Dual Language Detection using Machine Learning

[1]Shashank Simha B K, [2]Rahul M, [3]Jyoti R Munavalli and [4]Prajwal Anand
[1,2,3,4] Dept. of ECE, BNM Institute of Technology, Bengaluru, India.
[1]bk.shashanksimha@gmail.com, [2]rahul.gowda.76@gmail.com, [3]jyotirmunavalli@bnmit.in, [4]praj460@gmail.com

Article Info
Jenitta J and Swetha Rani L (eds.), *International Conference on VLSI, Communications and Computer Communication*, Advances in Intelligent Systems and Technologies,
Doi: https://doi.org/10.53759/aist/978-9914-9946-1-2_32
©2023 The Authors. Published by AnaPub Publications.

Abstract- There are number of languages around the world and knowing all the languages is very difficult for any person. At the same time, unawareness about the language will hinder communication. Language identification is the process where the identifying the language(s) in text form is performed based on the writing style and looking at the unique diacritics of each language. When a multitude of languages are spoken in any circumstances, the first step in communication is the identification of the language. There are several techniques used for language detection like machine learning and deep learning. These are used in detecting languages like German. In India, numerous languages are spoken by the people and thus we propose to develop a model that detects two languages: Kannada and Devanagari/Sanskrit. In this study, Support Vector Machines classifiers were used, for classification and an accuracy of 99% was achieved.

Keywords- Machine Learning, Support Vector Machine, Artificial Intelligence, Python, Language, Kannada, Devanagari, Sanskrit.

I. INTRODUCTION

Language is the basic way of communication that is required to exchange information between people. Each person acquires an ability, right from their childhood, to make use of this exchange of information either through sounds or gestures. As we progress, we tend to learn various languages to communicate among others. Language is not only limited to communicate thoughts and ideas, but it also builds friendship, career, strengthens the economic relationships of businesses and provides an understanding of diverse cultures scattered across the globe. Thousands of languages exist around the world which are used in everyday life. Most of the time, during visit to other places, a person would not be able to understand or communicate in their native language. So, here comes the role of Language Identification which determines the language of the written content given to it and help the user to communicate effectively.

Today, the world is moving towards Artificial Intelligence in every aspect of life. A generation of computers that are capable of depicting the thoughts and actions of human. One of the wonders of the world is the human visual system, which we are attempting to imitate. Any activity that a system undertakes must first be known to it in order for it to begin. Humans have ears, eyes, and a brain that they use to receive information and act on it. In a similar way, speech recognition and word recognition on their own would serve as a computer's eyes and ears, respectively. The most recent technology is used to recognize sign language and other languages.

In this paper, we develop a model that identifies both Kannada and Devanagari languages. The paper is structured as follows: Section II highlights the literature survey and the proposed methodology; Section III presents the results of the model and Section IV is conclusion.

II. MATERIALS AND METHODS

A literature survey was carried out for understanding language detection techniques and the languages that are detected. Scopus, Web of Science, and Google Scholar were the databases that were searched using the terms "language detection," "Machine Learning" and "Indian language." It was found that different research papers had detected different languages. In this paper, we propose to develop a language detection mechanism that is text based, for the two Indian Languages scripts, Kannada and Devanagari using Machine Learning techniques.

As a matter of fact, there are more than 15,000 spoken languages in India [1] of which only a handful of them are known to us (Fig. 1.).

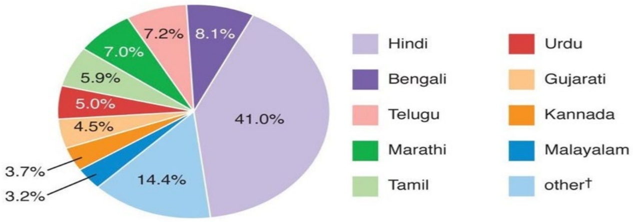


Fig 1. Statistics of Indian Language

As observed in literature, there exists a lot of studies on language detection and identification. Different languages like English, Italian, German, Dutch, Japanese and Indian languages like Hindi and Kannada are detected. Either text or handwritten [2]. To detect the handwritten languages techniques like Machine Learning and Deep Learning are used [3-6]. Deep Neural Networks (DNNs) are used for speech signals that automatically identify language at the acoustic frame levels. The designed DNNs architectures are compared with several state-of-the-art acoustic systems which are based on i-vectors, the results when tested against the two datasets i.e., NIST LRE 2009 and Google 5M LID, it was concluded that in most of the cases, the DNNs performed better than the current state-of-art approaches [3].

