RV Educational Institutions ®
RV College of Engineering ®

Autonomous Institution Affiliated to Visvesvaraya Technological University, Belagavi

Approved by AICTE, New Delhi, Accredited By NAAC, Bengaluru And NBA, New Delhi

Go, change the world

Major Project
16MCA61

on

Development Of An Efficient Data Extraction System For Rich Metadata In Video Streaming

Submitted by
MADHUSHREE M
USN: 1RV17MCA20

Under the Guidance
of

Internal Guide
Dr. Jasmine K S
Associate Professor
Department of MCA
RV College of Engineering®
Bengaluru – 560059

External Guide
Kirana Kumar
Senior Architect
Cognizant
Bengaluru – 560045

Submitted in partial fulfillment of the requirements for the award of degree of

MASTER OF COMPUTER APPLICATIONS

2019-2020

RV COLLEGE OF ENGINEERING®

(Autonomous Institution Affiliated to Visvesvaraya Technological University, Belagavi)

DEPARTMENT OF
MASTER OF COMPUTER APPLICATIONS

Bengaluru– 560059

CERTIFICATE

Certified that the project work titled **Development Of An Efficient Data Extraction System For Rich Metadata In Video Streaming** carried out by **Madhushree M, USN:1RV17MCA20**, a bonafide student of **RV College of Engineering®, Bengaluru** submitted in partial fulfilment for the award of **Master of Computer Applications** of **RV College of Engineering®, Bengaluru affiliated to Visvesvaraya Technological University, Belagavi** during the year **2019-20**.It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report deposited in the departmental library. The project report has been approved as it satisfies the academic requirement in respect of project work prescribed for the said degree.

Dr. Jasmine K S
Associate Professor
Department of MCA
RVCE, Bengaluru –59

Dr Andhe Dharani
Professor and Director
Department of MCA
RVCE, Bengaluru–59

Dr. K. N.Subramanya
Principal
RVCE, Bengaluru–59

RV COLLEGE OF ENGINEERING®

(Autonomous Institution Affiliated to Visvesvaraya Technological University, Belagavi)

DEPARTMENT OF
MASTER OF COMPUTER APPLICATIONS

Bengaluru– 560059

DECLARATION

I, **Madhushree M**, student of sixth semester MCA in **Department of Master of Computer Applications,** RV College of Engineering®, Bengaluru declare that the project titled **"Development of An Efficient Data Extraction System for Rich Metadata in Video Streaming"** has been carried out by me. It has been submitted in partial fulfilment of the course requirements for the award of degree in **Master of Computer Applications** of RV College of Engineering®, Bengaluru affiliated to Visvesvaraya Technological University, Belagavi during the academic year **2019-20**. The matter embodied in this report has not been submitted to any other university or institution for the award of any other degree or diploma.

Madhushree . M .

Date of Submission: 10-06-2020 **Signature of the Student**

Madhushree M
USN: 1RV17MCA20
Department of Master of Computer Applications
RV College of Engineering®
Bengaluru-560059

Dear Madhushree M,

Greetings from Cognizant !

Congratulations on completing your internship at Cognizant Technology Solutions in the period between **18th Jan 2020 and 15th May 2020**.

We appreciate the passion and professionalism you've exhibited during the internship. We take this opportunity to wish you the best in all your future endeavors.

Regards,
Cognizant

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the success of any work would be incomplete unless we mention the name of the people, who made it possible, whose constant guidance and encouragement served a beacon light and served our effort with success.

I express my wholehearted gratitude to **Dr.Subramanya K N**, Principal, RV College of Engineering® for providing me an opportunity to do the project.

I express my special thanks to **Dr. Andhe Dharani**, Professor and Director, Department of MCA, RV College of Engineering®, Bengaluru for her constant support and guidance.

I express my sincere thanks and wholehearted credit to my internal guide **Dr. Jasmine K S**, Associate Professor, Department of MCA, RV College of Engineering®, Bengaluru for her constant support, encouragement and guidance during the project work.

I am also thankful to all faculty of the department for their help and support during the project work and to **Mr. Kirana Kumar**, Senior Architect, Cognizant for his support and help during project work.

On a moral personal note, my deepest appreciation and gratitude to my beloved family, who have been a fountain of inspiration and have provided unrelenting encouragement and support.

Madhushree M
Master of Computer Applications
Department of MCA
RV College of Engineering®,
Bengaluru-59

ABSTRACT

Rich metadata is description of the media files that go beyond the basic attributes of the media file. Rich metadata associated with videos play a major role in recommending the videos to users. The metadata available with the videos are used by machine learning algorithms to recommend or formulate playlist for individual users. These metadata are also important for availability of the videos to the users. Dailymotion is a France based video sharing platform but lacks the Rich metadata required by the algorithms to enhance the searching and also availability of the videos. The project generates a consolidated JSON object of metadata for a given video using the video's content and also the already available metadata.

The functionalities of the project is, first extracting the already available metadata for the videos in Dailymotion using web scraping technique by passing the URL of the video. To overcome the problem of lack of metadata, transcription of audio is used to determine the content of the video. For the extracted transcription of the audio, natural language processing libraries is used to clean the text. The cleaned text is further used to identify various tags for the video such as language and profanity. The project is developed in python3 with Anaconda4 framework.. Initially, web scraping is performed with beautifulsoup library. For transcribing audio speech recognition library is used. The transcribed text is cleaned with nltk library to remove stop words and perform frequency distribution. sklearn and Keras libraries is used to build supervised machine learning models that identifies language and profanity in transcribed text.

The outcome of the project is to enhance the available metadata for each video in Dailymotion. The application provides the extracted and generated Rich metadata of a given video as JSON objects to any recommendation engine. The project helps Dailymotion to automate profiling their largely available videos.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Introduction

Rich metadata are attributes that go beyond basic attributes of a media file such as bit rate, duration. These attributes are derived from the content of the media files.

1.1 Project Description

With the increased availability of internet there is a sudden raise in content sharing platforms and so is the need to profile the content. One of them is video sharing platforms. In last few years there is an increase in generation and sharing of videos, hence there is a need of rich metadata that describes them. Most of the video sharing platform takes the required details from the creators. But few video sharing platforms have started this facility only recently so most of the old videos in their repository do not have the proper metadata that describe them.

Metadata are important for search engines to properly identify the content requested by the user. If a media file does not have the required metadata there is chance of the video never being available. With the use of machine learning algorithms by most search engines, based on the users lastly viewed video's metadata the next set of video's are recommended. Hence enhancing the metadata of the videos in video sharing platform to make it rich is the purpose of the project. The project concentrates on identifying language and profanity check to determine agegroup of the audience using the video's transcription. The project is developed such that it can be integrated with any video sharing platform that lacks rich metadata for their videos.

First the available metadata for the video is extracted using the technique of web scraping. The source file is fetched and the meta tags are extracted. The labels and content of theses meta tags are then saved in a dictionary. Next the audio of the video is transcribed to text using speechrecognition library. The text should be cleaned to remove all stop words and also words that are above or below a frequency threshold. This cleaned

text is used to identify the language-English, French, German. If the language is English then it is further is checked for profanity. Based on the value of profanity found the audience agegroup tag is set. The language tag and agegroup tag is also saved in the dictionary. The dictionary is then converted and saved as JSON object in JSON file.

To check the language of the text a machine learning model is developed using Keras library. Wikipedia articles of each language are taken as training/testing datasets. The dataset constitutes of 15 articles in each language- English, German, and French. The words in the dataset are converted into vectors, the labels i.e. language names are also converted into vectors. The size of the dataset after converting to vectors is around 44MB. These are then divided into training and testing data using sklearn library. The Keras sequential model with sigmoid activation function is then used to develop the model that will calculate the percentage of each language for a word. The language that is highest found in the text is then given as language tag for the video.

Once the language is determined if its English language then profanity check is done for the transcribed text. A SVM model is developed with twitter tweets as datasets. The dataset is labeled as 1 for tweets that has vulgar words and 0 for texts which has no bad words. The size of the dataset is 64MB with 184355 tweets. These texts are converted into vectors using sklearn library. The vectors will act like features that are divided into training and testing. The model is developed using sklearn library and is saved as file. This model's predict function is invoked by the profanity check for each word in the transcribed text. If the word is determined as profane it is set to 1.The number of 1's in transcription is counted and the agegroup tag for the video is set as above 12, above 18.

1.2 Company Profile

History

Cognizant Technology Solutions(CTS) is IT, consulting, and BPO service provider. Its main area of expertise includes consulting, Data warehouse, Data mining, Video labs, software development and maintenance, Testing, analytics, knowledge outsourcing. The company has four main category-financial services, health care, manufacturing, retail and logistics.

The company was found by Wijeyaraj Kumar Mahadev and Francisco D'Souza in 1994 .Its main headquarter is located in Teaneck,NewJeresy..Currently CTS is spread worldwide, the Chief Executive Office is currently headed by Brian Humphries.

Services

Financial Services: The Company serves traditional and commercial banks, financial institutions, brokers and other management firms

Health care: The Company serves global health care organizations and many pharmaceutical companies. Its area of focus is information management, billing and claims processing.

Manufacturing: The company serves manufactures of automobiles, chemicals, raw materials, energy utilities. Its main focus area is supply chain management, system integration and asset management.

Technology

The company works on various technologies such as Artificial Intelligence, Internet of Things, SD-WAN, and Salesforce.Globally, Cognizant has been focusing on strengthening its capabilities in four key areas: data, digital engineering, cloud, and the internet of things (IoT).

Chapter 2

LITERATURE REVIEW

2.1 Literature Survey

With growing internet users even education institutions now generate video lecturers. Even these videos need metadata to profile them. Manually identifying them is a long work so automatic metadata generation of video lecturers is now in practice .They follow the procedure of audio extraction, audio splitting, audio recognition to text. Once the text for the audio is obtained it is summarised. The methods used for summarization is extraction and abstraction. Hence metadata according to the SCORM standard where obtained such as language, keywords, title. [1].

Metadata of media file is required for availability when searched. If the videos are not tagged appropriately they will never be available to users. Metadata is needed for identifying relevant videos for the users so a database of metadata is created for each video. For a given URL the entire content of the URL like text, description, likes, comments all are tokenized and saved as metadata. Along with these the metadata are also updated based on the other videos that user will choose next. But these also have negative impact of increasing the size of the database with irrelevant data. The metadata collector here also works as features vector generator to feed to machine learning algorithms used by recommendation engines [2].

