iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44281

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Detectsy: A System for Detecting Language from the Text, Images, and Audio Files

Riya Menon
Department of Computer Engineering, VESIT, University of Mumbai

Abstract— Language detection is a natural language processing task where we need to identify the language of a text or document. As a human, we can easily detect the languages we know. However, it is not possible for an individual to identify many languages. This is where the language identification task can be used. The proposed solution is a complete system that detects language from the text, images, and audio files. Language identification task from text is carried out by training a Multinomial Naive Bayes classifier model. In the case of image and audio inputs, Python libraries are used to achieve the goal of language detection.

Keywords— Language identification, Natural Language Processing, Multinomial Naive Bayes classifier, Optical Character Recognition, Speech Recognition

I. INTRODUCTION

The task of identifying the language of text or utterances has a number of applications in natural language processing. It is a key step in the automatic processing of real-world data, where a multitude of languages may be present. Language detection is the first step towards achieving a variety of tasks like detecting the source language for machine translation, improving the search relevancy by personalizing the search results according to the query language [1], providing a uniform search box for a multilingual dictionary [2], etc. It is also a key component of many web services. For example, the language that a web page is written in is an important consideration in determining whether it is likely to be of interest to a particular user of a search engine, and automatic identification is an essential step in building language corpora from the web. It has practical implications for social networking and social media, where it may be desirable to organize comments and other user-generated content by language.

In this paper, a complete system for language identification is proposed. It has three distinct modules to detect language from text input, images, and audio files. The language identification is carried out by using a Multinomial Naive Bayes classifier in the case of text input. It is trained on a dataset containing 17 languages and has an accuracy of 97.87%. For images and audio files, a combination of Python libraries is used to achieve the goal of language identification.

The rest of the paper is organized as follows: Section II gives an overview of the related work on language detection. Section III explains the methodology adopted by the proposed system. Section IV explains the algorithms employed by the proposed approach. Section V discusses the results of the Detectsy along with the outputs obtained. Section VI describes the evaluation results obtained. Finally, the conclusion and future works are described in section VII.

II. LITERATURE SURVEY

Various papers were reviewed pertaining to different modules of the system. Each paper delved deeper into different methods of implementing functionality with each having its own shortcomings.

An off-the-shelf language identification tool, langid.py [3], is trained over a naive Bayes classifier with a multinomial event model over a mixture of byte n-grams. It is fast, unaffected by domain-specific features, is a single file with minimal dependencies, and has a flexible interface. However, it has an accuracy of 94%.

Support vector machines (SVMs) with n-gram counts as features are proposed for the language identification of very short texts such as proper nouns [4]. But language identification accuracy attained is 84% which is too low to be useful.

A graph-based N-gram approach for language identification (LI) called LIGA is proposed in [5] that allows learning elements of grammar besides using N-gram frequencies. To capture the ordering of words, a graph model is created on labelled data. Once the graph is created, it is used to classify unlabelled texts. It finds its application in language identification on relatively short texts typical for social media like Twitter.

Tesseract [6] is an open-source OCR engine that was developed at HP between 1984 and 1994. It assumes that its input is a binary image and makes use of an adaptive classifier for word recognition. Its key weakness is probably its use of a polygonal approximation as input to the classifier instead of the raw outlines.

A novel language identification (LID) method is proposed in [7] that accepts the architecture of time delay neural network (TDNN) followed by long short term memory (LSTM) recurrent neural network (RNN) to learn long-term phonetic patterns and

model the phonetic dynamics for different languages. It achieves better identification performance in both cases of long utterance and short utterance but is computationally expensive and resource intensive.

Mel Frequency Cepstral Coefficients (MFCC) have been used [8] to derive features of speech signals that can be used for identifying languages. For classification purposes, Support Vector Machines and Decision Tree classifiers were used with accuracies of 76% and 73% respectively.

Another approach is based on Linear Discriminant Analysis (LDA) [9] which is developed using the database of seven different Indian languages giving a maximum classification accuracy of 93.88%.

An approach based on Convolution Neural Networks (CNN) based on AlexNet for the automatic language identification of four Indian languages, Bengali, Gujarati, Tamil, and Telugu has been proposed in [10]. However, being a heavy model, its training time is more. Another weakness of this system is that most of the Tamil audio files are predicted as Telugu because Telugu and Tamil languages sound very similar to each other.