Class frequencies are used in a centroid-based classification method to determine the language. But, the success rate of centroid-based classification is generally lower when compared to the other methods. Hence, a new and different method which is known as Inverse Class Frequency (ICF) was developed which increases the quality of centroid values by updating the classical values that provided better successful results and also has lower time complexities when compared with other methods [7]. Using BLSTM-CTC based handwriting recognizers, script identification and character recognition are performed for English, Arabic, and French languages. The OCR result is then fed to a statistical language identifier to determine the language of the input handwritten document. The dataset was divided into training, validation, and test data at a ratio of 60%, 20%, and 20%, respectively. The testing outcomes demonstrated that language dependence in the same script for BLSTM-CTC based handwriting recognition. This framework has only been tested with two languages; more languages need to be added for it to be more reliable [8].

Marcos Zampieri et. al., proposed VarClass, an open-source Language Identification Tool that focused on the varieties of languages. The study deals with 27 language (English, French, Portuguese and Spanish) models. The average performance reported for a particular dataset was of 90.5% accuracy [9]. Three languages i.e., English, Kannada and Hindi are identified using an approach that is based on the characteristic features of the top and bottom profiles of the input text lines. Each text line column is scanned from top to bottom till a black pixel is found. It is possible to manage both regular and italic font kinds in the documents. If the text line is larger than the size of the images used in the training data set, good accuracy is obtained [10].

Methodology

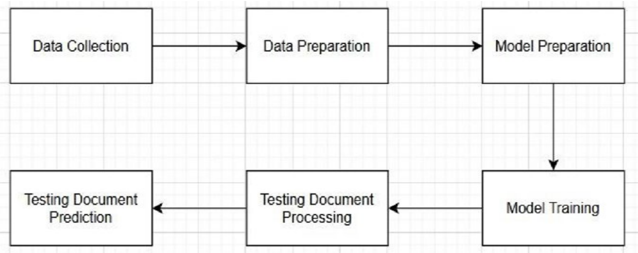


Fig 2. Methodology/Flow Diagram

The flow diagram of the proposed work is as shown in Fig. 2. To begin with, the datasets of Kannada and Devanagari are uploaded into the model. There are totally 3000 different images in the datasets. Furthermore, these dataset images are subjected to image processing techniques. In image processing, these images are read in 3-dimensional format and are resized to the favorable size. In our model, we have resized to 50x50 pixels. And then the 3-dimensional array is converted to 1-dimensional array. Then the model is trained using these datasets. Now, the model is ready for prediction. Now, for testing the model, we feed a random test image and the result is observed. A GUI is used that is aided by PyQt5 framework to design the output window. And finally at the output, the predicted language is displayed along with the test image.

Support Vector Machine (SVM) is a supervised machine learning method which can be applied to problems involving classification and regression. However, categorization issues are the context in which it's used most commonly. Each data point is represented by a point in n-dimensional space (n being the number of features we have) in the SVM method, with the value of each feature being the value of a certain coordinate. The next step in classification is to locate the hyper-plane that best separates the two classes (as demonstrated in Fig 3).

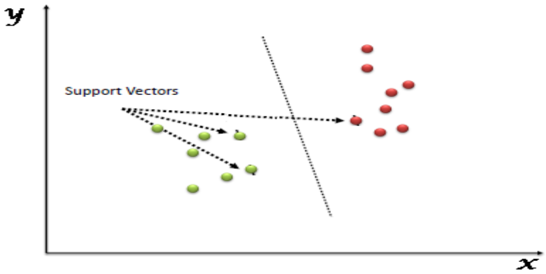
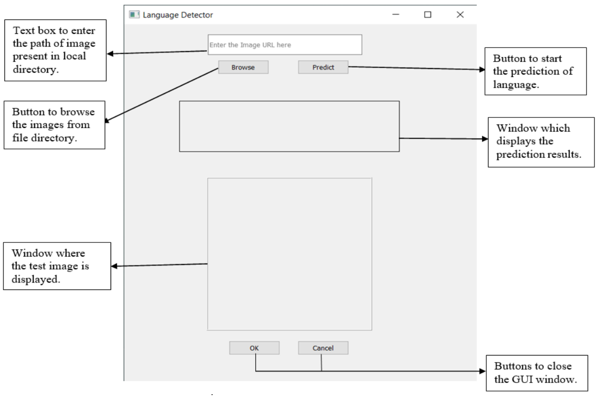


Fig 3. SVM hyper-plane demo

III. RESULTS AND DISCUSSION

The proposed work was implemented in Python and libraries like pandas, scikit learn, NumPy and PyQt5 were used. The output window of the proposed Language Detector model is as shown in Fig. 4. At the top, there is a text box wherein the path of the image containing the text present in the local directory is entered. Alternatively, the image can be browsed and selected from the file directory. The predict button starts the language identification process and displays the prediction results along with its accuracy in the window present below the Browse and the Predict buttons. Below this lies the window which displays the image selected from the local file directory. The OK and the Cancel buttons are used to close the GUI window and start the prediction process again.



Text box to enter the path of image present in local directory.

Button to start the prediction of language.