Content of a video is important source for automatic generation of video metadata. Content analysis in generation of the metadata is important since it can describe the video. A pipeline can be used for extracting metadata, first get the source of the video URL from which description of the video is taken and tokenized. The natural process algorithm for Entity recognition is used to identify the entities and is later saved as metadata for the videos. This method is suitable for videos where the content creator has given a brief description for the videos [3].

Transcription of a video can be reliable source to identify the agegroup tag for a given video ,by checking for profanity . Classifications methods are available to determine that text is profane or not. But this binary classification is not sufficient. A multi labeled multihead attention-based approach for detection of profane text can be used. Here a text is classified as intense toxic, hate speech, insult, obscene. The text is cleaned and stop words are removed. Further the text is encoded, to maintain the sequence positional encoder is employed [4].

Text in any language contains stop words, frequently used important words also infrequently used important words. The dependency between words is also an important factor to consider. Natural processing algorithms help in identifying all these words and build a vocabulary. One of the mechanisms is using Bag of Words but they don't preserve the dependency between words. Neural network such as CNN with multiple layers is also convenient for classifying text .Neural networks can be used to classify the test as hatespeech, bullying, aggressivespeech. First a vocabulary is built from training data and text is cleaned using TF_IDF. Once the training data is reduced CNN model is built, it is recommended as it has the property of local stationary and compositional structure. Local stationary means it establishes dependency with the preceding and next word [5].

Text is represented as vectors to be employed in a machine learning model. One of the traditional approach of representing a text document as vector is using Bag of Words. But two sentences semantically same but which differs in words are treated as two different vectors by Bag of Words. Hence an improvised version of BoW, is Fuzzy BoW. Fuzzy mapping enables words semantically similar trigger the BoW model. It also reduces sparse in vector, makes it more robust and also encode semantic information compared to the original BoW[6].

Transcription of videos can also be short so mechanism that effecietnly classify even short texts are needed. The algorithm used to create vectors for the words can be combination of Word2Vec and TF-ICF since Word2Vec alone is inefficient on short

texts.TF-ICF is derived from TF-IDF where a word occurs least number of times in a text but occurs more in other texts that uniquely identify this text. This weight of the word is introduced into the word2vec model to generate new word vector [7].

Lexical tokens are unique for each language and can serve as features to identify a language. Vocabulary of lexical tokens for each language is built with training and testing data of each language. The methodology is to analyze the distribution word length and build a matrix as part of scoring function. The matrix is again reduced using statistical methods and based on threshold values the language of a given text is identified [8].

Frequency of words, frequency of characters in each word, frequency of special characters is also features to identify language. In preprocessing steps the words should be separated based on space, and other delimiters. The drawback of character frequency algorithm is the characters can be common across different languages such as English, German and French. Hence word frequency algorithm is used where the combination of characters for words will be unique across the languages [9].

The output of the application requires data to be platform independent this makes data exchange easier. The format should be such that it must be accessible in Linux, Windows and Mac OS. Xml and JSON are the two data formats that are independent of the platform used. The choice between JSON and XML is dependent on the application developed. But use of XML is reducing because it has overhead of using tags while JSON consist of simple key value pairs [10].

Malicious comments can also be considered as flavor of profanity detection. Developers also consider developing a dictionary of words that are profane. The presence of these words will help in identifying that a statement is malicious. The dictionary will be updated at regular intervals by developers. Even if the solutions soundssimple it is a large overhead which requires maintaining a dictionary and going through the dictionary to check if a statement has bad words[11].

Web scraping has proved to be a very handy tool for acquiring data,also managing webpages based on their content. Web scraping is the process of going through a document for pattern matching to fetch and save all the matched content [12].

2.2 Existing and Proposed System

2.2.1 Existing System

Dailymotion is a video sharing platform which holds huge number of videos of various languages and also genres. The platform has new videos uploaded everyday by the content creators.

- Most of the old videos in this platform do not contain proper description of the video i.e., their metadata are poor
- Metadata such as language, agegroup, keyword tags are missing
- The metadata in the relatively newer videos are also not sufficient to help the recommendation algorithms
- The poor metadata affects the search engine as the videos searched by user cannot be properly identified
- The platform cannot recommend suitable videos according to their use due to lack of metadata

2.2.2 Proposed System

To make the platform efficient the metadata of the videos are made rich by adding more attributes that will describe the video. Manually going through these videos to enhance their metadata is infeasible hence mechanism to automate this process is required.

In the proposed system the available metadata of a video is taken and also the content of the video is analysed to add few extra attributes to the user. In this project the transcription of audio is used to identify language-English, German, French and use of profanity to decide the audience agegroup for a video. The final consolidated output of rich metadata of a video is saved in JSON file.

Videos in video sharing platform provide facility for uploaders to describe the video. This attributes must also be saved in the rich metadata file of a video. To get the available metadata the metatags from the source file is extracted. To anlayse the content,

transcription of video is obtained using speechrecognoition API. The text is then used to identify language; profanity of the text is also checked to decide the age of viewers.

- Using web scrapping technique to get available metadata
- Using speechrecognition API to transcribe audio
- Build a machine learning model to identify language using Keras
- Build a machine learning model to identify if a word is profane using sklearn library.

2.3 Tools and Technologies

beautifulsoup

beautifulsoup4 is a python library used to perform web scrapping on a source file. Web scraping is an approach used to collect information about a website. This is performed by fetching the source code of a website by passing its URL. Then using regular expressions, string matching functions the required information from it are scrapped and saved in a local file.

speechrecognition API

speechrecognition3.6 is a Google API to recognize the spoken words in audio. The recognized words are then saved in a python list. The API can be called on any .wav audio file. The API is powered with Neural Network model and can recognize words of up to 120 languages.

Text Vectorization

Text is one of the important data but machine learning algorithms only work with numerical data. Hence the text should be converted into numerical vectors to work as features. Text vectorization can be implemented through bag of words methodologies. Where each word in the document which is more than 2 characters length are assigned unique numbers. But this methodology will not maintain the ordering of the words. A NumPy array with 1D array for each line in the document is created, with count of occurrence of each word in the sentence . This will help the model learn what words are considered based on the labels of the sentences having those words.

Eg: x=[“ This is a nice place”,”this is is”,”place to live”]

This=1,is =2,nice=3,place=4,to=5,live=6

Text Vectors

[[1,1,1,1,0,0],

[1,2,0,0,0],

[0,0,0,1,1,1]]

Another way of text vectorization is assigning each letter a unique number and representing the characters in a word through those numbers.

Support Vector Machine

Support Vector machine is a supervised machine learning model that is used to find a hyper lane using existing features(X) and labels(Y) such that it can predict the label for a unobserved new values of X. SVM is useful for applications that has to classify X as belonging to C1 or C2 and doesnot have any multi dimensional features.

sklearn library

sklearn0.23.0 library is a popular python library used for building machine learning models. The library has facility to divide the datasets into training and testing. The library also provides CountVectorizer function for text vectorization. The Linear SVM library is used to build a classifier model based on training/testing data and also predict function to predict class of unobserved data.

Keras

Keras2.3.0 is neural network library written in python programming language. The basic model of Keras is sequence model which develops model layer by layer each layer has set of specified nodes.

JSON

JSON (JavaScript Object Notation) is a data exchange format. It is widely used for exchanging data between different platforms. It is compatible with all OS and also is preferred over XML.XML is equivalent to JSON but its tags are considered as overhead while JSON has simple key value pairs.

2.4 Hardware and Software Requirements

2.4.1 Hardware Requirements

- Process Clock speed: 1.5 to 2Ghz
- RAM:4GB
- Memory:400GB

2.4.2 Software Requirements

- Operating System: Windows 10
- Tools: Anaconda4 Distribution with jupyter notebook 6
- Language: python3
- Libraries: beuatifulsoup4.1.0, sklearn0.23.0, NumPy1.16.0, Keras2.3.0, speechrecognition3.6, nltk

Chapter 3

SOFTWARE REQUIREMENT SPECIFICATIONS

3.1 Introduction

SRS is document encompassed of set of features expected from a system under development. It also gives each requirement's constraints, expected outcomes. It is important for an SRS to be clear and concise because the developers cycle back to SRS for every decision to be made [17].

3.1.2 Definitions

Audio Transcription: It is a process of generating text of spoken words in an audio file.

Rich metadata: Metadata are set of attributes that describes a media file. Rich metadata goes beyond the basic descriptions and use the content of the media file to describe them.

Web scraping: In web scraping source file of a website is taken as input. From the input required information are scrapped and saved as output.

3.1.3 Acronyms

API: Application Programming Interface

GB: Giga Bytes

HLS: HTTP Live Streaming

JSON: JavaScript Object Notation

MB: Mega Bytes

RAM: Random Access Memory

sk learn: SciKit Learn

URL: Uniform Resource Locator

XML :eXtensible Markup Language

3.1.4 Overview

With increase in the number of internet users video sharing platforms today are improvising their recommendation engines to recommend better videos to the users.

Metadata of these videos play a vital role in recommending the right videos to the users. The metadata of the video files are termed as rich metadata as they also include attributes regarding the content of the videos along with the basic attributes. Dailymotion is a video sharing platform but lacks the rich metadata required by the recommendation engine to suggest right videos to the users. Metadata are also important for the availability of the videos.

The application should extract the existing metadata and alsoextract the metadata based on the content for a given Dailymotion video URL. The extracted metadata should be then consolidated and given as JSON object to the recommendation engine.

3.2 General Description

To generate rich metadata for a given URL with existing metadata and also attributes such as language and agegroup tag. The final output is JSON file with generated metadata for the video URL.

3.2.1 Product Perspective

The product developed will be used in video sharing platform to work with search engine and recommendation algorithms to suggest videos to users. The product developed should be standalone such that it can be easily integrated with any video sharing platform without any dependency.

3.2.2 Product Functions

The product's main goal is extracting rich metadata for the given a Dailymotion video URL.Since Dailymotion has less metadata associated with their videos to be used in recommendation engine the project overcomes this problem. First the attributes available with the videos are extracted using web scraping technique. Then audio of the video file is transcribed.The obtained transcribed text is cleaned using natural language processing

libraries. This cleaned text is used with machine learning algorithms to identify the language, profanity to decide agegroup tag.

3.2.3 User Characteristics

Developers:The application can be used by video sharing platform or streaming websites to profile their videos. The users of the application need only knowledge of using JSON data for their application. The users of the applications are mostly developers who request rich metadata for their video URLS as service from the application.