III.PROPOSED SYSTEM

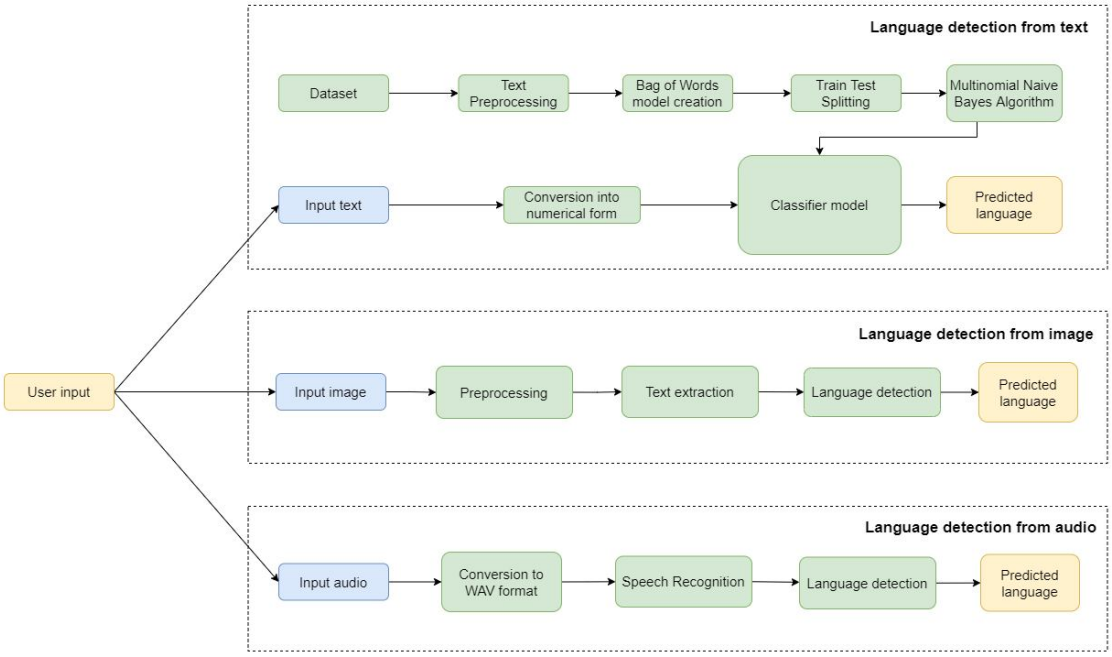


Fig. 1 Block diagram for Detectsy

Fig [1] represents the block diagram of the proposed system. The three major components of the proposed solution are as follows:

A. Language Identification from the Text

For identifying language from text input, the user can either type the text or paste it into the provided textbox. A Multinomial Naive Bayes classifier model is trained to detect the language. The algorithm is explained in detail in the next section. The dataset [11] is first pre-processed wherein many unwanted symbols, numbers are removed. Then, the text is converted into numerical form by creating a Bag of Words model. The next step is to create the training set, for training the model and the test set, for evaluation. The user input is passed to this trained classifier model to predict the language of the text.

B. Language Identification from Images

Language is detected from images containing printed text. The task is performed using Python-tesseract, a wrapper for Google's Tesseract-OCR (Optical Character Recognition) engine. It is available under the Apache 2.0 license. It can be used directly using an API to extract printed text from images. Further, Python's *langdetect* library [12] is used to identify the possible languages of the extracted text. The language with the highest probability is selected and provided as the output to the user.

C. Language Identification from Audio Files

In the case of audio files, the file taken as input from the user is first converted into WAV file format because *SpeechRecognition* [13] supports WAV (must be in PCM/LPCM format), AIFF, AIFF-C, FLAC (must be native FLAC format;

OGG-FLAC is not supported) file formats. Python's *SpeechRecognition* library acts as a wrapper for several popular speech APIs and is thus extremely flexible. One of the included APIs—the Google Web Speech API—supports a default API key that is hard-coded into the *SpeechRecognition* library. It is used to recognize speech from an audio source with the help of *recognize_*google*()*. Finally, the extracted content is passed through *langdetect's detect_lang()* to determine the language of the input audio file.

