Language Detection Using NLP

Saurabh Kumar
Computer Science and Engineering Department
Galgotias University
Greater Noida, Uttar Pradesh, India
saurabh5566kmr@gmail.com

Adarsh Pratap Singh
Computer Science and Engineering Department
Galgotias University
Greater Noida, Uttar Pradesh, India
shivasingh9393@gmail.com

Vatsal Pandey
Computer Science and Engineering Department,
Galgotias University
Greater Noida, Uttar Pradesh, India
vatsalpandey13@gmail.com

***Abstract*— A key task in natural language processing (NLP) is language detection, commonly referred to as language identification. This abstract examines language detection methods and uses in relation to contemporary information technology. Language detection, which identifies the language in which a text or document is written, is used in a variety of contexts, such as sentiment analysis, content filtering, machine translation, and search engine optimization. The importance of language detection in promoting cross-lingual communication in our increasingly globalized society is emphasized in this abstract, which highlights the fundamental methodologies, difficulties, applications, and future trends in language detection.**

Keywords: *Language detection, NLP, content filtering, search engines, multilingual chatbots, translation services, social media analysis.*

I. Introduction

Languages can be processed and transformed into forms that are easy for users to understand or interpret using a technology called natural language processing, or NLP. Pattern recognition is the foundation of NLP, a computer programming approach [1]. This system consists of Natural Language Understanding (NLU) and Natural Language Generation (NLG).

Whether the text is spoken or written, we can utilize NLU to ascertain the meaning of a particular phrase or section. NLG generates coherent phrases taken from a text representation or data set.

Language Detection functions on top of NLP. NLP is used for language identification and processing. Several word and language kinds can be identified with the use of NLP.

NLP helps to identify language and word meaning and analyzes text that is delivered. Business writing can be easily identified with the use of NLP. Once the datasets have been identified, proceed to ©2023 IEEE 979-8-3503-9737-6/23/$31.00

NLP helps us implement and detect several languages by identifying the language family to which each one belongs and analyzing the text to ascertain its meaning and intent. With the use of various datasets and libraries, NLP can be used to accomplish the same goal with greater assistance and coverage. Because NLP applications are language specific, most of them require monolingual data. Preprocessing and text filtering may be necessary. It is composed in languages other than that of the intended target, with the goal of developing an application in that language [2]. The exact language of each input, for example, needs to be declared. The processing stages of natural language encompass lexical (structural), syntactic, semantic, pragmatic, and discourse synthesis analysis. Linguistic communication frequently uses voice detectors, scanners, computational linguistics, and text chats. By analyzing vast examples of words written by humans (words used in conversation, keywords, and details), we use artificial intelligence (AI) approaches these days to operate tongue words [3]. By examining these patterns, training algorithms can perceive the "context" of written language, spoken language, and other human communication mediums. As natural language processing and language detection technologies advance, their uses are becoming more widespread in today's society.

II. Literature Review

While syntactic structures, the "Turing Test," and its rule-based system were developed in 1950 and 1957, respectively, the true beginning of NLP research was in the late 1940s. Up to 1990, progress was sluggish due to a lack of computer power, the usage of handwritten rule systems- based systems, and limited vocabulary. As the continued increase in computer power and the progress made in machine learning are responsible for the recent surge in interest in research and applications [15].

Speech recognition, dialogue systems, language processing, and the use of deep learning techniques are some of the most recent important NLP breakthrough areas.

The application of NLP approaches in digital transformation, robotics, and automation has become increasingly popular and has sparked a lot of research interest.

the obstacles that it still has to overcome, like those pertaining to HCI [3].

For the most part, machine translation and NLP concepts were studied before 1990. In the last few years, NLP research has made extensive use of statistical models, deep learning, and machine learning. Sometimes there are similarities between natural language processing and deep learning/artificial intelligence research. In order to complete NLP tasks as efficiently as possible, these techniques are now frequently used [1].

Speaking with a computer will eventually be just as easy as speaking with a human. To give unstructured data meaning for a machine, NLP continues to use it. Natural Language Processing (NLP) will remain beneficial to a number of industries, such as linked cars, smart homes, robotics, healthcare, and finance[2]. Early in machine translation in the twenty-first century between human languages was one of the earliest applications of natural language processing (NLP)[13].