General Public: General public who use streaming services or video sharing platforms to watch videos are also indirectly the users of the rich metadata system by giving the URL of the videos they want to watch. These users need not have any knowledge of the system or its technologies.

3.2.4 General Constraints

The application will be mostly used by developers of video sharing platform to request rich metadata for their video URL.The use of application has few constraints such as need of internet connection. The application needs internet connection to perform web scrapingand also access the speechrecognitionAPI.The platform using the application should adhere to the HLS format of content sharing..

3.2.5 Assumptions and Dependencies

- The profanity checking of transcribed text is only limited to English language
- The application assumes user has access to the internet to access API and the URL
- The application assumes that the Anaconda4 framework is installed in the system with all required libraries of latest version such as NumPy, pandas
- The application assumes that the videos sharing platform or streaming websites adhere to the HLS format of transferring videos

3.3 Functional Requirements

Module Name: Web Scraping

This module is for extracting the existing metadata associated with the video like bitrate,duration,title.

Input:URL of Dailymotion video

Processing:For the given URL the source code of the file is fetched. From the source code all the metatags are taken which will serve as one of the source of metadata for the videos.

Output: Available metadata of the video.

Module Name: Audio transcription

This module is for transcribing the spoken words in the video.

Input: Dailymotion video URL

Processing: For the given video a transcription of its audio is generated. The video is first converted to a .wav format from which words are recognized using speechrecognition library. The transcription is first cleaned with nltk libraries to remove stop words.

Output: Transcript for the video

Module Name: Language Detection

Input: Transcript of video

Processing: A deep learning model using Keras is built which uses articles from Wikipedia of various languages as its dataset to train and test. The words from transcription are passed into this model to identify the language such as English,French, and German.

Output: Language of the transcript.

Module Name: Profanity Check

Input: Clean text from previous module

Processing: The transcribed words are converted into vectors .The vector of words is used to identify bad words. Linear SVM is used to build model that is trained with set of data classified as good and bad. The input clean text is passed to model to check if the number of bad words is beyond a threshold.

Output: The output of the video is tag if the video is suitable for children below eighteen or not.

3.4 External Interface Requirements

- User Interface: The user interface is for passing the video URL to the application. The final output should be saved in file system
- Software Requirements: FFMPEG must be installed in the system to do video to audio conversion

3.5 Non Functional Requirements

- Reliability: The application should be reliable as the metadata generated is crucial for a search engine
- Interoperability: The solution must be such that it can be implemented on any platform with no dependencies
- Speed: The application should not slowdown the performance of the search engine

3.6 Design Constraint

3.6.1 Standard Compliance

- The application is developed with python3 programming language in Anaconda framework
- The application is used only for ethical purpose on the video sharing platform to which the user has access

3.6.2 Hardware Limitation

- The size of RAM should be more than 4GB for the application without slowing down the system
- For the project to build the machine learning models the processor should be i3 above or AMD equivalent with 8GB memory for smooth execution

Chapter 4

SYSTEM DESIGN

4.1 System Perspective

4.1.1 Problem Specification

The existing metadata of given video URL should be fetched and saved in a JSON file. Further to enhance the metadata the content of the video through its transcription should also be extracted such language and suitable audience age group for a video.

4.1.2 Block Diagram

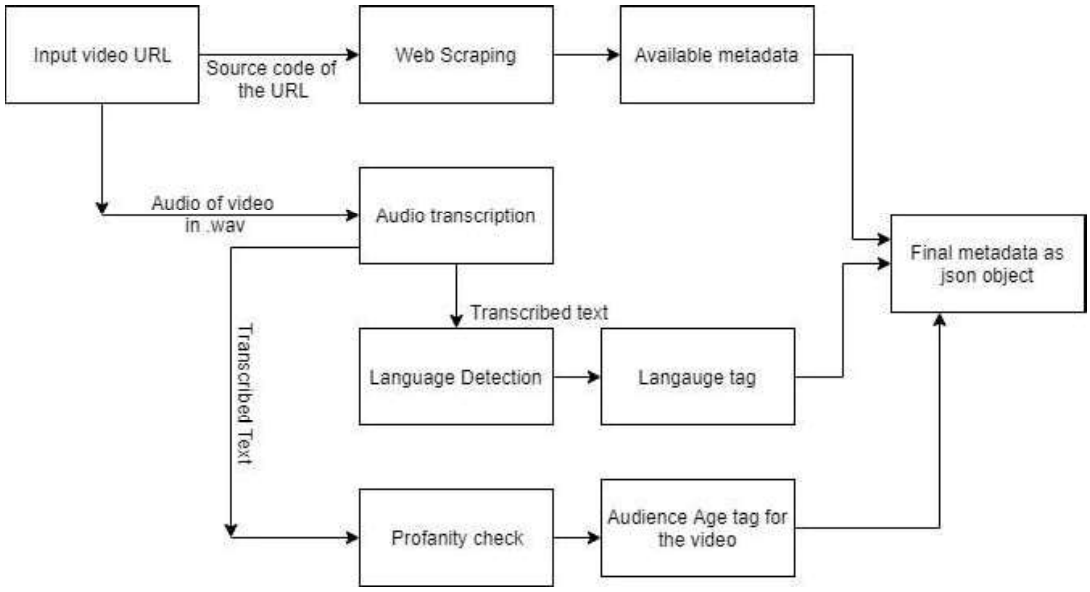


Figure 4.1: Block Diagram for extracting rich metadata extraction.

Figure 4.1 is a block diagram of the components involved in generating JSON file of rich metadata for a given video URL.

4.1.3 Module Specification

Module Name: Web Scraping

A video URL will have some metadata already available given by the content creator. These metadata may not be complete. But the available attributes are taken and saved in a JSON file.

Module Name:Audio Transcription

The spoken words in the video are the source considered for content analysis.The spoken words are transcribed into text using speechrecognitiom API.The text is saved in a python list. All stopwords in the text are discarded and the remaining words are converted to lowercase using nltk libraries.

Module Name: Language Detection

A model is built with keras that takes fifteen Wikipedia articles of English,French and German as datasets. The dataset is vectorised for each character and divided into training and testing datasets with sklearn library. The model is built with three dense layers and sigmoid activation function. This model is saved to be invoked later.

The transcribed words from previous module are iterated to determine its language using the model's predicting function. The count of language is calculated and language with highest count is saved as language tag for the video

Module Name:Profanity Check

A model is built with skleran's Linear SVC library which takes a twitter dataset with set of tweets. Each tweet is labeled as 1 for use of abusive words and 0 for safe tweets. The tweets are vectorized and sliced into training and testing features.The model is then saved to be invoked later.

The transcribed words if identified as English are iterated to check for abusive words using the model's predict function. Based on count of number of abusive

word if there more than 10 the agegroup tag is set as above 12,if the count is above 15 the agegroup tag is set as above 18.If the count is more than 50 the video is flagged as violent.

Chapter 5

DETAILED DESIGN

5.1 System Design

The system design diagrams aims in visualizing the underlying system's design in terms of object oriented concepts. The design diagrams help a reader to understand the interaction between the objects in the system to produce the final result [20].

5.1.1 Object Modelling

Class Diagram

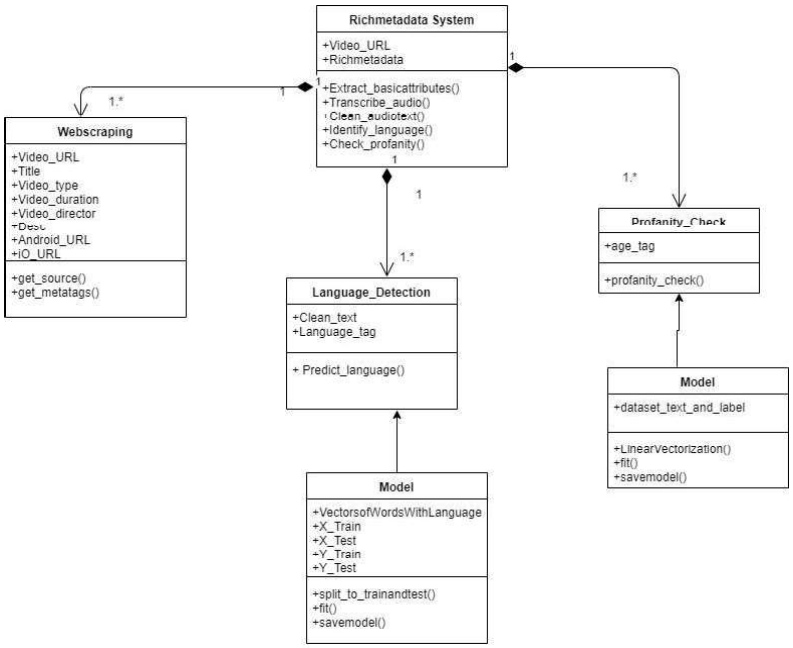


Figure 5.1: Class diagram of extracting rich metadata for a video URL.

Figure 5.1 depicts the relationship between various entities in the system. The language detection and profanity check use their respective models to predict the tags which aresaved in rich metadata system. Web scraping returns the available metadata to the rich metadata system.

