Cover Page

Automatic Subtitle Generation

Project report submitted for

5th Semester Minor Project-I

in

Department of Computer Science Engineering

By,

Abhilash Pani (16100004) & Akshay Gidwani (16101006)

Department of Computer Science Engineering

Dr. Shyama Prasad Mukherjee

International Institute of Information Technology, Naya Raipur

(A Joint Initiative of Govt. of Chhattisgarh and NTPC)

Email: iiitnr@iiitnr.ac.in, Tel: (0771) 2474040, Web: www.iiitnr.ac.in

CERTIFICATE

This is to certify that the project titled “Automatic Subtitle Generation” by “ABHILASH PANI” & “AKSHAY GIDWANI” has been carried out under my/our supervision and that this work has not been submitted elsewhere for a degree.

(Signature of Guide)

Dr. ANKIT CHAUDHARY

Assistant Professor

Department of CSE

Dr. SPM IIIT-NR

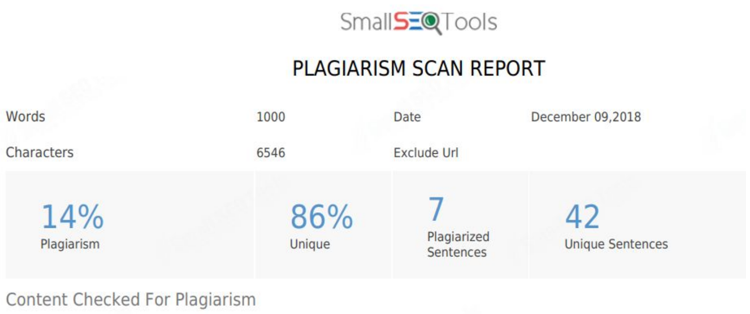
December, 2018-19

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I 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 of Author)	(Signature of Author)
___________________	___________________
Abhilash Pani	**Akshay Gidwani**
(16100004)	**(16101006)**
Date : ____________	**Date : ____________**

PLAGIARISM REPORT



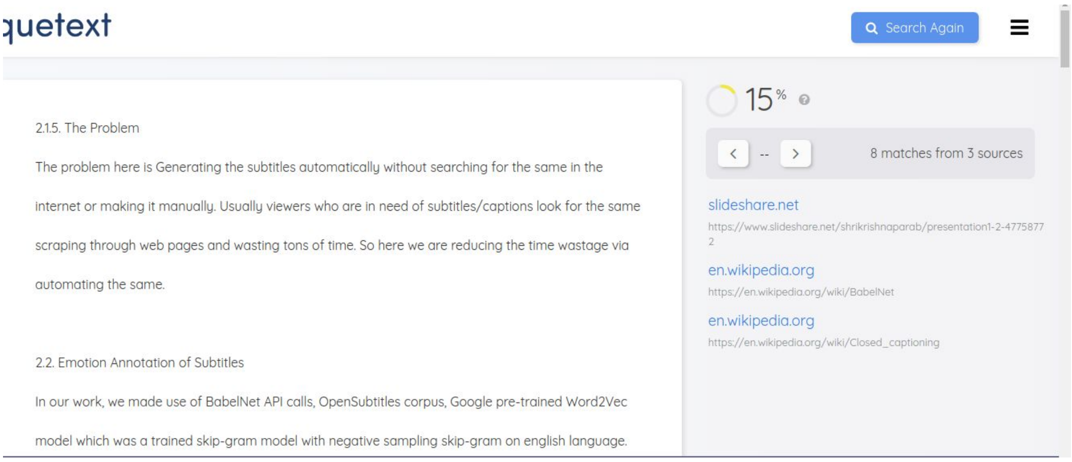
SmallSEOTools

PLAGIARISM SCAN REPORT

Words	1000	Date	December 09,2018
Characters	6546	Exclude Url	

14% Plagiarism	86% Unique	7 Plagiarized Sentences	42 Unique Sentences

Content Checked For Plagiarism



2.1.5. The Problem

The problem here is Generating the subtitles automatically without searching for the same in the internet or making it manually. Usually viewers who are in need of subtitles/captions look for the same scraping through web pages and wasting tons of time. So here we are reducing the time wastage via automating the same.