IV. ALGORITHMS

A. *Multinomial Naive Bayes*

The Multinomial Naive Bayes algorithm [14] is a Bayesian learning approach popular in Natural Language Processing (NLP). It is suitable for classification with discrete features. The classifier guesses the tag of a text using the Bayes theorem. It calculates each tag's likelihood for a given sample and outputs the tag with the greatest probability. The Multinomial Naive Bayes is widely used for assigning documents to classes based on the statistical analysis of their contents. In this case, the Multinomial Naive Bayes classifier model is trained on a dataset containing 17 languages including English, French, Spanish, Portuguese, Italian, Russian, Swedish, Malayalam, Dutch, Arabic, Turkish, German, Tamil, Danish, Kannada, Greek, and Hindi.

B. *LangDetect*

LangDetect [15] implements a Naive Bayes classifier, using a character n-gram-based representation without feature selection, with a set of normalization heuristics to improve accuracy. It is trained on data from Wikipedia and can be trained with user-supplied data. The language detection algorithm is non-deterministic, which means that if we try to run it on a text which is either too short or too ambiguous, we might get different results every time we run it. To enforce consistent results, it is recommended to set the *DetectorFactory* seed to some number.

V. RESULTS

The proposed system is developed in the form of a web application. Fig. [2] represents the home page screenshots through which users can navigate to text, image, or audio language detection pages.

Fig. 2 Homepage screenshots

For detecting language from text input, the user either types in or pastes the text into the textbox provided and the system would detect the language. Some of the results for text language identification (LID) are shown in Figs.[3-6].

Fig. 3 Malayalam language identification

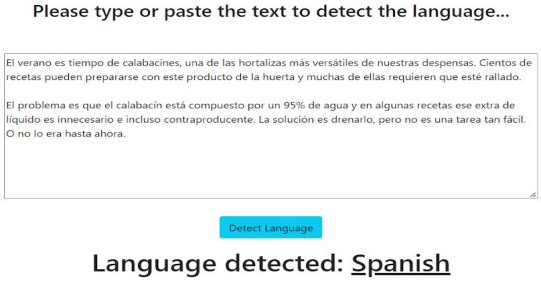


Fig. 4 Spanish language identification

Fig. 5 Dutch language identification

Fig. 6 Tamil language identification

For images, the user has to provide the image with printed text as input. After extraction of text, the language is identified and provided to the user as output as shown in Figs. [7-10].

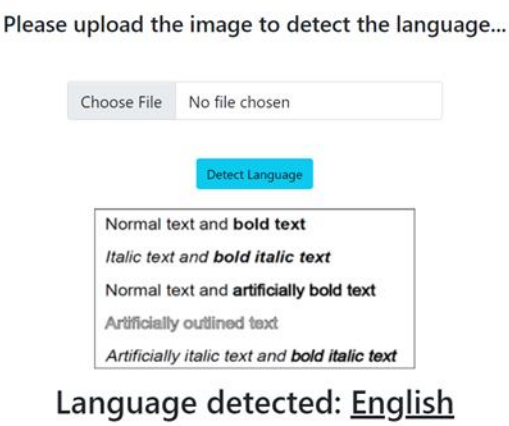


Fig. 7 English language identification

Fig. 8 Hindi language identification

Fig. 9 Arabic language identification

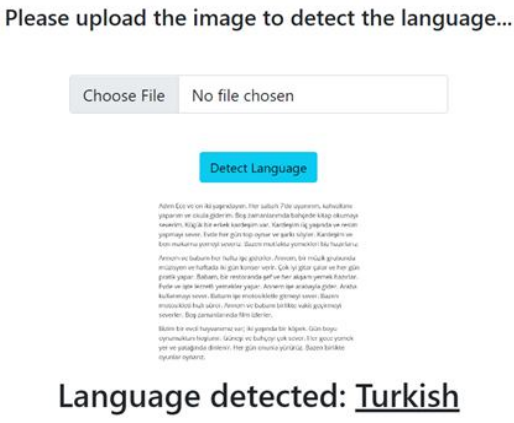


Fig. 10 Turkish language identification

In the case of audio files, the user uploads the audio file. The system processes it to detect the language of the speech. The results are shown in Figs.[11-14]. The audio files uploaded in the given examples have the following text:

- French: Je m'appelle Hugo et j'ai seize ans. Aujourd'hui, avec mes parents et ma sœur nous partons en voyage. Ma sœur s'appelle Laura, elle a treize ans. Nous sommes à l'aéroport : direction Barcelone en Espagne !