5.1.2 Dynamic Modelling

Use Case Diagram

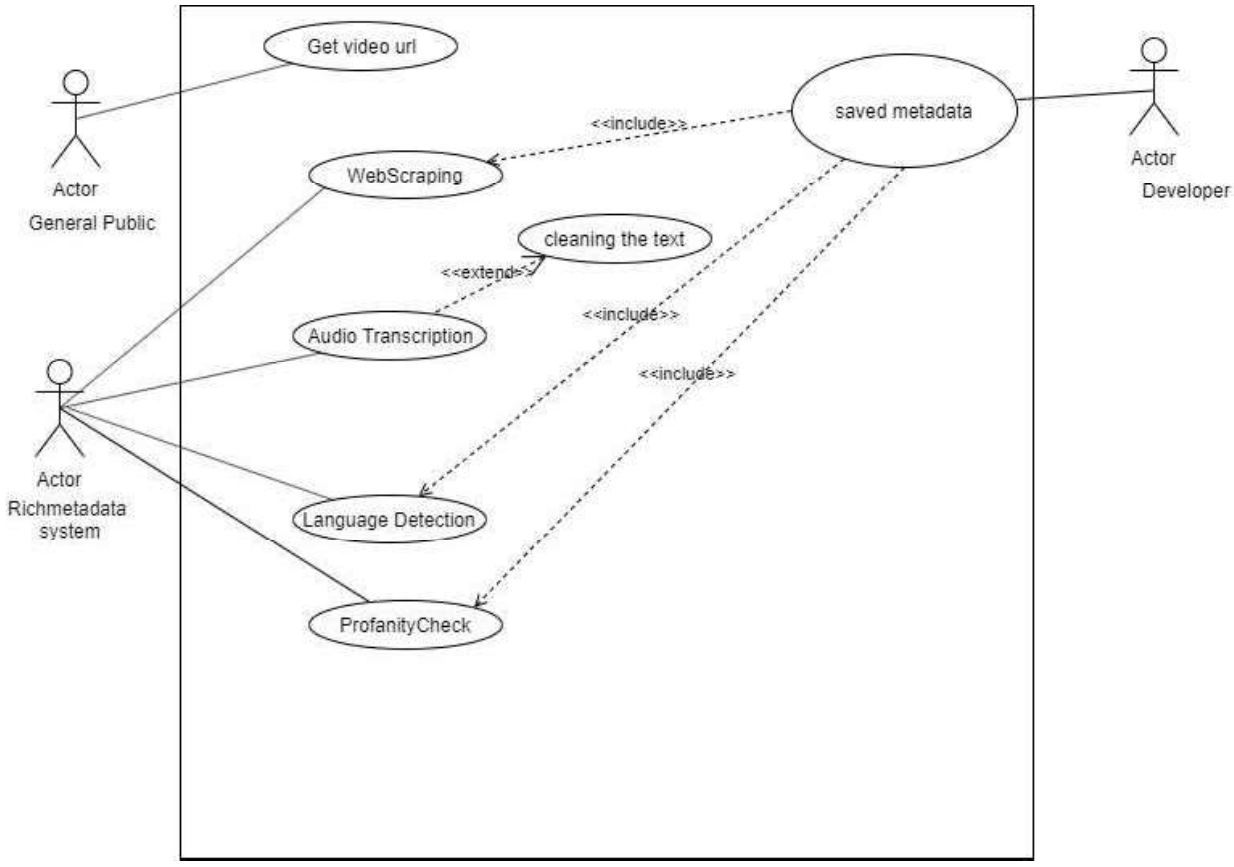


Figure 5.2: Use case diagram of extracting rich metadata

Figure 5.2 is a use case diagram depicting the various functionalities in the system of rich metadata extraction. The user gets a video URL, the applications executes web scraping, audio transcription, language detection and profanity check and saves all the collected metadata as JSON file in the file system.

Activity Diagram

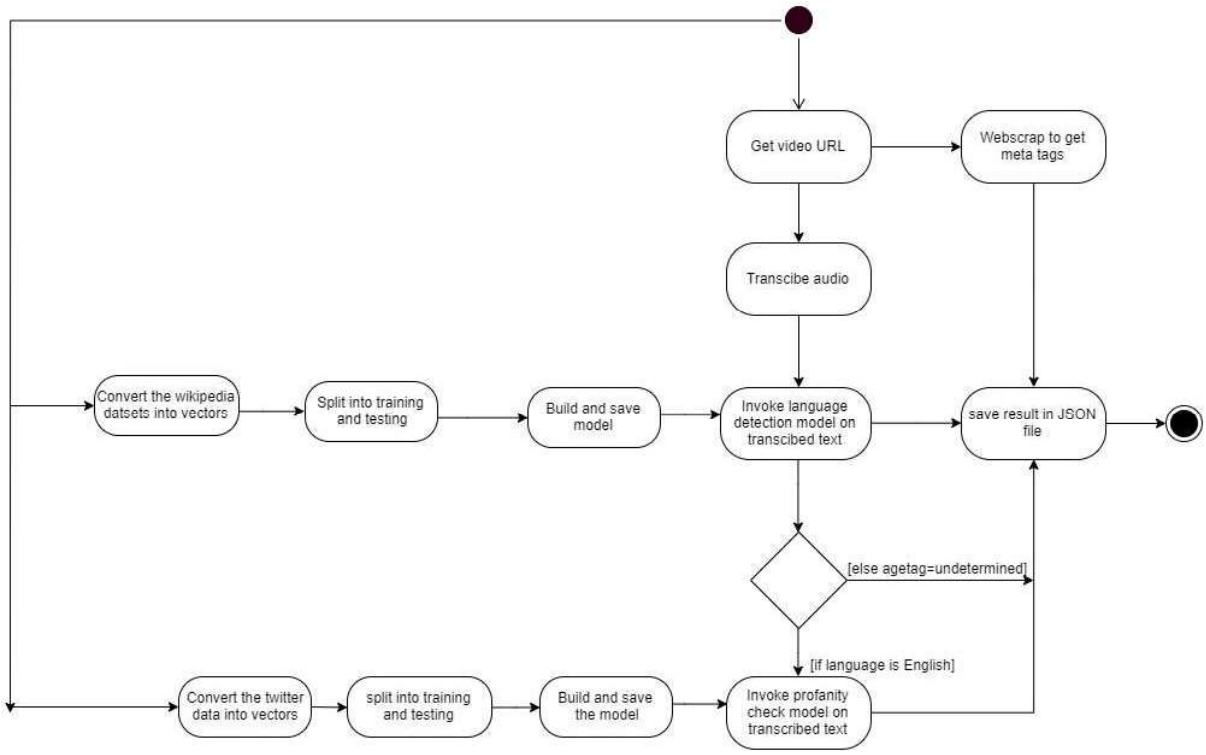


Figure 5.3: Activity Diagram of rich metadata extraction

Figure 5.3 is an Activity Diagram that depicts the flow activities in the system that leads to final output.

Sequence Diagram

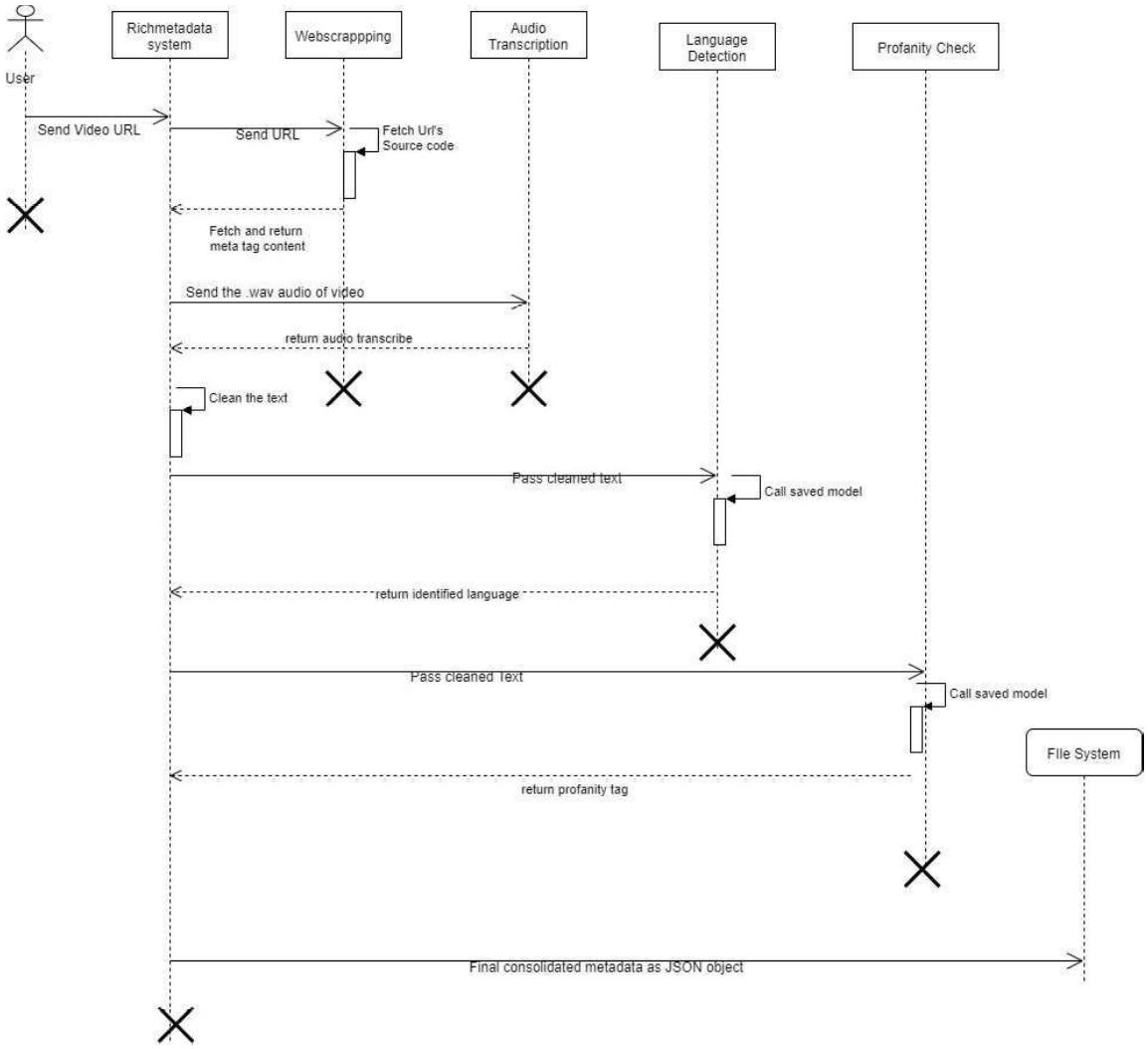


Figure 5.4: Sequence Diagram of rich metadata extraction

Figure 5.4 is a sequence diagram depicting how the various entities interact with each other one after the other to produce final output.

5.1.3 Functional Modelling

Data Flow Diagram

Zero Level DFD

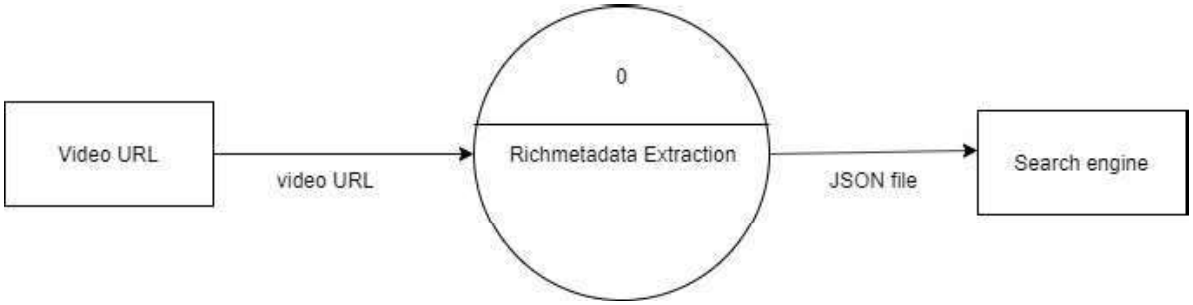


Figure 5.5: Zero level DFD of rich metadata extraction

Figure 5.5 is zero level DFD of the rich metadata extraction where for a passed URL its metadata is saved as JSON object in JSON file.

