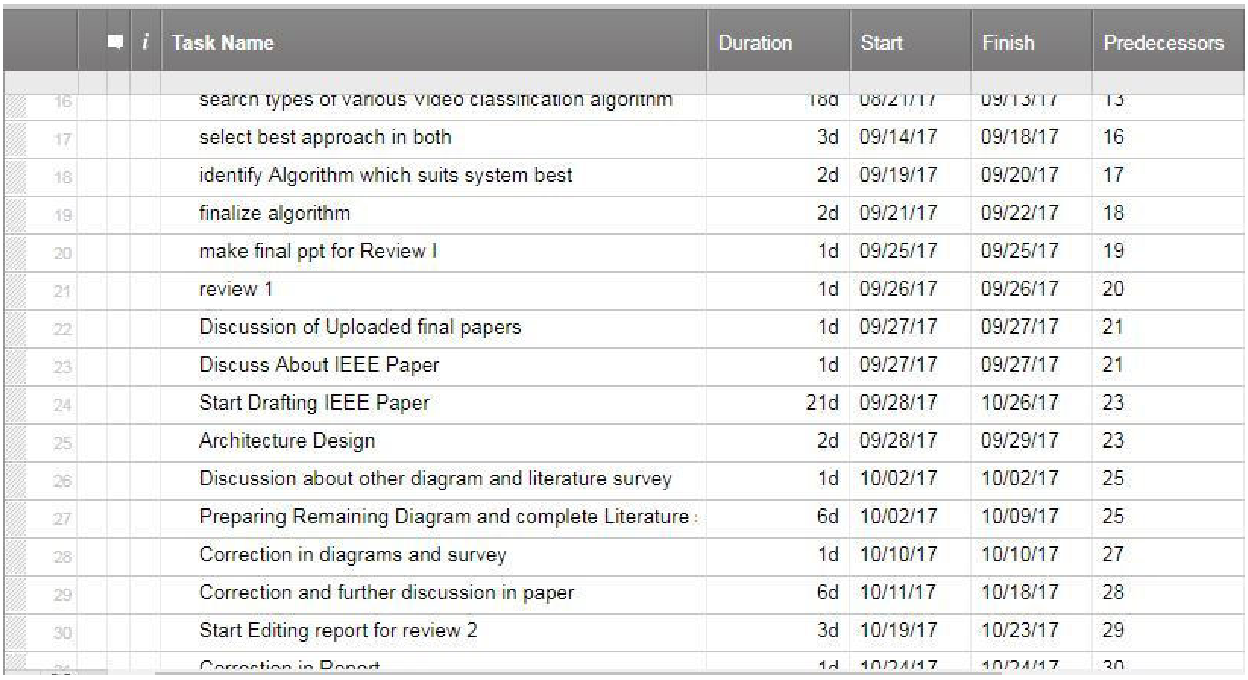
In project management, a schedule is a listing of a project's milestones, activities, and deliverables, usually with intended start and finish dates. Those items are often estimated by other information included in the project schedule of resource allocation, budget, task duration, and linkages of dependencies and scheduled events. A schedule is commonly used in the project planning.

Stage 1:



		i	Task Name	Duration	Start	Finish	Predecessors
1			create possible list of domain of our interest	1d	07/10/17	07/10/17	
2			searching for possible problem statement in domain	5d	07/11/17	07/17/17	1
3			search papers about domain	5d	07/11/17	07/17/17	1
4			choose text summarization	2d	07/18/17	07/19/17	2, 3
5			application searching	2d	07/20/17	07/21/17	4
6			finalizing project as video classification using text analyt	1d	07/24/17	07/24/17	5
7			rigorous searching about topic	5d	07/25/17	07/31/17	6
8			Identify Types of summarization	2d	07/25/17	07/26/17	6
9			Tensor Flow Approach Reading	3d	07/27/17	07/31/17	8
10			identify functionality and constraints	4d	08/01/17	08/04/17	9
11			synopsis Drafted	4d	08/07/17	08/10/17	10
12			correction in synopsis	3d	08/11/17	08/15/17	11
13			finalize synoposis	3d	08/16/17	08/18/17	12
14			papers search and identified algorithm	8d	08/21/17	08/30/17	13
15			search types of Various Text summarization algo	18d	08/21/17	09/13/17	13

Fig 8. Stage 1 event list



		i	Task Name	Duration	Start	Finish	Predecessors
16			search types of various video classification algorithm	18d	08/21/17	09/13/17	13
17			select best approach in both	3d	09/14/17	09/18/17	16
18			identify Algorithm which suits system best	2d	09/19/17	09/20/17	17
19			finalize algorithm	2d	09/21/17	09/22/17	18
20			make final ppt for Review I	1d	09/25/17	09/25/17	19
21			review 1	1d	09/26/17	09/26/17	20
22			Discussion of Uploaded final papers	1d	09/27/17	09/27/17	21
23			Discuss About IEEE Paper	1d	09/27/17	09/27/17	21
24			Start Drafting IEEE Paper	21d	09/28/17	10/26/17	23
25			Architecture Design	2d	09/28/17	09/29/17	23
26			Discussion about other diagram and literature survey	1d	10/02/17	10/02/17	25
27			Preparing Remaining Diagram and complete Literature :	6d	10/02/17	10/09/17	25
28			Correction in diagrams and survey	1d	10/10/17	10/10/17	27
29			Correction and further discussion in paper	6d	10/11/17	10/18/17	28
30			Start Editing report for review 2	3d	10/19/17	10/23/17	29
31			Correction in Report	1d	10/24/17	10/24/17	30

Fig 9. Stage 1 event list

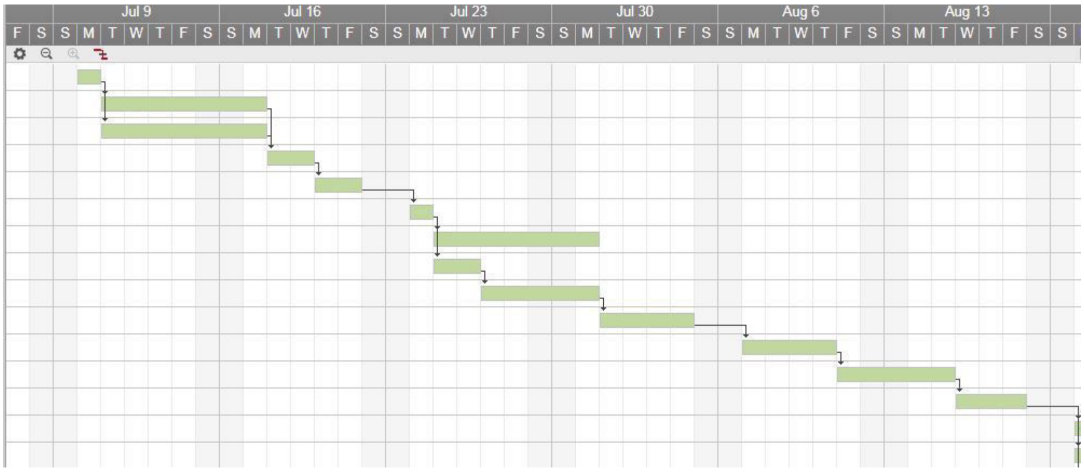

Jul 9
Jul 16
Jul 23
Jul 30
Aug 6
Aug 13
F S S M T W T F S S M T W T F S S M T W T F S S M T W T F S S M T W T F S S M T W T F S S

Fig 10. Stage 1 Timeline

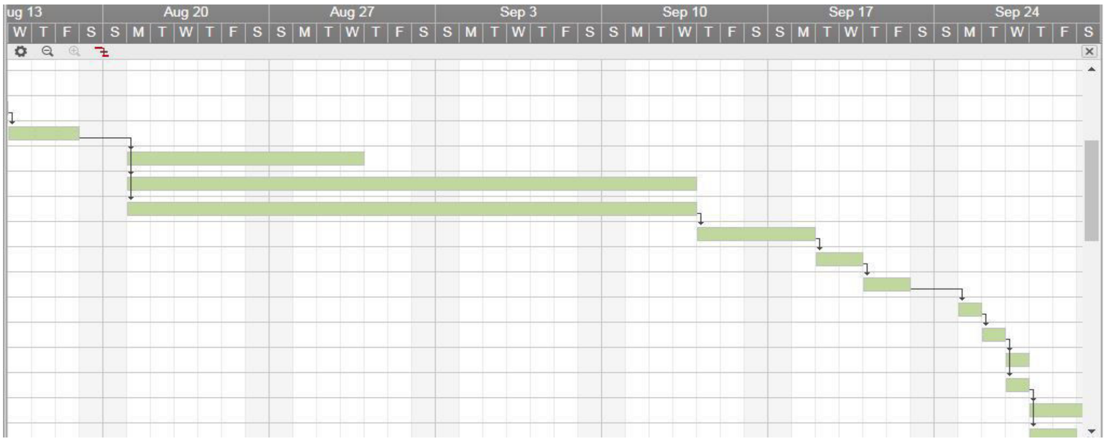

ug 13
Aug 20
Aug 27
Sep 3
Sep 10
Sep 17
Sep 24
W T F S S M T W T F S S M T W T F S S M T W T F S S M T W T F S S M T W T F S S M T W T F S

Fig 11. Stage 1 Timeline

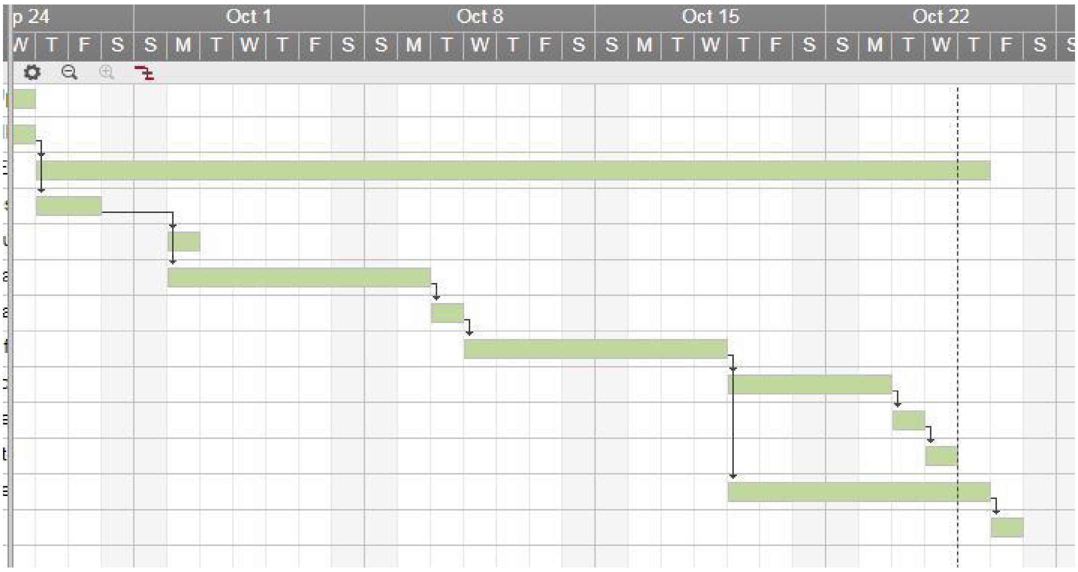
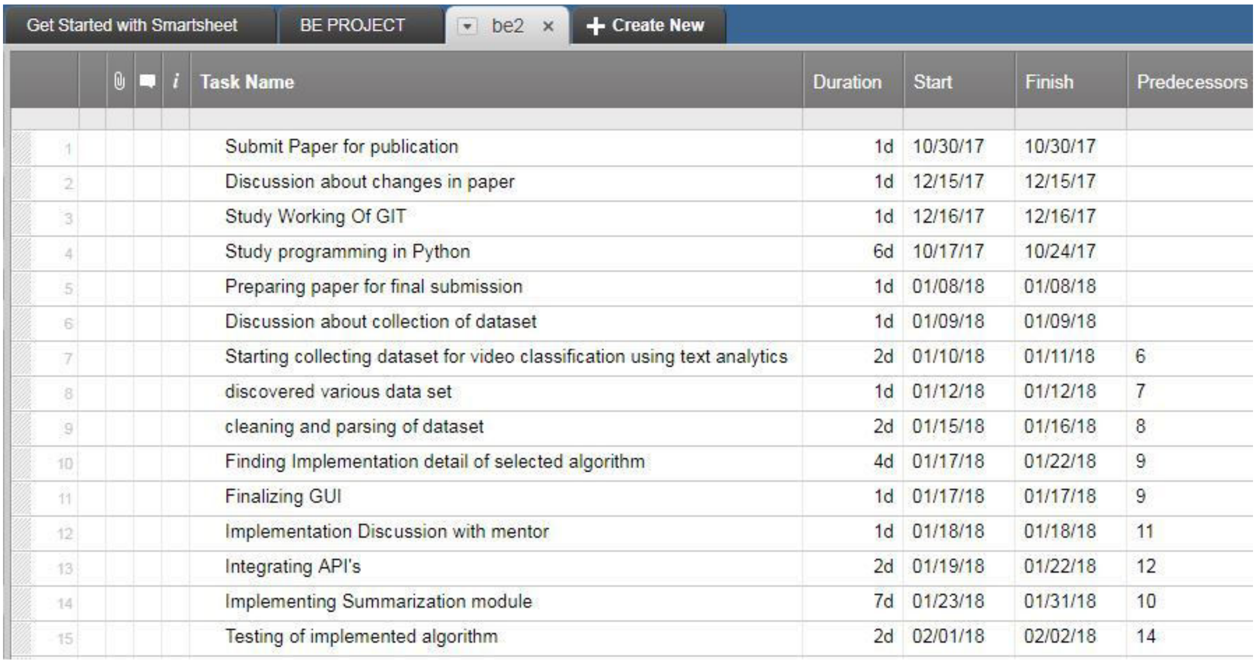

p 24
Oct 1
Oct 8
Oct 15
Oct 22
W T F S S M T W T F S S M T W T F S S M T W T F S S M T W T F S S

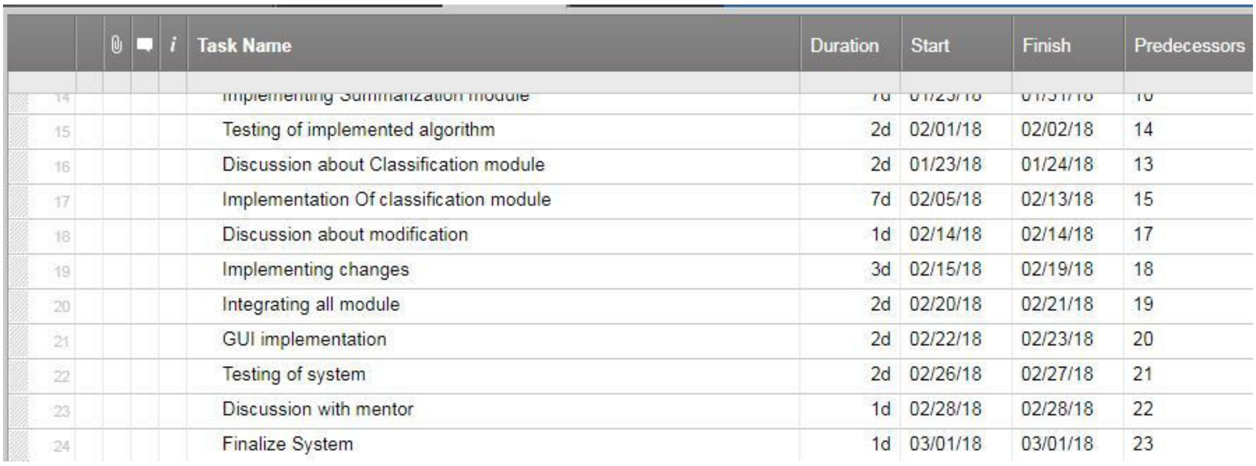
Fig 12. Stage 1 Timeline

Stage 2:



	Task Name	Duration	Start	Finish	Predecessors
1	Submit Paper for publication	1d	10/30/17	10/30/17	
2	Discussion about changes in paper	1d	12/15/17	12/15/17	
3	Study Working Of GIT	1d	12/16/17	12/16/17	
4	Study programming in Python	6d	10/17/17	10/24/17	
5	Preparing paper for final submission	1d	01/08/18	01/08/18	
6	Discussion about collection of dataset	1d	01/09/18	01/09/18	
7	Starting collecting dataset for video classification using text analytics	2d	01/10/18	01/11/18	6
8	discovered various data set	1d	01/12/18	01/12/18	7
9	cleaning and parsing of dataset	2d	01/15/18	01/16/18	8
10	Finding Implementation detail of selected algorithm	4d	01/17/18	01/22/18	9
11	Finalizing GUI	1d	01/17/18	01/17/18	9
12	Implementation Discussion with mentor	1d	01/18/18	01/18/18	11
13	Integrating API's	2d	01/19/18	01/22/18	12
14	Implementing Summarization module	7d	01/23/18	01/31/18	10
15	Testing of implemented algorithm	2d	02/01/18	02/02/18	14

Fig 13. Stage 2 event list



	Task Name	Duration	Start	Finish	Predecessors
14	Implementing Summarization module	7d	01/23/18	01/31/18	10
15	Testing of implemented algorithm	2d	02/01/18	02/02/18	14
16	Discussion about Classification module	2d	01/23/18	01/24/18	13
17	Implementation Of classification module	7d	02/05/18	02/13/18	15
18	Discussion about modification	1d	02/14/18	02/14/18	17
19	Implementing changes	3d	02/15/18	02/19/18	18
20	Integrating all module	2d	02/20/18	02/21/18	19
21	GUI implementation	2d	02/22/18	02/23/18	20
22	Testing of system	2d	02/26/18	02/27/18	21
23	Discussion with mentor	1d	02/28/18	02/28/18	22
24	Finalize System	1d	03/01/18	03/01/18	23