2.2. Emotion Annotation of Subtitles

In our work, we made use of BabelNet API calls, OpenSubtitles corpus, Google pre-trained Word2Vec model which was a trained skip-gram model with negative sampling skip-gram on english language.

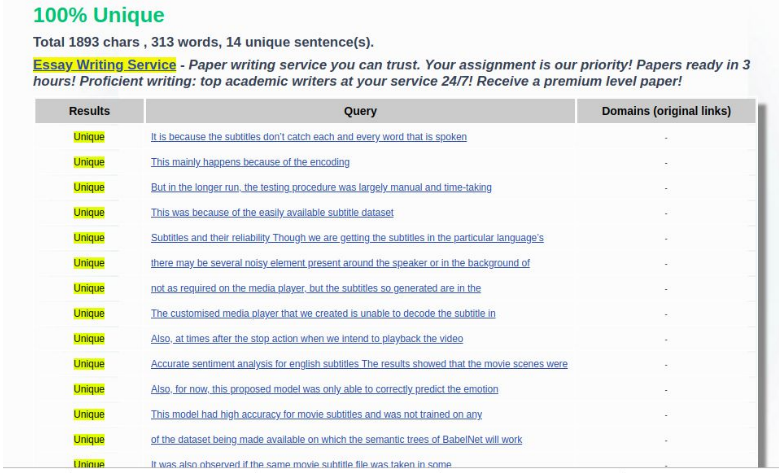
15%

< -- > 8 matches from 3 sources

slideshare.net
https://www.slideshare.net/shrikrishnaparab/presentation1-2-47758772

en.wikipedia.org
https://en.wikipedia.org/wiki/BabelNet

en.wikipedia.org
https://en.wikipedia.org/wiki/Closed_captioning



100% Unique

Total 1893 chars , 313 words, 14 unique sentence(s).

Essay Writing Service - ***Paper writing service you can trust. Your assignment is our priority! Papers ready in 3 hours! Proficient writing: top academic writers at your service 24/7! Receive a premium level paper!***

Results	Query	Domains (original links)
Unique	It is because the subtitles don't catch each and every word that is spoken	-
Unique	This mainly happens because of the encoding	-
Unique	But in the longer run, the testing procedure was largely manual and time-taking	-
Unique	This was because of the easily available subtitle dataset	-
Unique	Subtitles and their reliability Though we are getting the subtitles in the particular language's	-
Unique	there may be several noisy element present around the speaker or in the background of	-
Unique	not as required on the media player, but the subtitles so generated are in the	-
Unique	The customised media player that we created is unable to decode the subtitle in	-
Unique	Also, at times after the stop action when we intend to playback the video	-
Unique	Accurate sentiment analysis for english subtitles The results showed that the movie scenes were	-
Unique	Also, for now, this proposed model was only able to correctly predict the emotion	-
Unique	This model had high accuracy for movie subtitles and was not trained on any	-
Unique	of the dataset being made available on which the semantic trees of BabelNet will work	-
Unique	It was also observed if the same movie subtitle file was taken in some	-

Approval Sheet

This project report entitled “PROJECT TITLE” by “ABHILASH PANI” & “AKSHAY GIDWANI” is approved for 5th Semester Minor Project I.

(Signature of Examiner - I)

____

Name of Examiner -I

(Signature of Examiner - II)

____

Name of Examiner -II

(Signature of Chair)

____

Name of Chair

Date: _____________ Place: _____________

ABSTRACT

Speech recognition is one of the areas of modern day where still a lot has been undiscovered. Speech in its primitive form is just a form of energy but to do research and analysis on this field, machines need to be trained properly and a lot need to be done just to train and perform initial steps of speech recognition and natural language processing.