Nonetheless, the customer service sector quickly came to appreciate it. Virtual assistants are the most popular NLP customer service tool. or "Chatbot." Different industries employ different applications. The following is a list of these:

A. Systems for conversation

An automated system can have a natural-language conversation with us through a speech or text interface thanks to a conversational system [2]. Their assistance lies in helping businesses automate difficult tasks and provide client support available around-the-clock. Of all conversational devices, chatbots and virtual assistants are the most popular. In the modern era, self-service point-of-sale social media, banking, and e-commerce all use these two gadgets to offer their clients a variety of services.

B. Text Analytics

Extracting valuable information from text, whether it be in shorter texts like tweets and SMS texts or longer ones like emails and documents, is the aim of text analytics, also known as text mining [23]. One of the most prevalent applications social media in text analytics analysis.

C. Machine Translation

Automatically translating text from Machine translation aims to convert text from one natural language to another while preserving its intended meaning.

The most popular machine translation tool is Google Translate. Education and speech recognition both use machine translation software [14].

NLP is also utilized in the following industries: finance, retail, automotive, healthcare, manufacturing, and education.

Hospitals are using virtual assistants created by merging machine learning, computer vision, and natural language processing. When these virtual assistants communicate with patients, they will automatically create and gather patient histories [12][25]. Common duties like scheduling appointments and patient registration are handled by virtual assistants.

Among the most notable innovations in the manufacturing sector are self-driving automobiles.

In the banking sector, NLP-based solutions enable applications including document search, credit scoring, and sentiment analysis. Credit scoring programs help banks and other financial institutions determine an individual's creditworthiness and generate a credit score by using machine learning and natural language processing (NLP). Sentiment analysis applications automate named entity recognition and document categorization processes to choose the data most pertinent to investor needs [23]. Chatbot interfaces are employed by banks and other financial institutions to enable their customers to perform information searches and obtain straightforward transactional responses through document search applications [24].

Two very promising areas for NLP applications are robotics and process automation. Natural language processing (NLP) allows a robot on a production line to converse with a remote human operator to handle directions for putting machines and goods together and moving them [4].

Placing a retail virtual assistant in front of a store allows it to recognize Through the use of Natural Language, Computer Vision, and Machine Learning technologies, you can quickly ascertain what the customer needs and provide them information and special offers [10].

Students may be able to access a virtual classroom through an educational platform. because natural language processing and computer vision are integrated. Utilizing specialized information from online libraries, Assisting students with problem solving has already been done with digital assistants [9].

D. NLP Development Frameworks and Instruments

The global The current generation of Because open-source communities have demonstrated interest in development tools, they are readily available [6]. These frameworks and tools come with built-in libraries and can be tailored to meet particular industry standards.

Natural language knowledge is expressed through the natural language representation block, which uses structured, tree, or graph models [7]. A Natural Language database is an assortment of Natural Language data that machine learning algorithms use to complete additional NLP tasks, much like MNIST or other databases.

The representation and transformation blocks work with this database to accomplish their tasks. In order to extract useful and relevant tasks from NLP jobs, Many techniques for learning and extraction will be employed in natural language transformation [5]. Natural language communication refers to the way in which tasks facilitated by natural language processing (NLP) present the intended and desired behaviors [11]. Natural Language may be produced in the end, or computer activity such as a robot arm moving.

Conversations between people have led to the development of natural language processing. There is little doubt that the process will entail translating natural human language into a format that is comprehensible to machines. NLP could be used for the following tasks:

Word Sense Ambiguation: This method uses semantic analysis to identify the meaning of a word that has several meanings and choose the one that makes the most sense in a given situation.

Audio data is transformed into text data through a process called speech recognition.

Words are recognized as pertinent and helpful entities by Named Entity Recognition.

2) Segment of speech tagging: It determines a

segment of speech specific textual segment within a sentence or informational piece based on the most appropriate context.

NLP is divided into two categories: natural language generation (NLG) and natural language understanding (NLU).

NLU: It involves the following-

a. Lexical ambiguity: This occurs when a word's appropriate and pertinent meaning needs to be determined within a text.

b. Referential ambiguity: This occurs a word's multiple appearances within a sentence.

c. Syntactical Ambiguity: Perceiving multiple meanings within a text.

NLG: This method involves translating structured data into understandable language for humans [20]. It transforms a representation of data or text into meaningful sentences.

a. **That comprises:** Sentence planning is the process of choosing appropriate words and phrases for a given text passage.

b. **Text planning:** This allows us to pull pertinent data from a knowledge base, including facts and figures.

c. **Text Realization:** This is how sentence structure and sentence plan are mapped.

Sentiment analysis is another method of natural language processing that makes use of statistics to ascertain the emotional content's meaning and purpose.