- Portuguese: Rubens sempre quis ser jornalista. Desde quando era criança ele já escrevia para o jornal que havia na escola onde estudava. Quando entrou para a faculdade, ele achou que era o momento de organizar uma espécie de noticiário sobre os acontecimentos da própria universidade.
- Italian: I genitori di mio marito vivono lontano da qui, in città. I miei genitori invece abitano vicino a noi, nello stesso paese. Vogliono molto bene ai nostri tre figli e spesso si occupano di loro.
- German: Hansi Flick war nicht zufrieden, das konnte man seinem Gesichtsausdruck während des Spiels in Budapest deutlich ablesen. Der Bundestrainer hatte zwar vor dem Spiel in Ungarn gewarnt, es sei 'nach England das schwerste Spiel' und es warte eine 'ganz große Aufgabe', aber dass seine Elf mit so wenig Durchschlagskraft nach vorne agieren würde wie beim 1:1 (1:1) in Budapest, hätte er wohl nicht gedacht.

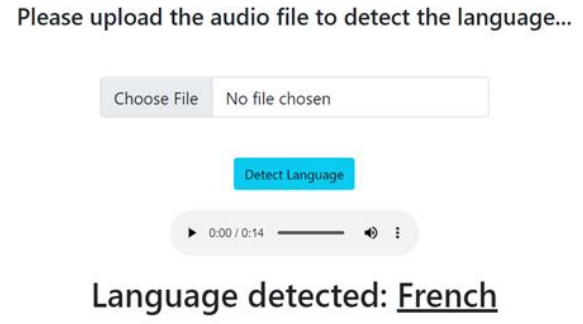


Fig. 11 French language identification

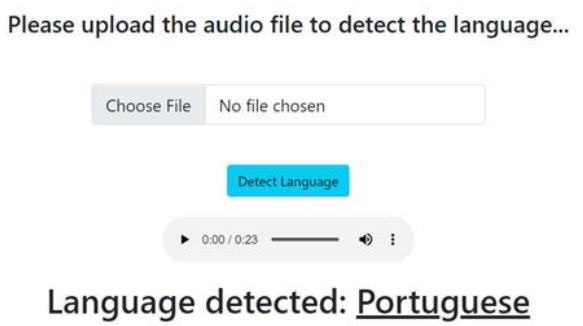


Fig. 12 Portuguese language identification

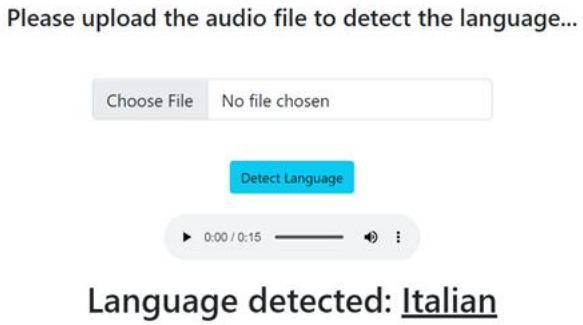


Fig. 13 Italian language identification

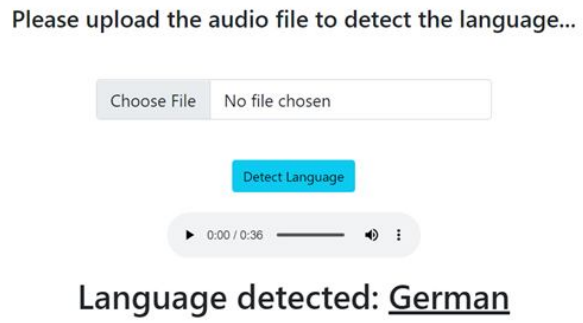


Fig. 14 German language identification

VI. EVALUATION

A dataset containing 17 languages is used for training the classifier model to detect languages from text input. Fig. [15] shows a snapshot of the dataset used and fig. [16] displays the value count of text for each language in the dataset. The trained Multinomial Naive Bayes classifier model is evaluated and the accuracy is computed to be 97.87%. Fig. [17] represents the confusion matrix for the same.

