Report On

Language Identification

Submitted in partial fulfillment of the requirements of the Course project in Semester VIII of Final Year Computer Engineering

by
Siddhant Vasaikar (Roll No. 81)
Sandesh Thakare (Roll No. 77)
Siddharth Vartha (Roll No. 80)

Mentor
Dr. Tatwadarshi Nagarhalli

University of Mumbai

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

(A.Y. 2021-22)

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

CERTIFICATE

This is to certify that the Mini Project entitled **“Language Identification”**is a bonafide work of **Siddhant Vasaikar (Roll no. 81), Sandesh Thakare (Roll no. 77), Siddharth Vartha (Roll no. 80)**submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **“Bachelor of Engineering”** in Semester VIII of Final Year **“Computer Engineering” .**

Dr. TatwadarshiNagarhalli
Mentor

Dr Megha Trivedi
Head of Department

Dr. H.V. Vankudre
Principal

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Course Project Approval

This Mini Project entitled **“Language Identification”**by **Siddhant Vasaikar (Roll no. 81), Sandesh Thakare (Roll no. 77), Siddharth Vartha (Roll no. 80)**is approved for the degree of **Bachelor of Engineering** in in Semester VIII of Final Year **Computer Engineering.**

Examiners

1……………………………………
(Internal Examiner Name & Sign)

2………………………………………
(External Examiner name & Sign)

Date:

Place:

Contents

ABSTRACT

Natural language processing, or NLP, is a type of artificial intelligence that deals with analyzing, understanding, and generating natural human languages so that computers can process written and spoken human language without using computer-driven language. Natural language processing, sometimes also called “computational linguistics,” uses both semantics and syntax to help computers understand how humans talk or write and how to derive meaning from what they say. This field combines the power of artificial intelligence and computer programming into an understanding so powerful that programs can even translate one language into another reasonably accurately. This field also includes voice recognition, the ability of a computer to understand what you say well enough to respond appropriately. This all together can be used to identify languages. Language identification (“LI”) is the problem of determining the natural language that a document or part thereof is written in. Automatic LI has been extensively researched for over fifty years. Today, LI is a key part of many text processing pipelines, as text processing techniques generally assume that the language of the input text is known. Research in this area has recently been especially active. This article provides a brief history of LI research, and an extensive survey of the features and methods used in the LI literature. We describe the features and methods using a unified notation, to make the relationships between methods clearer. We discuss evaluation methods, applications of LI, as well as off-the-shelf LI systems that do not require training by the end user. Finally, we identify open issues, survey the work to date on each issue, and propose future directions for research in LI.

1.1INTRODUCTION

Every Machine Learning enthusiast has a dream of building/working on a cool project, isn't it? Mere understandings of the theory aren't enough, you need to work on projects, try to deploy them, and learn from them. Moreover, working on specific domains like NLP gives you wide opportunities and problem statements to explore. Through this article, I wish to introduce you to an amazing project, the Language Detection model using Natural Language Processing. This will take you through a real-world example of ML(application to say). So, let's not wait anymore. Language identification ("LI") is the task of determining the natural language that a document or part thereof is written in. Recognizing text in a specific language comes naturally to a human reader familiar with the language. Table 1 presents excerpts from Wikipedia articles in different languages on the topic of Natural Language Processing ("NLP"), labeled according to the language they are written in. Without referring to the labels, readers of this article will certainly have recognized at least one language in Table 1, and many are likely to be able to identify all the languages therein.

1.2PROBLEM STATEMENT

Language identification can be an important step in a Natural Language Processing (NLP) problem. It involves trying to predict the natural language of a piece of text. It is important to know the language of text before other actions (i.e. translation/ sentiment analysis) can be taken. For instance, if you go to google translate the box you type in says 'Detect Language'. This is because Google is first trying to identify the language of your sentence before it can be translated. Development Time Phrasing Ambiguities, Misspellings, Language Differences Training Data, Innate Biases, Words with Multiple Meanings are some problems we try solving.