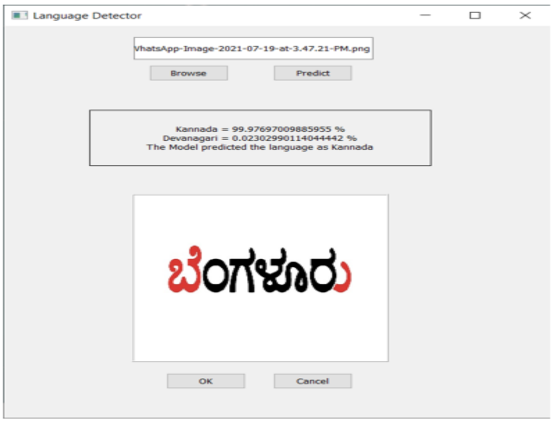
Button to browse the images from file directory.

Window which displays the prediction results.

Window where the test image is displayed.

Buttons to close the GUI window.

Fig 4. Output/GUI window



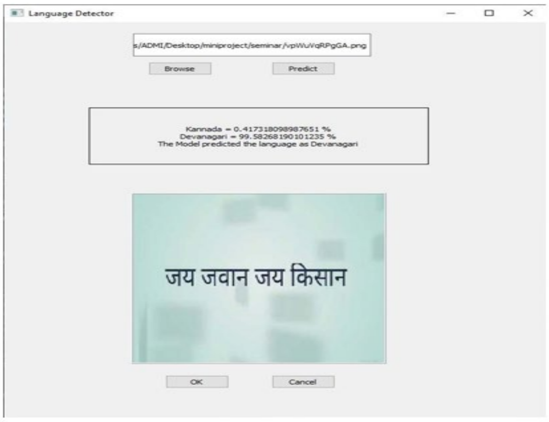
VhatsApp-Image-2021-07-19-at-3.47.21-PM.png

```
Kannada = 99.97697009885955 %
Devanagari = 0.02302990114044442 %
The Model predicted the language as Kannada
```

Fig 5. Output screenshot for Kannada

Fig 5 shows the output screenshot for an image in which the text is written in a language called Kannada which has been identified rightly by the model with an accuracy of about 99.97%.

Fig 6 shows the output screenshot for an image in which the text is written in a script called Devanagari which has been identified rightly by the model with an accuracy of about 99.58%.



s/ADMI/Desktop/miniproject/seminar/vpWuVqRPgGA.png

```
Kannada = 0.417318098987651 %
Devanagari = 99.58268190101235 %
The Model predicted the language as Devanagari
```

Fig 6. Output screenshot for Devanagari

IV. CONCLUSION

An intriguing issue in the present world is language identification. It functions as the first stage of several processes as well as standalone ones. When dealing with a collection of texts published in several languages, it offers the capability of employing background knowledge about the language and using specialist methodologies. Systems are needed for accurately recognizing the language of documents as global communication and trade grow (emails, letters, web pages etc.). Various aspects of natural language processing are involved in the task of language identification. Thus, this project focuses on eliminating this problem by using a model designed using Machine Learning algorithms powered by various Python libraries. In future, this language detector could also be integrated with Natural Language Processing based audio language detector.

References

[1] Sengupta, D. and G. Saha, *Study on Similarity among Indian Languages Using Language Verification Framework.* Advances in Artificial Intelligence, 2015. 2015: p. 325703. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Marco Lui, J.H.L., Timothy Baldwin, *Automatic Detection and Language Identification of Multilingual Documents.* Transactions of the Association for Computational Linguistics, 2014. 2: p. 27-40.

[3] Lopez-Moreno, I.;J. Gonzalez-Dominguez;O. Plchot;D. Martinez;J. Gonzalez-Rodriguez, and P. Moreno. Automatic language identification using deep neural networks. in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014.

[4] Jayanthi, N.;H. Harsha;N. Jain, and I.S. Dhingra. Language Detection of Text Document Image. in 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN). 2020.

[5] Rabby, A.K.M.S.A.;M.M. Islam;N. Hasan;J. Nahar, and F. Rahman. Language Detection using Convolutional Neural Network. in 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). 2020.

[6] Simões, A.;J.J. Almeida, and S.D. Byers. Language Identification: a Neural Network Approach. in SLATE. 2014.

[7] Takçı, H. and T. Güngör, A high performance centroid-based classification approach for language identification. Pattern Recognition Letters, 2012. 33(16): p. 2077-2084.

[8] Mioulet, L.;U. Garain;C. Chatelain;P. Barlas, and T. Paquet. Language identification from handwritten documents. in 2015 13th International Conference on Document Analysis and Recognition (ICDAR). 2015.

[9] Zampieri, M. and B.G. Gebre. VarClass: An Open-source Language Identification Tool for Language Varieties. in LREC. 2014.

[10] Padma, M.C.;P.A. Vijaya, and P. Nagabhushan. Language Identification from an Indian Multilingual Document Using Profile Features in 2009 International Conference on Computer and Automation Engineering, 2009.