Therefore, choosing speech recognition as the research topic was the first step in performing the title of the project ‘Automatic Subtitle Generation’ and further implementations such as ‘Emotion Annotation of Subtitle’.

In this implemented project, we generated subtitles for different languages which also included regional languages like Hindi, Bengali, Punjabi, Gujarati etc. , embedded those subtitles inside a built media player and further performed labelling of those subtitles into defined sets of emotion.

Table of Contents

Title **Page No.**

List of All Tables

List of All Figures

CHAPTER 1

INTRODUCTION

1.1. **Speech Recognition**

Speech recognition[10] is the ability of a machine or program to identify words & phrases in spoken language & convert them to a machine-understandable format. Fundamental speech recognition software, generally, has a confined vocabulary of words & phrases, and it may only key out these if they are spoken very clearly & precisely. More sophisticated software has the ability to accept natural speech.

Speech recognition works using algorithms through language & acoustic modeling. Language modeling compares sounds with word sequences to differentiate between words that sound similar; acoustic modeling depicts the relationship among linguistic units of speech & audio signals.

The most common applications of speech recognition include voice dialing(e.g. "call home") call routing(e.g. "I would like to make a collect call"), speech-to-text processing (e.g., word processors or emails), voice search and subtitle generation(e.g. generating subtitles for movies or tutorials).

1.2. **Emotion Annotation of Subtitles**

As the popularity of movie streaming and vast amount of online content available, it is highly important to find the taste of viewers.[1]

This might be the case of the sentiment of speech being delivered by a speaker to a large audience.

Another case could that be of a publication being used for reading.

For all of those cases, there is a need to analyse the sentiment.

Thus, emotion annotation of subtitle for the movie-watching crowd was chosen as the basis for continuing this project as the initial step resulted in generating subtitles for a video.

CHAPTER 2

LITERATURE REVIEW

2.1. **Speech Recognition**

2.1.1. History

Raj Reddy was the 1st person to take on continuous speech recognition as a graduate student at Stanford University in the late 1960s. Previous systems preferred the users to make a pause after completion of each word. Reddy's system was designed to issue spoken commands for the game of chess. Also around this time Soviet researchers invented the dynamic time warping (DTW) algorithm and used it to design a recognizer capable of operating on a 200-word vocabulary. The DTW algorithm processed the speech signal by dividing it into short frames, e.g. 10ms blocks, & processing each block as a single unit. Attaining speaker independence was a major unsolved aim of researchers during this time period.[10]

During the late 1960s Leonard Baum developed the mathematics of Markov chains at the Institute for Defense Analysis. About ten years later, at CMU, Raj Reddy's students James Baker and Janet M. Baker started employing the Hidden Markov Model (HMM) for speech recognition. The use of HMMs enabled researchers to combine variety of sources of knowledge, such as language, acoustics, & syntax, in a consolidated probabilistic model.[10]

2.1.2. Methods, Algorithms & Models

Both language & acoustic modeling are important parts of modern statistical-based speech recognition algorithms. HMMs are widely used in many systems. Language modeling is also used in many other natural language processing(NLP) applications such as document classification or statistical machine translation.

- Dynamic Time Warping (DTW): Dynamic time warping (DTW) is an algorithm for quantifying similarity among two temporal sequences, which may vary in their speed. In general, DTW is a technique that calculates optimal match between 2 given sequences with certain restriction and rules. The optimal match is denoted by that match which satisfies all the restrictions & the rules, which has the minimal cost, where the cost is computed as the sum of absolute differences for each matched pair of indices between their values.[10]
- Hidden Markov Model (HMM): A hidden Markov model may be portrayed as the simplest dynamic Bayesian network. The aim of the algorithm is to evaluate a hidden variable x(t) given a list of observations y(t). By using the Markov property, conditional probability distribution of hidden variable x(t) at time t, given the values of the hidden variable x at all times, depends only on the value of the hidden variable x[t − 1]. In the same way, the value of the observed variable, i.e., y(t) only depends on the value of the hidden variable x(t); both at time t.
- Artificial Neural Networks: Artificial neural network is based on a group of connected units called artificial neurons, which loosely build the neurons in a brain biologically. Each connection can transmit a signal from one neuron to another. An artificial neuron that receives a signal can process it and then signal additional artificial neurons connected to it. In common Artificial neural network(ANN) implementations, the signal at a connection between artificial neurons is a real

number, and the output of each artificial neuron is computed by some non-linear function of the sum of its inputs.[10]