One subset of NLP is called Language Detection (LD). As was previously mentioned, it operates on the foundation of NLP [19]. This is where a specific written work or knowledge base's language and linguistics are assessed and identified in their form. This is where it is determined which language the content is in [11]. Various statistical techniques are employed to solve this problem, which is viewed by computational approaches as a unique instance of text classification [21]. LD is an excellent method for quickly and effectively classifying and sorting data, as well as applying extra language-specific workflow layers [22]. It can assist us in recognizing and detecting spelling or grammar mistakes in a specific document. For instance, let's say we write a sentence in English with a specific spelling mistake [18]. The system can then assist us in analyzing the text and identifying the language that is written as "English" by employing the principle of Language Detection to help us find and fix spelling mistakes in words that are written incorrectly. NLP contains numerous libraries, including genism, spaCy, and NLTK [16].

III. METHODOLOGY

"Google Colab" Platforms are used for implementation. An already-prepared "Language Detection Using NLP" file is used to load the data. A dataset from Github and Kaggle is used to train a model. Based on the requirements, Out of all the languages included in the downloaded dataset, only a select few were selected. We'll walk through each implementation step-by-step and in detail.

STEP - I

The Importing all the libraries and packages required to complete the task is the first step. i.e.

STEP-II

Next, you mount a dataset to Google Drive from your local PC. Dataset is uploaded as a zip file to the Google Drive cloud storage service.Right now, Google Drive's "Google Colab" environment is where the dataset is mounted. On the distributed server of We can access about 80 GB of local storage in the Google Colab Environment.

STEP- III

The read_csv() function can be used to extract data from a CSV file into a data frame.

STEP -IV

We will now define the essential variable, which is crucial to the development of our machine learning model. The variable names and corresponding values are displayed in the picture.

Fig. 1. Defining the Variable

The variables in Fig. 1 have the following definitions: head

Python's head function displays the first five rows of the data frame by default. Its only parameter is the number of rows. We can make use of this parameter to see the number of rows that we desire.

To obtain the first n rows of the dataframe, use head(n). The number of rows you wish to obtain from the beginning, n, is one of its optional arguments.

- value_count

The function value_counts() returns an object containing the counts of unique values, or value counts. Subsequently, the object will manifest in descending order, starting with the element that occurs most frequently.

STEP - V

The Using the class label encoder from the sklearn module a description of its entire usage follows below:

The categorical feature levels are encoded into numerical values by Sklearn, which provides an incredibly powerful tool. In order to transform the labels into a format that a machine can read, a process known as label encoding must be performed. Following that, decisions about how to use the labels can be made more intelligently by machine learning algorithms. It is an important step in the structured dataset pre-processing for supervised learning.

With n being the number of unique labels, Label Encoder can encode labels with any value between 0 and n_classes-1. A label that appears more than once is given the value from the previous instance using the fit transform() method, Fitting the label encoder converts binary labels from multi-class labels. The 1-of-K coding scheme is the term used to refer to the outcome of this conversion. Figure 2 displays Fit_Trasform and LableEncoder.

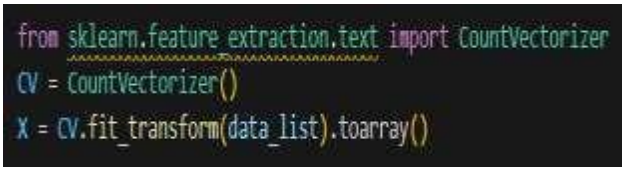

Fig. 2.Including Fit_Trasform and LableEncoder

STEP - VI

In Python code, the data_list array is created using the Regular Expressions (re) module, which also makes use of the re.sub() function.

When it returns a string, the replace string will take the place of every instance that matches the provided pattern. A string with replaced values is the result of using the re.sub() function to hold for a substring. With the use of a list and this function, we can swap out several elements.

Every alphabet can be converted to lower case using the lower function.

The append function is then used to add the text to the data_list array.

We'll refer to the usage guidelines provided below when using the sklearn.feature_extraction module class from the sklearn module. The sklearn.feature_extraction module is used to extract features. from datasets that include formats such as text and images in a way that is protected and supported by machine learning algorithms.

A count Vectorizer makes use of a technique that is a useful tool offered by the scikit-learn Python library. Depending on the frequency (count) or occurrence of each word that appears in the text, it can be used to turn a given text into a vector [8]. When there are several texts available and we want to translate every word in each one, this can be very useful.