Fig 14. Stage 2 event list

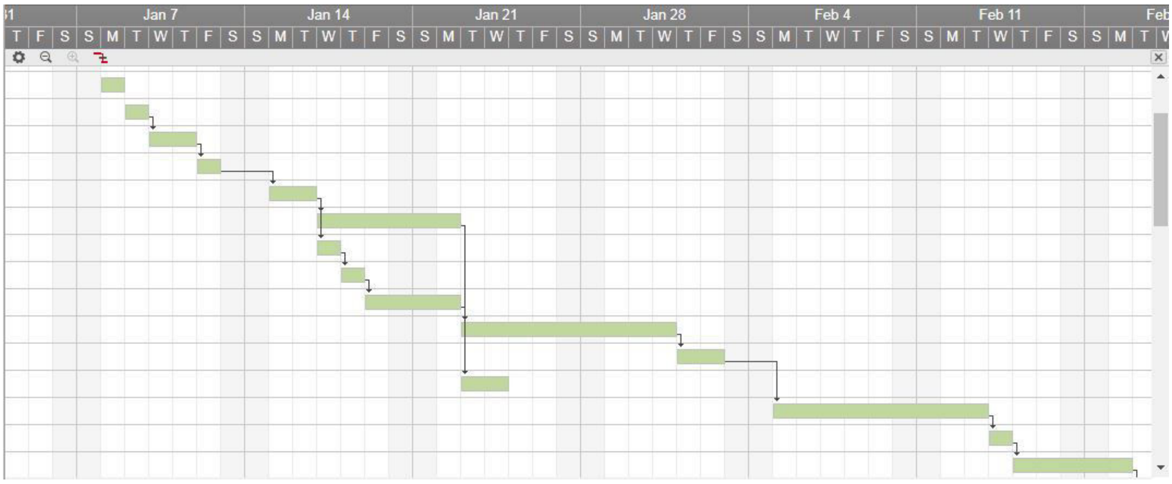


Fig 15. Stage 2 Timeline

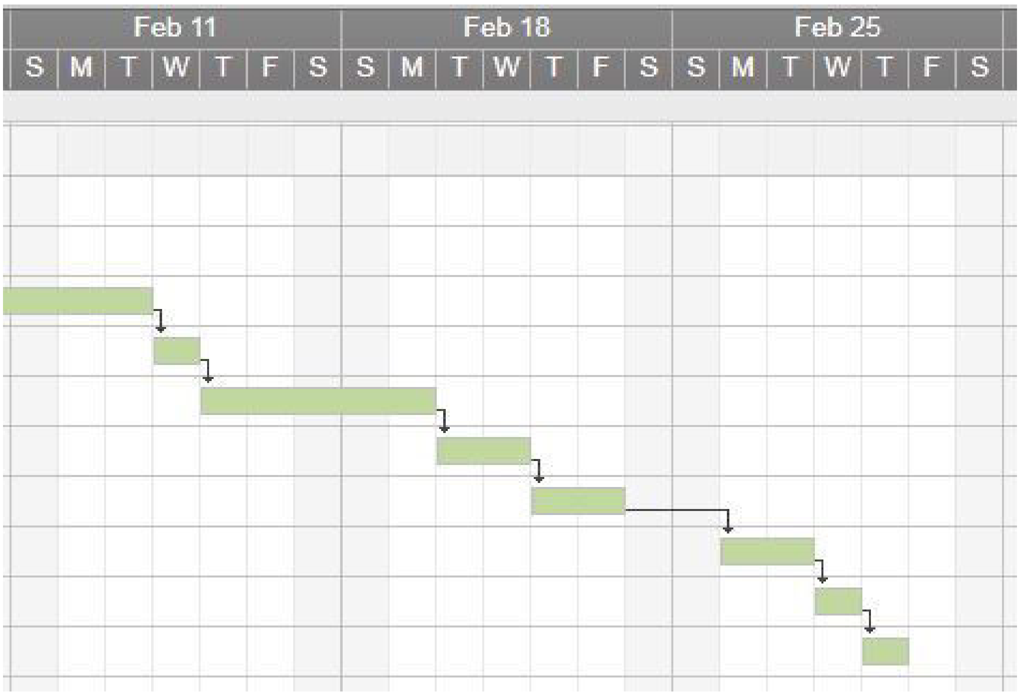


Fig 16. Stage 2 Timeline

CHAPTER - 5 : IMPLEMENTATION

A system is developed for users to generate videos in re-ranked form. This system allows users to search an query for videos as input and generates a relevant summary and videos in re-ranked form. The system can also be used for speech to text conversion. This chapter comprises of algorithm and evaluation for the system. Also comparison with other system is listed.

5.1 Algorithm:

1. User inserts query for which videos are required.
2. A query is buffered in database, if found output with videos link also summarized content of videos is displayed.
3. If the input query is unavailable in database, the query is redirected to Youtube and corresponding videos links are stored.
4. Videos are convert to text via speech conversion using youtube-dl for video to audio and Google speech translation for speech to text conversion.
5. Text obtained from above step is passed to preprocessed which contains following:
    a. Sentence Tokenization
    b. Word Tokenization
    c. Stemming of words
    d. Synonyms of words
    e. Synsets of words
6. After preprocessing followings factors are performed:
    a) Similarity with title: Words lists obtained from above step for each sentence is compared with words obtained in title of document to find similarity between title and a sentence to give a particular score to each sentence.

    b) Similarity with query: Words lists obtained from above step for each sentence is compared with words obtained in input query to find similarity between query and a sentence to give a particular score to each sentence.

    c) Similarity between sentences: Similarity between each sentence with other is calculated based on Synonyms, Hypernyms ,Hyponyms, Meronyms of words in both sentences and range is set to eliminate remaining sentences.

d) **Sentence Position:** Sentences in 5% start are given more weightage in order to obtain the importance of introduction about video from that so that it will be helpful for user to decide whether to watch a video or not.

e) **Sentence Length:** Sentences too short and too long are given less importance.

f) **TF-ISF:** The relevance of each sentence with the query can be calculated by using the TF-ISF algorithm, which considers the number of occurrences of every term in the query in individual sentences as well as the whole document. Again, sentences are scored by applying the TF-ISF algorithm.

g) **Unnecessary information:** Sentences containing words such as thus, because are giving less importance.

7. After that each sentence is scored and a threshold is set to get a summarized document.
8. After each document related to query is summarized re-ranking of documents takes places based on video content been summarized and on query being searched.
9. In this user gets a gist of videos links in re-ranked form followed by summarized content based on query being searched.

5.2 Comparison with existing Algorithms:

I. Speech Recognition Module:

	API's comparison				
Parameters	**Sphinx**	**WIT.AI**	**IBM Watson**	**Houndify**	**Google Cloud Speech to text**
Accuracy vs Ideal Speech	Precision - 0.85 Recall - 0.83	Precision - 0.95 Recall - 0.90	Precision - 0.96 Recall - 0.95	Precision - 0.95 Recall - 0.82	Precision - 0.96 Recall - 0.91
Advantages	Works Offline	No limitations on request rate	Automatically transcribe audio for 7 languages	Accuracy is very high.	Response time is very less.
Lacuna	Needs to add language module to improve accuracy	Speech to text conversion accuracy is very poor.	Terminate the conversion in the middle of a sentence.	Convert audio to text upto 1 min length.	Punctuation is not provided.

Table 1. Comparison of Speech to Text API

The input given is a twitter audio.Following are outputs obtained from different API's:

Ideal: the wise rabbit long ago a cruel lion ruled a forest he killed many animals so the animals decided to send him an animal each day for his meal one day it was a rabbit's turn he met him and said Your Majesty I meant a line on my way he wants to find you and rule his kingdom the angry lion asked a rabbit to take him there the clever rabbit took him to a lake the line saw his own reflection and thinking it to be the other line roared at it when he saw his own reflection roar back he was so furious that he jumped in and drowned worthy other animals happy with a rabbit yes or no leave a comment

Sphinx : the wise cracking long ago a cruel lying through the forest he killed many animals to the animals decided to send in an hour each day for is neal one day it was ravaged her he met him and said your majesty i'm airline on my way he wants to find you into his kingdom the angry lying asked about it to take him there the clever wrap it up into lake the line saw his own reflection and thinking it to the younger line where that it when he signs on reflection we're back to you so i'm furious that he jumped in and drowned when the animals happy with the rabbit yes are now being a comment.

Houndify : the wise rabbit long ago a cruel line rule the forest he killed many animals so the animals decided to send in an animal each day for his meal one day it was a rabbit's turn he met him instead your majesty I'm not a line on my way he wants to find you and rule the kingdom the angry line asked a rabbit to take him there the clever rabbit to come to a lake the line saw his own reflection and thinking it to be the other line ward at it when he saw his own reflection roar back he was so furious that he jumped

IBM Watson: the wise rabbit long ago a cruel line rule the forest he killed many animals so the animals decided to send in an animal each day for his meal one day it was a rabbit's turn he met him instead your majesty I'm not a line on my way he wants to find you and rule the kingdom the angry line asked a rabbit to take him there the clever rabbit to come to a lake the line saw his own reflection and thinking it to be the other line ward at it when he saw his own reflection roar back he was so furious that he jumped in and drowned where the other animals happy with the rabbit yes or no leave a comment

wit.ai: the wise Robin long ago accrual ruled Forest he killed many animals so the animals decided to spend of an animal each day for his meal one day it was a rabbit turn he met him and said your majesty I met a line on my way he wants to find you and Rudy Kingdom the angry line after rabbit to take him there the clever rabbit took him to a lake the line saw his own reflection and thinking it to be the other line Ward at it when he saw his own reflection War back he was so Furious that he jumped tin and drowned where the other animals have rabbit yes or no leave a comment

Google cloud speech to text: the wise rabbit long ago ruled Forest he killed many animals so the animals decided attended an animal each day for his one day it was a rabbit Stern he met him and said your majesty I met a line on my way he wants to fight you and ruled his kingdom at the rabbit to take him there the clever rabbit took him to a lake the line saw his own reflection and thinking it to be the other line roared at it he was so Furious that he jumped in where's the other animals happy with the rabbit yes or no leave a comment

II. Summarization Module:

Algorithm	TextRank	LexRank	TF-IDF	TF-ISF	Proposed System
Accuracy	41%	44%	61%	40%	70%
Lacuna	Does not consider related words	Similarity with user query or title is not considered	Relevant for multiple documents	Does not consider other factors such as sentence scoring, position, length, etc.	Eliminates lacunae in individual algorithms by employing an integrated approach.
Evaluation	ROUGE 0.4229	ROUGE 0.4443	ROUGE 0.3101	ROUGE 0.39925	ROUGE 0.70

Table 2. Comparison of Summarization Algorithms

5.3 Evaluation of Developed System:

CODE:

```
from PyRouge.pyrouge import Rouge
r = Rouge()
system_generated_summary = "I just say Jenn the director of effortless English.I have something exciting to talk about today.What a strange word!Well, what I'm doing is I'm doing small little podcast on Twitter, and I am doing these almost everyday, because it's so easy.I use my phone my iPhone and I record a short talk about some Topic in my daily life, my normal life, and then I put it on Twitter twitter.com, and this is a very easy way for you to get more easy.English listening about daily topics about normal daily life, so I am continuing the podcast, of course, but the podcast is usually about learning ideas, teaching ideas.How can you speak English better?Just now I did a short audio, tweet and audio little podcast, just two and a half minutes about the weather in San Francisco.So if you want to hear new English listening topics everyday, please follow me on Twitter, so you will get longer more serious listening here at the podcast and you can get almost daily listening about easy, simple topics on my Twitter page.Com, A J Hoge that / a j, a h, o g e twitter.com forward, slash AJ Hoge!There'S a button that says follow click the follow button.Since I can use my phone, my cell phone, when I'm walking around the city when I'm doing my normal life, sometimes I will stop and talk for one or two minutes about something happening in my life and put it on my Twitter page.So you can get a lot of new, easy, daily life English listening this way it's twitter.com flash AJ Hoge."
```

```
manual_summmary = "I just say Jenn the director of effortless English.I am now doing audio
tweets. I use my phone my iPhone and I record a short talk about some Topic in my daily life, my
normal life, and then I put it on Twitter twitter.com, and this is a very easy way for you to get more
easy.English listening about daily topics about normal daily life, so I am continuing the podcast, of
course, but the podcast is usually about learning ideas, teaching ideas. How can you speak English
better?It'S about very simple day today: topics, for example. Just now I did a short audio, tweet and
audio little podcast, just two and a half minutes about the weather in San Francisco.So how can
you follow me on Twitter?You can make a Twitter account and then follow me. There'S a button
that says follow click the follow button.So you can get a lot of new, easy, daily life English
listening this way it's twitter.com flash AJ Hoge."

print r.rouge_l([system_generated_summary], [manual_summmary])
[precision, recall, f_score] = r.rouge_l([system_generated_summary], [manual_summmary])
print("Precision is :"+str(precision)+"\nRecall is :"+str(recall)+"\nF Score is :"+str(f_score))
```

II. Results Obtained:
Precision : 0.608
Recall : 0.916
F Score : 0.701

CHAPTER - 6 : TESTING

Testing is the process of evaluating a system or its component(s) with the intent to find whether it satisfies the specified requirements or not. Testing is executing a system in order to identify any gaps, errors, or missing requirements in contrary to the actual requirements.This chapter describes the results obtained by performing different testing methodologies i.e Unit Testing, Integration Testing, Performance Testing, System Testing.

6.1 Unit Testing:

Unit Testing is a software testing method by which individual units of source code, sets of one or more computer program modules together with associated control data, usage procedures, and operating procedures, are tested to determine whether they are fit for use. Unit testing involves following module for testing.

- Youtube Video to audio downloader
- Video to text converter module
- Text Summarization module
- Re-ranked module

6.2 Integration Testing:

Integration testing is a software testing methodology used to test individual software components or units of code to verify interaction between various software components and detect interface defects. Components are tested as a single group or organized in an iterative manner.

- Each module require strong internet access otherwise failure occurs. User will get error message if any module is not working because of any failure in the system.
- Each module is dependent on the previous module. So all three modules integrated properly so as to give one complete integrated output to the user.

6.3 Performance Testing:

Performance testing is type of testing perform to determine the performance of system to major the measure, validate or verify quality attributes of the system like responsiveness, Speed, Scalability, Stability under variety of load conditions.

- The main focus of the system is summarization hence quality of original text should be very high for which powerful speech to text module is required for the system.
- Hence we tested Various API that are used for speech to text part which are Sphinx wit.ai, Google speech to text, IBM Watson, Houndify, Google Cloud Speech to text.
- Among which Google cloud speech to text shows highest accuracy in terms of response time and quality of generated text.
- Various algorithms were used for the summarization but implemented algorithm is hybrid approach which consider various factors such as sentence length, sentence position, TF-ISF, wordnet similarity etc which improves efficiency of text summarization.

6.4 System Testing:

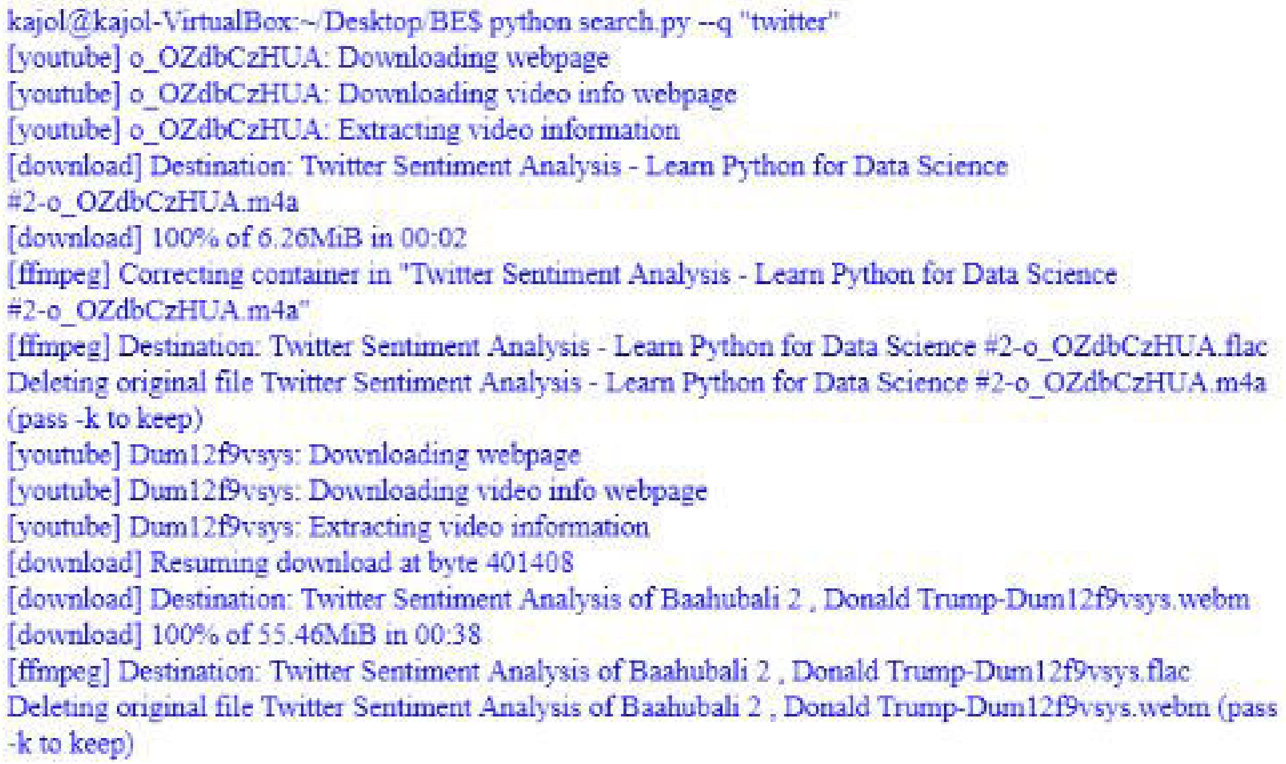
System testing is the type of testing to check the behaviour of a complete and fully generic integrated software product based on the software requirements specification(SRS) document. The main focus of this testing is to evaluate Business / Functional / End-user
requirements. In System testing, the functionalities of the system are tested from an end-to-end perspective. It is the final test to verify that the product to be delivered meets the specifications mentioned in the requirement document.

CHAPTER - 7 : RESULT ANALYSIS

7.1 Output Screenshots:

Step I : Youtube video download and audio extraction: In this step, according to the user's search query youtube's top videos are downloaded in local registry and audio is extracted from that.

