Kavi-Kannada Natural Language Processing System

Prasanna Kumar S Shivaraddi[1], Rudresha G T[2], Manoj H[3], Prajwal[4], Nagarjuna G[5]
Assistant Professor, Department of Computer Science[1]
Students, Department of Computer Science[2,3,4,5]
Rao Bahadur Y Mahabaleswarappa Engineering College, Bellary, Karnataka, India
rudresha.cse.rymec@gmail.com, manojh.cse.rymec@gmail.com,
prajwal.cse.rymec@gmail.com, nagarjuna.cse.rymec@gmail.com

Abstract: *Southwest Indian state of Karnataka is home to the majority of speakers of the ancient Dravidian language known as Kannada. The word models embeddings are employed in natural language processing for upcoming Indic NLP work. We convert Kannada voice data to Kannada text data using machine learning techniques and a specific analogy and similarity challenge, and then access our models for the subsequent task of text classification.*

Keywords: Natural Language Processing

I. INTRODUCTION

Even if they are able to speak Kannada, we are currently noticing that most Karnataka regions use other languages for business purposes despite the fact that they are capable of doing so. The illiterates are able to speak Kannada, but they are unable to compose sentences in Kannada for any of their employment-related purposes. Not just Kannadigas but even outsiders from other states have the same issue, hence they continue to do business in their own tongue. So even if they are able to speak Kannada, they will have trouble in writing the statement if they wish to send someone a conversation or message in Kannada.

A subfield of artificial intelligence called "natural language processing" is designed to receive, process, and modify human language. It is beneficial to simulate how the human mind processes natural language. Text and speech in natural language are recognized using NLP. In order to create an application that helps people better understand the information they encounter every day, mathematical, computational, and linguistic skills must be merged. Sentiment analysis, speech recognition, and other applications all employ NLP. Machine learning algorithm application. Therefore, by utilizing NLP toolkits, we are creating a programme that will print what we are providing to our system in Kannada. By doing so, it will assist in resolving the aforementioned issue by publishing the Kannada phrases supplied by the users. Illiterates and foreigners from Karnataka who desire to write in Kannada can benefit from this.

Textual papers have long been a crucial part of corporate procedures in general. These days, an increasing number of these documents are available online. This necessitates efficient means of automating their processing. Information retrieval is used to address the pertinent research issues (IR). The primary task at hand is discovering papers that meet the user's informational requirements. This necessitates accurate document representation, information demands (often expressed as a query), and suitable methods for matching them to one another. The automatic classification of texts into a number of specified categories, commonly referred to as text categorization, is another vital activity [1. The interest categories are frequently of a topical nature; hence, two papers fall under the same category if they are thematically connected. For the sake of a journal's editorial office, for instance, news stories may be categorized in this fashion, and materials provided by websites may be arranged according to their key subject areas. The practical uses of the text classification paradigm, however, may potentially contain other sources for the categories. For instance, poems can be categorized by their writers, and papers that a corporation receives can be categorized by to their type (ads, analyses, reports etc.), or according to Kannada language etc.

II. LITERATURE SURVEY

Natural Language Processing is an area under artificial Intelligence intended to accept process and manipulate the human language. It helps to model cognitive behavior of human mind to process natural language. The NLP is used to identify the natural language text and speech.

DOI: 10.48175/IJARSCT-5837

The primary mode of human-to-human communication is natural language and audio channels, although limited use between human and system. A Man can talk with his computer in the same way that with a friend which can be done through software tools. The presented work solely concentrated towards the development of a simple human-computer interaction system in Kannada language. Speech pattern recognition in the present study is related to Kannada language. Text classifications have made great performances in the NLP area. In the past few years, there were a lot of previous works about text classifications.

Recent years, more and more people have realized the importance of text classification. For example, text classification can help us work more efficiently and save a lot of times such as library classification, news classification, and so on. Therefore, there are much more applications of text classification in specific areas.

Alternatively, researchers put their efforts on classification techniques and analyze its advantages and disadvantages. Thanks to this research on text classification, there are much more applications of text classification in our life. Such as news classification, space knowledge management, emotional analysis, and commodity evaluation.

III. ARCHITECTURE DIAGRAM

In this voice-text design architecture (Fig.1), we are giving voice input to the system for the process. This process requires two specific libraries those are speech recognition and Py-audio etc. By recognizing the voice input it will provide us the output which is in text format in Kannada language.

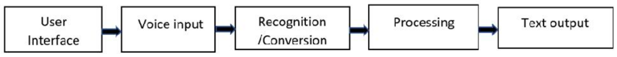


Fig 1. System architecture

In this text classification design architecture (Fig.2), large text documents have been given as input and given document has the preprocessed data set. Then feature extraction takes place in this feature extraction we are using machine learning algorithm and it learns the trained data set to produce the test data.

