[[1]](#footnote-0)

VIDEO: Translation, Sentiment Analysis & Summarisation

**Shankar R, Faisal Mukhtar, Laalithya A., Manvi Mehar N., Parth Jaiswal**

***Abstract*: *Data is the new oil. The proliferation of the internet and the development of technology making it available to people all over the world. This massive data is made available at our disposal, mostly free of cost. But what is more important is how this easily available data is used by the people. Language plays a very crucial role in the communication of data to the other end. Many websites and platforms have started supporting multiple languages and are replacing the convention of using English. This is to take the information to every corner of the world in this multilingual community. Information or data has no meaning if it can’t be understood and used efficiently by others. Hence, taking the information to the public in a language that they understand becomes essential. Education is vital in a person’s life no matter who the person is. The literacy rate in India is 74.04%, which means nearly 26% of the population can’t read and write in at least one language. YouTube is one of the most used platforms by people around the world, this is due to the seeping of technology and mobiles into the grassroots level. It is an entertainer for some and an earning job for others. This platform, which is widely used can be converted into an education platform by providing subtitles in native languages. And here comes the need for language translation into regional languages like Kannada. This huge data and information have no meaning if the humongous opinions generated aren’t used for the betterment. Today’s world is an open world where people don’t step back to express themselves and put their opinion across. And this valuable opinion has no value if it’s not analyzed and used for the betterment. That’s where Sentiment Analysis plays the role. It is here that the various opinions and sentiments of the people are analyzed. Lastly summarisation. Anyone would prefer a summary of a long video instead of wasting time watching it, to get a gist of the video. Encapsulation of huge information is always welcome.***

***Keywords* : *Transliteration, corpus, data-mining and subtitles.***

# **INTRODUCTION**

The project focuses on 3 different subjects, namely, Translation, Sentiment Analysis and Summarization.

Natural Language Processing(NLP) is the process of reading and understanding of spoken or written language through a machine. Specifically speaking, it reads, deciphers, understands and makes sense of the human languages.

Sentiment Analysis is the method of identifying and categorizing the expressed opinions in the given text, speech or a video. The opinions can be positive, negative or neutral in case of a text, or it can be happy, sad, anger, etc, in case of a video.

Summarization is the process of creating a condensed version of a large amount of data, in a meaningful way, covering only the main points of the data. TextRank Algorithm is a ranking algorithm which is used to summarise the given data, by computing the relation between sentences in the text.

# **LITERATURE SURVEY**

Translation: The authors of [2] have used a publicly available Support Vector Machine(SVM) for transliteration from English to Kannada. SVM is a trained machine learning algorithm that is mainly used for categorising with minimal margin. It belongs to a group of binary classifiers where the data is split into 2 classes. The transliteration takes place in 3 phases. The first is the pre-processing phase where the given file into a required format for the SVM. This stage consists of Romanisation, Segmentation, and Alignment where the English alphabets are converted to lower case first and then segmented and aligned by counting the number of units. The second stage is the training phase where the SVM is trained for the given input file. The last stage is the transliteration stage where the English words are mapped to the most anticipated Kannada word. The accuracy of this is 69%.

Sentiment Analysis: [4] uses a machine learning model that uses the Naive Bayes algorithm and takes feature extraction, tokenization, sentiment classification, and Parts of Speech(POS) tagging. In the initial preprocessing step, all the words do not give any meaning called the stop words are removed. Next is the POS tagging where the words are tagged with their part of the speech. In the next step, feature

extraction takes place which is done using the term frequency which gives the frequency of the words. Finally,

the opinion word extraction is done where the sentiment analysis or opinion mining is obtained for the text. The polarity of the words is used to analyze the sentiment.

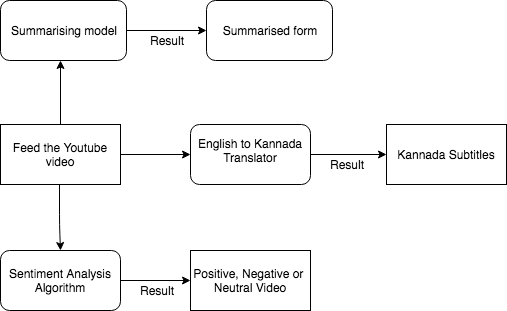
The proposed system in [5] works mainly on corpuses. Here mainly 2 corpuses are created: 1 is the positive corpus which contains positive words and the other is the negative corpus which contains negative words. The corpuses are collected from various Kannada sources like Wikipedia, blogs, journals, etc. The sentence for which the sentiment is to be found is split into words and this is called tokenization. Each word is compared for existence in the available corpus. If the number of positive corpus words is greater than the negative word corpus then the sentence is classified as positive, else negative and if the number of positive corpus is equal to the negative corpus, it is considered as a neutral sentence.