OUTPUT:

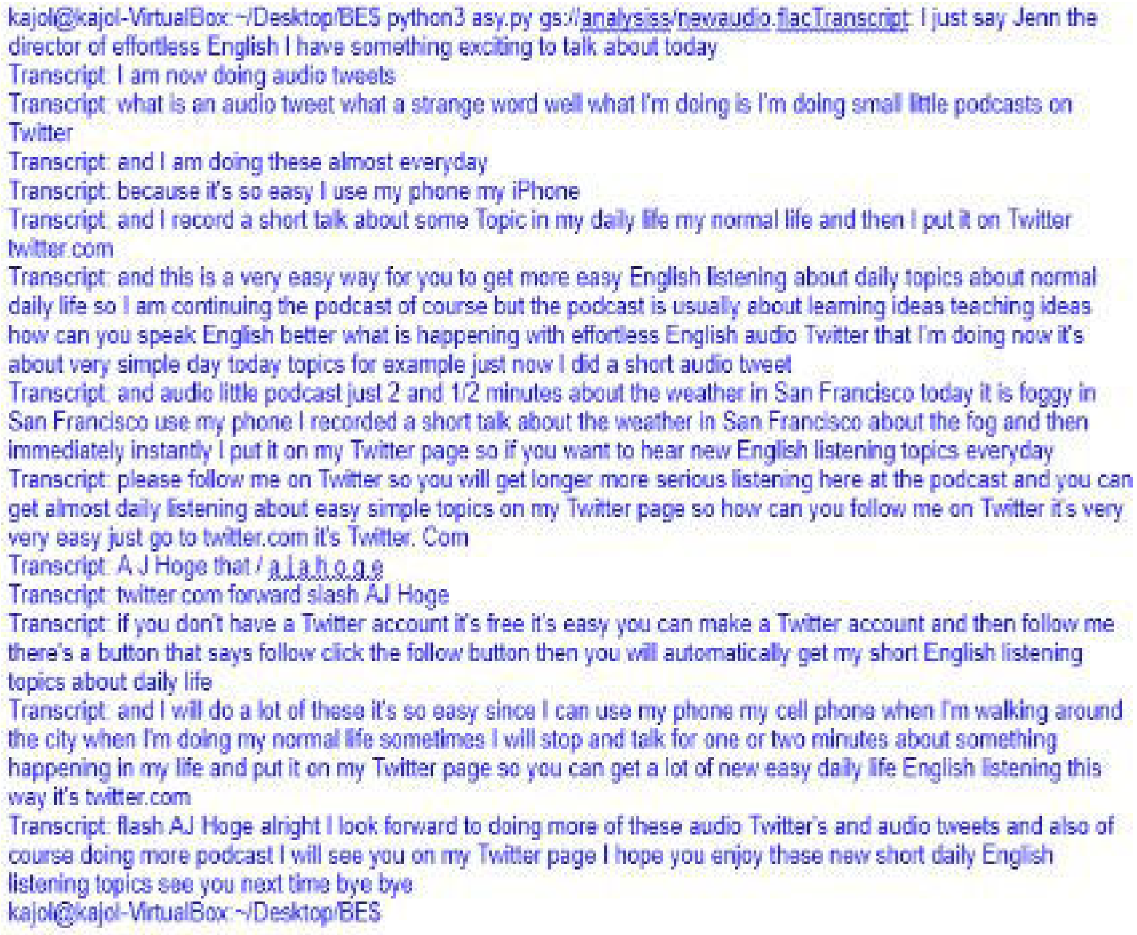


```
kajol@kajol-VirtualBox:~/Desktop/BE$ python search.py --q "twitter"
[youtube] o_OZdbCzHUA: Downloading webpage
[youtube] o_OZdbCzHUA: Downloading video info webpage
[youtube] o_OZdbCzHUA: Extracting video information
[download] Destination: Twitter Sentiment Analysis - Learn Python for Data Science
#2-o_OZdbCzHUA.m4a
[download] 100% of 6.26MiB in 00:02
[ffmpeg] Correcting container in "Twitter Sentiment Analysis - Learn Python for Data Science
#2-o_OZdbCzHUA.m4a"
[ffmpeg] Destination: Twitter Sentiment Analysis - Learn Python for Data Science #2-o_OZdbCzHUA.flac
Deleting original file Twitter Sentiment Analysis - Learn Python for Data Science #2-o_OZdbCzHUA.m4a
(pass -k to keep)
[youtube] Dum12f9vsys: Downloading webpage
[youtube] Dum12f9vsys: Downloading video info webpage
[youtube] Dum12f9vsys: Extracting video information
[download] Resuming download at byte 401408
[download] Destination: Twitter Sentiment Analysis of Baahubali 2 , Donald Trump-Dum12f9vsys.webm
[download] 100% of 55.46MiB in 00:38
[ffmpeg] Destination: Twitter Sentiment Analysis of Baahubali 2 , Donald Trump-Dum12f9vsys.flac
Deleting original file Twitter Sentiment Analysis of Baahubali 2 , Donald Trump-Dum12f9vsys.webm (pass
-k to keep)
```

Fig 17. Audio Extraction

Step II : Speech to text: In this step Google Cloud Speech to text API is used to convert audio into text , audio is uploaded in cloud so as to use the functionality of cloud. Fig 18. shows that to convert an audio to text need to access via gsutil i.e (gs://bucket_name/audio_name) .The text obtained is used in the next step.

OUTPUT:

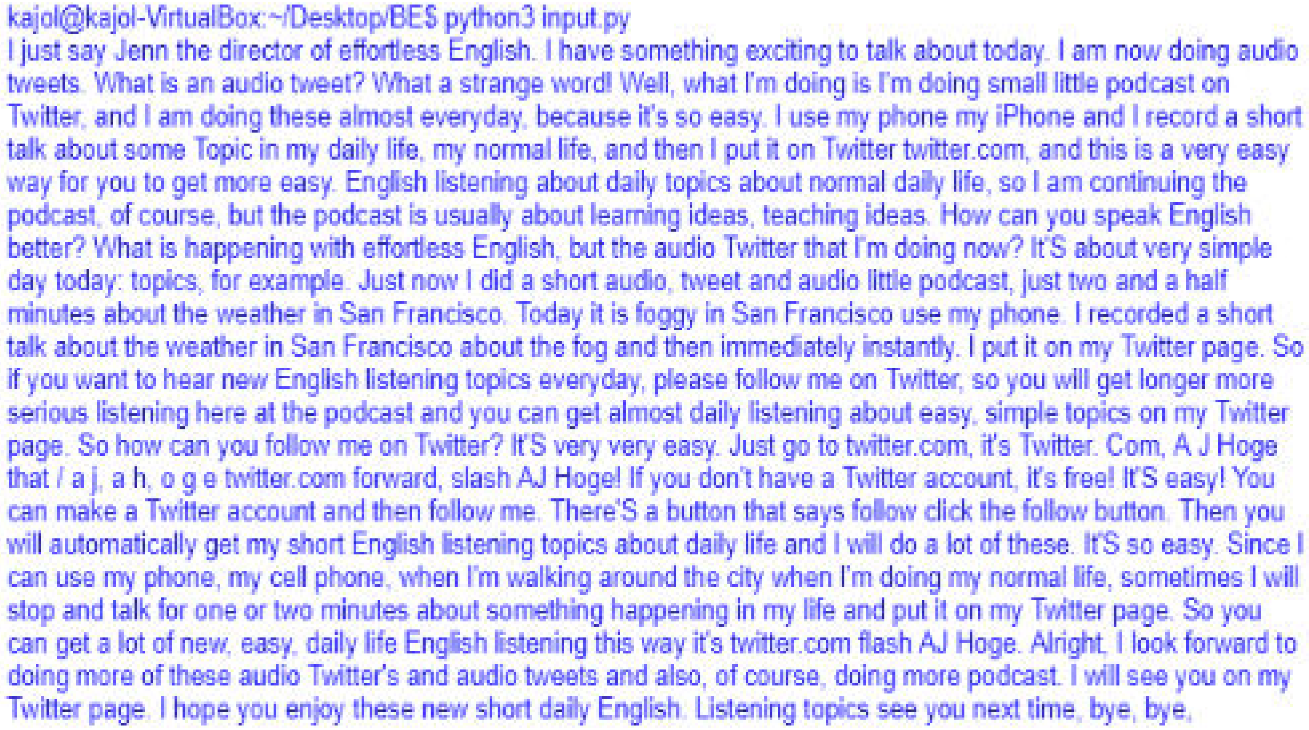


```
kajol@kajol-VirtualBox:~/Desktop/BES python3 asy.py gs://analysiss/newaudio.flacTranscript: I just say Jenn the
director of effortless English I have something exciting to talk about today
Transcript: I am now doing audio tweets
Transcript: what is an audio tweet what a strange word well what I'm doing is I'm doing small little podcasts on
Twitter
Transcript: and I am doing these almost everyday
Transcript: because it's so easy I use my phone my iPhone
Transcript: and I record a short talk about some Topic in my daily life my normal life and then I put it on Twitter
twitter.com
Transcript: and this is a very easy way for you to get more easy English listening about daily topics about normal
daily life so I am continuing the podcast of course but the podcast is usually about learning ideas teaching ideas
how can you speak English better what is happening with effortless English audio Twitter that I'm doing now it's
about very simple day today topics for example just now I did a short audio tweet
Transcript: and audio little podcast just 2 and 1/2 minutes about the weather in San Francisco today it is foggy in
San Francisco use my phone I recorded a short talk about the weather in San Francisco about the fog and then
immediately instantly I put it on my Twitter page so if you want to hear new English listening topics everyday
Transcript: please follow me on Twitter so you will get longer more serious listening here at the podcast and you can
get almost daily listening about easy simple topics on my Twitter page so how can you follow me on Twitter it's very
very easy just go to twitter.com it's Twitter. Com
Transcript: A J Hoge that / a.j.a.h.o.g.e
Transcript: twitter com forward slash AJ Hoge
Transcript: if you don't have a Twitter account it's free it's easy you can make a Twitter account and then follow me
there's a button that says follow click the follow button then you will automatically get my short English listening
topics about daily life
Transcript: and I will do a lot of these it's so easy since I can use my phone my cell phone when I'm walking around
the city when I'm doing my normal life sometimes I will stop and talk for one or two minutes about something
happening in my life and put it on my Twitter page so you can get a lot of new easy daily life English listening this
way it's twitter.com
Transcript: flash AJ Hoge alright I look forward to doing more of these audio Twitter's and audio tweets and also of
course doing more podcast I will see you on my Twitter page I hope you enjoy these new short daily English
listening topics see you next time bye bye
kajol@kajol-VirtualBox:~/Desktop/BES
```

Fig 18. Google Cloud Speech to text conversion

Step III: Adding Punctuation to the text: Punctuation is very important part of any document. None of the speech to text API provides punctuation in the output. So we need to add punctuation in output of Google cloud's Speech to text. We use Punctuator which adds effectively punctuation in the text.

OUTPUT:



```
kajol@kajol-VirtualBox:~/Desktop/BE$ python3 input.py
I just say Jenn the director of effortless English. I have something exciting to talk about today. I am now doing audio
tweets. What is an audio tweet? What a strange word! Well, what I'm doing is I'm doing small little podcast on
Twitter, and I am doing these almost everyday, because it's so easy. I use my phone my iPhone and I record a short
talk about some Topic in my daily life, my normal life, and then I put it on Twitter twitter.com, and this is a very easy
way for you to get more easy. English listening about daily topics about normal daily life, so I am continuing the
podcast, of course, but the podcast is usually about learning ideas, teaching ideas. How can you speak English
better? What is happening with effortless English, but the audio Twitter that I'm doing now? It'S about very simple
day today: topics, for example. Just now I did a short audio, tweet and audio little podcast, just two and a half
minutes about the weather in San Francisco. Today it is foggy in San Francisco use my phone. I recorded a short
talk about the weather in San Francisco about the fog and then immediately instantly. I put it on my Twitter page. So
if you want to hear new English listening topics everyday, please follow me on Twitter, so you will get longer more
serious listening here at the podcast and you can get almost daily listening about easy, simple topics on my Twitter
page. So how can you follow me on Twitter? It'S very very easy. Just go to twitter.com, it's Twitter. Com, A J Hoge
that / a j, a h, o g e twitter.com forward, slash AJ Hoge! If you don't have a Twitter account, it's free! It'S easy! You
can make a Twitter account and then follow me. There'S a button that says follow click the follow button. Then you
will automatically get my short English listening topics about daily life and I will do a lot of these. It'S so easy. Since I
can use my phone, my cell phone, when I'm walking around the city when I'm doing my normal life, sometimes I will
stop and talk for one or two minutes about something happening in my life and put it on my Twitter page. So you
can get a lot of new, easy, daily life English listening this way it's twitter.com flash AJ Hoge. Alright, I look forward to
doing more of these audio Twitter's and audio tweets and also, of course, doing more podcast. I will see you on my
Twitter page. I hope you enjoy these new short daily English. Listening topics see you next time, bye, bye,
```

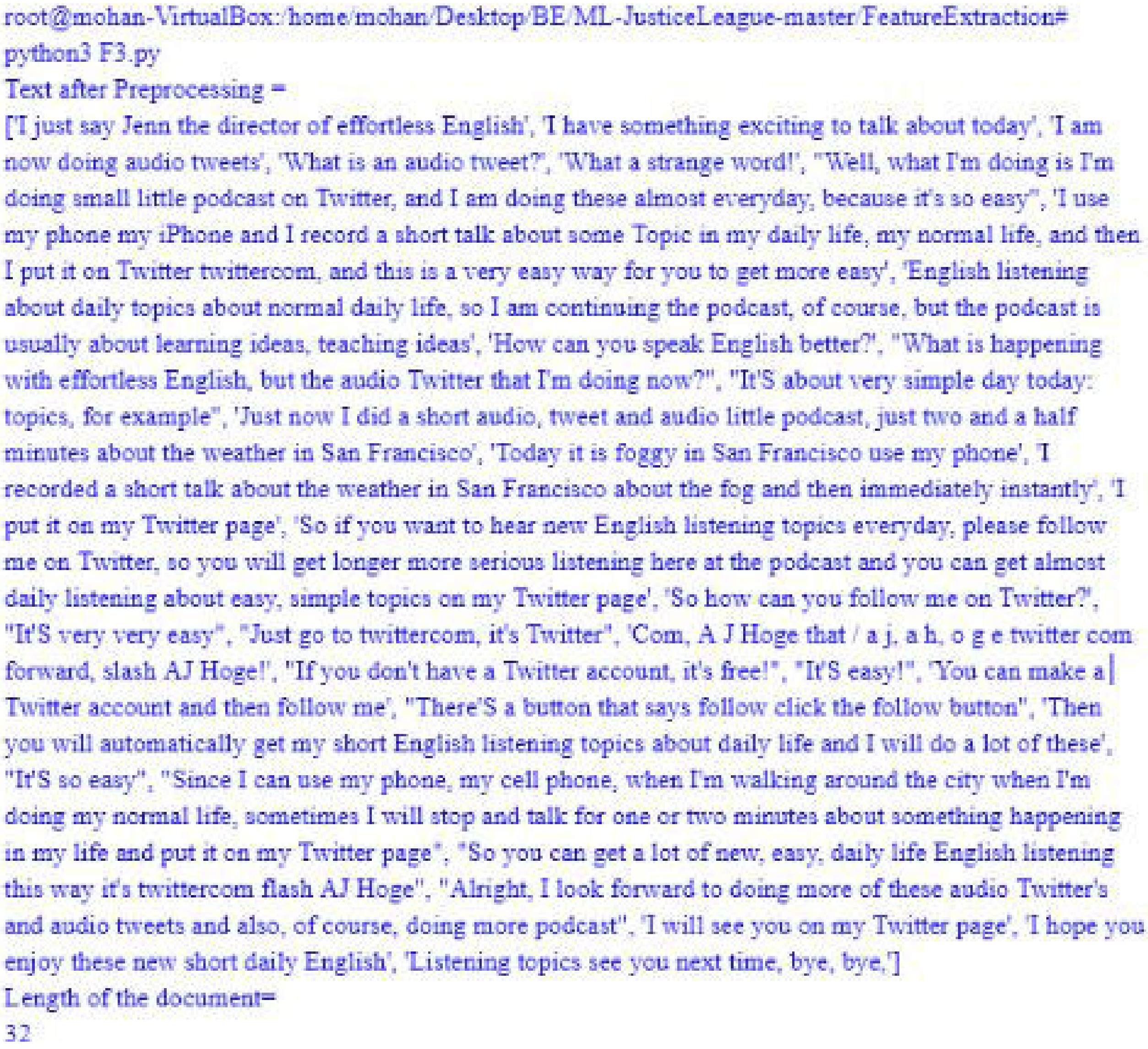
Fig 19. Punctuated text

Step IV: Summarization :

Input Text: twitter_input.txt

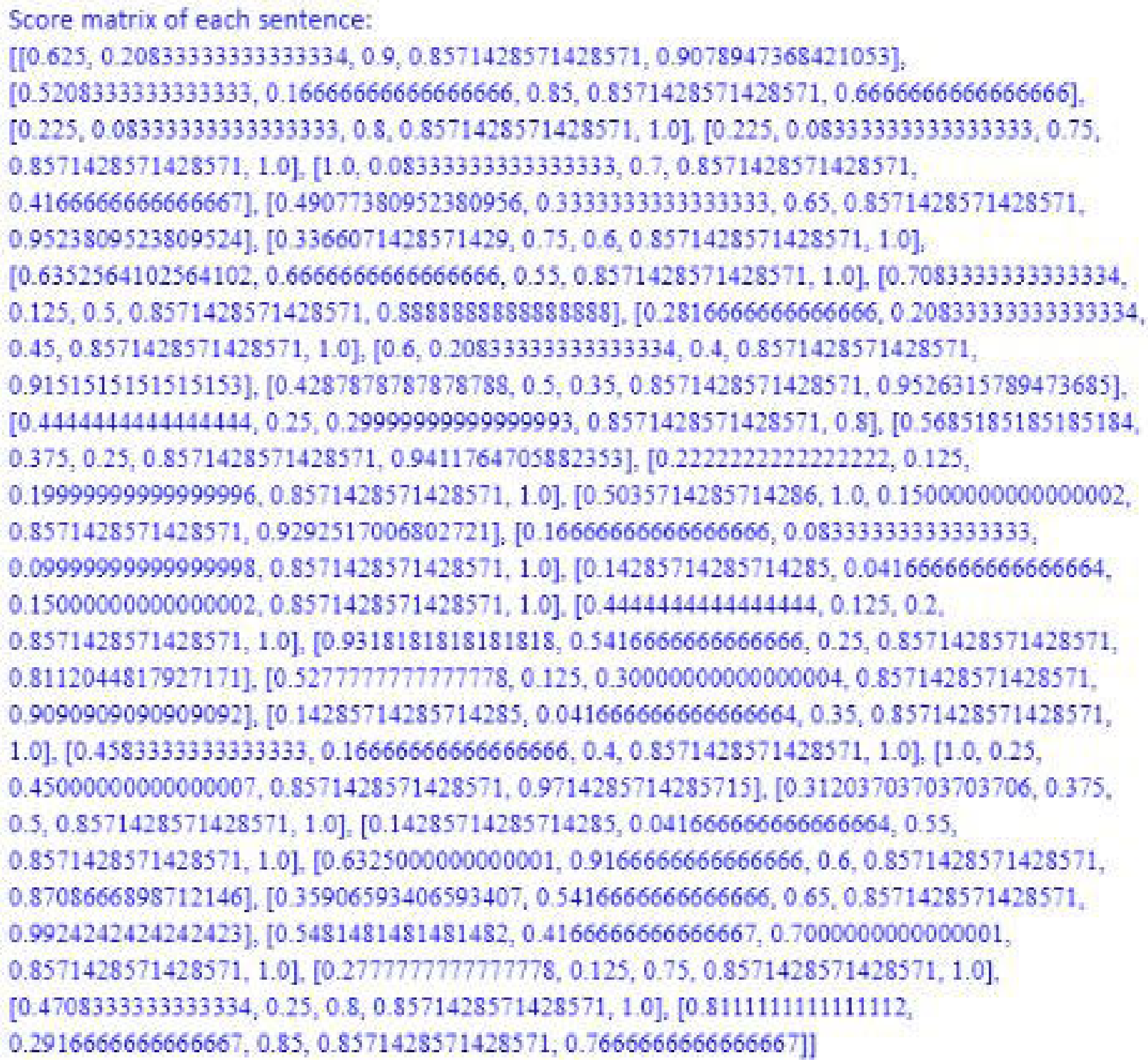
I just say Jenn the director of effortless English. I have something exciting to talk about today. I am now doing audio tweets. What is an audio tweet? What a strange word! Well, what I'm doing is I'm doing small little podcast on Twitter, and I am doing these almost everyday, because it's so easy. I use my phone my iPhone and I record a short talk about some Topic in my daily life, my normal life, and then I put it on Twitter twitter.com, and this is a very easy way for you to get more easy. English listening about daily topics about normal daily life, so I am continuing the podcast, of course, but the podcast is usually about learning ideas, teaching ideas. How can you speak English better? What is happening with effortless English, but the audio Twitter that I'm doing now? It'S about very simple day today: topics, for example. Just now I did a short audio, tweet and audio little podcast, just two and a half minutes about the weather in San Francisco. Today it is foggy in San Francisco use my phone. I recorded a short talk about the weather in San Francisco about the fog and then immediately instantly. I put it on my Twitter page. So if you want to hear new English listening topics everyday, please follow me on Twitter, so you will get longer more serious listening here at the podcast and you can get almost daily listening about easy, simple topics on my Twitter page. So how can you follow me on Twitter? It'S very very easy. Just go to twitter.com, it's Twitter. Com, A J Hoge that / a j, a h, o g e twitter.com forward, slash AJ Hoge! If you don't have a Twitter account, it's free! It'S easy! You can make a Twitter account and then follow me. There'S a button that says follow click the follow button. Then you will automatically get my short English listening topics about daily life and I will do a lot of these. It'S so easy. Since I can use my phone, my cell phone, when I'm walking around the city when I'm doing my normal life, sometimes I will stop and talk for one or two minutes about something happening in my life and put it on my Twitter page. So you can get a lot of new, easy, daily life English listening this way it's twitter.com flash AJ Hoge. Alright, I look forward to doing more of these audio Twitter's and audio tweets and also, of course, doing more podcast. I will see you on my Twitter page. I hope you enjoy these new short daily English. Listening topics see you next time, bye, bye,