- Deep feedforward & Recurrent Neural Networks: A deep feedforward neural network (DNN) is an artificial neural network with several hidden layers of units among the input & output layers. Similar to simple/shallow neural networks, DNNs can model complex relationships which generally prove out to be non-linear. DNN architectures produce compositional models, where extra layers help composition of features from bottom layers, giving a huge learning power & thus the prospect of modeling complex patterns of speech data.[10]

2.1.3. Major Applications Areas

- In-car systems
- Health Care
- Military
- Telephony and other domains
- Usage in education in daily life
- People with disabilities
- Home Automation and Virtual Assistants
- Closed Captioning(a.k.a Auto-Captioning, i.e. producing subtitles automatically)

2.1.4. Performance

Performance of speech recognition machines is generally calculated in terms of accuracy & speed. Speed is usually rated with the real time factor. A variety of factors can affect computer speech recognition performance, including accent, pronunciation, volume, pitch & background noise. Accuracy is measured using word error rate (WER). WER works at word level & identifies incorrectness in transcription, although it can't recognise how the error took place.

- Accuracy: There can be several factors that affect accuracy such as
    - Vocabulary is difficult to identify if it contains confusable words.
    - Acoustical signals are structured into a hierarchy of units, i.e. words, phrases, sentences, idioms, etc.
    - Individually levels result in additional constraints.
    - Dependency of the system over speaker(s).
    - Task and language restriction
- Security Issues: Speech recognition can become a way of attack, theft, or accidental operation. For e.g., triggering words like 'Alexa' spoken in a broadcast can cause devices in to start listening for input inappropriately, or may take an unwanted action. Voice-controlled devices are also easygoing to visitors to a building, or even those outside the building if they can be heard inside. Attackers may be able to gain access to personal information, like calendar, private messages, and documents.

2.1.5. **The Problem**

The problem here is Generating the subtitles automatically without searching for the same in the internet or making it manually. Usually viewers who are in need of subtitles/captions look for the same scraping through web pages and wasting tons of time. So here we are reducing the time wastage via automating the same.

2.2. **Emotion Annotation of Subtitles**

In our work, we made use of BabelNet API calls, OpenSubtitles corpus, Google pre-trained Word2Vec model which was a trained skip-gram model with negative sampling skip-gram on english language.

Algorithm Description -

2.2.1 Emotion Vectors

To obtain emotion vectors we have used BabelNet to construct Sense Trees with synset as nodes. Each Sense Tree represents one of the eight emotional states ('love' , 'happiness' , 'surprise' , 'emotionless' , 'sad' , 'disgust' , 'anger' , and 'fear') as root nodes. Finally we get Sense Tree with weights proportional to the distance from the root.

2.2.2 Sentence Vectors

For selected movies (or can use the generated subtitles from the described procedure) english subtitles were extracted with the help of OpenSubtitles2016.[3] The sentence vector was calculated by taking the average of the vectors of the words. Finally the closest emotion vector to a sentence vector using cosine distance as a measure. (cosine model available in Word2Vec package).