	Text	Language
0	Nature, in the broadest sense, is the natural...	English
1	"Nature" can refer to the phenomena of the phy...	English
2	The study of nature is a large, if not the onl...	English
3	Although humans are part of nature, human acti...	English
4	[1] The word nature is borrowed from the Old F...	English
5	[2] In ancient philosophy, natura is mostly us...	English
6	[3][4] \nThe concept of nature as a whole, the...	English

Fig. 25 Dataset

```
English       1385
French        1014
Spanish        819
Portugeese     739
Italian        698
Russian        692
Sweedish       676
Malayalam      594
Dutch          546
Arabic         536
Turkish        474
German         470
Tamil          469
Danish         428
Kannada        369
Greek          365
Hindi           63
Name: Language, dtype: int64
```

Fig.36Value count for each language

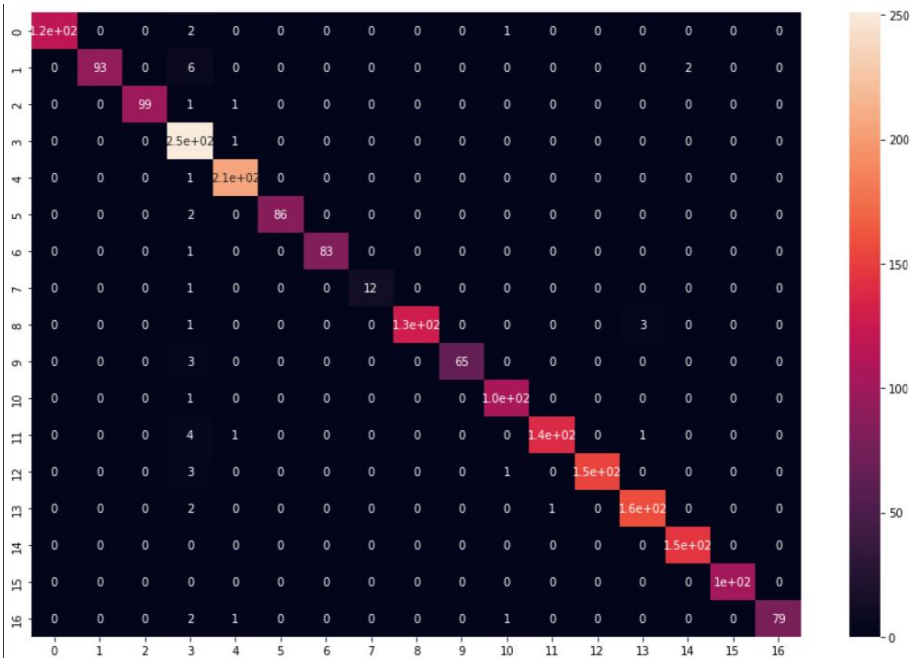


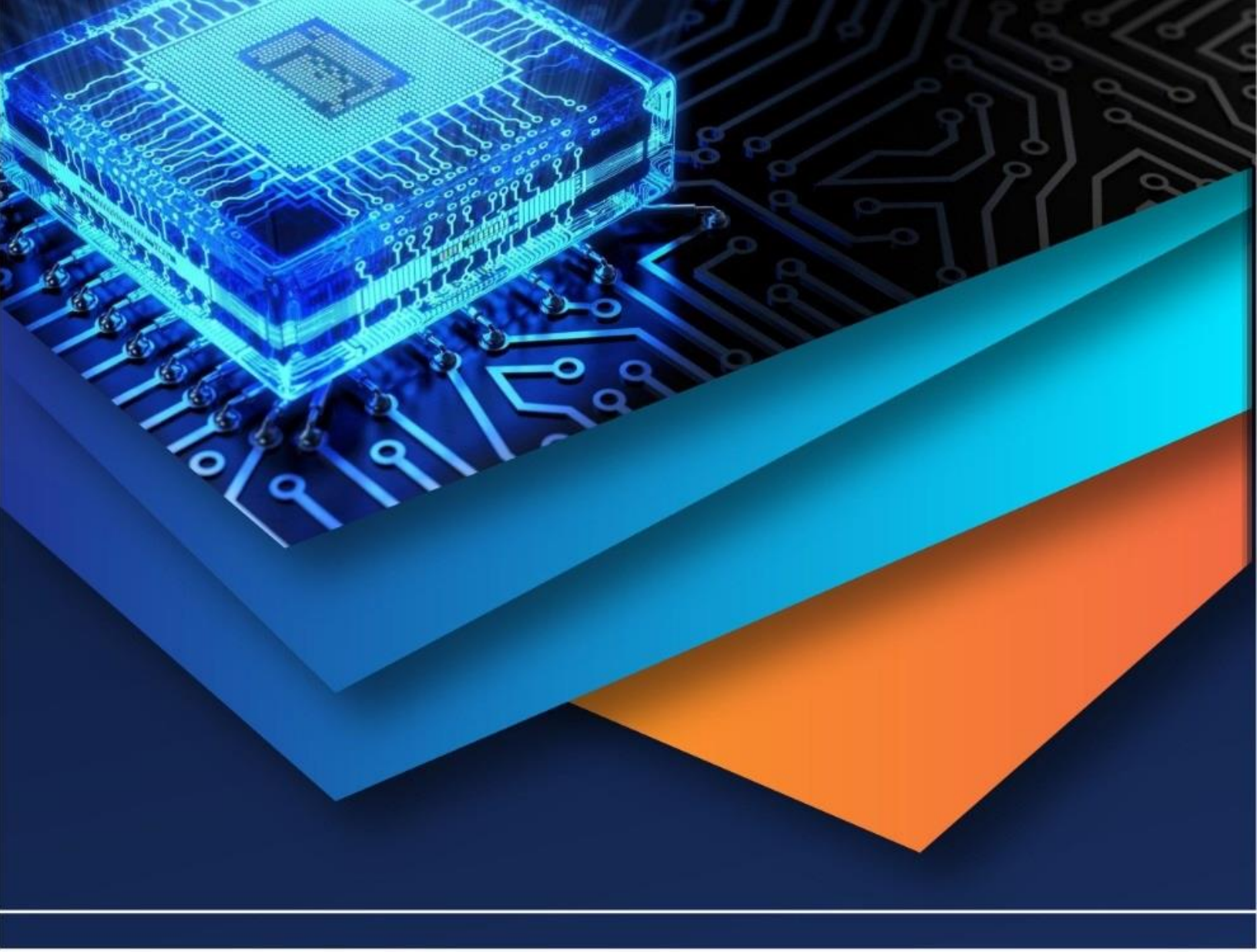
Fig. 47 Confusion matrix

VII. CONCLUSION AND FUTURE WORKS

In this paper, a complete language detection system is proposed having different modules for text input, image, and audio files. The trained Multinomial Naive Bayes classifier model has achieved an accuracy of 97.87% for 17 languages. The current work can be extended to include more languages. Currently, language can be detected from printed text. It can be extended to include handwritten text documents. Also, models can be trained for speech-language detection in order to achieve better performance.

REFERENCES

[1] Juliane Stiller, Maria Gade, and Vivien Petras. Ambiguity of queries ¨ and the challenges for query language detection. CLEF 2010 Labs and Workshops Notebook Papers, 2010.
[2] Dong Nguyen and a Seza Do. Word level language identification in online multilingual communication. 23(October):857–862, 2013.
[3] Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In Proceedings of the ACL 2012 System Demonstrations, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.
