Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Enhance the Effectiveness of Machine Learning Models f or Language Recognition and Translation for Multimodal Input

A. G. Chaure[1], A. T. Bhise[2], S. S. Ambike[3]

[1]Research Scholar, Department of CSE, Shri JJT University, Jhunjhunu, Rajasthan, India
[2]Research Guide, Department of CSE, Shri JJT University, Jhunjhunu, Rajasthan, India
[3]Research Co-Guide, Department of CSE, Shri JJT University, Jhunjhunu, Rajasthan, India
Email: avinash.choure@gmail.com

Communication is an important aspect of human beings and language is the heart of communication. The language preferred for communication plays a crucial role in any kind of conversation. There are a variety of languages in this Universe and these multiple languages are adapted by people geographically. In this world of digitization, a lot of innovations have taken place in the field of Language Identification and Translation. This particular feature has been adopted by several smart systems in today's date.

In this paper, authors have proposed an architecture to identify and translate different languages with the help of machine learning algorithms. Multimodal inputs are used in this system in the form of text and images for analysis. Machine learning techniques such as K-Nearest Neighbors(KNN), Multinomial Naive Bayes(MNB), Random Forest(RF), and Support Vector Machine (SVM) are used in this study for Language Identification and Translation. A particular language is translated into English, Hindi, or Marathi. The results provided by the proposed framework can be applied in a variety of disciplines, including politics, business, education, industry, psychology, and security.

Keywords: Machine Learning, Language Identification, Language Translation.

1. Introduction

The major mode of communication between humans is language. Because they can communicate through language, humans are superior to all other living beings. Because they possess both sight and hearing, humans are able to communicate through language [1].

Humans communicate information about a variety of topics, including their personal lives, businesses, jobs, and entertainment, in multilingual contexts. A Cross-cultural context is preferred nowadays because people are coming from different areas to work together. It is observed that people are more comfortable with their mother tongue than other languages. For smooth communication among them, identification and translation of a particular language became necessary without a mediator. People share the data on different social media

platforms using multiple languages. Unstructured data is produced by these platforms.

Translation is necessary for the global transmission of new concepts, information, and knowledge and to develop effective interactions between cultures. Translation has the power to alter history and facilitate the dissemination of recent knowledge.

Machine learning algorithms have dominantly contributed to the context of language identification and translation in this era. Machine Learning is a popular field in today's digital age. Machine Learning algorithms are employed in all aspects of the online world when browsing the Internet. This demonstrates how these professions have become an integral part of our lives. These strategies are used to classify the abundance of data generated by online mediums.

The goal of this paper is to shed light on the machine-learning algorithms used in language processing. In this paper, the authors have designed a proposed framework that can identify 21 distinct languages and which can translate them into English, Hindi, or Marathi.

This research paper is outlined as follows. Section II provides a quick summary of the related work. Information about Machine Learning algorithms implemented in the paper is given in Section III. Section IV describes the proposed framework in detail. In Section V, the authors evaluate and contrast the results of the proposed model. In the end, Section VI concludes the research work carried out here.

2. Related Work

I. This section has outlined previous research work done in the context of language identification using Machine Learning techniques.

Language identification involves properly categorizing text or documents depending on their language. However, the majority of the study concentrated on English language, with little attention paid to other South African official languages. Automatic language recognition in language-specific systems can help to bridge the digital gap in multilingual societies. Various machine learning methods can be used to recognize a document's natural language. This work describes a text-based language identification method that use individual proper names, especially surnames, in a South African setting. Three supervised machine learning approaches are used to conduct multiclass classification, including support vector machines and naive Bayes language models. Extensive trials were conducted to assess these algorithms for language identification in official languages of South Africa i.e. Tshivenda, Xitsonga, and Sepedi. All three machine learning approaches performed admirably in a 10-fold cross validation. The results show that a multinomial naïve Bayes technique outperformed other algorithms[2].

Text recognition is utilized in a variety of applications, including document analysis, picture labeling, and text analysis. The method of identifying text from images is extremely important. Maximally Stable Extremal Regions (MSER) and Optical Character Recognition (OCR) algorithms have significantly improved the accuracy and reliability of text identification from photographs. Our suggested model for text recognition combines the features of MSER and OCR algorithms to increase the accuracy and reliability of digital text extraction. OCR

employs cutting- edge techniques to recognize and identify individual characters, which serve as the fundamental building block for text recognition. The proposed approach uses a multistage process. The MSER method is used to extract the most likely text places from the input picture first. To improve OCR performance, these zones are fine-tuned with pre-processing techniques such as noise reduction and image enhancement. After cleaning the portions, the OCR system uses machine learning and pattern recognition to recognize the text in each location. The identified text is subsequently processed to improve accuracy and refine the results. The text recognition model with MSER and CNN (OCR) algorithm outperforms other models[3].

This research study focuses on several machine learning techniques that may be used to extract text from handwritten documents and photos, recognize them in digital format, and translate them based on the user's needs. Machine Learning is capable of learning on its own, without the assistance of humans or explicit programming, based on experience and knowledge. Machine learning is used in a variety of real-time applications in our daily lives. Text detection and extraction is an essential application that extracts vital information from photos acquired from a variety of sources. Text with different variants varies in size, direction, alignment, style, low brightness, and contrast in photos with complicated backgrounds. Many people have difficulty reading owing to differences in text on pictures. As a result, text identification and extraction are more crucial and difficult tasks. The goal here is to assist people who speak different languages from all around the world in reading and understanding any language that is written. Researchers employ a variety of machine learning techniques and tools to detect handwritten text and text collected from photos and convert it to digital representation. Optical Character Recognition (OCR) is a machine learning approach that allows us to recognize and extract text data or information from documents, converting it into editable and searchable data[4].

Language identification is the process of recognizing a language(s) in text form based on its writing style and distinctive diacritics. When many languages are spoken in any situation, the first step in communicating is to identify the language. Language detection uses a variety of approaches, including machine learning and deep learning. These are used to detect languages such as German. People in India speak a variety of languages, thus we suggest developing a model that can recognize two: Kannada and Devanagari/Sanskrit. In this work, Support Vector Machines classifiers were utilized for classification, and an accuracy of 99% was obtained[5].

In this paper, researchers demonstrate that self-supervised pre-trained representations learned from unlabeled audio data are useful for improving language recognition. We showed that cross-lingual representations are especially useful in low-resource settings with little labeled data and that pre- training is more successful than training LID models only on labeled data. Because pre-trained models are still rather large, researchers may examine strategies to make these models more useful for inference[6].

II. Machine Learning Models Implemented for Language Identification and Translation

In conventional programming, the system receives input, and the program generates output depending on the logic. In Machine Learning, the system receives both input and output, and models are created. That model is used to create predictions and solve complicated problems such as data analysis problems, business problems, and real-world issues[11].

Machine Learning computer programs learn from experience and examples and then do related tasks. The Machine Learning algorithm is used to generate a model from a training data set. When the Machine Learning algorithm is supplied with new input data (test data), it predicts results using the model. The prediction is checked for accuracy, and if it fulfills the expectations, the Machine Learning algorithm is applied. If the Machine Learning algorithm does not perform as predicted, it is trained using a larger training data set[4].

Machine learning is equipped with a diverse set of algorithms and features. In this study, the authors evaluated the following four Machine Learning methods for developing models for language recognition and translation from text and image data.

K-Nearest Neighbors (KNN)

K-Nearest Neighbour is a basic machine-learning method based on the Supervised Learning approach. The K-NN method assumes a resemblance between the new case/data and the existing cases and assigns the new case to the category that is most similar to the extant categories. The K- NN algorithm maintains all existing data and classifies new data points based on their similarity. This implies that when fresh data comes, it may be quickly sorted into a suitable category using the K-NN method. K-NN algorithms may be used for both regression and classification, however, they are more commonly utilized for classification issues [7].

Multinomial Naive Bayes

The Naive Bayes classifier has several forms, including multinomial NB. It operates on the basis of multinomial distribution and term frequency. Multinomial NB considers each model feature independently, and the count of a specific word is taken into account throughout the classification process.[1]

The equation followed by Multinomial NB is

$$C_{NB} = \arg\max_{c \in C} P(c_j) \prod_{x \in X} P(x|c)$$

Where P(xi|cj) is an independent feature probability for class c and document d. The formula that can be used for the text classification is

$$C_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_{i \in positions} P(x_i|c_j)$$

Random Forest

Random Forest is a supervised method that also solves regression problems. Forest consists of trees and decision trees. Random forest involves creating a decision tree from data samples, making predictions, and selecting the best answer by voting. It eliminates over-fitting by averaging the best solution results. In an RF, the algorithm generates a specific number of trees in the forest and uses the product result. Using additional trees for analysis improves accuracy. Random forests and decision trees are not the same. Random Forest generates a tree and divides its characteristics[12].

Support Vector Machine

Support Vector Machine (SVM) is a classification machine learning technique. This is a supervised learning strategy for solving regression issues. The primary goal of the SVM machine is to establish a decision boundary. The SVM classification method plots each data point in n dimensions. SVM draws hyperplanes to represent two classes: linear and non-linear. SVM uses individual observations for analysis. SVM covers extreme points for hyper-planes. The extreme point is sometimes called a support vector. The goal is to design a hyperplane in k-space to categorize data points based on distinct properties. The hyperplane in Support Vector Machine is used to distinguish two types of data points. The term "maximal perimeter" refers to the greatest distance between two samples of a class. Margin distance takes into account reinforcement, so that future data points may be categorized [8-10].

3. Proposed Methodology for Language Identification and Translation

This architecture is enhanced with the use of machine learning techniques, which are used to predict of variety of spoken languages and further multilingual translation of the same with reference to text and text-image. The proposed model has been implemented using various Python libraries.

The data set required for the model building is furnished with national and international languages which is obtained from the Kaggle website, which is a repository of a vast collection of various types of datasets.

The steps followed for model implementation in Figure 1.1 are:

- The multimodal data is used in the form of text and text-image which is extracted from the dataset.
- After data extraction data cleaning process is carried out in which noisy data in the form of punctuation marks, special characters, or HTML tags are removed to improve model accuracy.
- The next step is feature extraction, which turns the text into a matrix of token counts that indicate how frequently each word appears in the dataset.
- Data encoding and labeling are done afterward to input for the model.
- The datasets are divided into a training dataset and a testing dataset in an 80/20 pattern, which means that 80 percent of the dataset is used for model training and 20% for model testing.
- After dividing the dataset into training and testing sets, the model is trained on a training dataset. When the model-building process is complete, the model is evaluated against the test dataset.
- At this point predictions are made and their accuracy is verified. All of these models' outcomes are displayed in the form of a confusion matrix. Accuracy, precision, recall, and the F1-score are used to illustrate more precise results in the evaluation.

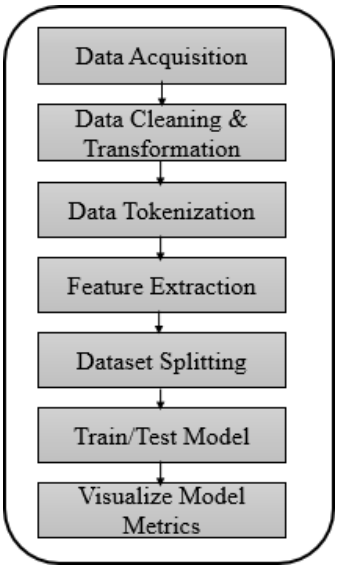

Figure 1.1 Proposed Architecture of Language Identification using Machine Learning techniques using text input

4. Results and Discussion

This study has implemented different Machine Learning algorithms KNN, MNB, RF, and SVM using Python by referring datasets related to text which is extracted from "Kaggle". The results disclosed by all machine learning models are sketched in this section.

The implemented algorithms have predicted the 21 different languages. These languages include English, Malayalam, Hindi, Tamil, Kannada, French, Spanish, Portuguese, Italian, Russian, Swedish, Dutch, Arabic, Turkish, German, Danish, Greek, Sanskrit, Gujarati, Telugu and Oriya. The data in the text-image format is first converted into the text and then the language predicted for it. Once the language prediction is done these languages can be translated into English, Hindi, and Marathi.

The performance of the Machine Learning models in this experiment is evaluated using four measures: accuracy, precision, recall, and F1-Score. The Confusion matrix is used to represent the results revealed by each model. The accuracy given by each model is outlined in Table 1.1

Table 1.1 Accuracy of Machine Learning Models

Machine Learning Model	Accuracy
K-Nearest Neighbors	49%
Multinomial Naive Bayes	98%
Random Forest	92%
Support Vector Machine	95%

5. Conclusion

Language identification is the challenge of accurately categorizing text in a variety of languages. Machine Learning is a subfield of Artificial Intelligence that is being studied extensively across the world. It can mostly make its own judgments or forecast outcomes while executing certain tasks based on the expected input dataset and training set. In this research paper authors have employed various machine learning algorithms with the intention of language prediction from text and text-image data. The accuracy given by the machine learning algorithms KNN, MNB, RF, and SVM is 49%, 98%, 92%, and 95% respectively. So finally, the study concludes that Multinomial Naive Bayes has the maximum accuracy of 98% for language prediction analysis of text data. In the future, the author intends to increase model performance by experimenting with different parameters like batch sizes, epochs, and datasets throughout the training process.

References

1. Ovishake Sen , Mohtasim Fuad , Md. Nazrul Islam , Jakaria Rabbi Mehedi Masud, Md. Kamrul Hasan , Md. Abdul Awal, Awal Ahmed Fime , Md. Tahmid Hasan Fuad , Delowar Sikder,Md. Akil Raihan IFTEE(2022). Bangla Natural Language Processing: A Comprehensive Analysis of Classical, Machine Learning, and Deep Learning-Based Methods. IEEE Access, DOI: 10.1109/ACCESS.2022.3165563
2. Tshephisho Joseph Sefara, Madimetja Jonas Manamela and Promise Tshepiso Malatji(2016). Text-based Language Identification for Some of the Under-resourced Languages of South Africa. In International Conference on Advances in Computing and Communication Engineering (ICACCE), DOI: 10.1109/ICACCE.2016.8073765, pp 303-307.
3. Chaitanya U , Emmanuel Alisetti , Harsitha Ballam, Maneesha Dodda(2023). Digital Image Text Recognition Using Machine Learning Algorithms. International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 11, Issue VI, PP 3583-3589
4. Shivani Surana; Komal Pathak; Mehul Gagnani; Vidhan Shrivastava; Mahesh T R; Sindhu Madhuri G (2022). Text Extraction and Detection from Images using Machine Learning Techniques: A Research Review. In International Conference on Electronics and Renewable Systems (ICEARS), DOI: 10.1109/ICEARS53579.2022.9752274
5. Shashank Simha B K, Rahul M, Jyoti R Munavalli, Prajwal Anand(2023). Dual Language Detection using Machine Learning. 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, 177-180
6. Tjandra, A., Choudhury, D. G., Zhang, F., Singh, K., Conneau, A., Baevski, A., Sela, A., Saraf, Y., & Auli, M. (2021). Improved Language Identification Through Cross-Lingual Self- Supervised Learning. http://arxiv.org/abs/2107.04082, Electronic ISBN:978-1-6654-0540-9 https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning
7. Ming Li, Hongbin Suo, Xiao Wu, Ping Lu, Yonghong Yan(2007). Spoken Language Identification Using Score Vector Modeling and Support Vector Machine, 350-353.
8. Muhammad Pervez Akhter, Zheng Jiangbin, Irfan Raza Naqvi, Mohammed AbdelMajeed, Tehseen Zia(2021). Abusive language detection from social media comments using conventional machine learning and deep learning approaches. Multimedia Systems, Springer, https://doi.org/10.1007/s00530-021-00784-8
9. Ming Li, Hongbin Suo, Xiao Wu, Ping Lu and Yonghong Yan, "Spoken Language Identification Using Score Vector Modeling and Support Vector Machine", proc. 8th annual conference of the international speech communication association, pp. 351, 2007.
10. Kumar, V., Recupero, D. R., Riboni, D., & Helaoui, R. (2021). Ensembling Classical Machine Learning and Deep Learning Approaches for Morbidity Identification from Clinical Notes. IEEE Access, 9, 7107–7126. https://doi.org/10.1109/ACCESS.2020.3043221 https://www.javatpoint.com/machine-learning-random-forest-algorithm