.

OBJECTIVE

1. To predict different languages.

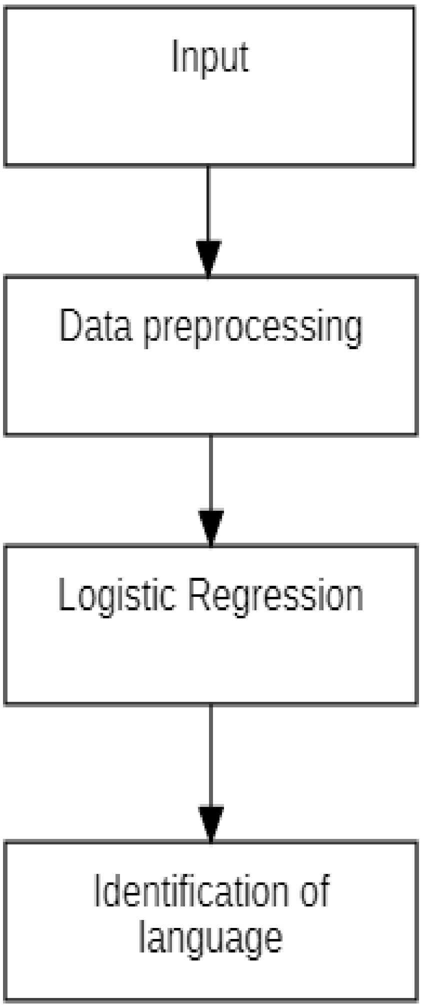
1.3SCOPE

NLP has a broad scope, with so many uses in customer service, grammar check software, business marketing, etc. If you are interested in computing and languages, then NLP is a good career option for you. You can consider career options like NLP Engineer, NLP Architect, etc. This can we widely used for identifying which actual language is used. It involves trying to predict the natural language of a piece of text. It is important to know the language of text before other actions (i.e. translation/ sentiment analysis) can be taken.

2.1Introduction

Research into LI aims to mimic this human ability to recognize specific languages. Over the years, a number of computational approaches have been developed that, through the use of specially-designed algorithms and indexing structures, are able to infer the language being used without the need for human intervention. The capability of such systems could be described as super-human: an average person may be able to identify a handful of languages, and a trained linguist or translator may be familiar with many dozens, but most of us will have, at some point, encountered written texts in languages they cannot place. However, LI research aims to develop systems that are able to identify any human language, a set which numbers in the thousands (Simons and Fennig, 2017). In a broad sense, LI applies to any modality of language, including speech, sign language, and handwritten text, and is relevant for all means of information storage that involve language, digital or otherwise. However, in this survey we limit the scope of our discussion to LI of written text stored in a digitally-encoded form. Research to date on LI has traditionally focused on monolingual documents (Hughes et al., 2006) (we discuss LI for multilingual documents in Section 10.6). In monolingual LI, the task is to assign each document a unique language label. Some work has reported nearperfect accuracy for LI of large documents in a small number of languages, prompting some researchers to label it a “solved task” (McNamee, 2005). However, in order to attain such accuracy, simplifying assumptions have to be made, such as the aforementioned monolinguality of each document, as well as assumptions about the type and quantity of data, and the number of languages considered. The ability to accurately detect the language that a document is written in is an enabling technology that increases accessibility of data and has a wide variety of applications. For example, presenting information in a user’s native language has been found to be a critical factor in attracting website visitors (Kralisch and Mandl, 2006). Text processing techniques developed in natural language processing and Information Retrieval (“IR”) generally presuppose that the language of the input text is known, and many techniques assume that all documents are in the same language. In order to apply text processing techniques to real-world data, automatic LI is used to ensure that only documents in relevant languages are subjected to further processing