First Level DFD

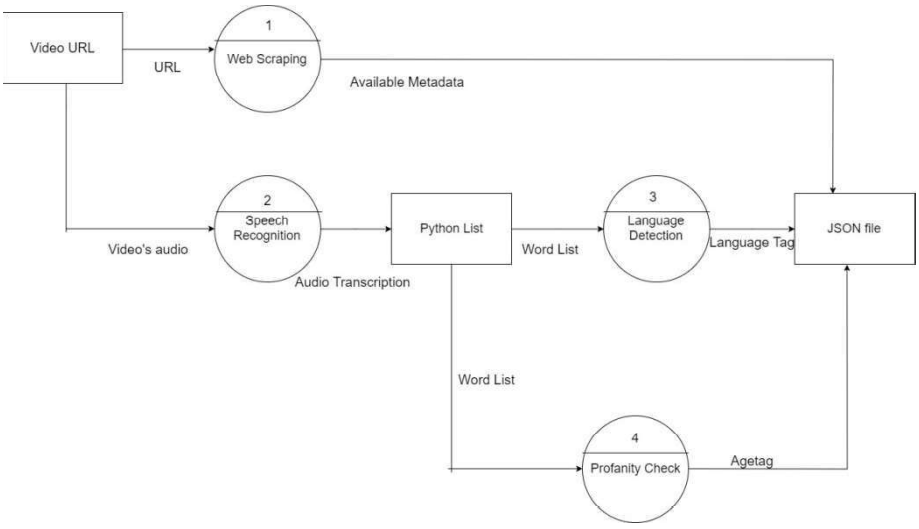


Figure 5.6: First Level DFD of rich metadata Extraction

Figure 5.6 is a first level DFD where the flow of data through all the modules to produce JSON object of rich metadata for given video URL is depicted.

Second Level DFD

Module Name: Web Scraping

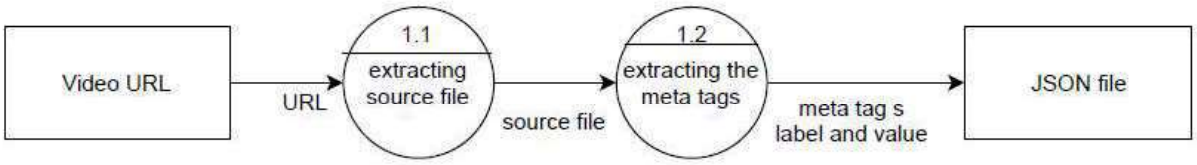


Figure 5.7: Second Level DFD of Web scraping

Figure 5.7 is a second level DFD of web scraping module where the source file of the URL is taken and the meta tags key and value is saved in JSON object

Module Name: Audio Transcription

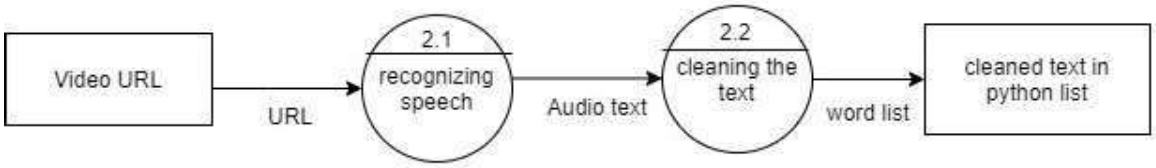


Figure 5.8: Second Level DFD of Audio Transcription

Figure 5.8 is a second level DFD of audio transcription where the audio of video is taken and converted to text. The text is then cleaned and saved in a python list.

Module Name: Language Detection

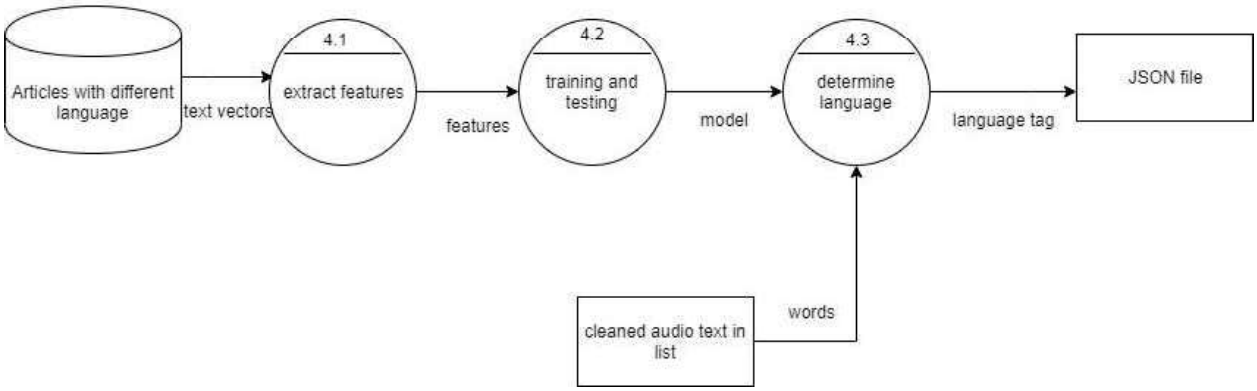


Figure 5.9: Second Level DFD of Language Detection

Figure 5.9 depicts the flow of data for model creation and also using the model to check language of transcribed words and saving in the JSON file

Module Name: Profanity Check

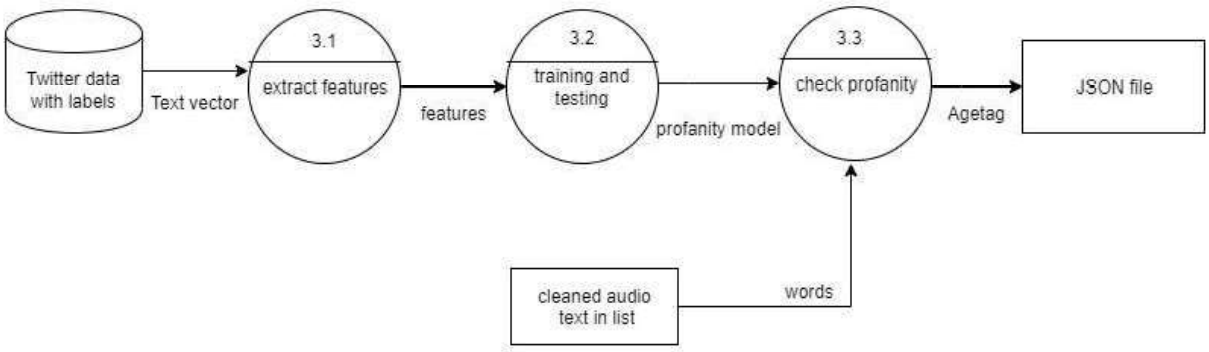


Figure 5.10: Second Level DFD of Profanity Check

Figure 5.10 flow of data in model creation and using the model on the transcribed text to identify profanity and saving the result in JSON file

5.2 Detailed Design

5.2.1 Design Decisions

- The datasets used for creation of models are saved as data frames
- The transcribed text is saved in python list
- The final output should be saved as JSON object in JSON file

5.2.2 Logic Design

Module Name: Web Scraping

Insert a Dailymotion video URL as input.

If URL valid

Source file of URL extracted

Pattern matching for meta tags

Save key and value of meta tags in a dictionary

Else

Invalid URL

End If

Module Name: Audio Transcription

If URL is valid

Extract .wav of the video

Invoke speechrecognition API

Save the recognized words in a list

Else

Invalid URL

End IF

Module Name: Language Detection

I. Building Model

Input Wikipedia datasets with language labels

Foreach word **in** language:

Convert into vector and save in list

End For

Split the list into Training and Testing

Build model using keras, sigmoid activation function

Save the model

II. Using Model

Pass the python list of transcribed words

Invoke the language detection model

While python list **NOT** Empty

Identify language for each word

Maintain count of each language

End While

languagetag=language with highest count

Save languagetag in JSON file

Module Name: Profanity Check

I. Building Model

Input Twitter dataset with labels for profanity

Foreach word **in** tweet:

Convert into vector and save in list

End For

Split the list into Training and Testing

Split the list into Training and Testing

Build model using sklearn library's Linear SVC and Count Vectors

Save the model

II. Using Model

Pass python list of transcribed words
While python list **NOT** Empty

Check profanity for each word

Maintain count

End While

If count >5

Agetag=above 12

Else if count > 25

Agetag=above 18

Else If count > 50

Tag=abusive language

Else

Agetag: Suitable for all

End If

Save tag in JSON file

Chapter 6

IMPLEMENTATION

Output of Audio Transcription

Figure 6.1:Audio transcription

Figure 6.1 is the screenshot of audio transcription obtained for a passed video URL

Output of model built for profanity check

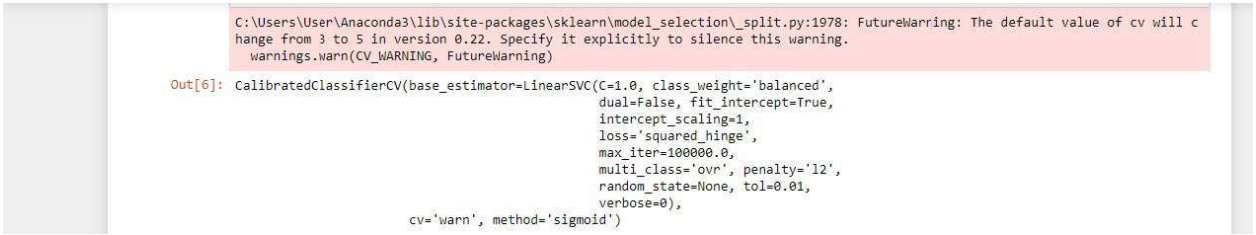


Figure 6.2:Model built for profanitycheck

Figure 6.2 is the screenshot of profanity check model that is built with sklearn's Linear SVC library using twitter text vectors as features for training and testing