Result after summarization:



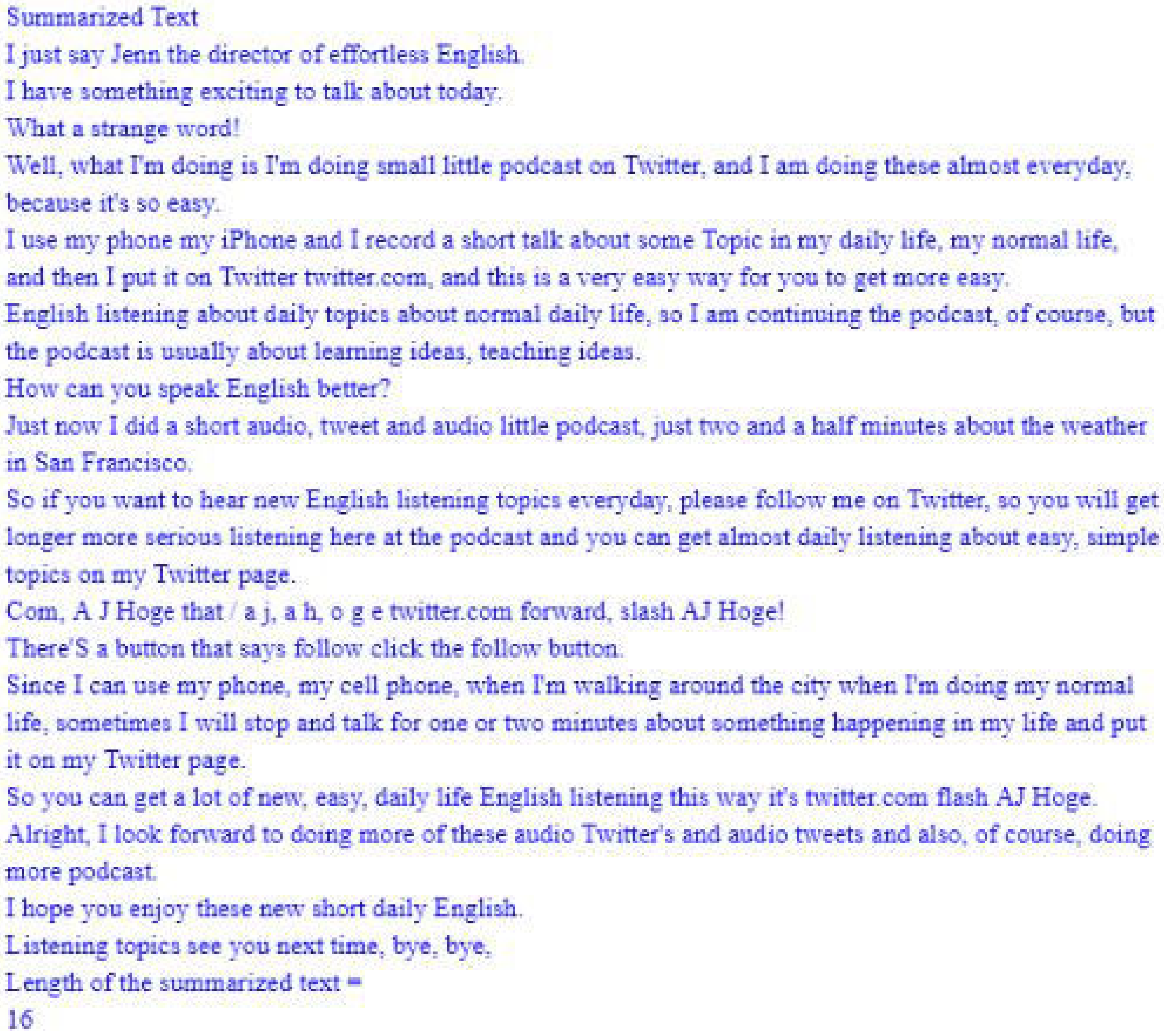
```
root@mohan-VirtualBox:/home/mohan/Desktop/BE/ML-JusticeLeague-master/FeatureExtraction#
python3 F3.py
Text after Preprocessing =
['I just say Jenn the director of effortless English', 'I have something exciting to talk about today', 'I am
now doing audio tweets', 'What is an audio tweet?', 'What a strange word!', "Well, what I'm doing is I'm
doing small little podcast on Twitter, and I am doing these almost everyday, because it's so easy", 'I use
my phone my iPhone and I record a short talk about some Topic in my daily life, my normal life, and then
I put it on Twitter twittercom, and this is a very easy way for you to get more easy', 'English listening
about daily topics about normal daily life, so I am continuing the podcast, of course, but the podcast is
usually about learning ideas, teaching ideas', 'How can you speak English better?', "What is happening
with effortless English, but the audio Twitter that I'm doing now?", "It'S about very simple day today:
topics, for example", 'Just now I did a short audio, tweet and audio little podcast, just two and a half
minutes about the weather in San Francisco', 'Today it is foggy in San Francisco use my phone', 'I
recorded a short talk about the weather in San Francisco about the fog and then immediately instantly', 'I
put it on my Twitter page', 'So if you want to hear new English listening topics everyday, please follow
me on Twitter, so you will get longer more serious listening here at the podcast and you can get almost
daily listening about easy, simple topics on my Twitter page', 'So how can you follow me on Twitter?',
"It'S very very easy", "Just go to twittercom, it's Twitter", 'Com, A J Hoge that / a j, a h, o g e twitter com
forward, slash AJ Hoge!', "If you don't have a Twitter account, it's free!", "It'S easy!", 'You can make a
Twitter account and then follow me', "There'S a button that says follow click the follow button", 'Then
you will automatically get my short English listening topics about daily life and I will do a lot of these',
"It'S so easy", "Since I can use my phone, my cell phone, when I'm walking around the city when I'm
doing my normal life, sometimes I will stop and talk for one or two minutes about something happening
in my life and put it on my Twitter page", "So you can get a lot of new, easy, daily life English listening
this way it's twittercom flash AJ Hoge", "Alright, I look forward to doing more of these audio Twitter's
and audio tweets and also, of course, doing more podcast", 'I will see you on my Twitter page', 'I hope you
enjoy these new short daily English', 'Listening topics see you next time, bye, bye,']
Length of the document=
32
```

Fig 20. Preprocessed Text Required For summarization



```
Score matrix of each sentence:
[[0.625, 0.20833333333333334, 0.9, 0.8571428571428571, 0.9078947368421053],
[0.5208333333333333, 0.16666666666666666, 0.85, 0.8571428571428571, 0.6666666666666666],
[0.225, 0.08333333333333333, 0.8, 0.8571428571428571, 1.0], [0.225, 0.08333333333333333, 0.75,
0.8571428571428571, 1.0], [1.0, 0.08333333333333333, 0.7, 0.8571428571428571,
0.41666666666666667], [0.49077380952380956, 0.3333333333333333, 0.65, 0.8571428571428571,
0.9523809523809524], [0.3366071428571429, 0.75, 0.6, 0.8571428571428571, 1.0],
[0.6352564102564102, 0.6666666666666666, 0.55, 0.8571428571428571, 1.0], [0.7083333333333334,
0.125, 0.5, 0.8571428571428571, 0.8888888888888888], [0.2816666666666666, 0.20833333333333334,
0.45, 0.8571428571428571, 1.0], [0.6, 0.20833333333333334, 0.4, 0.8571428571428571,
0.9151515151515153], [0.4287878787878788, 0.5, 0.35, 0.8571428571428571, 0.9526315789473685],
[0.4444444444444444, 0.25, 0.29999999999999993, 0.8571428571428571, 0.8], [0.5685185185185184,
0.375, 0.25, 0.8571428571428571, 0.9411764705882353], [0.2222222222222222, 0.125,
0.19999999999999996, 0.8571428571428571, 1.0], [0.5035714285714286, 1.0, 0.15000000000000002,
0.8571428571428571, 0.9292517006802721], [0.16666666666666666, 0.08333333333333333,
0.09999999999999998, 0.8571428571428571, 1.0], [0.14285714285714285, 0.041666666666666664,
0.15000000000000002, 0.8571428571428571, 1.0], [0.4444444444444444, 0.125, 0.2,
0.8571428571428571, 1.0], [0.9318181818181818, 0.5416666666666666, 0.25, 0.8571428571428571,
0.8112044817927171], [0.5277777777777778, 0.125, 0.30000000000000004, 0.8571428571428571,
0.9090909090909092], [0.14285714285714285, 0.041666666666666664, 0.35, 0.8571428571428571,
1.0], [0.4583333333333333, 0.16666666666666666, 0.4, 0.8571428571428571, 1.0], [1.0, 0.25,
0.45000000000000007, 0.8571428571428571, 0.9714285714285715], [0.31203703703703706, 0.375,
0.5, 0.8571428571428571, 1.0], [0.14285714285714285, 0.041666666666666664, 0.55,
0.8571428571428571, 1.0], [0.6325000000000001, 0.9166666666666666, 0.6, 0.8571428571428571,
0.8708666898712146], [0.3590659340659340 7, 0.5416666666666666, 0.65, 0.8571428571428571,
0.9924242424242423], [0.5481481481481482, 0.41666666666666667, 0.7000000000000001,
0.8571428571428571, 1.0], [0.2777777777777778, 0.125, 0.75, 0.8571428571428571, 1.0],
[0.4708333333333334, 0.25, 0.8, 0.8571428571428571, 1.0], [0.8111111111111112,
0.2916666666666667, 0.85, 0.8571428571428571, 0.7666666666666667]]
```

Fig 21. Score Matrix For each Sentence



```
Summarized Text
I just say Jenn the director of effortless English.
I have something exciting to talk about today.
What a strange word!
Well, what I'm doing is I'm doing small little podcast on Twitter, and I am doing these almost everyday,
because it's so easy.
I use my phone my iPhone and I record a short talk about some Topic in my daily life, my normal life,
and then I put it on Twitter twitter.com, and this is a very easy way for you to get more easy.
English listening about daily topics about normal daily life, so I am continuing the podcast, of course, but
the podcast is usually about learning ideas, teaching ideas.
How can you speak English better?
Just now I did a short audio, tweet and audio little podcast, just two and a half minutes about the weather
in San Francisco.
So if you want to hear new English listening topics everyday, please follow me on Twitter, so you will get
longer more serious listening here at the podcast and you can get almost daily listening about easy, simple
topics on my Twitter page.
Com, A J Hoge that / a j, a h, o g e twitter.com forward, slash AJ Hoge!
There'S a button that says follow click the follow button.
Since I can use my phone, my cell phone, when I'm walking around the city when I'm doing my normal
life, sometimes I will stop and talk for one or two minutes about something happening in my life and put
it on my Twitter page.
So you can get a lot of new, easy, daily life English listening this way it's twitter.com flash AJ Hoge.
Alright, I look forward to doing more of these audio Twitter's and audio tweets and also, of course, doing
more podcast.
I hope you enjoy these new short daily English.
Listening topics see you next time, bye, bye.
Length of the summarized text =
16
```

Fig 22. Summarization Output

7.2 Observation and Analysis:

- Based on query, videos in re-ranked form are displayed.
- Summary of videos is obtained.
- Videos links is obtained

Table 3. shows summary of system with other existing system

	Analysis of summary obtained by different system					
Parameters	**Text Summarizer**	**Free Summarizer**	**Tools for noobs**	**Auto Summarizer**	**Text Comparator**	**Proposed System**
Precision	0.82	0.63	0.48	0.56	0.54	0.61
Recall	0.31	0.72	0.59	0.70	0.77	0.92
F Score	0.46	0.67	0.53	0.63	0.64	0.70

Table 3. Analysis of Summary obtained by System

Observation of Proposed system summary w.r.t Manual Summary:

Ideal Summary :

I just say Jenn the director of effortless English.I am now doing audio tweets.I use my phone my iPhone and I record a short talk about some Topic in my daily life, my normal life, and then I put it on Twitter twitter.com, and this is a very easy way for you to get more easy.English listening about daily topics about normal daily life, so I am continuing the podcast, of course, but the podcast is usually about learning ideas, teaching ideas. How can you speak English better?It'S about very simple day today: topics, for example. Just now I did a short audio, tweet and audio little podcast, just two and a half minutes about the weather in San Francisco.So how can you follow me on Twitter?You can make a Twitter account and then follow me. There'S a button that says follow click the follow button.So you can get a lot of new, easy, daily life English listening this way it's twitter.com flash AJ Hoge.

Proposed Summary:

I just say Jenn the director of effortless English.I have something exciting to talk about today. What a strange word!Well, what I'm doing is I'm doing small little podcast on Twitter, and I am doing these almost everyday, because it's so easy.I use my phone my iPhone and I record a short talk about some Topic in my daily life, my normal life, and then I put it on Twitter twitter.com, and this is a very easy way for you to get more easy.English listening about daily topics about normal daily life, so I am continuing the podcast, of course, but the podcast is usually about learning ideas, teaching ideas.How can you speak English better?Just now I did a short audio, tweet and audio little podcast, just two and a half minutes about the weather in San Francisco.So if you want to hear new English listening topics everyday, please follow me on Twitter, so you will get longer more serious listening here at the podcast and you can get almost daily listening about easy, simple topics on my Twitter page.Com, A J Hoge that / a j, a h, o g e twitter.com forward, slash AJ Hoge!There'S a button that says follow click the follow button.Since I can use my phone, my cell phone, when I'm walking around the city when I'm doing my normal life, sometimes I will stop and talk for one or two minutes about something happening in my life and put it on my Twitter page.So you can get a lot of new, easy, daily life English listening this way it's twitter.com flash AJ Hoge.

CHAPTER - 8 : CONCLUSION

This chapter aims at summarizing all the chapters included in the book. Also this chapter aims at specifying the limitations and future scope of project.

8.1 Limitations

- Multilingual- Restricted to only English language.
- Abstract Summary- As a lot of CPU size is required.
- Music Video not Summarized- Only videos containing content are preprocessed.

8.2 Conclusion

The system aims to provide user with gist of video in the form of text i.e an informative summary(not detailed summary). For generating this summary, aim is to get the context of the video in minimum no. of important sentences, to produce such informative summary various parameters are considered like sentence length, sentence position, similarity with title, similarity with query, TF ISF, sentence similarity. Important sentences are extracted using weighted score which is dependent on each of parameter.The less important sentences are the identified using the individual sentence score. Thus,a sentence with length too long and similar with other sentences will get less score than a sentence short and less similar with other sentences.This is because long sentences may contain words conjunctions such as because, thus ,therefore, which makes them explanatory sentences and need not to be included in summary.

Thus this procedure gives relevance of video with the query as well as the gist of video. The score generated is then used to re-rank the videos. Thus the re-ranked videos are arranged based on their relevance with the query and content of video.

8.3 Future Scope

- Multilingual Videos - This involves videos of various languages as input. This also acts as a communication medium for people from different cultures.
- Detecting terrorist activities - Terrorists use videos for training. However, often these are in a language different than english. Multilingual summary generation can be used to detect such activities from amongst hundreds of videos.
- Tagged Classification - This involves classification of videos based on tags given to each of the video. This improves the classification of videos as the tags represent keywords generated from summary of videos.Thus a single video can be classified into multiple classes.

REFERENCES:

[1] Tandel, A., Modi, B., Gupta, P., Wagle, S., & Khedkar, S. (2016, March). Multi-document text summarization-a survey. In *Data Mining and Advanced Computing (SAPIENCE), International Conference on* (pp. 331-334). IEEE.

[2] Shimpikar, Sheetal, and Sharvari Govilkar. "A Survey of Text Summarization Techniques for Indian Regional Languages." *International Journal of Computer Applications* 165.11 (2017).

[3] Wang, Shuai, et al. "Integrating Extractive and Abstractive Models for Long Text Summarization." *Big Data (BigData Congress), 2017 IEEE Internatonal Congress on*. IEEE, 2017.

[4]Doko, Alen, Maja Stula, and Darko Stipanicev. "A recursive TF-ISF Based Sentence Retrieval Method with Local Context." *International Journal of Machine Learning and Computing* 3.2 (2013): 195.

[5]Moawad, Ibrahim F., and Mostafa Aref. "Semantic graph reduction approach for abstractive Text Summarization." *Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on*. IEEE, 2012.

[6]Mihalcea, R., & Tarau, P. (2004, July). TextRank: Bringing Order into Text. In *EMNLP* (Vol. 4, pp. 404-411).

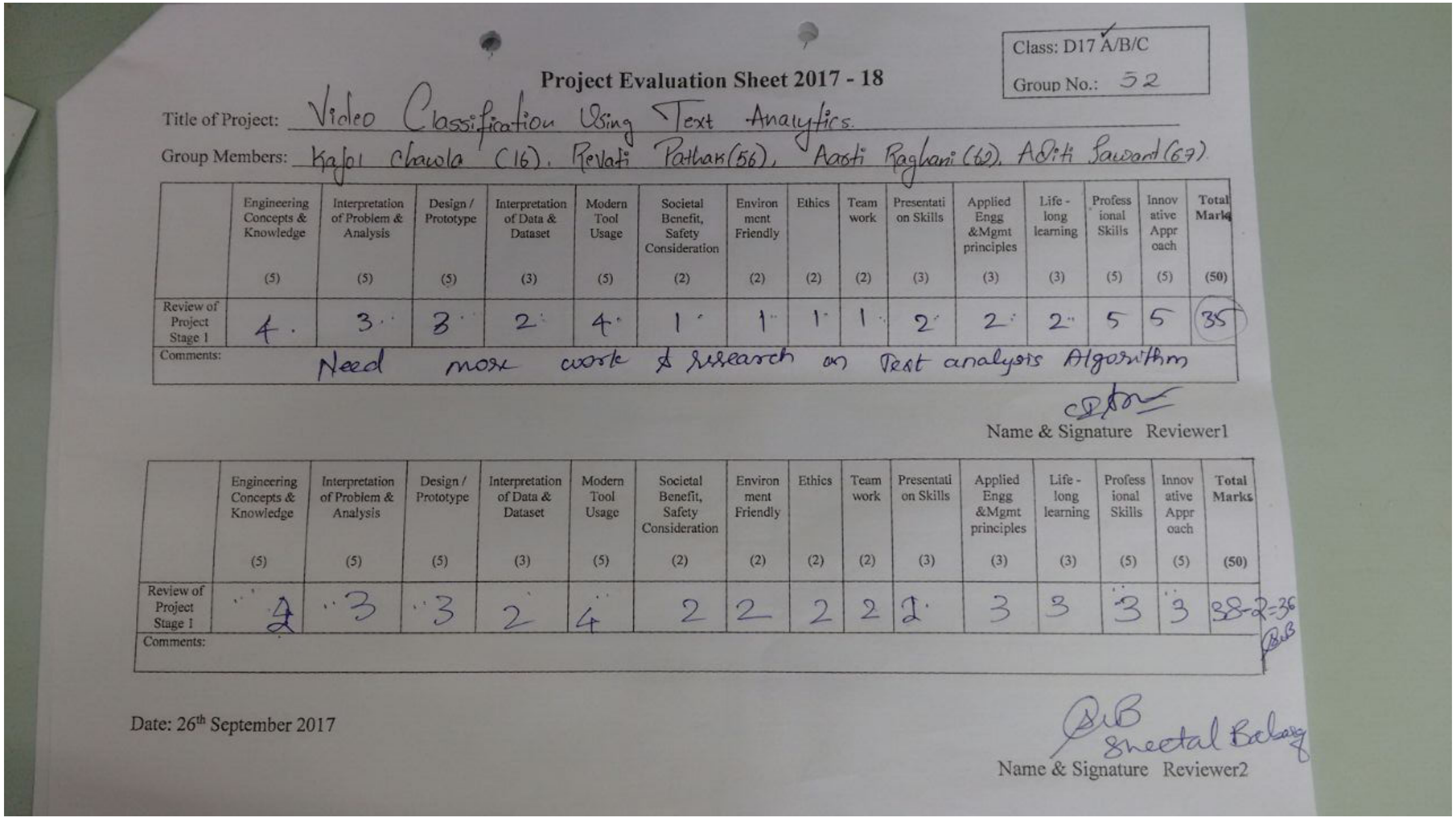
[7]Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 457-479.