2.2Architecture/ Framework/Block diagram



2.3Algorithm and Process Design

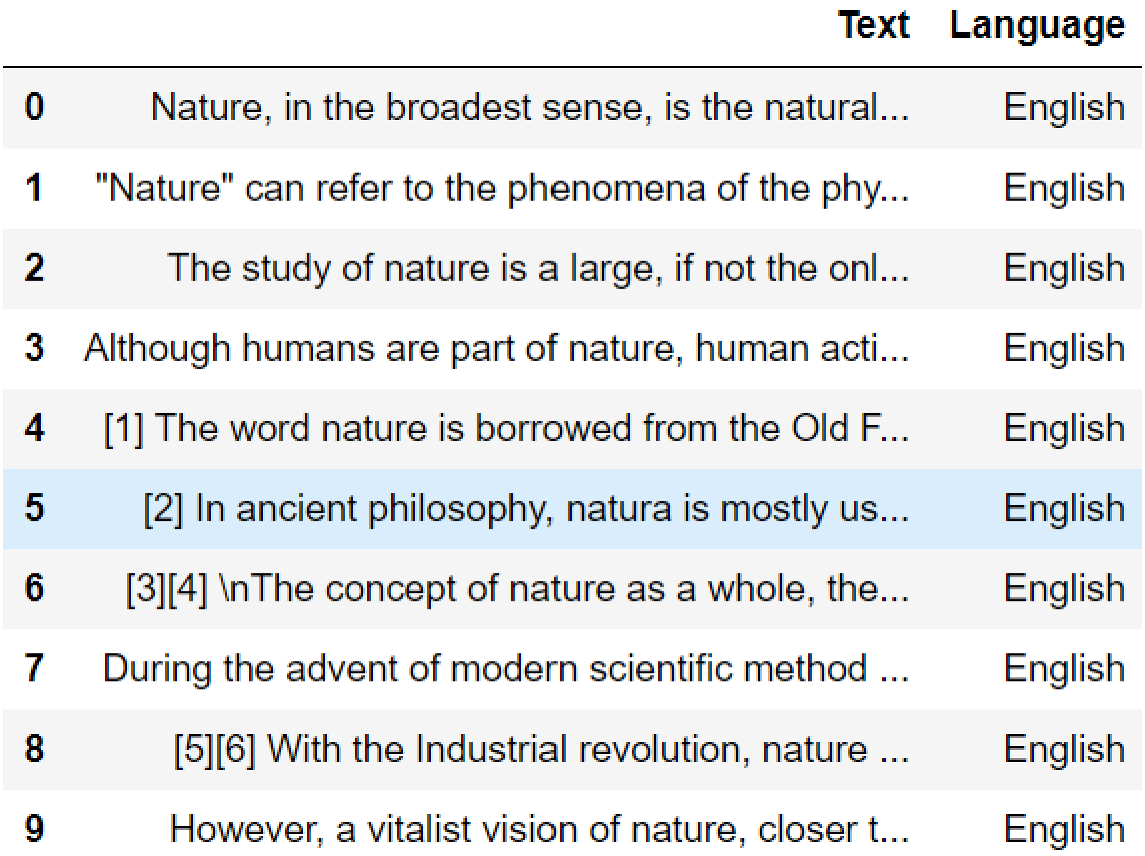
Importing libraries and dataset

So let’s get started. First of all, we will import all the required libraries.

```
import pandas as pd
import numpy as np
import re
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.simplefilter("ignore")
```

Now let’s import the language detection dataset

```
data = pd.read_csv("Language Detection.csv")
data.head(10)
```



	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
7	During the advent of modern scientific method ...	English
8	[5][6] With the Industrial revolution, nature ...	English
9	However, a vitalist vision of nature, closer t...	English

As I told you earlier this dataset contains text details for 17 different languages. So let’s count the value count for each language.

```
data["Language"].value_counts()
```

Output :

```
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
```

Separating Independent and Dependent features

Now we can separate the dependent and independent variables, here text data is the independent variable and the language name is the dependent variable.

```
X = data["Text"]
y = data["Language"]
```

Label Encoding

Our output variable, the name of languages is a categorical variable. For training the model we should have to convert it into a numerical form, so we are performing label encoding on that output variable. For this process, we are importing LabelEncoder from sklearn.