Output of model built for language detection

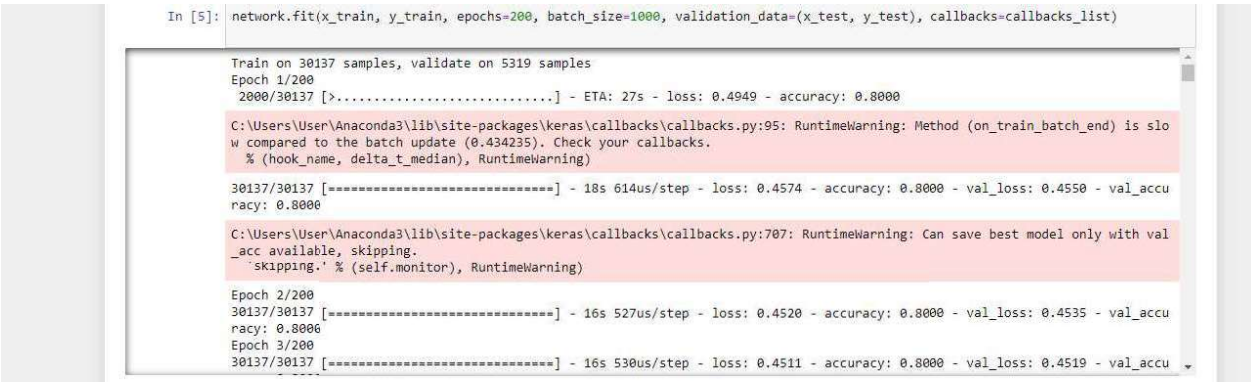


Figure 6.3: Model built for language detection

Figure 6.3 is a screenshot of language detection sequence model developed using Keras library that uses Wikipedia article's word vectors for training and testing.

Output metadata extracted and saved in JSON file

Figure 6.4: rich metadata saved in JSON file

Figure 6.4 is a screenshot of the final metadata that is stored in a JSON file. The output includes-Title, videoduration, agetag, language, url, videodirector, video duration ,video type,autoplay,secureurltag,description.

Chapter 7

SOFTWARE TESTING

7.1 Test Cases

Testing is important to confirm that the developed product is reliable and meets all the requirements [40].

Unit Testing is done for the smallest unit i.e., code snippets during development. Unless the unit testing for small units are not successful it is not possible for units to interact with one another.

Integration testing is done to check for compatibility between different units. A system requires the units to interact smoothly to produce desired output.

System testing is performed at end and works like a checklist checking if the application meets all requirements listed in SRS.

7.1.1 Unit Testing

Following are the modules where testing is performed

- Web Scraping
- Audio Transcription
- Language Detection
- Profanity Check

Unit Test cases for Web Scraping

Table 7.1: Unit Test cases for Web Scraping source file

Test Id	Features Tested	Input	Expected Output	Output
UT11	Check if the URL exists	URL of a Dailymotion video	The code to execute web scraping should not throw error	Fail
UT12	Check if the web scraping returns a list of key value pairs as dictionary	Valid URL of a Dailymotion video	A valid JSON object with available metadata	Pass

Unit Test cases for Audio Transcription

Table 7.2: Unit Test cases for Audio Transcription

TestId	Features Tested	Input	Expected Output	Output
UT21	Test if the speechrecognition library is imported	Import speechrecognition statement	The statement should run without error	Pass
UT22	Test if the audio is transcribed	.wav audio of the video	The spoken words are saved as list	Pass
UT23	Test if all stop words are removed	The transcribed list of words	Cleaned list of words	Fail

Unit Test cases for Language Detection

Table 7.3: Unit Test cases for Language Detection

TestId	Features Tested	Input	Expected Output	Output
UT31	Test if the Wikipedia articles are loaded for each language- English, French, German	Fifteen Wikipedia articles in 3 languages each	The data set is loaded	Pass
UT32	Test if each word is of specified maximum length	All words in the articles	Words with length greater than specified length are discarded and rest are saved in dictionary of a language	Pass
UT33	Test if all the words are vectorized	Dictionary of a language	The vectors are saved as NumPy array	Fail
UT34	Test if the vectors are divided into testing and training data	NumPy array of vectors	The data is divided into training and testing data	pass
UT35	Test if the Sequence model is fitted and saved	Training and testing data NumPy array	Sequence model to predict language	Pass
UT36	Test for a sample word	A word	Language predicted	pass

Unit Test cases for Profanity Check

Table 7.4: Unit Test cases for Profanity Check

TestId	Features Tested	Input	Expected Output	Output
UT41	Test if the twitter dataset is loaded	The twitter data with labels	The twitter data is loaded from .csv into dataframes.	Fail
UT42	Test if the words in dataframes are converted to vectors of features	The dataframes of data	The words in the dataframes are converted to vectors of features	Pass
UT43	Test if the vector dataframe is divided into testing and training	The data frame with vector data	The dataframe is divided into training and testing data	Pass
UT44	Test if the LinearSVM model is fitted and the model is saved	Training dataframe and testing dataframe	A Linear SVM model is created and saved.	Pass
UT45	Test with sample word	A profane word	Output was 1(bad word)	Pass

7.1.2 Integration Testing

Table 7.5: Integration testing test cases of all modules

TestId	Features Tested	Input	Expected Output	Output
IT01	Check if the webscrap, audio transcription,profanity, languagedetection modules are imported	Importing modules	All modules are imported	Pass
IT02	Test if the JSON object returned by the module is received when the function is invoked with a valid URL	Webscrap function with valid URL	A valid JSON object with available metadata	Fail
IT03	Test if the transcription of audio is returned as clean text in list	Audio transcription function with valid URL	List of cleantranscribed words	Pass
IT04	Test if the profanitycheck function is returning correct label for each word	A list of transcribed words	A list of labels	Pass
IT05	Test if the threshold for profanity and return the agegroup tag	List of labels with 0 or 1	Agegroup tag	Pass
IT06	Test if the language detection function determines correct language	A list of transcribed words	Language tag	Pass

7.1.3 System Testing

Table 7.6: System testing test cases for rich metadata extraction

TestId	Features Tested	Input	Expected Output	Output
ST01	Check if the main function returns a dictionary all metadatafrom webscraping, languagetag,Agegrouptag	URL of a Dailymotion video	Dictionary of metadata	Pass
ST02	Convert the final dictionary into JSON object and save in a JSON file	Metadata dictionary	A JSON filewith JSON objects.	Pass

7.2 Testing and Validations

Figure7.1: Test for invalid URL

Figure 7.1 Test to check that invalid URL is not accepted

Figure 7.2: Test for accepting valid URL

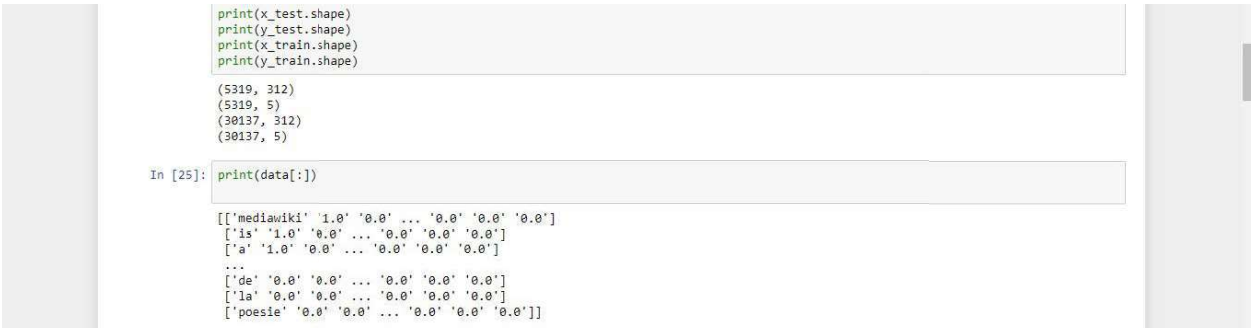
Figure 7.2 Test to verify that the given video URL is valid.

Figure 7.3: Test for generating Source code extracted from passed URL

Figure 7.3 is Screenshot of the source code successfully extracted for web scraping

Figure7.4:Test for request to speechrecognition API

Figure 7.4 Test to check if the speechrecognition API request was successful or not



```
print(x_test.shape)
print(y_test.shape)
print(x_train.shape)
print(y_train.shape)
```

```
(5319, 312)
(5319, 5)
(30137, 312)
(30137, 5)
```

```
In [25]: print(data[:])
```

```
[['mediawiki' '1.0' '0.0' ... '0.0' '0.0' '0.0']
 ['is' '1.0' '0.0' ... '0.0' '0.0' '0.0']
 ['a' '1.0' '0.0' ... '0.0' '0.0' '0.0']
 ...
 ['de' '0.0' '0.0' ... '0.0' '0.0' '0.0']
 ['la' '0.0' '0.0' ... '0.0' '0.0' '0.0']
 ['poesie' '0.0' '0.0' ... '0.0' '0.0' '0.0']]
```

Figure 7.5: Splitting dataset into training and testing datasets for language detection model

Figure 7.5 is the verification to check if the dataset for language detection was loaded and split into testing and training

```
try:
    network.load_weights('weights.hdf5')
except:
    print("Called model couldnot be found")
```

```
Called model couldnot be found
```

Figure 7.6: Saving language detection model

Figure 7.6 is to check if the language detection built model is saved

```
try:
    data = pd.read_csv('clean_data.csv')
    texts = data['text'].astype(str)
    y = data['is_offensive']
except:
    print("File not Found.Check if datafile exists")
```

```
File not Found.Check if datafile exists
```

Figure 7.7: Loading of twitter data for profanity check model

Figure 7.7 is to check if Twitter dataset for profanity check is loaded

```
In [7]: text=['hello',' [redacted] ',' [redacted] ']
        predict(list(text))

Out[7]: array([0, 1, 1], dtype=int64)
```

Figure 7.8: Labelling of sample words

Figure 7.8 is to check if the words are labelled appropriately(1 for profane words and 0 for other words).

```
mainprg()

Enter urlhttps://www.dailymotion.com/video/x2297kt
exception occured
webscraping failed
```

Figure 7.9: Returned result of web scraping to main module

Figure 7.9 Test to check if web scraping module returns the result to main module

File Edit View Language JSON

{"Title": "Tenet (French Trailer 2 Subtitled) - Vid\u00e9o Dailymotion", "Url": "https://www.dailymotion.com/video/x7u2kzl", "video_url": "https://www.dailymotion.com/embed/video/x7u2kzl?autoplay=1", "video_secureurl": "https://www.dailymotion.com/embed/video/x7u2kzl?autoplay=1", "video_type": "text/html", "video_width": "640", "video_height": "360", "video_duration": "172", "video_director": "https://www.dailymotion.com/IVA_MovieTrailers_French", "release_date": "2020-05-22T16:36:16+02:00", "Tag": "cinema", "Description": "Trailer for 'Tenet' starring Aleksei Podlesnov, Kenneth Wolf Andersen Haugen, Marcel Sabat, Ronald Pelin, Carina Velva", "image": "https://s2.dmcdn.net/v/SFJ311Uo0oNTVS_Ls/526x297", "image_secureurl": "https://s2.dmcdn.net/v/SFJ311Uo0oNTVS_Ls/526x297", "age": "all age", "languagee": "english"}

Figure 7.10: Final rich metadata saved

Figure 7.10 Test to check if final result is saved in the JSON file for a given video URL.