The CountVectorizer tool generates a matrix where each row corresponds to a sample of text from the document and each column represents a unique word. Word count in the provided partial text sample determines the value of each individual cell.

The fit_transform () method will be utilized, which is essentially the transform() counterpart of the fit method plus the transform method. fit_(). This method converts data points by simultaneously applying transform and fit to the input data.

Next, we use the.shape method to set the array X's dimension. In Fig. 3, an array data_list is created.

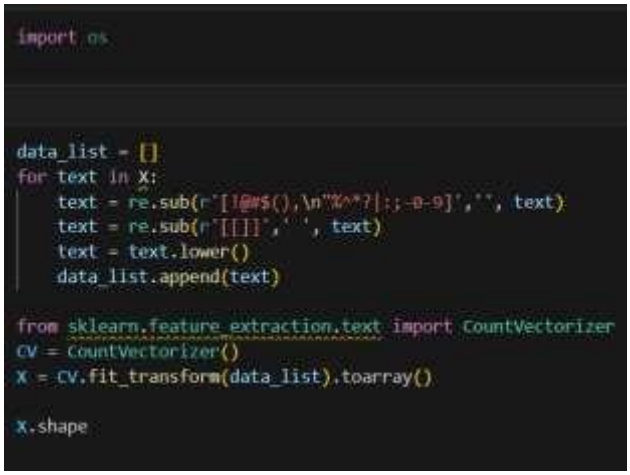

Fig. 3. Creating an array data_list

STEP - VII

The training set and testing set of the list are separated in this snippet of code. Regarding the construction of machine learning models, it is one of the key ideas [17]. Train_test_split, a function in the sklearn module, takes three parameters: X, Y, and test_size. Fig. 4 illustrates how the dataset was divided into training and testing datasets.

Four variables are separated into the training and testing dataset as follows:

1. X_train
2. Y_train
3. X_test
4. Y_test

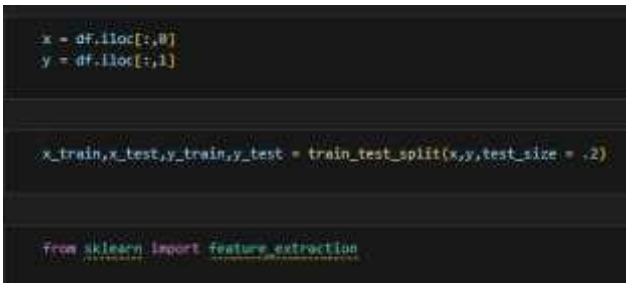

Fig. 4. separating a dataset into two categories: testing and training

STEP - VIII

The MultinomialNb module's model.fit() is used in this step to construct a neural network model.

It's an extra useful Naïve Bayes classification. This assumes that a simple multinomial distribution is used to extract the features. Scikit-Learn provides To implement the Multinomial Naïve Bayes classification algorithm, use sklearn.naive_bayes.MultinomialNB.

We are now using MultinomalNB's fit method, which takes x and y as input. The training vectors, or training data, should now be represented by x, and the target values, or y. Fig. 5 shows the building neural network model.

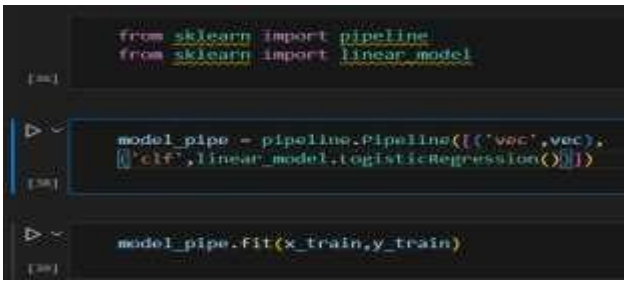

Fig. 5. Building Neural Network Model

Concluding Remarks: We stress the importance of precise language detection in the modern, globalized world and stress the necessity of continued study and advancement in this area in our concluding remarks.

3. We have improved our knowledge of natural language processing for language identification (NLP) by carrying out this research project, which opens the door to more reliable and accurate language identification systems in a range of practical applications.

The model's accuracy is determined in this step. In Fig. 6, the model accuracy is displayed.

Fig. 6. Find the Model Accuracy

STEP-X

We are going to verify our model's accuracy. Fig. 7 displays the model's accuracy.