2.3 Word2Vec

To put in simple words, it is just the numerical/vectorial representation of pure language words for real-life machine learning models to work upon.[4]

2.2.4 BabelNet

It is a multilingual lexicalized semantic network and ontology. It was automatically created by linking wikipedia to the most popular computational lexicon of the English language, WordNet. It can be termed as "encyclopedic dictionary" with huge amounts of semantic relations.[2]

CHAPTER 3

PROPOSED SOLUTION

3.1. **Automatic Subtitle Generation**

Subtitle and Closed captioning (CC)[8] both are processes of displaying text on screen to provide additional interpretive information. Both are typically used as transcriptions of the audio portion of a program as it occurs, sometimes including descriptions of non-speech elements. Other uses have been to provide a textual means of language translation of a presentation's main audio language that is usually burned-in (or "open") to the video. HTML5 interprets subtitles as a "transcription/translation of dialogues ... when sound is present but not perceived" by the viewer (for e..g, dialogue in a foreign language) and captions as a "transcription/translation of dialogues, sound effects, applicable musical cues, & other material audio information ... when sound is unavailable or not clearly audible" (for e.g., when audio is muted or the viewer is deaf /hard of hearing).[7]

The solution to our problem for generating subtitles for leisure use where subtitle searching in the web is a heavy task for a infrequent internet user, our system generates the subtitles with just few clicks of buttons on the media player for the video.

Here, we have build a personalised media player using python programming language and the python-vlc[6] module of the VLC Media Player. The Media Player has 4 buttons - Playback-toggle, Stop, Hindi, English. The playback-toggle button avails the user with the facility of pausing/playing a video, while the stop button just stops the running video. The 'Hindi' and 'English' buttons generates the subtitle of the video in their respective scripts.

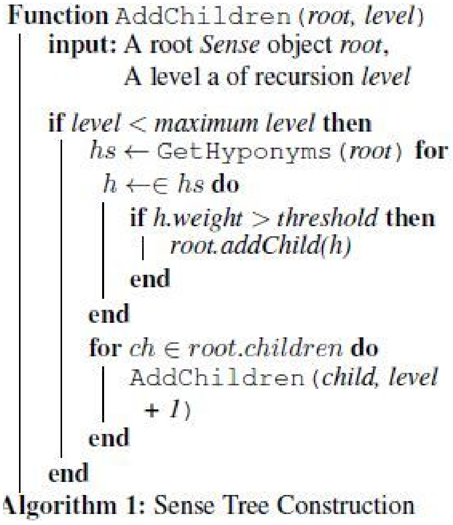
The subtitles are generated using the Autosub utility of python and google's resources(database, hardware machines, etc.). Google's speech recognition[9] is powered by Machine Learning where it applies the most advanced form of Deep-Neural Network to audio. Autosub uses Google API for speech recognition and hence produces timed-texts(i.e., storing formats for subtitles). With a fully paid google cloud service, this utility can be further harnessed to generate subtitles in different languages, i.e., for a video that is burned-in a language say english the subtitles can be generated in a translated form to another language say hindi, with the required scripts of that particular language. The utility can only convert one-to-one subtitles, i.e. for a particular language in can convert the speech into a script of that particular language only precisely, if the user uses free google-api-client.

3.2. **Automatic multilingual emotion annotation of subtitles**

In this proposed solution of ours, we take advantage of semantic networks provided by BabelNet as well as the distributed word representations trained by skip-gram model with negative sampling in an effort to obtain vector representations of 8 emotional states.

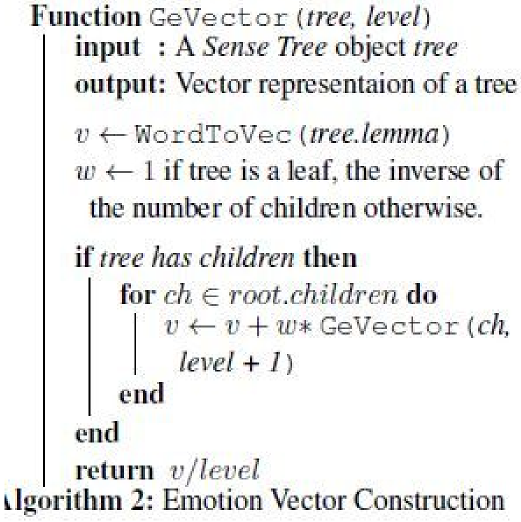
After that, we employed above mentioned emotion vectors to identify the emotion content scenes in various movies of different languages.