Summarisation: In [12], a text-mining based algorithm (TextRank) is used to perform automatic summarization which utilizes a weighted undirected graph to represent the sentences in an article, and uses Google PageRank to order the sentences. TextRank is an algorithm that uses graphs to connect the various sentences in a given document and the dependence of each sentence on the other by analysing the frequently used words and connecting them.

Another method for summarization is by creating vector representation of the document which is attempted in [14]. This process takes place in 3 steps: Sentence vector representations, Projecting documents into semantic sentence spaces, Sentence extraction using synthetic document vector. In the first step, each sentence is represented as a vector splitting the document and mapping them to a matrix for representation. Then, each line is projected into the semantic space to understand the sentence contains core and important information. Finally, the important sentences are extracted measuring the similarity with the Synthetic Document Vector which is created by taking the mean of individual document vectors. The sentences are ranked, the basis for which is their rank and the selected sequentially using their rank.

# **PROPOSED METHODOLOGY**

Given any video with English audio and the corresponding Kannada subtitles: Firstly, the dataset given to the machine as input for translation to the Kannada language is the subtitles of the video. The output from the machine is then merged with the corresponding audio of the video and displayed. Secondly, the video is closely analysed according to its audio and video, and hence deciding the sentiment of the overall video as happy, sad, angry,etc,. Lastly, using the TextRank algorithm, the video’s subtitles are summarised in a few lines.

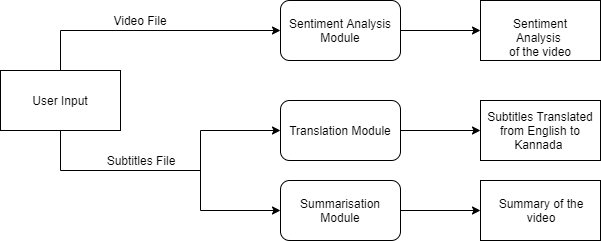


**A. Translation**: Given a text (subtitles) file, the translation sub-system converts it into kannada. This is done using Google API.

**B. Sentiment Analysis (SA):** Given the video with its corresponding audio, the SA module provides the closest sentiment that matches the input. In this case, the outputs for each input were, neutral, happiness, sadness, anger, fear, surprise and disgust. A score is given to each of the feelings when they were expressed. The feeling which was reciprocated the most throughout the video, i.e., the sentiment with the highest score is decided to be the sentiment analysed for the video.

**C. Summarization:** Firstly, the text (subtitles) provided to the sub-system is split into sentences and word embeddings for each of the sentences is created.. Before processing this text, the text is cleaned, i.e, stopwords like is, am, the ,of ,etc,. From the previously created word embeddings, we create vectors for the sentences, i.e., vectors for each word in a sentence are calculated and then the mean/average of those vectors is taken to arrive at a consolidated vector for the sentence. Then, a likeness network is made (utilizing cosine) considering the similitudes between each sentence, i.e., we use cosine similarity module to compute similarity between a couple of sentences and initialize the the matrix with cosine similarity scores. This matrix is then converted into a graph, where the nodes speak to the sentences and the edges speak to the closeness score between the sentences. This text is ranked based on the similarity score, i.e., high similarity score corresponds to higher rank. Based on this ranking, the top ranked sentences are extracted for summary generation.