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
y = le.fit_transform(y)
```

Text Preprocessing

This is a dataset created using scraping the Wikipedia, so it contains many unwanted symbols, numbers which will affect the quality of our model. So we should perform text preprocessing techniques.

```
# creating a list for appending the preprocessed text
data_list = []
# iterating through all the text
for text in X:
    # removing the symbols and numbers
     text = re.sub(r'[!@#$(),n"%^*?:;~`0-9]', ' ', text)
     text = re.sub(r'[[]]', ' ', text)
     # converting the text to lower case
     text = text.lower()
     # appending to data_list
     data_list.append(text)
```

Bag of Words

As we all know that, not only the output feature but also the input feature should be of the numerical form. So we are converting text into numerical form by creating a Bag of Words model using CountVectorizer.

```
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer()
X = cv.fit_transform(data_list).toarray()
X.shape # (10337, 39419)
```

Train Test Splitting

We preprocessed our input and output variable. The next step is to create the training set, for training the model and test set, for evaluating the test set. For this process, we are using a train test split.

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(X, y, test_size = 0.20)
```

Model Training and Prediction

And we almost there, the model creation part. We are using the naive_bayes algorithm for our model creation. Later we are training the model using the training set.

```
from sklearn.naive_bayes import MultinomialNB
model = MultinomialNB()
```

```
model.fit(x_train, y_train)
```

So we’ve trained our model using the training set. Now let’s predict the output for the test set.

```
y_pred = model.predict(x_test)
```

Model Evaluation

Now we can evaluate our model

```
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
ac = accuracy_score(y_test, y_pred)
cm = confusion_matrix(y_test, y_pred)

print("Accuracy is :",ac)
# Accuracy is : 0.9772727272727273
```

The accuracy of the model is 0.97 which is very good and our model is performing well. Now let’s plot the confusion matrix using the seaborn heatmap.

2.4 Details of Hardware & Software

Hardware

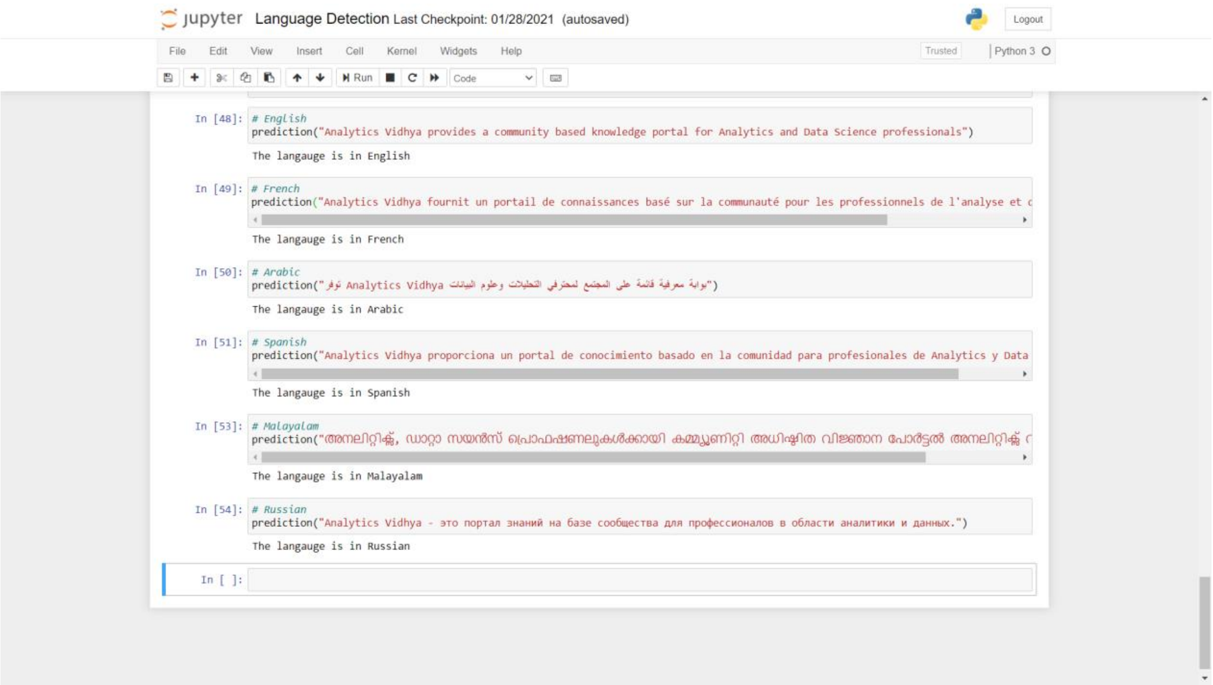
- Intel i5 processor
- RAM – 8GB
- Hard disk – 1TB
- Web browser
- Internet Connection