Chapter 8

CONCLUSION

Today with availability of internet to general public at cheap price there is a increase in generation and consumption of videos. New video sharing platforms or streaming services are coming up and these platforms need a way to profile their videos to meet their consumer's needs. This profiling of videos can be done using rich metadata of the videos. This metadata will be used by search engines and also machine learning algorithms in the platforms to recommend videos to their users.

The project has generatedand saved rich metadata related to given Dailymotion video URL. The available metadata is first extracted and saved in a JSON file. Further based on the transcription of video the language of video is determined and saved as language tag in the same JSON file. Next profanity check is done on the transcribed text to check use of abusive language and tag the video appropriately.

The project is developed as standalone application that will be integrated withrecommendation engine to generate rich metadata for the videos that the user hassearched or.This will help the recommendation engine to enhance its performance.The project can be used in any website or mobile device that implements video searching in their product.

Chapter 9

FUTURE ENHANCEMENTS

Video sharing platform requires data about their videos for better recommendation of videos and thus increase their consumers. The data of the videos are rich metadata that describes it content along with other basic attributes such as bit rate,title,date of creation etc.

In the project description of videos is predicted from its audio transcription, tags such as language and age group are identified. These tags are important to classify videos based on language, age group. But more tags are needed to describe the videos.

- The transcribed text can also be used to summarise the videos and identify keywords that describes its content. This is called summarization techniques and machine learning algorithms are available to implement them
- The language detection module in the project is limited to only three languages.This can be extended to all languages such as Hindi, Kannada,Tamil,Chinese,Korean etc whose script is different from English
- The profanity check module in the project is limited to only English. This facility to can be extended to other languages
- The frames of videos can be used for content analysis to describe the video. Frames of a video too can have text, objects, scenes that can describe the video

BIBILOGRAPHY

[1] Maratea, Antonio &Petrosino, Alfredo & Manzo, Mario. (2016). Generation of description metadata for video files. ACM International Conference Proceeding Series. 767. 262-269. 10.1145/2516775.2516795.

[2] G. Mathew, S. T. Smith and J. Passarelli, "Large Scale Open Source Video Recommender Tool Using Metadata Surrogates," *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, 2018, pp. 1974-1977.

[3] Rangaswamy, Shanta & Ghosh, Shubham & Jha, Srishti & Ramalingam, Soodamani. (2016). Metadata extraction and classification of YouTube videos using sentiment analysis. 1-2. 10.1109/CCST.2016.7815692.

[4] P. Ratadiya and D. Mishra, "An Attention Ensemble Based Approach for Multilabel Profanity Detection," 2019 International Conference on Data Mining Workshops (ICDMW), Beijing, China, 2019, pp. 544-550.

[5] S. V. Georgakopoulos, S. K. Tasoulis, A. G. Vrahatis, and V. P. Plagianakos, ”Convolutional neural networks for toxic comment classification”, arXiv preprint arXiv:1802.09957, 2018.

[6] R. Zhao and K. Mao, "Fuzzy Bag-of-Words Model for Document Representation," in IEEE Transactions on Fuzzy Systems, vol. 26, no. 2, pp. 794-804, April 2018.

[7] D. Chen and H. Wang, "Research on Short Text Classification Algorithm Based on Neural Network," 2018 2nd IEEE Advanced Information Management,Communicates,Electronic and Automation Control Conference (IMCEC), Xi'an, 2018, pp. 1726-1729.

[8] Word-length algorithm for language identification of under-resourced languages Ali Selamata,*, Nicholas Akosu

[9] Effective language identification of forum texts based on statistical approaches KheireddineAbainia,SihamOuamour,HalimSayoud

[10] G. Goyal, K. Singh and K. R. Ramkumar, "A detailed analysis of data consistency concepts in data exchange formats (JSON & XML)," 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, 2017, pp. 72-77.

[11] Y. Wang and Y. Shi, "The research of sin malicious comments detection based on semantic information and stop-words table," 2017 7th IEEE International Conference on Electronics Information and Emergency Communication (ICEIEC), Macau, 2017, pp. 444-447.

[12] D. PRATIBA, A. M.S., A. DUA, G. K. SHANBHAG, N. BHANDARI and U. SINGH, "Web Scraping And Data Acquisition Using Google Scholar," 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 2018, pp. 277-281

[13] R. Diouf, E. N. Sarr, O. Sall, B. Birregah, M. Bousso and S. N. Mbaye, "Web Scraping: State-of-the-Art and Areas of Application," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 6040-6042.

[14] D. Dai, P. Carns, R. B. Ross, J. Jenkins, N. Muirhead and Y. Chen, "An asynchronous traversal engine for graph-based richmetadata management", *Parallel Comput.*, vol. 58, pp. 140-156, 2016.

[15] E. N. SARR, S. A. L. L. Ousmane and DIALLO A, "FactExtract: Automatic Collection and Aggregation of Articles and Journalistic Factual Claims from Online Newspaper", 2018 Fifth International Conference on Social Networks Analysis Management and Security (SNAMS), pp. 336-341, 2018, October.

[16] N. R. Haddaway, "The use of web-scraping software in searching for grey literature", Grey J, vol. 11, no. 3, pp. 186-90, 2015.

[17] M. Asif, I. Ali, M. S. A. Malik, M. H. Chaudary, S. Tayyaba and M. T. Mahmood, "Annotation of Software Requirements Specification (SRS), Extractions of Nonfunctional Requirements, and Measurement of Their Tradeoff," in IEEE Access, vol. 7, pp. 36164-36176, 2019.

[18] Sood, Sara &Antin, Judd & Churchill, Elizabeth. (2012). Profanity use in online communities. Conference on Human Factors in Computing Systems - Proceedings. 10.1145/2207676.2208610.

[19] J. Cook and J. Ranstam, "Overfitting", *Brit. J. Surgery*, vol. 103, no. 13, pp. 1814-1814, 2016.

[20] J. Yang, Y. Lee, D. Gandhi and S. G. Valli, "Synchronized UML diagrams for object-oriented program comprehension," 2017 12th International Conference on Computer Science and Education (ICCSE), Houston, TX, 2017, pp. 12-17.

[21] G. Boeing and P. Waddell, "New insights into rental housing markets across the United States: web scraping and analyzing craigslist rental listings", *Journal of Planning Education and Research*, vol. 37, no. 4, pp. 457-476, 2017.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the inception architecture for computer vision", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2818-2826, 2016.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", *Proc. Int. Conf. Mach. Learn.*, pp. 448-456, 2015.

[24] H. Jiang, K. Huang and R. Zhang, "Field support vector regression", *Proc. Int. Conf. Neural Inf. Process.*, 2017.

[25] G. Boeing and P. Waddell, "New insights into rental housing markets across the United States: web scraping and analyzing craigslist rental listings", *Journal of Planning Education and Research*, vol. 37, no. 4, pp. 457-476, 2017.

[26] Victor Zhou “Building a Better Profanity Detection Library with scikit-learn Internet:www.towardsdatascience.com/building-a-better-profanity-detection-library-with-scikit-learn-3638b2f2c4c2.com [Feb 4, 2019]

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the inception architecture for computer vision", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2818-2826, 2016.

[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", *Proc. Int. Conf. Mach. Learn.*, pp. 448-456, 2015.

[29] H. Jiang, K. Huang and R. Zhang, "Field support vector regression", *Proc. Int. Conf. Neural Inf. Process.*, 2017.

[30] TomHam“Language Recognition Using Deep Neural Networks“ Internet: www.medium.com/coinmonks/language-prediction-using-deep-neural-networks-42eb131444a5.com **[**Aug 9, 2018**]**

[31] L. Ma and Y. Zhang, "Using Word2Vec to process big text data", *Proceeding of IEEE International Conference on Big Data. IEEE*, pp. 2895-2897, 2015.

[32] M. Tang, L. Zhu and X. Zhou, "A text vector representation based on Word2Vec model", *Computer Science*, vol. 43, no. 6, pp. 214-217, 2016.

[33] Hafiz Saeed, Khurram Shahzad and Faisal Kamiran, "Overlapping Toxic Sentiment Classification using Deep Neural Architectures", *International Conference on Data Mining Workshops (ICDMW)*, 2018.

[34] S. V. Georgakopoulos, S. K. Tasoulis, A. G. Vrahatis and V. P. Plagianakos, "Convolutional neural networks for toxic comment classification", *arXiv preprint arXiv:1802.09957*, 2018.

[35] Vikas Chavan and Shylaja SS, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network", *International Conference on Advances in Computing Communications and Informatics (ICACCI)*, 2015.

[36] W. A. Qader, M. M. Ameen and B. I. Ahmed, "An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges," 2019 International Engineering Conference (IEC), Erbil, Iraq, 2019, pp. 200-204.

[37] Y. Li, T. Li and H. Liu, "Recent advances in feature selection and its applications", *Knowledge and Information Systems*, vol. 53, pp. 551-577, 2017.

[38] H.K. Kim, H. Kim and S. Cho, "Bag-of-concepts: Comprehending document representation through clustering words in distributed representation", *Neurocomputing*, vol. 266, pp. 336-352, 2017.

[39] N. Passalis and A. Tefas, "Learning bag-of-features pooling for deep convolutional neural networks", *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5755-5763, 2017.

[40] J. Gaur, A. Goyal, T. Choudhury and S. Sabitha, "A walk through of software testing techniques," 2016 International Conference System Modeling& Advancement in Research Trends (SMART), Moradabad, 2016, pp. 103-108.