The pseudo codes for construction of Sense Tree and Emotion Vector are shown as-



```
Function AddChildren (root, level)
    input: A root Sense object root,
           A level a of recursion level
    if $level < maximum\ level$ then
        $hs \leftarrow$ GetHyponyms (root) for
        $h \leftarrow\in hs$ do
            if $h.weight > threshold$ then
                root.addChild(h)
            end
        end
        for $ch \in root.children$ do
            AddChildren (child, level + 1)
        end
    end
Algorithm 1: Sense Tree Construction
```

Fig. 1 Pseudo Code for Construction of Sense Tree



```
Function GeVector (tree, level)
    input  : A Sense Tree object tree
    output: Vector representaion of a tree
    $v \leftarrow$ WordToVec (tree.lemma)
    $w \leftarrow 1$ if tree is a leaf, the inverse of
     the number of children otherwise.
    if tree has children then
        for $ch \in root.children$ do
            $v \leftarrow v + w*$ GeVector (ch, level + 1)
        end
    end
    return $v/level$
Algorithm 2: Emotion Vector Construction
```

Fig. 2 Pseudo Code for Construction of Emotion Vector

CHAPTER 4

RESULTS

4.1. **Subtitles for the videos**

Following are the screenshots of the results obtained by following the above mentioned procedures:

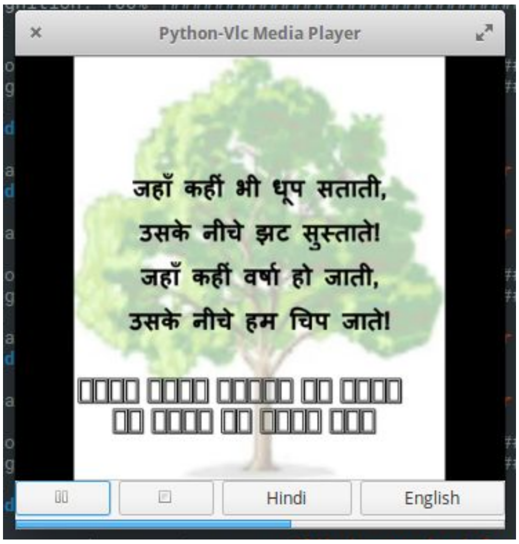


Fig. 3 Screenshot of the Hindi Subtitles on Media Player

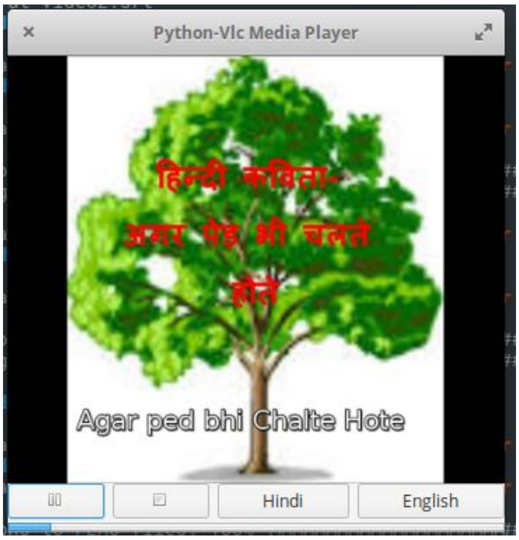
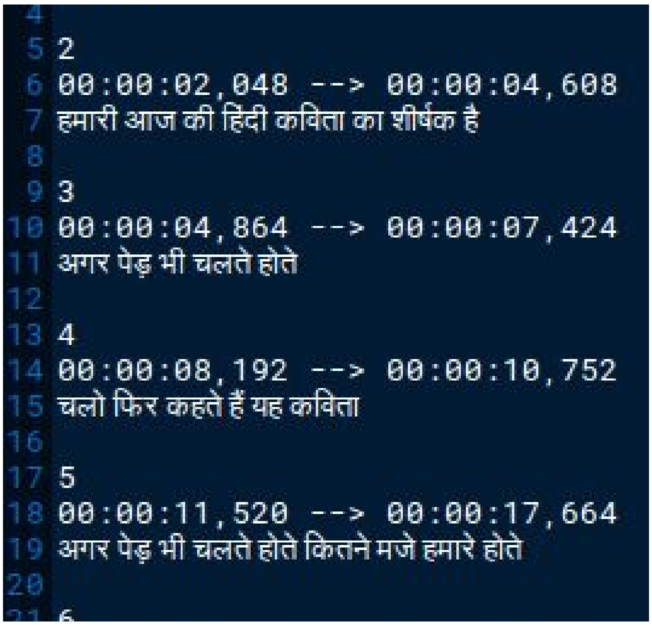