Software

- Jupyter Notebook
- Python — a programming language
- Pandas — data manipulation and analysis library
- NumPy — scientific computing library
- Matplotlib — plotting library

2.5Experiment and Results for Validation and Verification

Visualizing data gives a clearer picture of what we are doing. Here, we put different languages for identification



```
In [48]: # English
         prediction("Analytics Vidhya provides a community based knowledge portal for Analytics and Data Science professionals")

         The langauge is in English

In [49]: # French
         prediction("Analytics Vidhya fournit un portail de connaissances basé sur la communauté pour les professionnels de l'analyse et d

         The langauge is in French

In [50]: # Arabic
         prediction("توفر Analytics Vidhya بوابة معرفية قائمة على المجتمع لمحترفي التحليلات وعلوم البيانات")

         The langauge is in Arabic

In [51]: # Spanish
         prediction("Analytics Vidhya proporciona un portal de conocimiento basado en la comunidad para profesionales de Analytics y Data

         The langauge is in Spanish

In [53]: # Malayalam
         prediction("അനലിറ്റിക്സ്, ഡാറ്റാ സയൻസ് പ്രൊഫഷണലുകൾക്കായി കമ്മ്യൂണിറ്റി അധിഷ്ഠിത വിജ്ഞാന പോർട്ടൽ അനലിറ്റിക്സ്

         The langauge is in Malayalam

In [54]: # Russian
         prediction("Analytics Vidhya - это портал знаний на базе сообщества для профессионалов в области аналитики и данных.")

         The langauge is in Russian

In [ ]:
```

From the above figure, it is clear that accurate languages are identified.

2.6 Conclusion .

This article has presented a comprehensive way on language identification of digitally encoded text. We have shown that LI is a rich, complex, and multi-faceted problem that has engaged a wide variety of research communities. LI accuracy is critical as it is often the first step in longer text processing pipelines, so errors made in LI will propagate and degrade the performance of later stages. Under controlled conditions, such as limiting the number of languages to a small set of Western European languages and using long, grammatical, and structured text such as government documents as training data, it is possible to achieve near-perfect accuracy. Modern approaches to LI are generally data-driven and are based on comparing new documents with models of each target language learned from data.

2.7 References

[1] J. Hakkinen and Jilei Tian, "n-gram and decision tree based language identification for written words," IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU '01., 2001, pp. 335-338, doi: 10.1109/ASRU.2001.1034655.
[2 J. Tang, X. Chen and W. Liu, "Efficient Language Identification for All-Language Internet News," 2021 International Conference on Asian Language Processing (IALP), 2021, pp. 165-169, doi: 10.1109/IALP54817.2021.9675270.
[3] C. Sreejith, M. Indu and P. C. R. Raj, "N-gram based algorithm for distinguishing between Hindi and Sanskrit texts," 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 2013, pp. 1-4, doi: 10.1109/ICCCNT.2013.6726777.