[8 Ragunath, R., & Sivaranjani, N. (2015). Ontology based text document summarization system using concept terms. *ARPN Journal of Engineering and Applied Sciences*, *10*(6), 2638-2642.

[9]Blake, Catherine. "A comparison of document, sentence, and term event spaces." Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006.

[10]Azhar-ul-Haq, R. A Review: Ranking documents using Ranking Algorithms & Techniques.

Project Progress Review Sheet:



Class: D17 A/B/C

Group No.: 52

Project Evaluation Sheet 2017 - 18

Title of Project: Video Classification Using Text Analytics.

Group Members: Kajol Chawla (16), Revati Pathak (56), Aarti Raghani (62), Aditi Sawant (67).

	Engineering Concepts & Knowledge	Interpretation of Problem & Analysis	Design / Prototype	Interpretation of Data & Dataset	Modern Tool Usage	Societal Benefit, Safety Consideration	Environment Friendly	Ethics	Team work	Presentation Skills	Applied Engg &Mgmt principles	Life - long learning	Professional Skills	Innovative Approach	Total Marks
	(5)	(5)	(5)	(3)	(5)	(2)	(2)	(2)	(2)	(3)	(3)	(3)	(5)	(5)	(50)
Review of Project Stage 1	4.	3	3	2	4	1	1	1	1	2	2	2	5	5	35

Comments: Need more work & research on Text analysis Algorithm

Name & Signature Reviewer1

	Engineering Concepts & Knowledge	Interpretation of Problem & Analysis	Design / Prototype	Interpretation of Data & Dataset	Modern Tool Usage	Societal Benefit, Safety Consideration	Environment Friendly	Ethics	Team work	Presentation Skills	Applied Engg &Mgmt principles	Life - long learning	Professional Skills	Innovative Approach	Total Marks
	(5)	(5)	(5)	(3)	(5)	(2)	(2)	(2)	(2)	(3)	(3)	(3)	(5)	(5)	(50)
Review of Project Stage 1	2	3	3	2	4	2	2	2	2	2	3	3	3	3	38-2=36

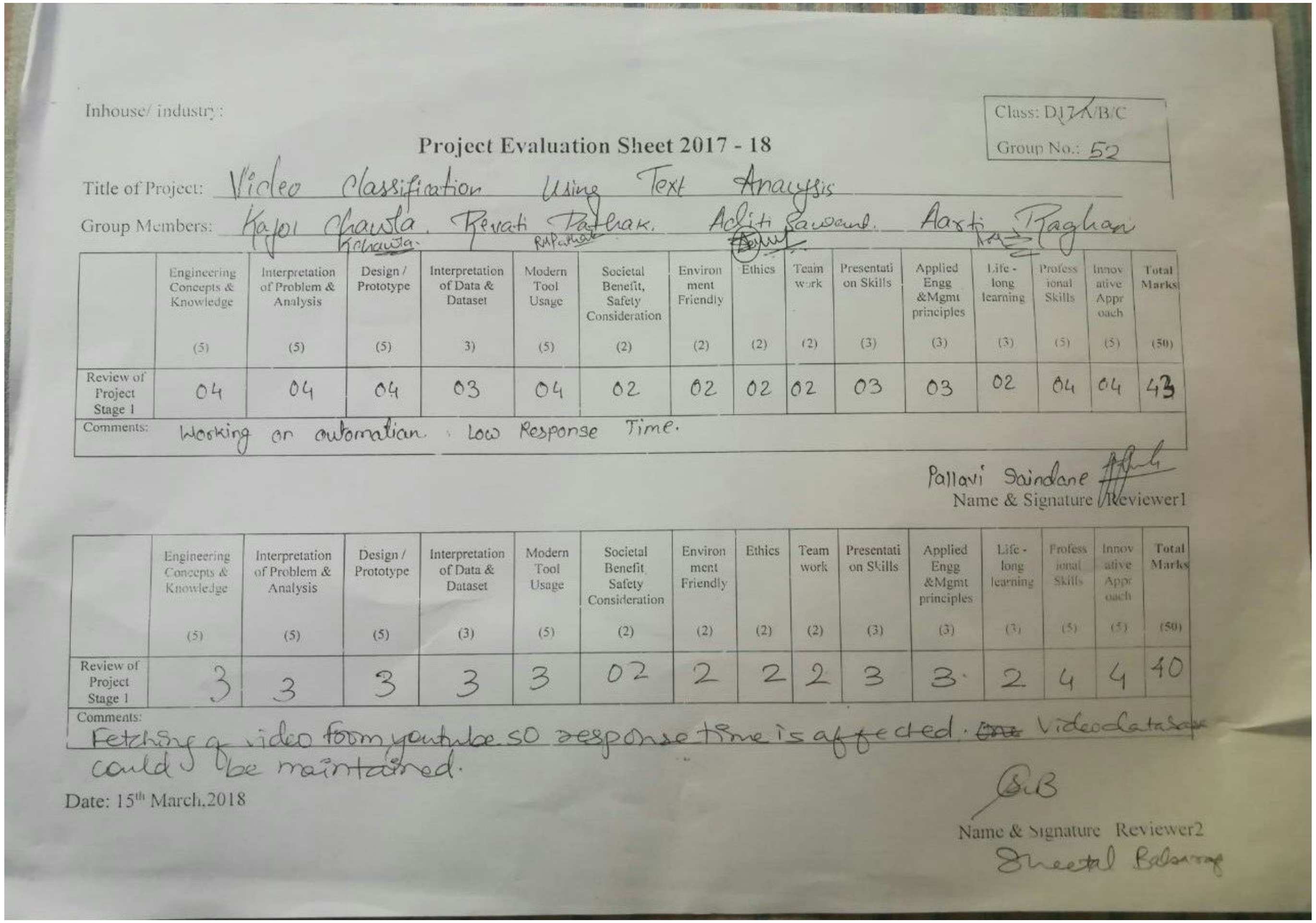
Comments:

Date: 26th September 2017

Sheetal Bahaeg

Name & Signature Reviewer2

Fig 23. Review 1 Sheet



Inhouse/ industry :

Class: DJ7A/B/C

Group No.: 52

Project Evaluation Sheet 2017 - 18

Title of Project: Video classification Using Text Analysis

Group Members: Kajol Chawla, Revati Pathak, Aditi Sawant, Aarti Raghav

	Engineering Concepts & Knowledge	Interpretation of Problem & Analysis	Design / Prototype	Interpretation of Data & Dataset	Modern Tool Usage	Societal Benefit, Safety Consideration	Environ ment Friendly	Ethics	Team work	Presentati on Skills	Applied Engg &Mgmt principles	Life - long learning	Profess ional Skills	Innov ative Appr oach	Total Marks
	(5)	(5)	(5)	3)	(5)	(2)	(2)	(2)	(2)	(3)	(3)	(3)	(5)	(5)	(50)
Review of Project Stage 1	04	04	04	03	04	02	02	02	02	03	03	02	04	04	43

Comments: Working on automation. Low Response Time.

Pallavi Saindane

Name & Signature Reviewer1

	Engineering Concepts & Knowledge	Interpretation of Problem & Analysis	Design / Prototype	Interpretation of Data & Dataset	Modern Tool Usage	Societal Benefit Safety Consideration	Environ ment Friendly	Ethics	Team work	Presentati on Skills	Applied Engg &Mgmt principles	Life - long learning	Profess ional Skills	Innov ative Appr oach	Total Marks
	(5)	(5)	(5)	(3)	(5)	(2)	(2)	(2)	(2)	(3)	(3)	(3)	(5)	(5)	(50)
Review of Project Stage 1	3	3	3	3	3	02	2	2	2	3	3.	2	4	4	40

Comments: Fetching a video from youtube. so response time is affected. Video database could be maintained.

Date: 15th March,2018

Name & Signature Reviewer2

Sheetal Balsarog

Fig 24. Review 2 Sheet

List of Figures

List of Tables

APPENDIX

A. Paper I

VIDEO CLASSIFICATION USING TEXT ANALYTICS

Aarti Raghani
Dept. of Computer Engineering
V.E.S.I.T
Chembur, Mumbai, India

Aditi Sawant
Dept. of Computer Engineering
V.E.S.I.T
Chembur, Mumbai, India

Kajol Chawla
Dept. of Computer Engineering
V.E.S.I.T
Chembur, Mumbai, India

Revati Pathak
Dept. of Computer Engineering
V.E.S.I.T
Chembur, Mumbai, India

Mrs. Lifna C.S.
Assistant Professor
Dept. of Computer Engineering
V.E.S.I.T
Chembur, Mumbai, India

Abstract—**With the advent of the Information Age, there has been an exponential growth in the number of videos available for any given topic. Hence, internet users are relentlessly in search for foolproof Video Classification Algorithms. The existing state-of-the-art techniques do not provide a satisfactory summary for ranking videos. The objective of the paper is to rank videos using Text Summarization techniques. This application will provide the users with the most relevant videos and their summaries, thereby saving time and increasing efficiency.**

Keywords—Text Summarization; Video Ranking; Video Classification.

I. Introduction

In today's digital era, the internet is the reliable source for obtaining videos. There has been an enormous growth in videos available for any given topic. A query for any topic yields a heap of results and, it is not easy to manually discard irrelevant results. Hence, it is essential to automatically summarize the videos to get a gist of their content.

The available video analytics systems analyze videos by using Image Processing techniques. The efficiency of these Image-based Summarization techniques heavily depends on the quality of the videos. Hence, these techniques have the following drawbacks: (1)Blurred Images; (2) Poor Illumination Effect; (3) Need for large bandwidth; (4) Storage requirements; (5) Output is also in the form of video. These constraints degrade the quality of the summaries generated by the systems. This led us to consider Text-based Summarization techniques for summarizing the videos and evaluating the efficiency.

In short, our proposal can be summarized as follows: (1) Search for topmost videos using input keyword; (2) Extract audio from the videos; (3) Convert the audio files to text; (4) Generate document summaries; and (5) Re-rank the videos based on the summaries.

II. Literature Survey

Towards finalizing Text-based summarization techniques, a thorough survey was performed. As a part of the survey, the following conclusions were extracted from the respective papers.

Paper [1] discusses two kinds of Text Summarization approaches; (1) Extractive Summarization - extract the most relevant sentences from the given text document and include them in the summary. (2) Abstractive Summarization - original sentence from the given document is replaced with a sentence with similar meaning. While abstractive summaries are closer to the summaries generated by a domain expert, extractive summarization techniques are faster and more efficient for summarizing documents with multiple sentences. Since an exhaustive study is required to validate summaries of large documents, extractive summarization approach was selected and further studied.

In paper [2], an Extractive Text Summary is generated from the list of top-ranked sentences by using Sentence Scoring Method. For Sentence scoring, the following factors are considered: (1) Word frequency; (2) Sentence Position; (3) Cue words; (4) Title Similarity; (5) Sentence length; (6) Proper noun; and (7) Sentence reduction. After each sentence is scored, the sentences are arranged in the descending order of their score value. From the top-ranked sentences, a number of sentences are chosen to be included in the summary.

In paper [3], lexical chains are used as source representation for summarization. For this, WordNet is used to find relatedness among words. Words of the same category are linked through semantic relations like synonymy and hyponymy. The algorithm differs from other approaches as it creates chains for words using relatedness criteria and merges different segments using strong criterion, i.e. chains will be merged if they contain a common word with the same sense. The problem of similarity between sentences can be solved by using the chain representation approach. The final summary can be extracted based on different heuristics functions i.e. Heuristic 1- For each chain in the summary representation, choose the sentence that contains the first appearance of a chain member in the text. Heuristic 2-The sentence containing the first appearance of a representative chain member is chosen. Heuristic 3- A frequently discussed topic may have its chain spread all over the document.

In this paper [4], an indexing structure based on the context of the document is proposed. The WordNet database helps in identifying similarities between different sentences in the document. Later, TextRank Algorithm is used to find a context-sensitive indexing weight of each term. On generating a graph of a document, each edge gives the lexical association between the terms corresponding to the vertices.The similarity between words is found using context-based indexing weights. Then this similarity between words is used to find similarity between sentences. Thus, for each sentence in the document, the sentence vector is built. Later sentences are extracted based on higher sentence score. So, this approach can be used to retrieve results within a short span.

This is also a graph-based unsupervised algorithm for extractive summarization [5]. It uses TF-IDF (Term Frequency- Inverse Document Frequency) to calculate how important a word is in multiple documents. A vector is defined for every sentence where the entry in the vector for every word is the value of its frequency in the sentence multiplied by its idf. The dimensionality of this vector is equal to the number of all words in the targeted language. An idf-modified-cosine function is defined to compute the similarity between two sentences. A cosine similarity matrix is formed with entries for the similarity calculated above. A threshold value is chosen to choose only the significantly similar sentences and discard the rest.

In paper [6], a recursive TF-ISF based method that takes into account the local context of a sentence is proposed. The context is defined as the previous and next sentence of the current sentence. On comparing this method to the TF-ISF baseline, statistically significant improvements in the results were found.

In this paper [7], TF-IDF method is used to generate a summary. Here, features of the document are extracted by obtaining the scores for the sentences in the text document based on their importance with a value between zero and one. Thirdly, it includes sentence selection and assembly, where the sentences are stored in descending order of the rank, and the highest ranked sentences are considered for the summary.

TF-IDF finds the importance of a word in the document and to control this value, the frequency of the word in the document is considered. The frequency term is the number of occurrences of a term in a document. Inverse Document Frequency is calculated by dividing the total number of documents by the number of documents in which the term occurs.

TABLE I. SUMMARY OF LITERATURE SURVEY

	Pros	**Cons**
[2]	Sentence Scoring method is used so accuracy is much higher than when considering only Tf-Idf or sentence ranking. Any number of extensions for scoring techniques can easily be added.	Query based summarization is not supported. It involves summarization with title which involves that if title word is present in sentence, then that sentence will get higher priority rather than considering the relevance of title with sentence.
[3]	The problem of similarity between sentences produced in the summary is overcome. Also, the sense of the word based on its position in the sentence is considered which is done using lexical chains.	Large sentences are more likely to be a part of summary.This approach does not concern the length and detail of the summary produced, which may act as noise.
[4]	It builds the index by considering context of the document. This is different compared to earlier methods that consider the terms for building the index. This approach is found to give better results than previously used approaches.	It fails to consider related terms i.e verbs, adjectives of words. It uses a linear function to calculate similarity between two sentences. It is found that a cosine function produces better results.
[5]	Sentences subsuming the information of other sentences get higher scores than individual sentences; thus a compact summary is formed. Prevents unnatural boosting of sentence score by an irrelevant topic.	Basic LexRank uses threshold to establish links between the sentences. Thus, improper threshold values may result into information loss or inclusion of irrelevant details in the summary making it bigger.
[6]	Summary is generated by considering relevance between previous and next sentence. It is achieved by defining recursive ranking function.	It does not consider factors for sentence scoring. Recursive TF-ISF algorithm is purely query based approach. This algorithm not considers factors such as Combination of word frequency in document, Sentence positional value ,Sentence length and Proper noun etc
[7]	The result of this research produces 67% accuracy with three data samples which are higher compared to the other online summarizers.	Relevance with the title factor is not considered while generating summaries. There is a need for involving more respondents to evaluate the system by determining the number of correct, wrong, or missed sentences within the summary.

III. PROPOSED SYSTEM

The proposed system has the following modules.(1) Search Module; (2) Audio Extraction Module; (3) Speech-to-Text Conversion Module; (4) Text Summarization Module and (5) Re-ranking Module.

A. Search Module

The query input by user is searched for in the database. If found, relevant video links and generated summaries are displayed to the user. Else, the query is redirected to YouTube.

B. Audio Extraction Module

The video for which summary is to be generated is the input and the audio is extracted from it.

C. Speech-to-Text Conversion Module

The audio extracted in the preceding step is converted to textual format. This text document forms the input to the Summarization Module.

D. Text Summarization Module

The text document obtained is summarized using extractive summarization technique. The factors that we consider for sentence scoring are: (1) Position of sentence; (2) Length of Sentence; (3) Similarity with the Title; (4) Proper Nouns; (5) Lexical Chains; (6) Term Frequency-Inverse Sentence Frequency. These scores are combined and the best scored sentences are ranked better.

E. Re-ranking Module

After summaries are generated for the videos, they are re-ranked on the basis of their summaries' relevance to the query.

The diagram shown in Fig. 1.describes the proposed system. The user inputs the search query over the internet, query is searched for in the database and if the corresponding query is accessible in the database, the results, that is, the relevant re-ranked videos, along with the videos links and the summary of those videos is displayed to user. But if query is not available, it is redirected to YouTube and the listed videos are pre-processed by initially converting them to text document, summarizing their content and finally re-ranking them based on the video content and the input query.