# **SYSTEM DESIGN**



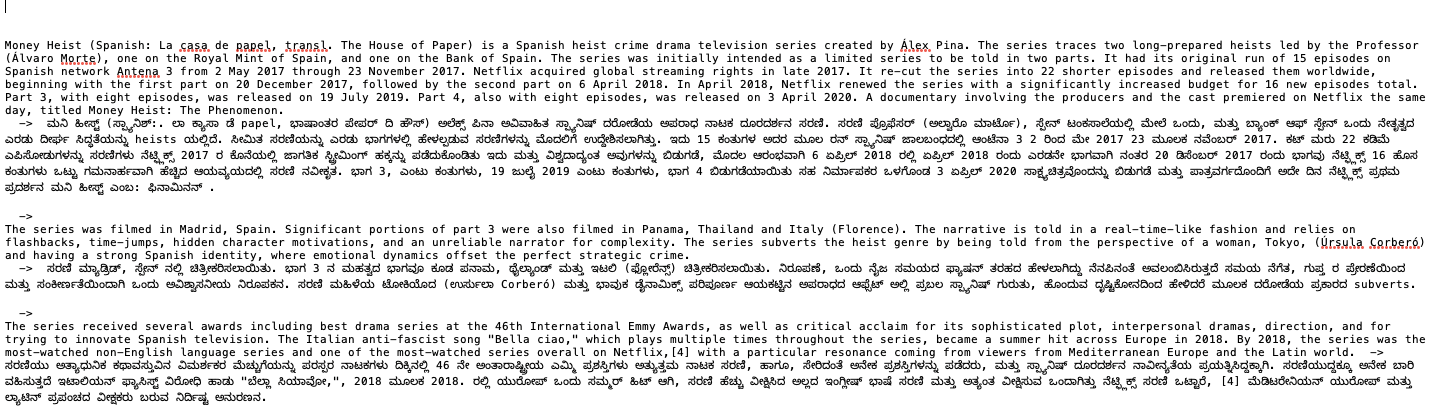
The system mainly has 3 modules: Sentiment Analysis Module, Translation Module, and Summarisation Module. It takes 2 inputs from users. They are the video file and the subtitles file. The Sentiment Analysis Module performs SA on the video file. After analyzing the various emotions and sentiments in the video, it gives the percentages of the different emotions in terms of percentage. The emotion with the highest percentage indicates the emotion of the video as to whether it is positive, negative, or neutral. The Translation Module works on the subtitles file which is given as input by the user. This module mainly translates the subtitles file which is fed in English into Kannada using a Google Translator. Finally, the Summarisation Module is used to summarise the video and the input to this is the subtitles file that is fed in by the user. This module works on the Text Rank Algorithm to summarise the video and give the summarised format to the user.

# **RESULTS & DISCUSSIONS**

The results from each sub-system were as follows:

**A. Translation**:

For translation, we have used the most used translator: Google Translator. Google Translator is the most accurate of all the available translators. BLEU (bilingual assessment understudy) is a calculation which is utilized for assessing the nature of text which has been machine-interpreted starting with one common language then onto the next. The main idea behind BLEU is "the closer a machine translation is to a professional human translation, the better it is". Google Translator has a BLUE score of more than 0.5.



**B. Sentiment Analysis (SA):**

We have tested our modules for multiple input files (.mp4) for assessing the accuracy of our modules. The following table represents the confusion matrix.

|  |  |  |
| --- | --- | --- |
| **SA** | **True** | **False** |
| **Positive** | True Positive: 9 | False Positive: 3 |
| **Negative** | True Negative: 7 | False Negative: 3 |

Positive = (Happy, Surprise, Neutral)

Negative=(Anger, Disgust, Sadness, Fear)

Accuracy = (TP + TN)/(TP + TN + FP + FN)

= (9+7)/(9+7+3+3) = 0.727

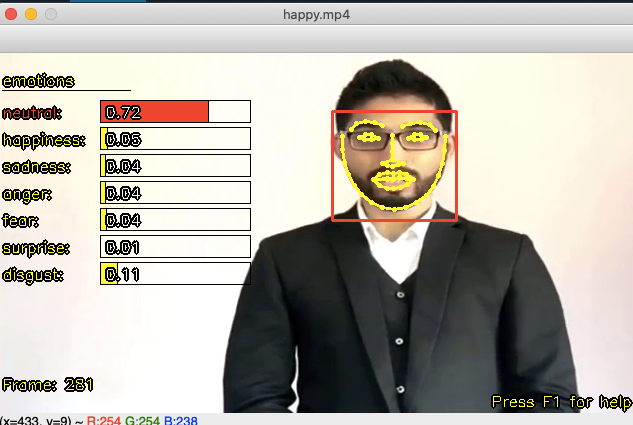
Precision = TP/(TP+FN)

= 9/(9+3) = 0.75

Recall = TP/(TP+FN)