[41] P. Ratadiya and D. Mishra, "An Attention Ensemble Based Approach for Multilabel Profanity Detection," 2019 International Conference on Data Mining Workshops (ICDMW), Beijing, China, 2019, pp. 544-550.

[42] Z. Zhang, D. Robinson and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network", *European Semantic Web Conference*, pp. 745760, 2018.

[43] S. Lai, L. Xu, K. Liu and J. Zhao, "Recurrent convolutional neural networks for text classification", *AAAI*, vol. 333, pp. 22672273, 2015.

[44] X. Zhang, J. Zhao and Y. LeCun, "Character-level convolutional networks for text classification", *Advances in neural information processing systems*, pp. 649657, 2015.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, et al., "Attention is All you Need", *Advances in Neural Information Processing Systems(NIPS)*, vol. 30, 2017.

[46] T. Dybå and T. Dingsøyr, "Agile Project Management: From Self-Managing Teams to Large-Scale Development," 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering, Florence, 2015, pp. 945-946.

[47] D. Roshan and T. H. Reddy, "Parts of Speech tagging mechanism to unravel positive and negative patterns in an unstructured text document," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018, pp. 1-6.

[48] R. Sharma and P. Kaushik, "Literature survey of statistical, deep and reinforcement learning in natural language processing," 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, 2017, pp. 350-354, doi: 10.1109/CCAA.2017.8229841.

[49] Y. You, H. Noh, J. Park, Y. Kim, Y. KwaK and Y. Kim, "A development of a speech data transcription tool for building a spoken corpus," 2018 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, 2018, pp. 1437-1439, doi: 10.1109/ICTC.2018.8539450.

[50] ÖmerUludag, Martin Kleehaus, Christoph Caprano, Florian Matthes, "Identifying and Structuring Challenges in Large-Scale Agile Development Based on a Structured Literature Review", *Enterprise Distributed Object Computing Conference (EDOC) 2018 IEEE 22nd International*, pp. 191-197, 2018.

CERTIFICATE
OF PUBLICATION

International Journal of Innovative Research in Computer and Communication Engineering

Website: www.ijircce.com Email: ijircce@gmail.com

This is hereby Awarding this Certificate to

MADHUSHREE M

PG Student, Department of MCA, RV College of Engineering, Bangalore, India

Published a paper entitled

Overview of Differential Privacy in Machine Learning Algorithms

in IJIRCCE, Volume 8, Issue 4, April 2020

e-ISSN: 2320-9801
p-ISSN: 2320-9798

INTERNATIONAL STANDARD SERIAL NUMBER NEPAL

INNO SPACE SJIF Scientific Journal Impact Factor

Impact Factor 7.488

Editor-in-Chief

ORIGINALITY REPORT

16%	10%	13%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Pratik Ratadiya, Deepak Mishra. "An Attention Ensemble Based Approach for Multilabel Profanity Detection", 2019 International Conference on Data Mining Workshops (ICDMW), 2019 Publication	1%
2	Submitted to University of Sheffield Student Paper	1%
3	Habib Ullah, Zahid Ullah, Shahid Maqsood, Abdul Hafeez. "Web Scraper Revealing Trends of Target Products and New Insights in Online Shopping Websites", International Journal of Advanced Computer Science and Applications, 2018 Publication	1%
4	Submitted to University of Queensland Student Paper	1%
5	Rabiyatou Diouf, Edouard Ngor Sarr, Ousmane Sall, Babiga Birregah, Mamadou Bousso, Seny Ndiaye Mbaye. "Web Scraping: State-of-the-Art	<1%

and Areas of Application", 2019 IEEE International Conference on Big Data (Big Data), 2019
Publication

---

6 www.osapublishing.org <1%
Internet Source

---

7 Submitted to University of Southampton <1%
Student Paper

---

8 Wisam A. Qader, Musa M. Ameen, Bilal I. Ahmed. "An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges", 2019 International Engineering Conference (IEC), 2019 <1%
Publication

---

9 Chunzi Wu, Bai Wang. "Extracting Topics Based on Word2Vec and Improved Jaccard Similarity Coefficient", 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC), 2017 <1%
Publication

---

10 Muhammad Ajmal, Farooq Ahmad, Mudasser Naseer, Mona Jamjoom. "Recognizing Human Activities From Video Using Weakly Supervised Contextual Features", IEEE Access, 2019 <1%
Publication

Submitted to Victorian Institute of Technology

11	Student Paper	<1%
12	"Implementation of K-Means Technique in Data Mining to Cluster Researchers Google Scholar Profile", International Journal of Engineering and Advanced Technology, 2019 Publication	<1%
13	section.iaesonline.com Internet Source	<1%
14	Submitted to University of Lancaster Student Paper	<1%
15	Dong Dai, Yong Chen, Philip Carns, John Jenkins, Wei Zhang, Robert Ross. "Managing Rich Metadata in High-Performance Computing Systems Using a Graph Model", IEEE Transactions on Parallel and Distributed Systems, 2019 Publication	<1%
16	Submitted to Kingston University Student Paper	<1%
17	export.arxiv.org Internet Source	<1%
18	astesj.com Internet Source	<1%
19	Jieming Ma, Haochuan Jiang, Ziqiang Bi,	

	Kaizhu Huang, Xingshuo Li, Huiqing Wen. "Maximum Power Point Estimation for Photovoltaic Strings Subjected to Partial Shading Scenarios", IEEE Transactions on Industry Applications, 2019 Publication	<1%
20	ceur-ws.org Internet Source	<1%
21	Submitted to Korea University Student Paper	<1%
22	Submitted to Indian Institute of Science, Bangalore Student Paper	<1%
23	Submitted to Visvesvaraya Technological University Student Paper	<1%
24	Randy Arifanto, Yudistira D.W. Asnar, M.M.Inggriani Liem. "Domain Specific Language for Web Scraper Development", 2018 5th International Conference on Data and Software Engineering (ICoDSE), 2018 Publication	<1%
25	hal.archives-ouvertes.fr Internet Source	<1%
26	Submitted to Glasgow Caledonian University Student Paper	<1%

27	Submitted to Wenatchee Valley College Student Paper	<1%
28	Ziwei Dong, Jun Sun, Jingming Sun, Meiyan Pan. "Marine Weak Moving Target Detection Using Sparse Learning Dictionary", 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), 2019 Publication	<1%
29	Ivan Tsmots, Vasyl Rabyk, Oleg Riznyk, Yurii Kynash. "Method of Synthesis and Practical Realization of Quasi-Barker Codes", 2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT), 2019 Publication	<1%
30	Kaizhu Huang, Haochuan Jiang, Xu-Yao Zhang. "Field Support Vector Machines", IEEE Transactions on Emerging Topics in Computational Intelligence, 2017 Publication	<1%
31	Submitted to University of Kent at Canterbury Student Paper	<1%
32	soar.wichita.edu Internet Source	<1%
33	Ali Selamat, Nicholas Akosu. "Word-length algorithm for language identification of under-	<1%

resourced languages", Journal of King Saud University - Computer and Information Sciences, 2016
Publication

34	Mingyuan Lin, Rundi Zhou, Qing Yan, Xin Xu. "Automatic Pavement Crack Detection Using HMRF-EM Algorithm", 2019 International Conference on Computer, Information and Telecommunication Systems (CITS), 2019 Publication	<1%
35	www.wins.or.kr Internet Source	<1%
36	link.springer.com Internet Source	<1%
37	www.aclweb.org Internet Source	<1%
38	Submitted to National College of Ireland Student Paper	<1%
39	Submitted to The Hong Kong Polytechnic University Student Paper	<1%
40	Submitted to University of Bradford Student Paper	<1%
41	www.ijert.org Internet Source	<1%

42	researchprofiles.herts.ac.uk Internet Source	<1%
43	J. L. Linn. "All example of using pseudofields to eliminate version shuffling in horizontal code compaction", ACM SIGMICRO Newsletter, 8/1/1989 Publication	<1%
44	eprints.hsr.ch Internet Source	<1%
45	journals.plos.org Internet Source	<1%
46	www.businessinsider.in Internet Source	<1%
47	dblp.dagstuhl.de Internet Source	<1%
48	Submitted to HELP UNIVERSITY Student Paper	<1%
49	www.fmcsa.dot.gov Internet Source	<1%
50	abainia.net Internet Source	<1%
51	Submitted to Queen Mary and Westfield College Student Paper	<1%
	Submitted to SVKM International School	

52	Student Paper	<1%
53	Submitted to University of Nottingham Student Paper	<1%
54	circle.ch Internet Source	<1%
55	Submitted to Manipal University Student Paper	<1%
56	Submitted to University of Melbourne Student Paper	<1%
57	Submitted to INTI University College Student Paper	<1%
58	alexandria.tue.nl Internet Source	<1%
59	Submitted to University of Nebraska, Lincoln Student Paper	<1%
60	"ICDSMLA 2019", Springer Science and Business Media LLC, 2020 Publication	<1%
61	D. Roshan, T. Hanumantha Reddy. "Parts of Speech tagging mechanism to unravel positive and negative patterns in an unstructured text document", 2018 International Conference on Computational Techniques, Electronics and	<1%

Mechanical Systems (CTEMS), 2018
Publication

62	Submitted to Glasgow Clyde College Student Paper	<1%
63	Submitted to University of Central England in Birmingham Student Paper	<1%
64	"Advanced Machine Learning Technologies and Applications", Springer Science and Business Media LLC, 2021 Publication	<1%
65	easytra.com Internet Source	<1%
66	"Machine Learning Paradigms", Springer Science and Business Media LLC, 2019 Publication	<1%
67	Zhu, Fang, Zongtian Liu, Juanli Yang, and Ping Zhu. "Chinese event place phrase recognition of emergency event using Maximum Entropy", 2011 IEEE International Conference on Cloud Computing and Intelligence Systems, 2011. Publication	<1%
68	Submitted to Upper Iowa University Student Paper	<1%
69	theses.gla.ac.uk Internet Source	<1%

70	ethesis.nitrkl.ac.in Internet Source	<1%
71	Submitted to Colorado Technical University Online Student Paper	<1%
72	Rui Zhao, Kezhi Mao. "Fuzzy Bag-of-Words Model for Document Representation", IEEE Transactions on Fuzzy Systems, 2018 Publication	<1%
73	Submitted to University of Westminster Student Paper	<1%
74	Submitted to University College London Student Paper	<1%