TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.02 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.42004

Language Detection: Using Natural Language Processing

Mukesh prajapati, Alok Mishra, Pintu Verma, Abhishek Yadav, Bibhuti Kumar Bhusan

[1,2,3,4]Scholar B tech Final Year, Department of Computer Science & Engineering, Goel Institute of Technology & Management, Lucknow, Uttar Pradesh, India
[5]Associate Professor, Department of computer Science & Engineering, Goel Institute of Technology & Management, Lucknow, Uttar Pradesh, India

KEYWORD

Natural Language Processing, Language Detection, Virtual Assistants, Text Analytics, Machine Learning , artificial intelligence

ABSTRACT

Natural language processing (NLP) is a method for correctly identifying text based on the provided content or topic matter. An extensive study will make it simple to interpret any language and comprehend what is being said. Despite the fact that NLP is a challenging technique, notable examples include Siri and Alexa. Natural language detection allows us to determine the language being used in a given document. A Python-written model that has been utilised in this work can be used to analyse the basic linguistics of any language. The "words" that make up sentences are the essential building blocks of knowledge and its expression. Correctly identifying them and comprehending the situation in which they are used are essential. NLP steps in to help us in this circumstance by making it easier for us to identify the linguistics used in a particular piece of information, whether it be written or vocal. NLP gives computers the ability to understand human language and respond correctly, performing language detection for us. The current paper provides a summary of developments in tongue process, including analysis, establishment, various areas of rapid advancement in natural language processing research, development tools, and techniques.

1. Introduction

Natural Language Processing (NLP) is a technique for processing languages and transforming them into forms that the user can readily process or interpret. NLP is a method of computer programming that is based on pattern learning [1]. It consists of two parts i.e., Natural Language Understanding (NLU) and Natural Language Generation (NLG). We can use NLU to determine the meaning of a specific word or passage of text, whether it is written or spoken. Using a representation of text or data, NLG creates meaningful sentences. NLP is the foundation of how Language Detection operates. Language is processed and identified using NLP. With the aid of

Corresponding Author: Pintu Verma, Department of Computer Science & Engineering, Goel Institute of Technology & Management, Lucknow, Uttar Pradesh, India
Email: verma.pintu950@gmail.com

NLP, different word and language types can be detected. NLP aids in analyzing presented text and identifies language and word meaning. NLP makes it simple to recognise business writings. By identifying the datasets to which each language belongs and evaluating the text to determine its meaning and intent, NLP assists us in implementing numerous languages and detecting them. The same can be implemented using NLP with the use of numerous datasets and libraries for assistance and a wider scope. The majority of NLP applications require data that is monolingual because they are language specific. It can be essential to perform preprocessing and filter out text that is written in languages other than the target language in order to develop an application in the target language [2]. For instance, we must declare each input's precise language. Lexical (structural) analysis, syntactic analysis, semantic analysis, discourse synthesis, and pragmatic analysis are all included in the processing processes of natural language. Voice detector, Scanner, computational linguistics, and text chats are common applications in linguistic communication. These days, we employ artificial intelligence (AI) techniques to operate tongue words by analysing enormous samples of human-written words (conversation, keywords, and details) [3]. Training algorithms can comprehend the "context" of writing, human speech, and other forms of human communication by looking at these patterns. Algorithms for deep learning and machine learning are frequently used to build NLP frameworks and efficiently complete typical NLP tasks [1]. The application of language detection and natural language processing is expanding significantly in the current world as it develops.

2. Literature Review

The work on NLP truly started in the late 1940s, even though the "Turing Test," syntactic structures, and its system that was based on rules were developed in 1950 and 1957, respectively. Up until 1990, growth was sluggish because to inadequate computer power, the use of systems that relied on complex handwritten rule systems, and a narrow vocabulary. Due to the advancement of machine learning and the ongoing expansion of computer power, interest in research and applications has recently surged [15]. The recent major NLP breakthrough areas include speech recognition, dialogue systems, language processing, and the application of deep learning techniques.

NLP has generated a great deal of research interest and opened up many opportunities for using its techniques in automation, robotics, and digital transformation despite the challenges it still faces (such as those related to human computer interfaces) [3]. Prior to 1990, the majority of the research on NLP concepts and machine translation was done. Deep learning, machine learning, and statistical models have been used to great effect in the most recent NLP research. Research in deep learning and artificial intelligence occasionally overlaps with research in natural language processing. Today, these techniques are commonly employed to do NLP tasks in the most efficient way possible [1].

One day, conversing with a machine will be as simple as conversing with a person. NLP continues to use unstructured data to give it meaning for a machine. Industries including robotics, healthcare, finance, linked autos, and smart homes will continue to benefit from NLP [2]. One of the first uses of NLP in the early years of the twenty-first century was machine translation from one human language to another[13]. However, it immediately became well-liked in the customer service industry. The most well-known NLP customer service tool is a virtual assistant, also known as a "Chatbot." Different applications are used in various sectors. These are listed below:

A. Systems for conversation

A conversational system enables us to hold a naturallanguage conversation with an automated system using a speech or text interface [2]. They help businesses automate challenging activities and offer round-the-clock service to their customers. The two most common varieties of conversational devices are chatbots and virtual assistants. Today, e-commerce, social media, banking, and other self-service point-of-sale systems use these two devices to provide a range of services to its customers.

B. Text Analytics

The goal of text analytics, sometimes referred to as text mining, is to extract useful information from text, whether it be in longer texts like emails and documents or in shorter ones like SMS texts and tweets [23]. Social media analysis is one of the most common use cases for text analytics.

C. Machine Translation

The objective of machine translation is to automatically translate material from one natural language to other also ensuring maintenance of the intended meaning.

Google Translate is the most widely used machine translation tool. In speech translation and education, other machine translation software is also employed [14]. NLP is also used in manufacturing, healthcare, customer service, automotive, retail, finance, and education. Virtual assistants that were developed by combining machine learning, computer vision, and natural language processing are being used by hospitals. These virtual assistants will automatically develop and obtain patient histories by interacting with patients [12][25]. Virtual assistants manage common tasks including patient registration and appointment scheduling. Self-driving cars are one of the most remarkable developments in the manufacturing sector. which are enabled by NLP and are becoming in popularity in the industry.

In banking sector NLP-based solutions are used to create applications such as sentiment analysis, document search, and credit scoring. Credit scoring programmes let banks and financial institutions determine a person's creditworthiness and provide a credit score by using NLP and machine learning. Applications for sentiment analysis automate the procedures of document categorization and named entity recognition to select the information that is most relevant to investors' demands [23]. Banks and other financial organisations utilise chatbot interfaces to let their consumers conduct information searches and get simple transactional answers in document search apps [24].

Robotics and process automation are two incredibly potential NLP application topics. In order to process instructions for assembling and moving products and machines, a robot on a manufacturing line can use natural language processing (NLP) to communicate with a human operator who is stationed remotely [4].

Using Natural Language, Computer Vision, and Machine Learning technologies, a retail virtual assistant that is placed in front of a retail business can detect and know what the customer requires and provides them with quick information and promotional offers [10]. Because computer vision and natural language processing are integrated, a platform in the education industry can provide students a virtual classroom. Digital assistants have already been used to help students solve problems using specialised information from online libraries [9].

D. Frameworks and Tools for NLP Development

Today's development tools are readily accessible due to the worldwide interest that open-source communities have shown in them [6]. These frameworks and tools contain built-in libraries and can be customised to fit specific industry standards. The natural language representation block uses structured, tree or graph models to express the knowledge of natural language [7]. A Natural Language database is a set of Natural Language data that machine learning algorithms use to do extra NLP tasks, similar to MNIST or other databases. This database is used by representation and transformation blocks to perform their tasks. Natural language transformation will employ a range of learning and extraction techniques to gain meaningful and pertinent activities from the NLP jobs [5]. Natural language communication is the presentation of the behaviours that are intended and desired to occur as a result of tasks aided by NLP [11]. The end result might either be computer activity, like a robot arm moving, or it could be Natural Language [27].

Natural language processing has developed as a result of human conversation. The procedure will undoubtedly involve the conversion of human natural language into a machine-understandable format. The following tasks could be included in NLP:

1) Word Sense Ambiguation- In this, a meaning of a word with multiple meanings is selected with the help of semantic analysis through which the word that is most suitable in a particular context is selected.

2) Speech Recognition- This is a process in which voice data is converted into text data.

3) Named Entity Recognition- It identifies words as relevant and useful entities.

4) Part of speech tagging- It determines the part of speech of a particular piece of text in a sentence or piece of information according to the most suitable context.

There are two components of NLP i.e. Natural Language Understanding (NLU) and Natural Language Generation (NLG)

Referential Ambiguity: It comes into picture English language and it has a particular spelling error [18]. Then, using the principle of Language Detection in the system, we can identify and correct the errors in the spelling of the word that is written incorrectly and also, the system can help us analyze the text and recognize the language in which the text is written as ‘English’. NLP has many libraries such as NLTK, spaCy, genism, etc [16]. These libraries help in accessing the features of NLP and in the creation of NLP models through their use. These help widely and vastly in Language Detection models and therefore, serve their purpose.

3. METHODOLOGY

For implementation, "Google Colab" Platforms are utilised. The data is loaded using a "Language Detection Using NLP" file that has been prepared. To train a model, a dataset from

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.02 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.42004

Kaggle and Github is used. Only a few of the many languages in the downloaded dataset were picked based on the requirements. We'll go over every implementation in depth, step by step.

- STEP - I

First step is importing all the libraries and packages which are needed to accomplish the task.