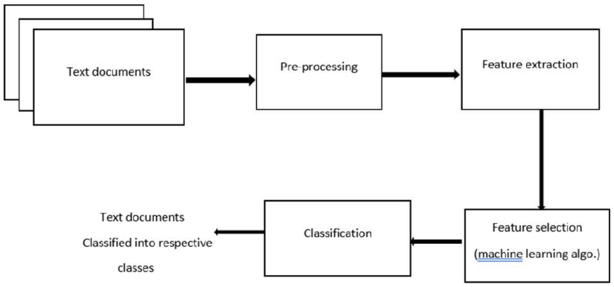


Fig 2. System architecture

IV. PROPOSED SYSTEM

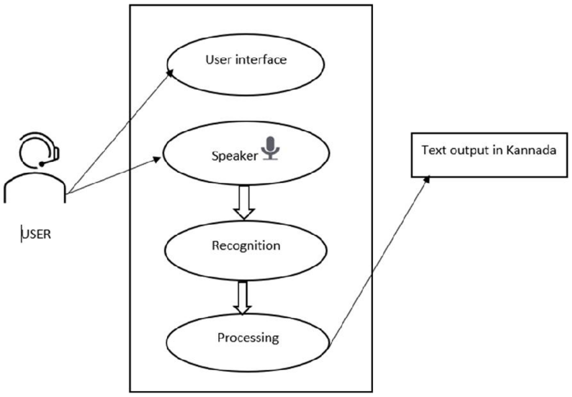
Our method offers a workable option to those who are unable to write Kannada by producing text in Kannada using voice as the input.

It also aids in communication is easier - no more difficult-to-read handwriting quick turnaround on documents, the ability to work from home or outside of the workplace, time saved through improved productivity and less paperwork, The time it takes to type or create a text can be cut in half by using speech recognition software.

When it comes to difficult NLP categorization jobs, machine learning is typically significantly more accurate than rule systems created by humans. Additionally, machine learning classifiers are simpler to maintain, and you can always tag fresh instances to pick up new skills. However, under the three categories of our recently adopted recommended system, they are

V. METHODOLOGY

5.1 Use Case Diagram



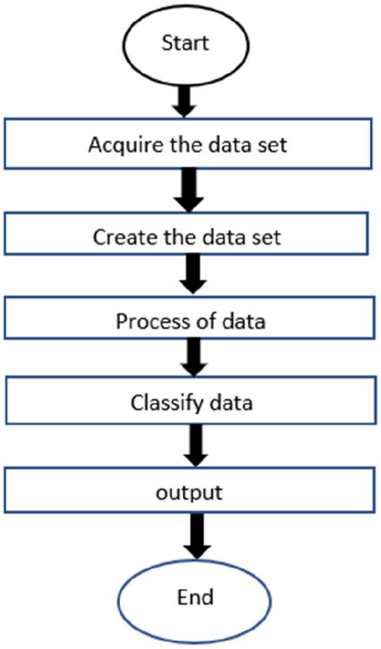
A use case diagram is a way to summarize details of a system and the users within that system. It is generally shown as a graphic depiction of interactions among different elements in a system.
User interface provides microphone logo to speak, after providing voice input it will recognize speech using some standard libraries then conversion process takes place finally it provide the text format in Kannada that we are giving as voice input and also it showing options such as copy, notepad, word document it will save on that particular folder.

5.2 Speech Recognition Module

Speech recognition helps the system to recognize the voice, which is given as input, and enables a program to process the speech into written format.

5.3 Py-Audio Module

Py-audio provides python binding for port, the cross-platform audio input/output library. We can easily use python to play and record the audio on a various platform.



DOI: 10.48175/IJARSCT-5837

In this text classification process, we will be acquiring a data through the web server or the training data. After acquiring the data we create the data that data will be having the content which is based on sports, entertainment etc.
All the data which we have in the data that will processed on the basis of the given input.
The data will be classified after processing the input, it will be classified on the basis of machine learning theorem in machine learning theorem we have naive bayes theorem and k-nearest algorithms which will help to classify the data and provide the output.

5.4 Machine Learning Algorithm

The amount, speed, and variety of data have increased, necessitating the automation of text processing methods. In some circumstances, creating a set of logical rules based on professional judgement and knowledge engineering methodologies to categorise documents aids in automating the categorization operation. Based on the learning process, three kinds of text classification might be created: semi-supervised text classification, unsupervised text classification, and supervised text classification.

5.5 Naive Bayes

As an alternative to the Bag of Words strategy supported by Cardinality of the Intersection, a Naive Bayesian classifier with an estimation was utilized. Based on the practice data, a word vector is produced. The vector's dimensions showed that the word was there, and no particular weight age parameter was applied to the categorization.

VI. CONCLUSION

The main aim of this project is to help the people to get Kannada text document, sometimes people they are able speak Kannada, but not able to write. So, it is very useful t such people and reduce the time, also it is very useful in and out of the office works which are in Kannada text documents.
Text classification helps to classify/categorise the content based on trained data set. Text classification of NLP techniques are used to topic detection, text getting particular data from large collection of documents. In this project we have categorized and implemented mainly in three basis they are 1] sports, 2] technology, 3]entertainment.

REFERENCES

[1]. Akshata K Shinde1, Anjali H , Deepika N Karanth, Gouthami K , Vijetha T S, “Development of Automatic Kannada Speech Recognition System” ,Department of Electronics and Communication Assistant Professor, Department of Electronics and Communication, Alva’s Institute of Engineering and Technology, Mijar.

[2]. Yuhan Zheng Southwest , “An Exploration on Text Classification with Classical Machine Learning Algorithm”, Petroleum University, Chengdu, Sichuan, China.

[3]. Shivakumar K.M, Aravind K.G, Anoop. “Kannada Speech to Text Conversion Using CMU Sphinx”, Amrita Vishwa Vidyapeetham University, Mysore campus, Amrita University, Mysuru, India.

DOI: 10.48175/IJARSCT-5837