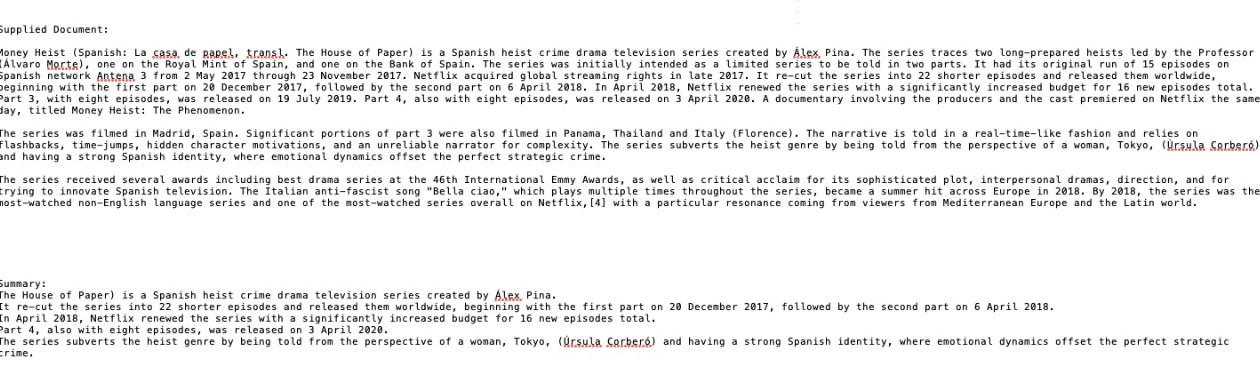
= 9/(9+3) = 0.75

F-measure = 2 \* Precision \* Recall/(Precision + Recall)

= 2 \* 0.75 \* 0.75/(0.75 +0.75) = 0.75 **C. Summarization:**

Comparison of TextRank algorithm with RAKE algorithm using [18] as reference.

|  |  |  |
| --- | --- | --- |
| Measure | TextRank Algorithm | RAKE algorithm |
| Accuracy | 0.8207 | 0.7698 |
| Precision | 0.4234 | 0.3815 |
| Recall | 0.7436 | 0.8800 |
| F measure | 0.5395 | 0.5323 |
| Error Rate | 0.1793 | 0.2302 |



# **CONCLUSION & FUTURE SCOPE**

Translation plays a significantly important role in pulling masses towards the boundless information available online in a language which cannot be interpreted by them. In future, the concept can be further implemented by the usage of different languages for translation and not just from English to Kannada.

The enormous amount of user activity online, where plenty of people share their opinion everyday has led to the increase in the usage of sentiment analysis on such information. Sentiment Analysis is mostly performed in language English due to its immense amount of resources, hence, a development on the same could be seen in the future where it can be performed in other languages with lesser resources.

The sentiment analysis system developed here is based on the user’s expressions and not on the content the user is narrating. This mostly works when there is one human figure in the video and fails when there are multiple faces in the video.

With increasing amount of online information and other resource text, text summarization is implemented to preserve and show the main purpose of textual information. Since it is impossible to manually summarise large documents of text, the TextRank algorithm is used to ease the process. Although TextRank algorithm is a better option available, it is still naive and it needs future development to be better for summarisation.

**ACKNOWLEDGMENT**

The authors of this paper are thankful to the editor and reviewing committee for their invaluable advice which significantly improved the paper. Students are grateful to Professor Shankar R, Assistant Professor, Department of CSE, for being a pillar and giving us the mind-boggling backing and direction to compose this paper. Finally, thanking parents, friends, and our Principal, HOD, Teaching and Non- Teaching faculties of Department of CSE, B M S Institute of Technology & Management, Bangalore.

1. **Revised Manuscript Received on July 22, 2019**.

   **\*** Correspondence Author

   **Shankar R.\***, Department of CSE, BMSIT&M, Bangalore, India. Email: [shankar@bmsit.in](mailto:shankar@bmsit.in)

   **Faisal Mukhtar**, Department of CSE, BMSIT&M, Bangalore, India. Email: [1by16cs032@bmsit.in](mailto:1by16cs032@bmsit.in)

   **Laalithya A.**, Department of CSE, BMSIT&M, Bangalore, India. Email: [1by16cs043@bmsit.in](mailto:1by16cs043@bmsit.in)

   **Manvi Mehar N.**, Department of CSE, BMSIT&M, Bangalore, India. Email: [1by16cs046@bmsit.in](mailto:1by16cs046@bmsit.in)

   **Parth Jaiswal**, Department of CSE, BMSIT&M, Bangalore, India. Email: [1by16cs056@bmsit.in](mailto:1by16cs056@bmsit.in) [↑](#footnote-ref-0)