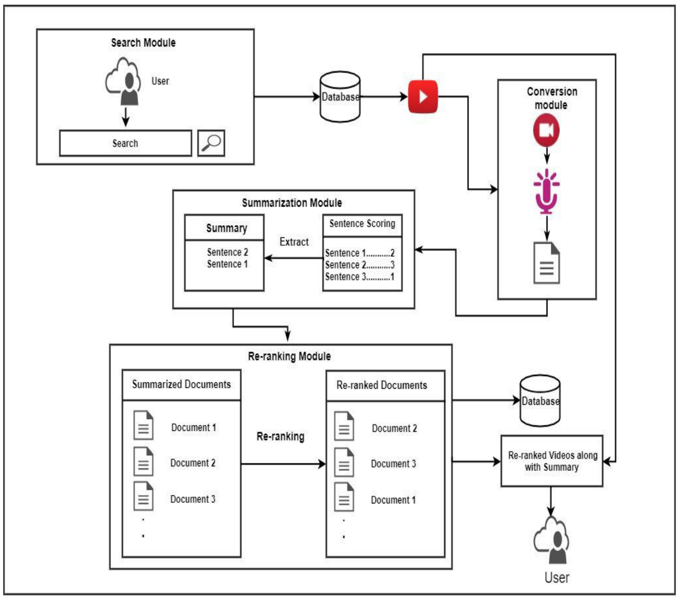


Fig. 1. Proposed System Diagram

IV. Conclusions

Our system includes scoring sentences in the document based on the sentence position, similarity with the title, sentence length, proper nouns, etc. The approach proposed in a paper [1] doesn’t take into account the similarity between different sentences in a document. This drawback can be overcome, by using lexical chains approach, which gives scores to sentence based on chains form- extra strong chains, strong chains, medium strong chains.

Instead of considering the frequency of words and cue words, proposed system consider the local context of a sentence using recursive TF-ISF. This method improves upon the approach proposed in paper [1] and gives a more efficient summary.

The system can be tuned further as an application for Cyber Security Cell to address Terrorism by incorporating multilingual module

Acknowledgment

We are sincerely thankful to Mrs. Sujata Khedkar, Associate Professor, Dept. of Computer Engineering, V.E.S.I.T. for her valuable insight and guidance in developing this approach.

References

[1] Shimpikar, Sheetal, and Sharvari Govilkar. "A Survey of Text Summarization Techniques for Indian Regional Languages." *International Journal of Computer Applications* 165.11 (2017).

[2] Raju, T. Sri Rama, and Bhargav Allarpu. "Text Summarization using Sentence Scoring Method." (2017).

[3] Barzilay, Regina, and Michael Elhadad. "Using lexical chains for text summarization." *Advances in automatic text summarization* (1999): 111-121.

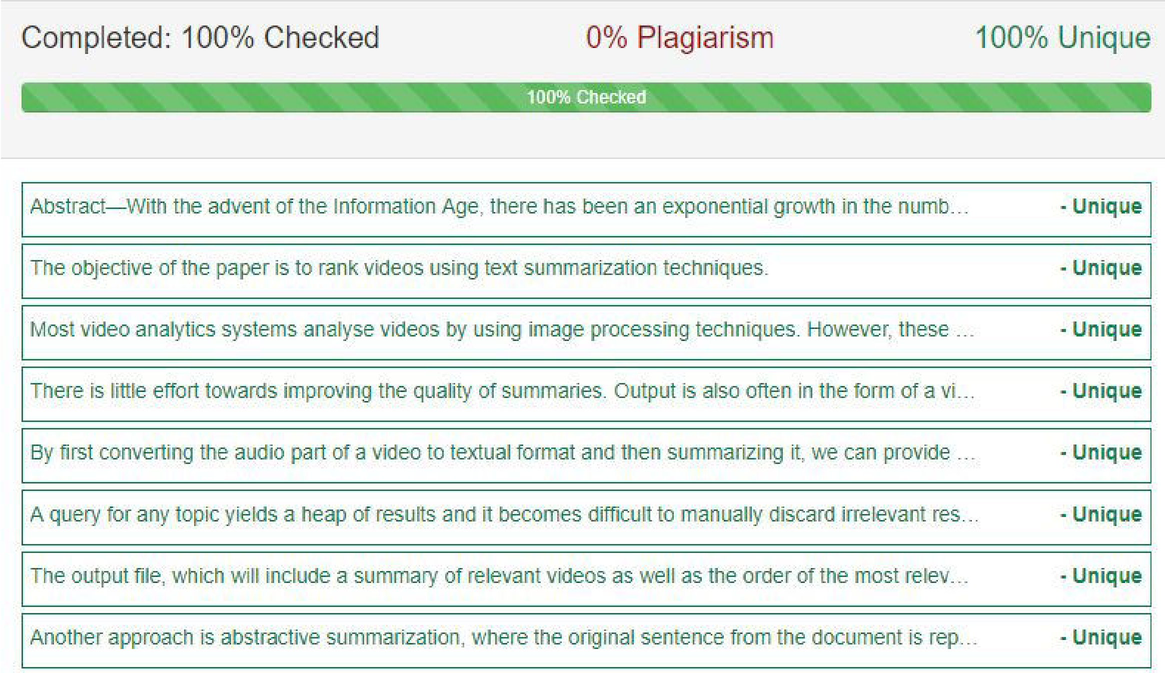
[4] Pawar, Dipti D., M. S. Bewoor, and S. H. Patil. "Text Rank: A novel concept for extraction based text summarization." *International Journal of Computer Science and Information Technologies* 5.3 (2014): 3301-3304.

[5] Erkan, Günes, and Dragomir R. Radev. "Lexrank: Graph-based lexical centrality as salience in text summarization." *Journal of Artificial Intelligence Research* 22 (2004): 457-479.

[6] Doko, Alen, Maja Stula, and Darko Stipanicev. "A recursive TF-ISF Based Sentence Retrieval Method with Local Context." *International Journal of Machine Learning and Computing* 3.2 (2013): 195.

[7] Christian, Hans, Mikhael Pramodana Agus, and Derwin Suhartono. "Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF)." *ComTech: Computer*

B.Plagiarism Report of Paper I



Completed: 100% Checked 0% Plagiarism 100% Unique

100% Checked

Abstract—With the advent of the Information Age, there has been an exponential growth in the numb... - Unique

The objective of the paper is to rank videos using text summarization techniques. - Unique

Most video analytics systems analyse videos by using image processing techniques. However, these ... - Unique

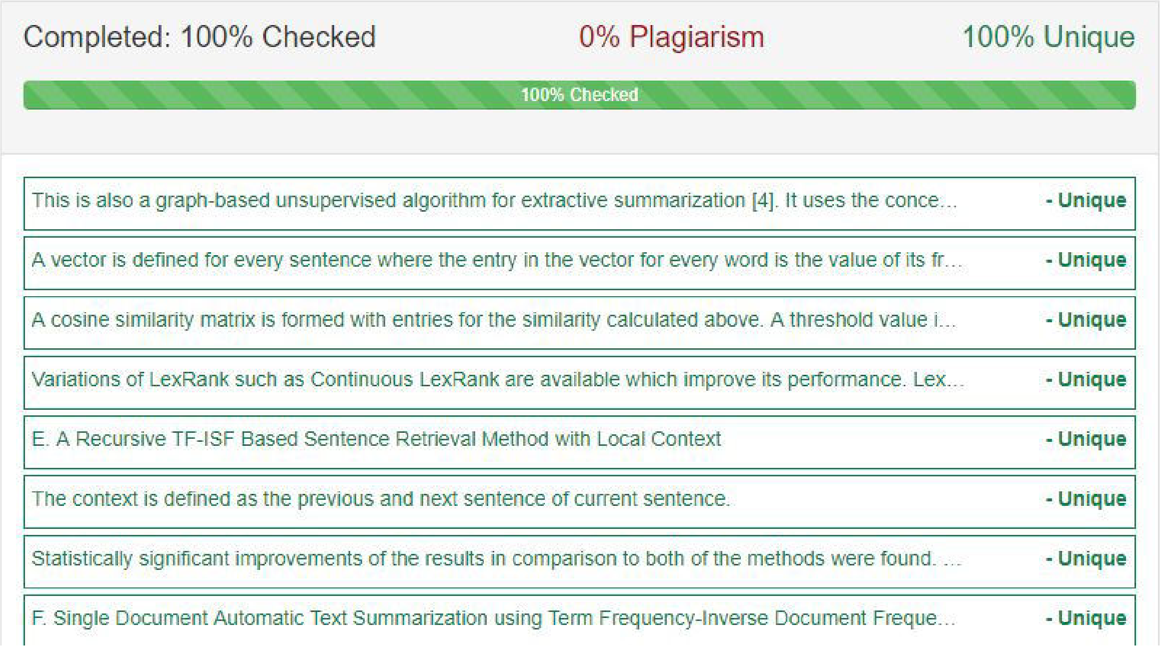
There is little effort towards improving the quality of summaries. Output is also often in the form of a vi... - Unique

By first converting the audio part of a video to textual format and then summarizing it, we can provide ... - Unique

A query for any topic yields a heap of results and it becomes difficult to manually discard irrelevant res... - Unique

The output file, which will include a summary of relevant videos as well as the order of the most relev... - Unique

Another approach is abstractive summarization, where the original sentence from the document is rep... - Unique



Completed: 100% Checked 0% Plagiarism 100% Unique

100% Checked

This is also a graph-based unsupervised algorithm for extractive summarization [4]. It uses the conce... - Unique

A vector is defined for every sentence where the entry in the vector for every word is the value of its fr... - Unique

A cosine similarity matrix is formed with entries for the similarity calculated above. A threshold value i... - Unique

Variations of LexRank such as Continuous LexRank are available which improve its performance. Lex... - Unique

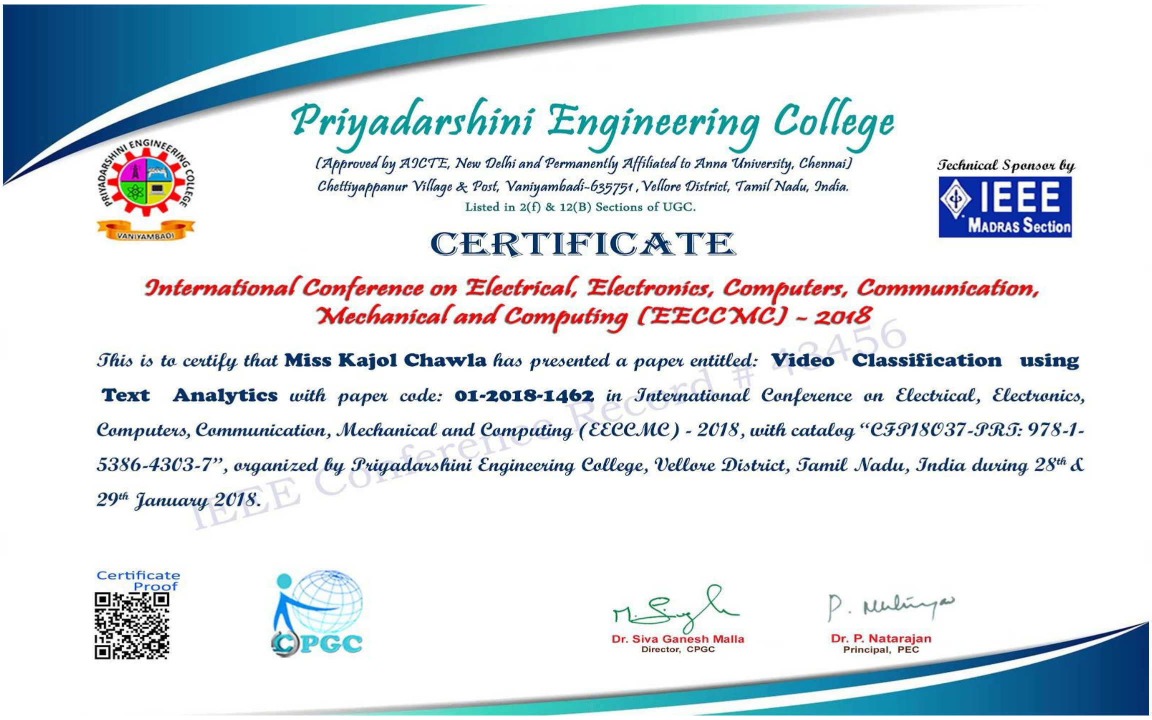
E. A Recursive TF-ISF Based Sentence Retrieval Method with Local Context - Unique

The context is defined as the previous and next sentence of current sentence. - Unique

Statistically significant improvements of the results in comparison to both of the methods were found. ... - Unique

F. Single Document Automatic Text Summarization using Term Frequency-Inverse Document Freque... - Unique

C.Certification of publication and presentation for Paper I:



Priyadarshini Engineering College

(Approved by AICTE, New Delhi and Permanently Affiliated to Anna University, Chennai)
Chettiyappanur Village & Post, Vaniyambadi-635751, Vellore District, Tamil Nadu, India.
Listed in 2(f) & 12(B) Sections of UGC.

Technical Sponsor by IEEE MADRAS Section

CERTIFICATE

International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing (EECCMC) - 2018

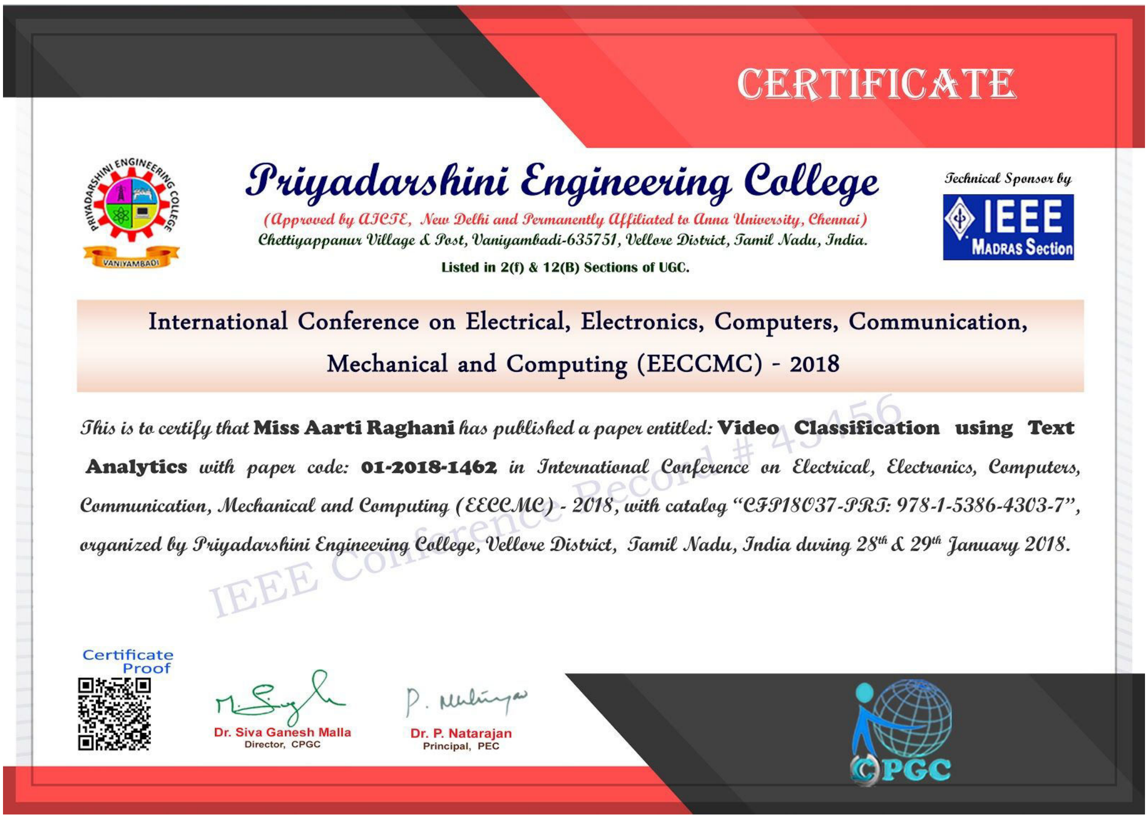
This is to certify that **Miss Kajol Chawla** has presented a paper entitled: **Video Classification using Text Analytics** with paper code: **01-2018-1462** in International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing (EECCMC) - 2018, with catalog "CFP18037-PRT: 978-1-5386-4303-7", organized by Priyadarshini Engineering College, Vellore District, Tamil Nadu, India during 28th & 29th January 2018.

Certificate Proof

CPGC

Dr. Siva Ganesh Malla
Director, CPGC

Dr. P. Natarajan
Principal, PEC



CERTIFICATE

Priyadarshini Engineering College

(Approved by AICTE, New Delhi and Permanently Affiliated to Anna University, Chennai)
Chettiyappanur Village & Post, Vaniyambadi-635751, Vellore District, Tamil Nadu, India.
Listed in 2(f) & 12(B) Sections of UGC.

Technical Sponsor by IEEE MADRAS Section

International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing (EECCMC) - 2018

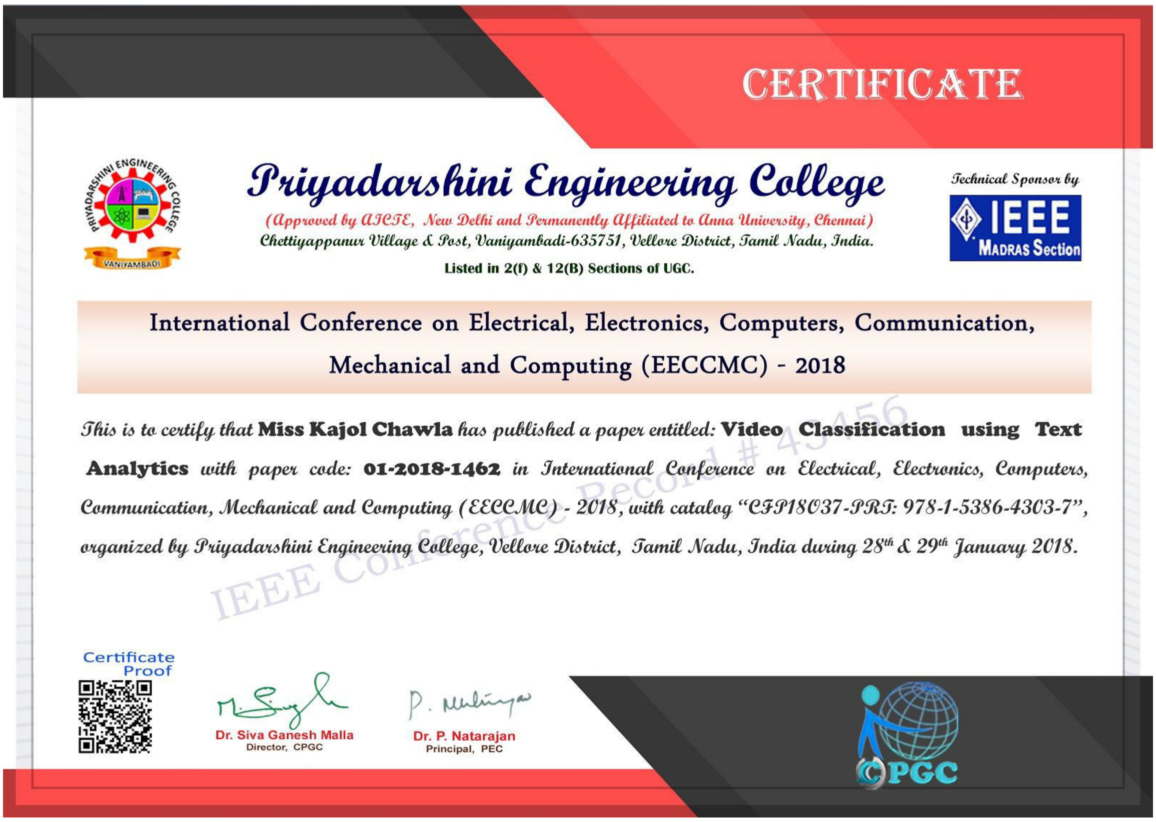
This is to certify that **Miss Aarti Raghani** has published a paper entitled: **Video Classification using Text Analytics** with paper code: **01-2018-1462** in International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing (EECCMC) - 2018, with catalog "CFP18037-PRT: 978-1-5386-4303-7", organized by Priyadarshini Engineering College, Vellore District, Tamil Nadu, India during 28th & 29th January 2018.