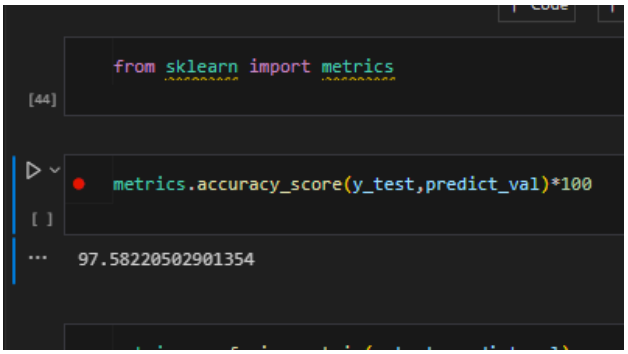

Fig. 7. Accuracy of the model

IV. RESULTS

In this experiment, an overall precision of 0.98 was achieved. The model can be regarded as a good fit for this kind of analysis given its 98% accuracy.

V. CONCLUSION

We highlight the most important discoveries and learnings from our study endeavor in the section that follows:

1. 1 Overview of Results: We provide a summary of the effectiveness of various language detection models, the significance of data preprocessing, and the effect of training data volume.

2. Implications and Future Work: We address how our research may affect the use of content filtering, machine translation, and multilingual search engines, among other applications. Further research directions that we propose to investigate are zero-shot language detection and the application of more sophisticated pre-trained models.

REFERENCES

1. Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. 2018. A Survey of the Usages of Deep Learning in Natural Language Processing. 1, 1

(July 2018), 35 pages.
2. ROBERT DALE. "The commercial NLP Landscape in 2017", Article in Natural Language Engineering, July 2017
3. ACL 2018: 56th Annual Meeting of Association for Computational Linguistics https://acl2018.org
4. Predictive Analytics Today: www.predictiveanalyticstoday.com[accessed in Dec 2018]
5. Ali Shatnawi, Ghadeer Al-Bdour, Raffi Al-Qurran and Mahmoud Al- Ayyoub 2018. A Comparative Study of Open Source Deep Learning Frameworks. 2018 9th International Conference on Information and Communication Systems (ICICS)
6. Intelligent automation: Making cognitive real Knowledge Series I Chapter 2. 2018, EY report.
7. Jacques Bughin, Eric Hazan, SreeRamaswamy, Michael Chui , TeraAllas, Peter Dahlström, Nicolaus Henke, Monica Trench, 2017. MGI ARTIFICIAL INTELLIGENCE THE NEXT DIGITAL FRONTIER? McKinsey & Company McKinsey & Company report July 2017
8. Svetlana Sicular, Kenneth Brant 2018, Hype Cycle for Artificial Intelligence, 2018 Gartner report July 2018.
9. Oshin Agarwal, Funda Durupinar, Norman I. Badler,and Ani Nenkova. 2019. Word embeddings (also) encode human personality stereotypes. In Proceedings of the Joint Conference on Lexical and Computational Semantics, pages 205–211, Minneapolis, MN.
10. Quarteroni, Silvia. (2018). Natural Language Processing for Industry: ELCA's experience. Informatik-Spektrum. 41.10.1007/s00287-018- 1094-1.
11. Young, Tom &Hazarika, Devamanyu&Poria, Soujanya& Cambria, Erik. (2018). Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. IEEE Computational Intelligence Magazine. 13.55-75.10.1109/MCI.2018.2840738.
12. Amirhosseini, Mohammad Hossein, Kazemian, Hassan, Ouazzane, Karim and Chandler, Chris (2018) Natural language processing approach to NLP meta model automation. In: International Joint Conference on Neural Networks (IJCNN), 8-13 July 2018, Rio de Janeiro,Brazil.
13. Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. Proceedings of the 28th International Conference on Computational Linguistics, pages 6838–6855.
14. Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges.
15. Garrett Wilson and Diane J Cook. 2020. A survey of unsupervised deep domain adaptation. ACM Transactions on Intelligent Systems and Technology (TIST), 11(5):1–46.
16. Artem Abzaliev. 2019. On GAP coreference resolution shared task: insights from the 3rd place solution.In Proceedings of the Workshop on Gender Bias in Natural Language Processing, pages 107–112, Florence, Italy.
17. Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong,and Quoc Le. 2020. Unsupervised data augmentation for consistency training.
18. Advances in Neural Information Processing Systems, 33
19. Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender Bias inNeural Natural Language Processing, pages 189–202. Springer International Publishing, Cham.George A. Miller. 1995. Wordnet: a lexical database for english. Communications of the ACM,38(11):39–41.
20. Su Lin Blodgett, Solon Barocas, Hal Daume, III, and ´Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In Proc. of ACL.