Fig. 4 Screenshot of the English Subtitles on Media Player



```
2
00:00:02,048 --> 00:00:04,608
हमारी आज की हिंदी कविता का शीर्षक है

3
00:00:04,864 --> 00:00:07,424
अगर पेड़ भी चलते होते

4
00:00:08,192 --> 00:00:10,752
चलो फिर कहते हैं यह कविता

5
00:00:11,520 --> 00:00:17,664
अगर पेड़ भी चलते होते कितने मजे हमारे होते

6
```

Fig. 5 Screenshot of the generated Hindi Subtitles

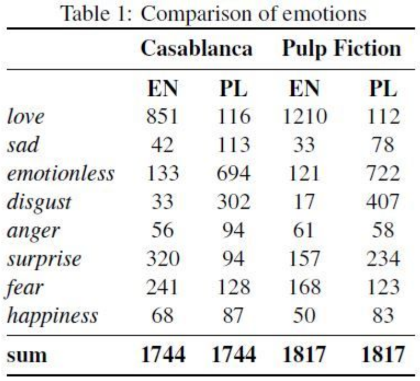
4.2. **Plots and Emotions Table of Subtitle**

Obviously labeled (8-emotion labelled subtitles) subtitles were not publicly available, we decided to check the results with the help of visual plots and weightage of each emotion vector in a movie subtitle file carefully compiled into a table.

The emotion-weightage table was calculated for two movies - Casablanca & Pulp Fiction while plots were drawn for three movies - 12 Angry Men, Casablanca & Pulp Fiction.

Plots and table are as shown -

Table 1: Comparison of emotions



	Casablanca		**Pulp Fiction**	
	EN	**PL**	**EN**	**PL**
love	851	116	1210	112
sad	42	113	33	78
emotionless	133	694	121	722
disgust	33	302	17	407
anger	56	94	61	58
surprise	320	94	157	234
fear	241	128	168	123
happiness	68	87	50	83
sum	**1744**	**1744**	**1817**	**1817**