Certificate Proof

Dr. Siva Ganesh Malla
Director, CPGC

Dr. P. Natarajan
Principal, PEC

CPGC



CERTIFICATE

Priyadarshini Engineering College

(Approved by AICTE, New Delhi and Permanently Affiliated to Anna University, Chennai)
Chettiyappanur Village & Post, Vaniyambadi-635751, Vellore District, Tamil Nadu, India.
Listed in 2(f) & 12(B) Sections of UGC.

Technical Sponsor by IEEE MADRAS Section

International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing (EECCMC) - 2018

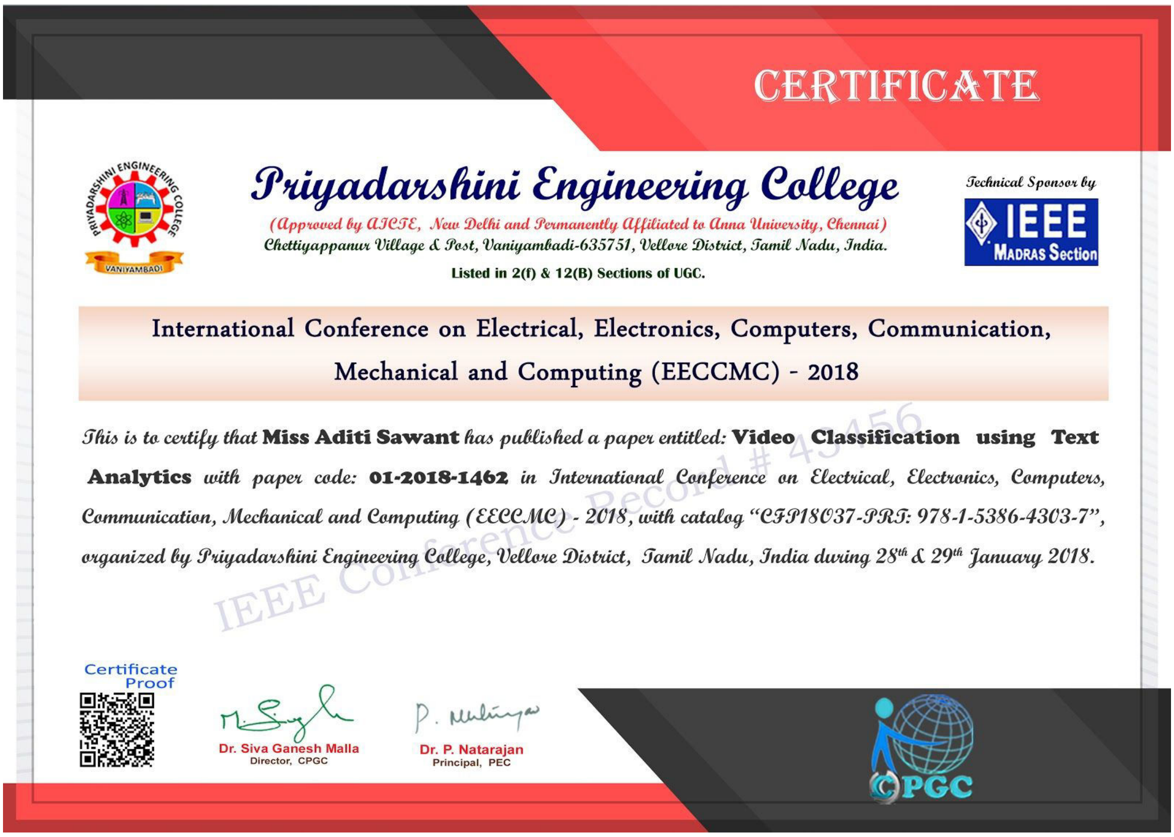
This is to certify that **Miss Kajol Chawla** *has published a paper entitled:* **Video Classification using Text Analytics** *with paper code:* **01-2018-1462** *in International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing (EECCMC) - 2018, with catalog "CFP18037-PRT: 978-1-5386-4303-7", organized by Priyadarshini Engineering College, Vellore District, Tamil Nadu, India during 28th & 29th January 2018.*

Certificate Proof

Dr. Siva Ganesh Malla
Director, CPGC

Dr. P. Natarajan
Principal, PEC

CPGC



CERTIFICATE

Priyadarshini Engineering College

(Approved by AICTE, New Delhi and Permanently Affiliated to Anna University, Chennai)
Chettiyappanur Village & Post, Vaniyambadi-635751, Vellore District, Tamil Nadu, India.
Listed in 2(f) & 12(B) Sections of UGC.

Technical Sponsor by IEEE MADRAS Section

International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing (EECCMC) - 2018

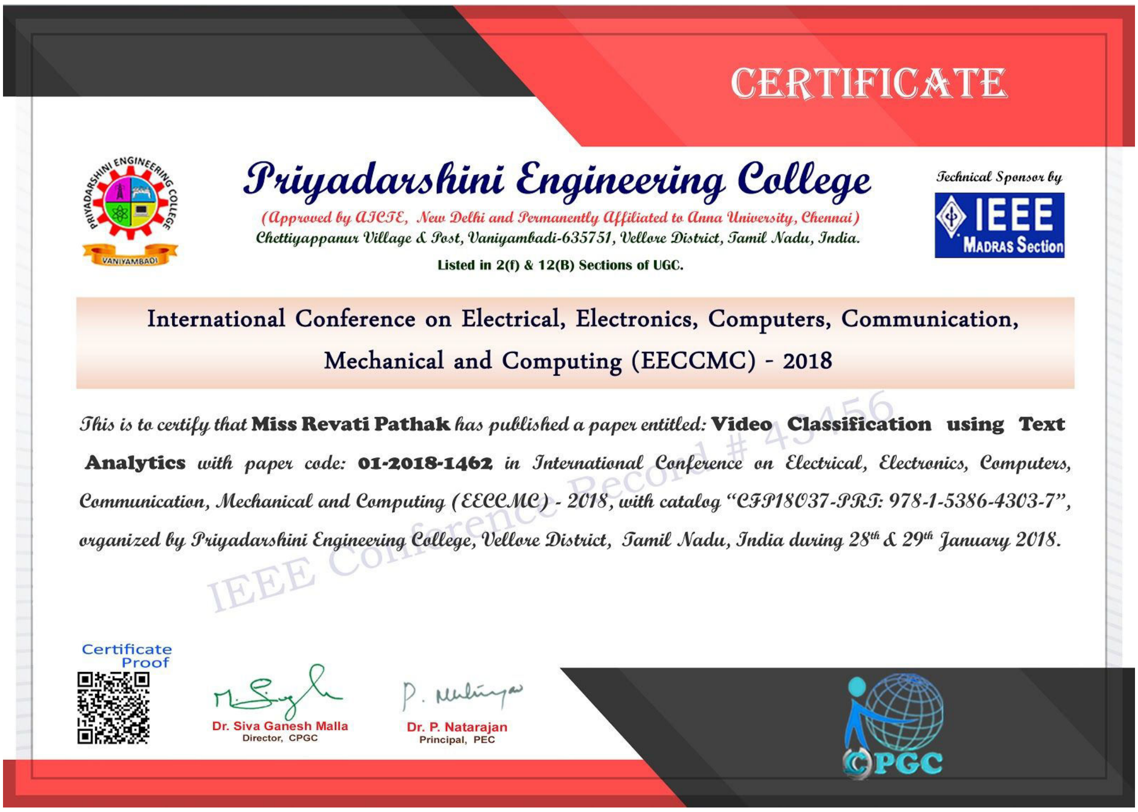
This is to certify that **Miss Aditi Sawant** *has published a paper entitled:* **Video Classification using Text Analytics** *with paper code:* **01-2018-1462** *in International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing (EECCMC) - 2018, with catalog "CFP18037-PRT: 978-1-5386-4303-7", organized by Priyadarshini Engineering College, Vellore District, Tamil Nadu, India during 28th & 29th January 2018.*

Certificate Proof

Dr. Siva Ganesh Malla
Director, CPGC

Dr. P. Natarajan
Principal, PEC

CPGC



CERTIFICATE

Priyadarshini Engineering College

(Approved by AICTE, New Delhi and Permanently Affiliated to Anna University, Chennai)
Chettiyappanur Village & Post, Vaniyambadi-635751, Vellore District, Tamil Nadu, India.
Listed in 2(f) & 12(B) Sections of UGC.

Technical Sponsor by
IEEE MADRAS Section

International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing (EECCMC) - 2018

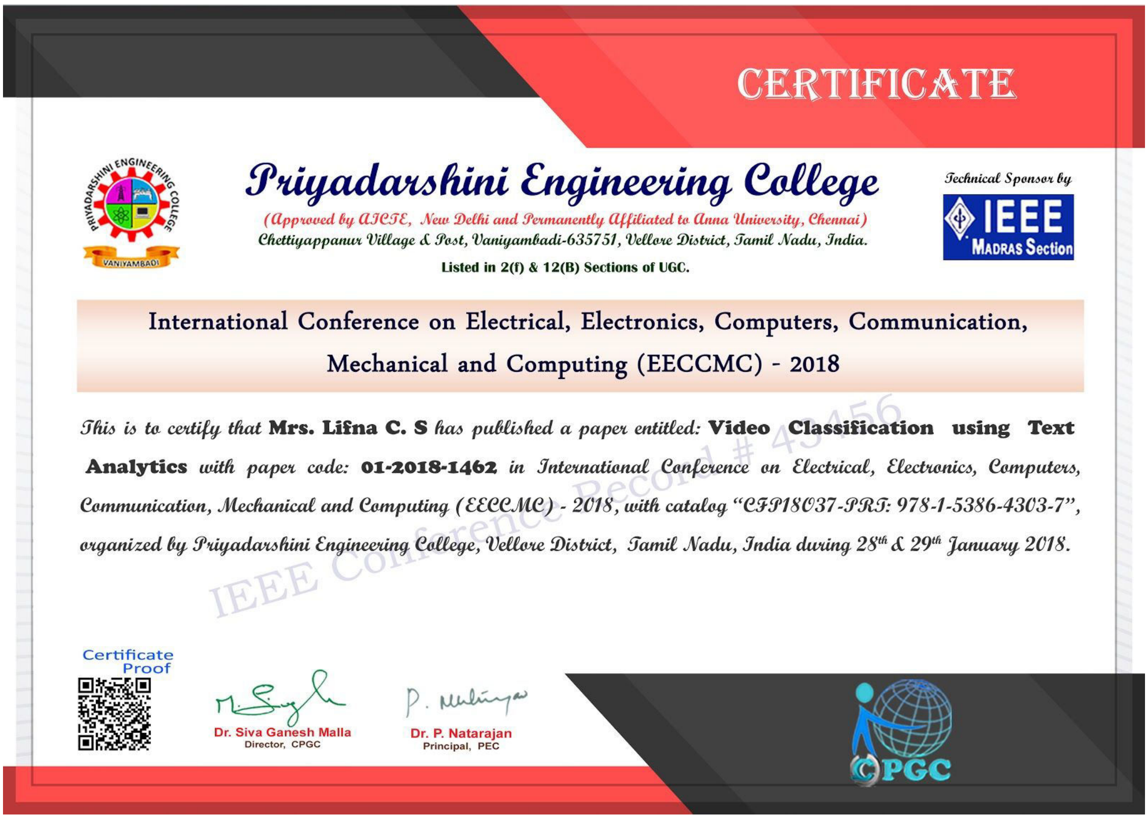
This is to certify that **Miss Revati Pathak** *has published a paper entitled:* **Video Classification using Text Analytics** *with paper code:* **01-2018-1462** *in International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing (EECCMC) - 2018, with catalog "CFP18037-PRT: 978-1-5386-4303-7", organized by Priyadarshini Engineering College, Vellore District, Tamil Nadu, India during 28th & 29th January 2018.*

Certificate Proof

Dr. Siva Ganesh Malla
Director, CPGC

P. Natarajan
Dr. P. Natarajan
Principal, PEC

CPGC



CERTIFICATE

Priyadarshini Engineering College

(Approved by AICTE, New Delhi and Permanently Affiliated to Anna University, Chennai)
Chettiyappanur Village & Post, Vaniyambadi-635751, Vellore District, Tamil Nadu, India.
Listed in 2(f) & 12(B) Sections of UGC.

Technical Sponsor by
IEEE MADRAS Section

International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing (EECCMC) - 2018

This is to certify that **Mrs. Lifna C. S** *has published a paper entitled:* **Video Classification using Text Analytics** *with paper code:* **01-2018-1462** *in International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing (EECCMC) - 2018, with catalog "CFP18037-PRT: 978-1-5386-4303-7", organized by Priyadarshini Engineering College, Vellore District, Tamil Nadu, India during 28th & 29th January 2018.*

Certificate Proof

Dr. Siva Ganesh Malla
Director, CPGC

P. Natarajan
Dr. P. Natarajan
Principal, PEC

CPGC

D. PAPER II

VIDEO SUMMARIZATION AND RE-RANKING

Aarti Raghani
Dept. of Computer Engineering
V.E.S.I.T.
Chembur, Mumbai, India

Aditi Sawant
Dept. of Computer Engineering
V.E.S.I.T.
Chembur, Mumbai, India

Kajol Chawla
Dept. of Computer Engineering
V.E.S.I.T.
Chembur, Mumbai, India

Revati Pathak
Dept. of Computer Engineering
V.E.S.I.T.
Chembur, Mumbai, India

Mrs. Lifna C.S.
Assistant Professor
Dept. of Computer Engineering
V.E.S.I.T.
Chembur, Mumbai, India

***Abstract*—With the advent of the Information Age, there has been an exponential growth in the number of videos available for any given topic. Hence, internet users are relentlessly in search for foolproof Video Classification Algorithms. The existing state-of-the-art techniques do not provide a satisfactory summary for ranking videos. The objective of the paper is to rank videos using Text Summarization techniques. This application will provide the users with the most relevant videos and their summaries, thereby saving time and increasing efficiency.**

Keywords—Text Summarization; Video Re-ranking; Natural Language Processing

I.INTRODUCTION

In today's digital era, the internet is the reliable source for obtaining videos. There has been an enormous growth in videos available for any given topic. A query for any topic yields a heap of results, and it is not easy to manually discard irrelevant results. Hence, it is essential to automatically summarize the videos to get a gist of their content.

The available video analytics systems analyze videos by using Image Processing techniques. The efficiency of these Image-based Summarization techniques heavily depends on the quality of the videos. Hence, these techniques have the following drawbacks: (1)Blurred Images; (2) Poor Illumination Effect; (3) Need for large bandwidth; (4) Storage requirements; (5) Output is also in the form of video.

These constraints degrade the quality of the summaries generated by the systems. This led us to consider Text-based Summarization techniques for summarizing the videos and evaluating the efficiency.

In short, our proposal can be summarized as follows: (1) Search for topmost videos using input keyword; (2) Extract audio from the videos; (3) Convert the audio files to text; (4) Generate document summaries; and (5) Re-rank the videos based on the summaries.

II. MOTIVATION

For finalizing our approach towards the re-ranking of videos based on their content, we conducted extensive research. First, we looked at individual algorithms and how they performed. Paper [1] discusses the TF-ISF approach, which is a modified form of TF-IDF for single documents.

Using this approach, we can identify sentences that have a greater number of important words and lesser number of unnecessary words.

Next, we considered the sentence scoring approach discussed in paper [2]. The sentence scoring approach works to eliminate the deficiencies in individual approaches and, at the same time, gives the benefit of different approached clubbed into one. Using this as our foundation, we have then carefully chosen the factors that provide significant contribution towards creating a concise and rich summary.

Paper [3] discusses the application of machine learning to improve the summary. Improvement in summary means that we endeavour to bring it as close to human- generated summaries as possible.

III.Implemented System

The implemented system has the following modules.(1) Search Module; (2) Audio Extraction Module; (3) Speech-to-Text Conversion Module; (4) Text Summarization Module and (5) Re-ranking Module.

A. Search Module

The query input by user is searched for in the database. If found, relevant video links and generated summaries are displayed to the user. Else, the query is redirected to YouTube.

B. Audio Extraction Module

The video for which summary is to be generated is the input
and the audio is extracted from it. Since we are downloading videos from YouTube, we make use of the API provided by YouTube for downloading the top videos that appear in the search results.
The -dl command can be used for the same.

C. Speech-to-Text Conversion Module

The audio extracted in the preceding step is converted to textual format. This text document forms the input to the Summarization Module. We assessed the accuracy of several speech-to-text APIs available, including Sphinx, Google cloud speech recognition, WIT.AI, IBM Watson and Houndify.

Of these, the Google cloud speech recognition service was able to provide the most accurate and coherent results. It converts videos of greater length, too Hence, we have chosen this API for our system. However, this API does not provide punctuation, which is necessary for extractive text summarization, where two sentences need to be distinguishable to extract a sentence. To overcome this lacuna, we have used the Punctuator API. The Punctuator uses machine learning on a large corpus and provides sufficiently accurate punctuation for any non-punctuated text document.

D. Text Summarization Module

The text document obtained is summarized using extractive summarization technique. The factors that we consider for sentence scoring are: (1) Position of sentence; (2) Length of Sentence; (3) Similarity with the Title; (4) Similarity with the Query; (5) Similarity between sentences; (6) Term Frequency-Inverse Sentence Frequency; (7) Unnecessary information. These scores are combined and the best scored sentences are ranked better.

(1)]Position of sentence: The sentences in the introductory part of a text document contain more information. Hence, the starting few sentences are given more weightage than the rest of them. The scores obtained are in the range from 0 to 1.

For sentences in the first half,
$Score_i$ = 0.9 - 1.6 * i / no_of_sentences
Else,
$Score_i$ = 0.1 + 1.6 * (min ([i - no_of_sentences / 2, no_of_sentences / 2])) / no_of_sentences

(2) Length of Sentence: The lengths for different sentences are calculated and the maximum length is identified. The score awarded to each sentence is a ratio of the length of the sentence to the length of the longest sentence.

It can be given by:
$Score_i$ = len_i / max_length

(3) Similarity with the Title: We want to give a higher score to those documents that have their content strongly related to their titles. This ensures that the document delivers what it promises. Hence, sentences with title words are scored higher than other sentences. The normalised scores range from 0 to 1.

(4) Similarity with the Query: Many of the current applications do not consider similarity with the query. This can lead to higher scores for documents that are coherent and relevant as a whole, but are strongly related to what the user is looking for. Keeping this requirement in mind, we award higher scores to the documents that contain keywords, that is, words appearing in the query input by the user. The scores are normalised.

(5) Similarity between sentences: Another important factor to consider is the cohesion between the sentences within a document. For this, we use the powerful WordNet database, which provides semantic relations(like synonymy, hyponymy, etc.) between words in the document. Sentences with higher cohesion are assumed to represent the same concept and hence given higher score. This is used to weed out irrelevant sentences.