[4] Aditya Bhargava and Grzegorz Kondrak. 2010. Language identification of names with SVMs. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 693–696, Los Angeles, California. Association for Computational Linguistics.
[5] Erik Tromp and Mykola Pechenizkiy. Graph-based n-gram language identification on short texts. ”Proceedings of the 20th annual BelgianDutch Conference on Machine Learning”, pages 27–34, 2011.
[6] R. Smith, "An Overview of the Tesseract OCR Engine," Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 2007, pp. 629-633, doi: 10.1109/ICDAR.2007.4376991.
[7] L. Sun, "Language Identification with Unsupervised Phoneme-like Sequence and TDNN-LSTM-RNN," 2020 15th IEEE International Conference on Signal Processing (ICSP), 2020, pp. 341-345, doi: 10.1109/ICSP48669.2020.9320919.
[8] H. Venkatesan, T. V. Venkatasubramanian and J. Sangeetha, "Automatic Language Identification using Machine learning Techniques," 2018 3rd International Conference on Communication and Electronics Systems (ICCES), 2018, pp. 583-588, doi: 10.1109/CESYS.2018.8724070.
[9] J. S. Anjana and S. S. Poorna, "Language Identification From Speech Features Using SVM and LDA," 2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2018, pp. 1-4, doi: 10.1109/WiSPNET.2018.8538638.
[10] L. R. Arla, S. Bonthu and A. Dayal, "Multiclass Spoken Language Identification for Indian Languages using Deep Learning," 2020 IEEE Bombay Section Signature Conference (IBSSC), 2020, pp. 42-45, doi: 10.1109/IBSSC51096.2020.9332161.
[11] https://www.kaggle.com/datasets/basilb2s/language-detection
[12] https://pypi.org/project/langdetect/
[13] https://pypi.org/project/SpeechRecognition/
[14] https://www.upgrad.com/blog/multinomial-naive-bayes-explained/
[15] https://towardsdatascience.com/benchmarking-language-detection-for-nlp-8250ea8b67c



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY