Language Detection using Natural Language Processing

Yashashvini Rachamallu

Abstract

Since the dawn of the internet, the world has come closer unlike never. Internet has completely removed the barrier in communication due to long distance between two people in different corners of the world. With increase in the power of internet and social media, people from different language background have started to interact. With various people with different language, they have started to put content on internet in their own language which is verry difficult for person who needs that information to read since he has no clue of that language. This has given rise to new form of data called multilingual data, where you have tons of data in different language. With the growing volume of multilingual data on the web, language detection has become a crucial task in various NLP applications, such as machine translation and sentiment analysis. However, for the case of language detection, the previous works predominantly use machine learning models and very few variations in text vectorization. In our project, we intend to explore the use of deep learning for the classification of languages and the effect of different text-vectorization methods. The dataset was taken as a mixture of 4 different datasets, to facilitate better training procedures. We will also conduct experiments to evaluate the impact of different parameters on the performance of the models. We aim to demonstrate the effectiveness of deep learning techniques for language detection in NLP and provide insights into the performance of various deep learning models and text vectorizers for this task.

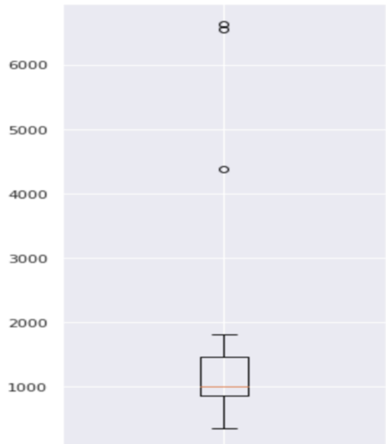
1 Introduction

Identification of a language typical refers to a method that aims to classify text into a preset set of accessible languages in a language. It is critical in many NLP applications to be able to accurately identify the language in which the current input is written. There can be variety of applications, like identifying the language on sign board, translation of language and many more. This capability, for example, is required to load the appropriate autocorrection lexicon and language model for predictive and multilingual typing, machine translation, information retrieval, summarization, and question answering, etc., language identification is a vast study topic. Properly implemented it will be highly beneficial in breaking the barriers of unknown language while communicating. Say for example you visit Quebec in Canada where French is dominant language, and you don't know it. Well with the help of NLP you can not only identify what is written on menu in French in French Restaurant but also translate it into English. It can also be used to translate what a person is speaking in French into English. Before processing, they must first identify the language.

There are two ways to identify languages: computational and non-computational. Non-computational approaches necessitate that authors have sufficient linguistic expertise to recognize diacritics, symbols, the most frequently used character combinations, and other elements. Statistical methods, as opposed to linguistic understanding, are used in computational ways to handle similar challenges, an essential NLP approach, especially when working with text that rely on classification and language. Unlike non-computational approach in statistical approach, you can use your language identification supported by NLP to easily carry with you to place where's language is unknown for you and help yourself out for better communication. One best thing of statistical approach is you got one model that can do translation for numerous numbers of language. You don't need a assistance of person who known's language for which you need help. The goal of this project is to use the larger datasets which contains more unique languages, and able to make a classification model using custom neural network, and compare the results with the existing ones, by using different text-vectorization methods.

2 Dataset

We decided to work on a merge of 4 different datasets. First, we combined 3 smaller datasets. The dataset (4) has two files, the CSV with sentences and a JSON which has the corresponding language for abbreviation in the dataset. We retrieved the language from JSON and filled in the code in the data frame. The final dataset has 10^7 sentences in it, which was extremely challenging for our systems to process. Hence, to cut down on data, using the below box plot,



The links from where the datasets were taken are :

1. https://www.kaggle.com/datasets/basilb2s/language-detection
2. https://www.kaggle.com/code/martinkk5575/language-detection/data
3. https://www.kaggle.com/datasets/lailaboullous/language-detection-dataset
4. https://www.kaggle.com/datasets/chazzer/big-language-detection-dataset?select=sentences.csv

The first three datasets contain 10267, 21859 and 12646 unique rows respectively. Languages that have number of rows between 1000 and 6000 are considered from the fourth dataset to increase the overall median and reduce the gap to outliers.

As a result, we now have a final data frame with 88 distinct languages.

3 Previous Works

Language detection is a crucial task in NLP that has received significant attention from researchers in recent years. In previous works, machine learning models have been predominantly used for language detection tasks. For example, Canvar and Trenkle [2] used n-gram based text categorization, while Ahmed et al. [3] used n-gram based cumulative frequency addition for language identification from text.

Sarma et al. [1] focused on word level language identification in code-mixed social media text of Assamese, Bengali, Hindi, and English languages. They used a machine learning approach based on Support Vector Machine (SVM) and achieved a maximum accuracy of 97.5%. A recent work by H. Singh and P. Singh [4] used machine learning algorithms such as Decision Tree (DT), K-Nearest Neighbor (KNN), Random Forest (RF), and Support Vector Machine (SVM) for language identification. They used a dataset of five languages (Hindi, Punjabi, Bengali, Telugu, and Tamil) and achieved a maximum accuracy of 97.6% using the Random Forest algorithm.

Another recent work by A. Singh and A. Gupta [5] gives insights of the various important terminologies of NLP and NLG and can be useful for the readers interested to start their early career in NLP and work relevant to its applications. It also focuses on the history, applications, and recent developments in the field of NLP. The third objective is to discuss datasets, approaches and evaluation metrics used in NLP.

[6] One more work was published which used language identification system (LID). LID helps us to identify spoken language say an audio file in any language. This makes it first choice to use in technologies like Speech Recognition. Mathur et al. [7] proposed a robust language identifier Stanford Language Identification Engine (SLIDE) to identify languages such as Nepali, Bengali etc used in micro blogging websites.

The authors of [8] studied code switching between Hindi and English multilingual speakers in twitter. They proposed a neural stacking model to uses parts of speech efficiently during parsing. A work on language identification of English and Gujarati code mix data[9] shows how a word with same set of characters in English and Gujarati have different set of meanings and how it will be difficult for the model to identify.

4 Methodology

In this work, we used a dataset that contains data from multiple languages like English, French, German, Spanish etc. The following steps have been implemented:

Text Preprocessing: We have processed the text data by removing special characters, punctuation, htmls, email addresses, and stop words.

Classification Modeling:

1. Machine Learning based classification: We have intended to use different classification algorithms and different text vectorization methods. We have started with text vectorization of the sentences, to allow classification algorithms to process the data. Two text vectorization methods, Bag of Words and TF-IDF vectorizer have been used to convert the text data into vectors. Following the text vectorization, we have worked with 2 machine learning classification algorithms namely, Naive Bayes text classifier and Decision tree classifier.

2. Deep Learning based classification: We are experimenting with various text vectorization techniques to represent the text data as numerical features that can be used as input to our deep learning models. We experimented with various vectorization techniques, includes tokenizer, and a pretrained model BERT. We implemented a neural network with 4 variations. The variations include LSTM layers, dropout layers, spatial dropout layer, Bi-LSTM layers. We have trained our models on 5 epochs with different learning rates and used Adam optimizer and the loss function as categorical cross entropy.

Training & Testing: The dataset is divided into 80% for training and 20% for testing.

5 Results And Analysis

Below table represents the classification accuracies of models by varying embedding styles on the test data:

Model	Accuracy
BoW + Naïve Bayes	0.871
BoW + Decision Trees	0.785
TF-IDF + Naïve Bayes	0.85
TF-IDF + Decision Trees	0.76
BERT embedding + NN	0.72
LSTM + 1 Dense	0.870
LSTM + 2 Dense	0.867
Bi-LSTM + 2 Dense	0.856

Table 1: Results

From the above table, the models that used bag of words as text vectorizer method has higher an accuracy compared to models that used TF-IDF as text vectorizer. The BoW + naïve Bayes and LSTM + 1 Dense models have almost similar accuracy and outperformed other models.

It is interesting to note that the accuracy of model using BERT embeddings is quite less compared to other models. The reason for such an output is due to less amount of training data for few languages and model's inability to perform identification. We will explore on how we can improve the ability of the model to identify these languages by getting more data, make changes to preprocessing.

6 Next Steps

We would be exploring the following to improve the performance of models:

- Add more training data to languages with less performance.
- Understand the data by performing sentimental analysis, and n-gram analysis.

- Improve neural networks models by adding more layers, change hyper parameters, change optimizers etc.

7 Workload Distribution

Below is the workload distribution of the team members for the entire project.

Student	Task
Yashashvini Rachamallu	Proposal, Datasets collection, Preprocessing, classification models implementation, Training & Testing
Bhanu Kanamarlapudi	Preprocessing, vectorization, classification models implementation, Training & Testing , Mid report
Neel Joshi	Vectorization, classification models implementation, Training & Testing, Final report

Table 2: Work Distribution

References

[1] Sarma, N., Singh, S.R., & Goswami, D. (2018). Word Level Language Identification in Assamese-Bengali-Hindi-English Code-Mixed Social Media Text. 2018 International Conference on Asian Language Processing (IALP), 261-266.

[2] W. B. Canvar and J. M. Trenkle. N-gram based Text Categorization. Proceedings of Symposium on Document Analysis and Information Retrieval, University of Nevada, Las Vegas, pp. 161-176, 1994.

[3] B. Ahmed, S.-H. Cha, and C. Tappert. Language Identification from Text Using N-gram Based Cumulative Frequency Addition. Proceedings of Student/Faculty Research Day, CSIS, Pace University, 2004.

[4] Pujeri*, B. P., & Sai D, J. (2020). An anatomization of language detection and translation using NLP techniques. *International Journal of Innovative Technology and Exploring Engineering*, *10*(2), 69–77. https://doi.org/10.35940/ijitee.b8265.1210220.

[5] Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4.

[6] Christian Bartz, Tom Herold, Haojin Yang, Christoph Meinel(2017). Language Identification Using Deep Convolution Recurrent Neural Networks. https://arxiv.org/abs/1708.04811.

[7] Priyanka Mathur, Arkajyoti Misra, Emrah Budur(2015). Language Identification from Text Document. https://cs229.stanford.edu/proj2015/324_report.pdf

[8] Bhat, Irshad Ahmad, et al. "Universal Dependency parsing for Hindi-English code-switching." *arXiv preprint arXiv:1804.05868* (2018).

[9] D. Patel and R. Parikh, "Language Identification and Translation of English and Gujarati code-mixed data," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-4, doi: 10.1109/ic-ETITE47903.2020.410.