(6) Term Frequency- Inverse Sentence Frequency: This is a modified form of TF-IDF, used for a single text document. The number of occurrences of a term in a single sentence are calculated and called as term frequency. Next, the inverse sentence frequency is calculated by dividing the term frequency by the number of sentences that the term appears in. This method gives higher scores to words occurring greater number of times in a single sentence and, at the same time, diminishes the score for the words occurring in many sentences. This is effective to prevent higher scoring to common words.

(7) Unnecessary information: Conjunctions, adverbs, etc. appear many times in a sentence. However, they do not contribute to the meaningful content in a sentence. Hence, the sentences containing higher number of such words are likely to provide little to no information. Since we are aiming for a compact summary, we must ensure that such sentences receive lower scores. This is done by first using a parts-of-speech tagger to identify word type. Next, sentences having unnecessary words are scored lower to prevent them from being included in the summary.

E. Re-ranking Module

After summaries are generated for the videos, they are re-ranked on the basis of their summaries' relevance to the Query. The total score for each document is calculated as a linear combination of sentence scores. This score is then normalised to account for different number of sentences in summaries. The normalised scores are compared and the summary documents are re-ranked on this basis, with highest scores being ranked first.

The diagram shown in Fig. 1. describes the proposed system. The user inputs the search query over the internet, query is searched for in the database and if the corresponding query is accessible in the database, the results, that is, the relevant re-ranked videos, along with the videos links and the summary of those videos is displayed to user. But if query is not available, it is redirected to YouTube and the listed videos are pre-processed by initially converting them to text document, summarizing their content and finally re-ranking them based on the video content and the input query.

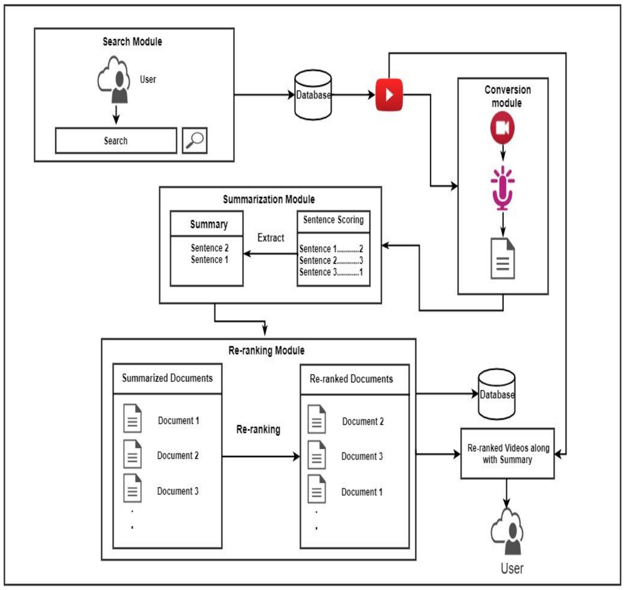


Fig. 1. Implemented System Diagram

IV.Algorithm Performance

We compared the performance of our implemented algorithm with that of other individual algorithms. The metrics used for performance comparison were Accuracy and ROUGE evaluation. Both the metrics showed significant superiority of the implemented algorithm over other individualistic approaches.Table I depicts the results obtained.

TABLE I. COMPARISON WITH OTHER ALGORITHMS

Algorithm	**TextRank**	**LexRank**	**TF-IDF**	**TF-ISF**	**Proposed System**
Accuracy	41%	44%	61%	40%	70%
Lacuna	Does not consider related words	Similarity with user query or title is not considered	Relevant for multiple documents	Does not consider other factors such as sentence scoring, position, length, etc.	Eliminates lacunae in individual algorithms by employing an integrated approach.
Evaluation	ROUGE 0.4229	ROUGE 0.4443	ROUGE 0.3101	ROUGE 0.39925	ROUGE 0.70

V.Results

We compared the summary generated by the implemented system with the summaries generated by some existing systems. We have used the Rouge metrics to compare the results.Table II describes the results obtained by performing Precision, Recall and F Score measures to the systems.

TABLE II. COMPARISON WITH OTHER SYSTEMS

	Analysis of summary obtained by different system					
Parameters	**Text Summarize r**	**Free Summarizer**	**Tools for noobs**	**Auto Summarizer**	**Text Comparator**	**Proposed System**
Precision	0.82	0.63	0.48	0.56	0.54	0.61
Recall	0.31	0.72	0.59	0.70	0.77	0.92
F Score	0.46	0.67	0.53	0.63	0.64	0.70

Summary generated by obtained Proposed System:

Input Text:

I just say Jenn the director of effortless English. I have something exciting to talk about today. I am now doing audio tweets. What is an audio tweet? What a strange word! Well, what I'm doing is I'm doing small little podcast on Twitter, and I am doing these almost everyday, because it's so easy. I use my phone my iPhone and I record a short talk about some Topic in my daily life, my normal life, and then I put it on Twitter twitter.com, and this is a very easy way for you to get more easy. English listening about daily topics about normal daily life, so I am continuing the podcast, of course, but the podcast is usually about learning ideas, teaching ideas. How can you speak English better? What is happening with effortless English, but the audio Twitter that I'm doing now? It'S about very simple day today: topics, for example. Just now I did a short audio, tweet and audio little podcast, just two and a half minutes about the weather in San Francisco. Today it is foggy in San Francisco use my phone. I recorded a short talk about the weather in San Francisco about the fog and then immediately instantly. I put it on my Twitter page. So if you want to hear new English listening topics everyday, please follow me on Twitter, so you will get longer more serious listening here at the podcast and you can get almost daily listening about easy, simple topics on my Twitter page. So how can you follow me on Twitter? It'S very very easy. Just go to twitter.com, it's Twitter. Com, A J Hoge that / a j, a h, o g e twitter.com forward, slash AJ Hoge! If you don't have a Twitter account, it's free! It'S easy! You can make a Twitter account and then follow me. There'S a button that says follow click the follow button. Then you will automatically get my short English listening topics about daily life and I will do a lot of these. It'S so easy. Since I can use my phone, my cell phone, when I'm walking around the city when I'm doing my normal life, sometimes I will stop and talk for one or two minutes about something happening in my life and put it on my Twitter page. So you can get a lot of new, easy, daily life English listening this way it's twitter.com flash AJ Hoge. Alright, I look forward to doing more of these audio Twitter's and audio tweets and also, of course, doing more podcast. I will see you on my Twitter page. I hope you enjoy these new short daily English. Listening topics see you next time, bye, bye,

Summary Generated by Implemented System:

I just say Jenn the director of effortless English.I have something exciting to talk about today. What a strange word!Well, what I'm doing is I'm doing small little podcast on Twitter, and I am doing these almost everyday, because it's so easy.I use my phone my iPhone and I record a short talk about some Topic in my daily life, my normal life, and then I put it on Twitter twitter.com, and this is a very easy way for you to get more easy.English listening about daily topics about normal daily life, so I am continuing the podcast, of course, but the podcast is usually about learning ideas, teaching ideas.How can you speak English better?Just now I did a short audio, tweet and audio little podcast, just two and a half minutes about the weather in San Francisco.So if you want to hear new English listening topics everyday, please follow me on Twitter, so you will get longer more serious listening here at the podcast and you can get almost daily listening about easy, simple topics on my Twitter page.Com, A J Hoge that / a j, a h, o g e twitter.com forward, slash AJ Hoge!There'S a button that says follow click the follow button.Since I can use my phone, my cell phone, when I'm walking around the city when I'm doing my normal life, sometimes I will stop and talk for one or two minutes about something happening in my life and put it on my Twitter page.So you can get a lot of new, easy, daily life English listening this way it's twitter.com flash AJ Hoge.

VI. Conclusion

As we can see from the results, the summary generated from the implemented methodology gives good accuracy and is closer to human-generated summaries. This is because the generated summary is a comprehensive indication of all the factors considered. The factors were chosen thoughtfully so as to eliminate the deficiencies created by using a single factor.This model can be extended to generate summaries for multilingual videos. This needs an efficient and accurate translation module.Such an application can prove to be quite useful in detecting terrorist attacks and opening up a world of information for people from different linguistic backgrounds.

VII.References

[1] Doko, Alen, Maja Stula, and Darko Stipanicev. "A recursive TF-ISF Based Sentence Retrieval Method with Local Context." *International Journal of Machine Learning and Computing* 3.2 (2013): 195.

[2] Raju, T. Sri Rama, and Bhargav Allarpu. "Text Summarization using Sentence Scoring Method." (2017).

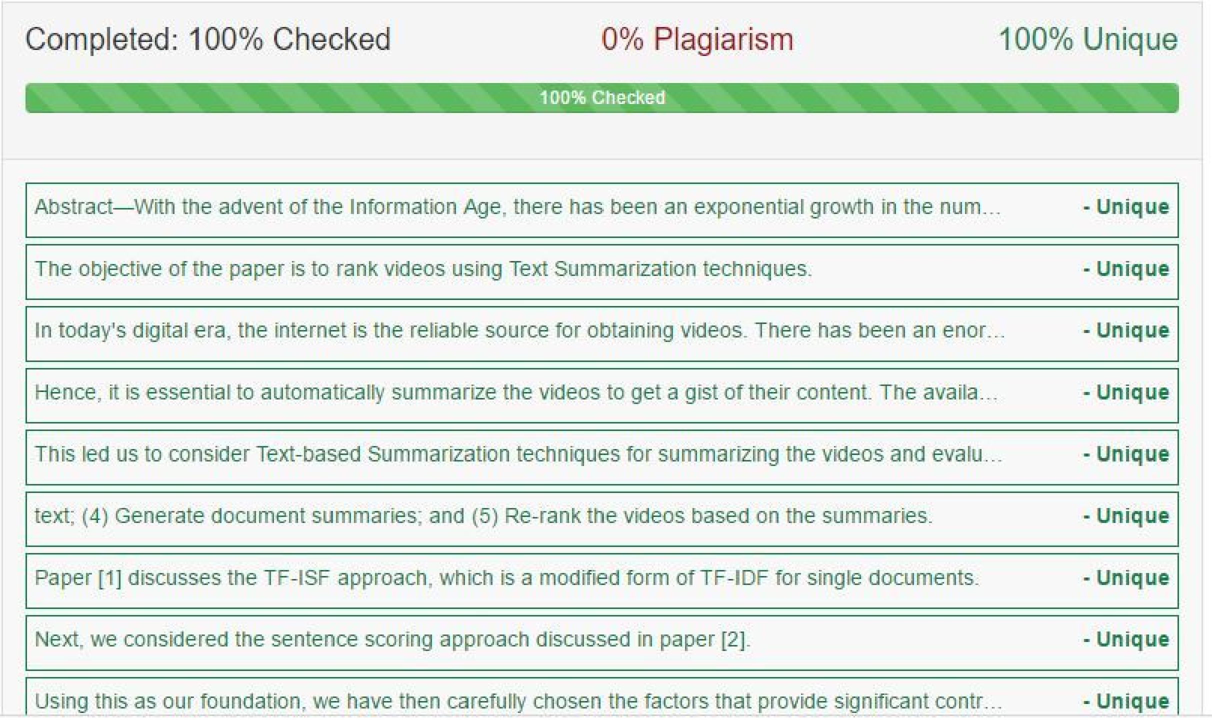
[3] Neto, Joel Larocca, Alex A. Freitas, and Celso AA Kaestner. "Automatic text summarization using a machine learning approach." *Brazilian Symposium on Artificial Intelligence*. Springer, Berlin, Heidelberg, 2002.

[4] Allahyari, Mehdi, et al. "Text summarization techniques: A brief survey." *arXiv preprint arXiv:1707.02268* (2017).

Acknowledgement

We are sincerely thankful to Mrs. Sujata Khedkar, Associate Professor, Dept. of Computer Engineering, V.E.S.I.T. for her valuable insight and guidance in developing this approach

E.Plagiarism Report of Paper II



Completed: 100% Checked 0% Plagiarism 100% Unique

100% Checked

Abstract—With the advent of the Information Age, there has been an exponential growth in the num... - Unique

The objective of the paper is to rank videos using Text Summarization techniques. - Unique

In today's digital era, the internet is the reliable source for obtaining videos. There has been an enor... - Unique

Hence, it is essential to automatically summarize the videos to get a gist of their content. The availa... - Unique

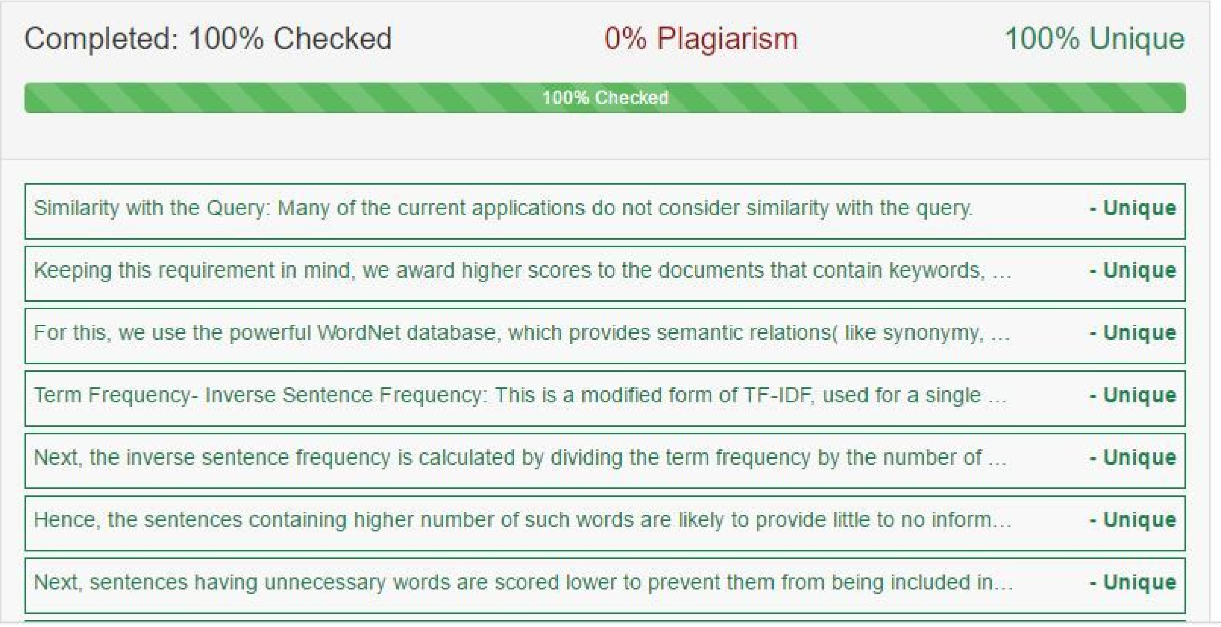
This led us to consider Text-based Summarization techniques for summarizing the videos and evalu... - Unique

text; (4) Generate document summaries; and (5) Re-rank the videos based on the summaries. - Unique

Paper [1] discusses the TF-ISF approach, which is a modified form of TF-IDF for single documents. - Unique

Next, we considered the sentence scoring approach discussed in paper [2]. - Unique

Using this as our foundation, we have then carefully chosen the factors that provide significant contr... - Unique



Completed: 100% Checked | 0% Plagiarism | 100% Unique

Similarity with the Query: Many of the current applications do not consider similarity with the query. - Unique

Keeping this requirement in mind, we award higher scores to the documents that contain keywords, ... - Unique

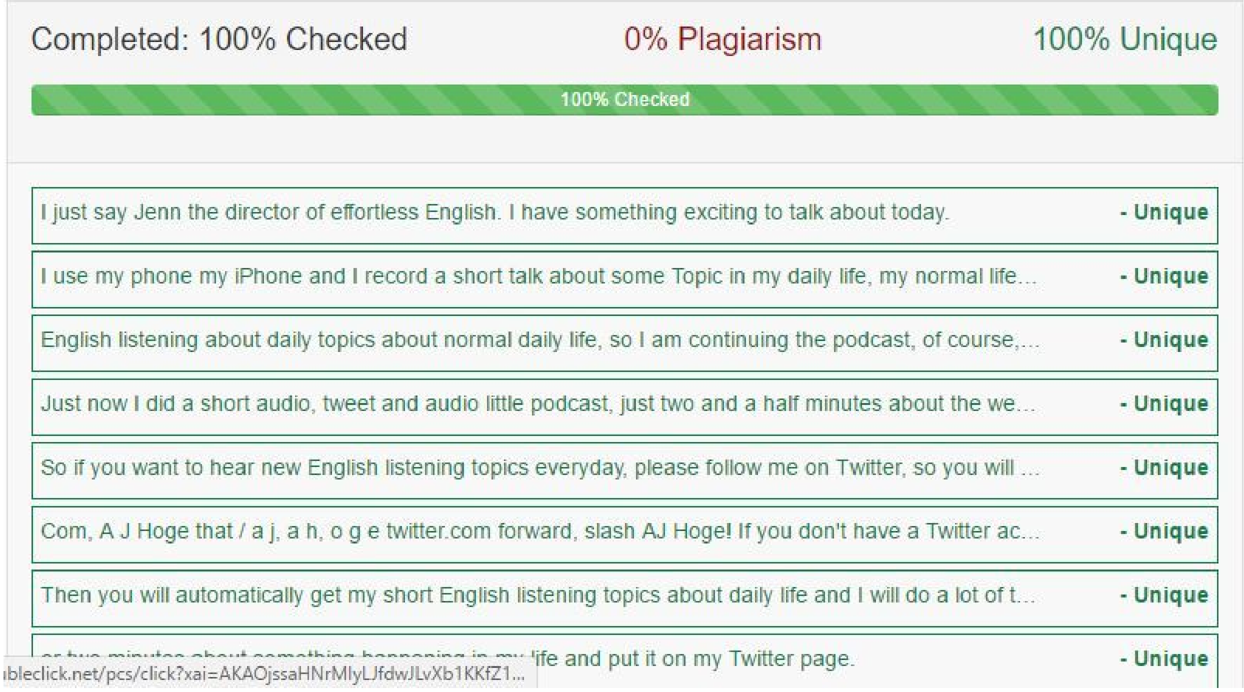
For this, we use the powerful WordNet database, which provides semantic relations(like synonymy, ... - Unique

Term Frequency- Inverse Sentence Frequency: This is a modified form of TF-IDF, used for a single ... - Unique

Next, the inverse sentence frequency is calculated by dividing the term frequency by the number of ... - Unique

Hence, the sentences containing higher number of such words are likely to provide little to no inform... - Unique

Next, sentences having unnecessary words are scored lower to prevent them from being included in... - Unique



Completed: 100% Checked | 0% Plagiarism | 100% Unique

I just say Jenn the director of effortless English. I have something exciting to talk about today. - Unique

I use my phone my iPhone and I record a short talk about some Topic in my daily life, my normal life... - Unique

English listening about daily topics about normal daily life, so I am continuing the podcast, of course,... - Unique

Just now I did a short audio, tweet and audio little podcast, just two and a half minutes about the we... - Unique

So if you want to hear new English listening topics everyday, please follow me on Twitter, so you will ... - Unique

Com, A J Hoge that / a j, a h, o g e twitter.com forward, slash AJ Hoge! If you don't have a Twitter ac... - Unique

Then you will automatically get my short English listening topics about daily life and I will do a lot of t... - Unique

or two minutes about something happening in my life and put it on my Twitter page. - Unique

ıbleclick.net/pcs/click?xai=AKAOjssaHNrMlyLJfdwJLvXb1KKfZ1...