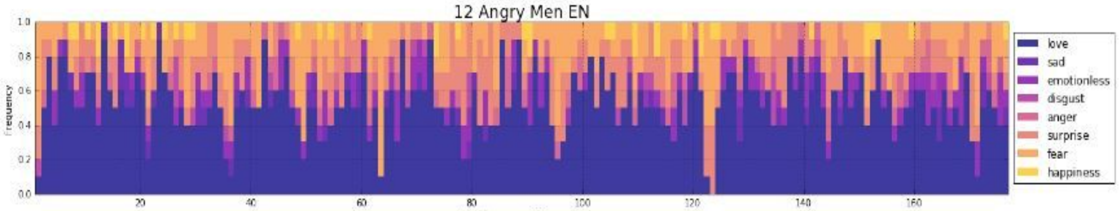


Fig. 6 Plot for Angry Men Movie

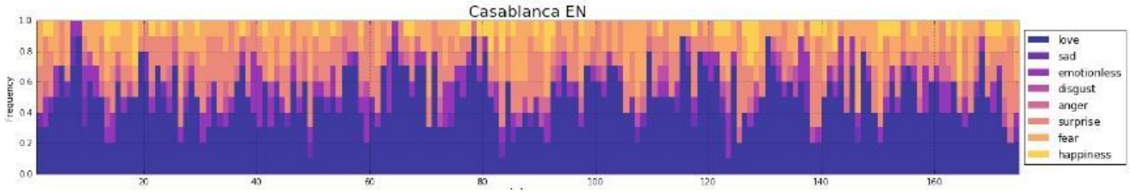


Fig. 7 Plot for Casablanca Movie

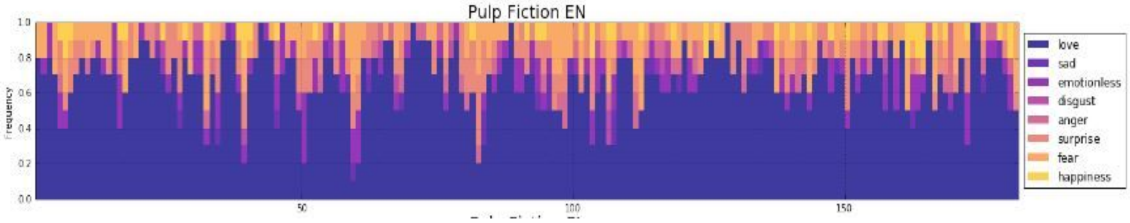


Fig. 8 Plot for Pulp Fiction Movie

CHAPTER 5

CONCLUSIONS & FURTHER WORK

5.1. **Subtitles and their reliability**

Though we are getting the subtitles in the particular language's scripts but we can not fully be dependable on them. It is because the subtitles don't catch each and every word that is spoken.

Being unable to catch every word the possibilities are that it may not be clearly spoken-off, the accent, pronunciation may differ, the speed at which the sentences/dialogues are delivered, there may be several noisy element present around the speaker or in the background of the speaker.

5.2. **Media Player defects**

Though the results obtained for the languages other than english are not as required on the media player, but the subtitles so generated are in the required scripts. This mainly happens because of the encoding. The customised media player that we created is unable to decode the subtitle in required format and hence we are unable to recognise the same.

Also, at times after the stop action when we intend to playback the video the video doesn't play back and we need to restart the application.

5.3. **Accurate sentiment analysis for english subtitles**

The results showed that the movie scenes were almost correctly able to identify emotion content. But in the longer run, the testing procedure was largely manual and time-taking. Also, for now, this proposed model was only able to correctly predict the emotion content for movies only. This was because of the easily available subtitle dataset. This model had high accuracy for movie subtitles and was not trained on any random video.

5.4. **Extension to other regional language subtitles**

The proposed solution largely depends on the precision of the dataset being made available on which the semantic trees of BabelNet will work upon. It was also observed if the same movie subtitle file was taken in some other language, the majority of the emotion state was different as compared for english language.

REFERENCES

Journal

1. Wojciech Stokowiec, "Whose line is it anyway? Automatic Multilingual Emotion Annotation of Movie Dialogue" , National Information Processing Institute
2. Roberto Navigli and Simone Paolo Ponzetto, "BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network". in: Artificial Intelligence 193 (2012) pp. 217–250.
3. Jog Tiedemann. "Parallel Data, Tools and Interfaces in OPUS". In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). May 2012.
4. Tomas Mikolov et al. "Efficient Estimation of Word Representations in Vector Space" in:CoRR abs/1301.3781 (2013).

URL

5. https://github.com/1o0ko/Emotion-vectors
6. https://wiki.videolan.org/python_bindings
7. https://support.google.com/youtube/answer/2734796?hl=en
8. https://en.wikipedia.org/wiki/Closed_captioning
9. https://cloud.google.com/speech-to-text/
10. https://en.wikipedia.org/wiki/Speech_recognition