- STEP-II

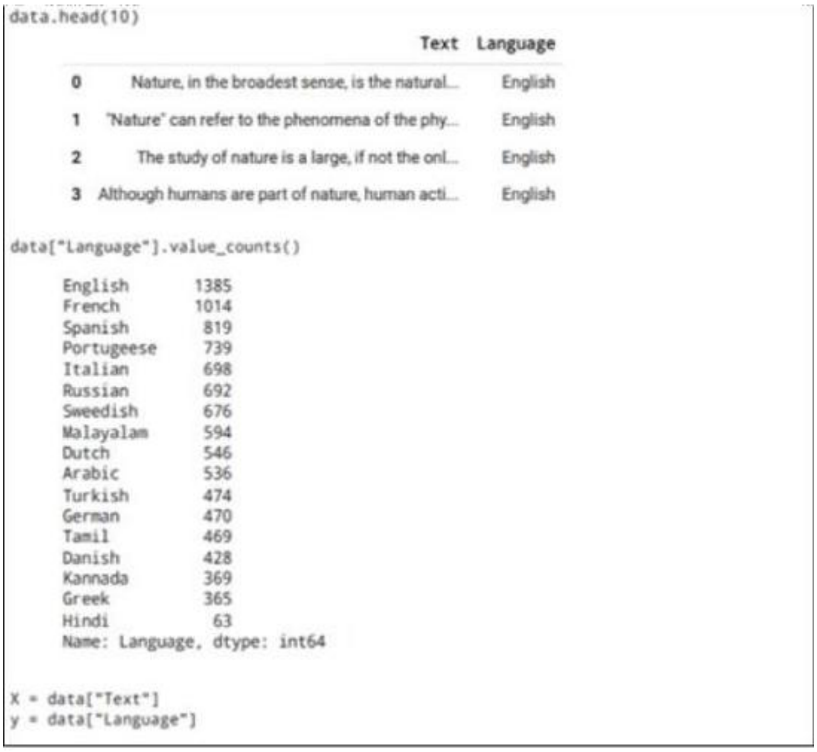
Mounting a dataset from the local computer to Google this problem look at this as a special case of text categorization that is solved with the help of various statistical methods [21]. LD is a great way to easily and efficiently sort as well as categorize information and apply additional layers of workflows that are language specific [22]. It can help us in identifying and detecting errors in a particular document, be it grammatically or with the spelling. For example, if we write a sentence in Drive platform as a zip file. The dataset is currently mounted to the "Google Colab" environment on Google Drive. We have access to about 80 GB of local storage on the distributed server of the Google Colab Environment[20].

- STEP- III

To obtain data from a csv file, use the function read_ csv (), which will extract data in the form of a data frame.

- STEP -IV

Now, we will define the necessary variable which plays a very important role to build our machine learning model. The picture tells us the variable names and their respective values.



```
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

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

    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


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

Fig. 1. Defining the Variable

Definition of the variables in the Fig.1 are as follows: head

The first five rows of the data frame are displayed in head function in python by default. It includes a single parameter: the number of rows. We can use this parameter to display the number of rows of our choice.

.head(n) is used to get the first n rows of the dataframe. It includes one optional argument n (number of rows you want to get from the start).

- value_count

The function, value_counts(), returns the object that comprises of counts of values that are unique. The object obtained as a result, will appear in an order that is descending so that the element occurring most often is the first element.

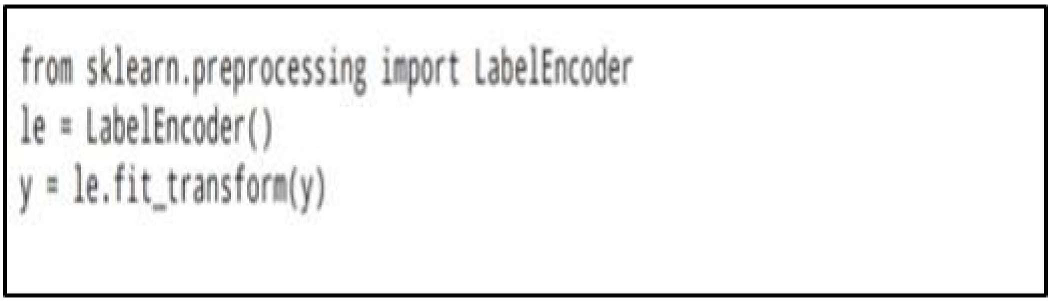
- STEP - V

Class Label Encoder from the sklearn module is used, and all the use of this module is given below:

Sklearn gives an immensely effective tool as it encodes into values that are numeric, the categorical features levels. Label Encoding means to convert the labels into a form that is numeric in order to get them converted into a form that can be read by the machine. Then, Machine Learning algorithms can decide better how the labels should be operated. It is an important step for pre-processing of the structured dataset in supervised learning.

LabelEncoder does encode labels with any value between 0 and n_classes-1, where n refers to the labels whose numbers are distinct. If there is a repeatition in a label, then the same value is assigned to as assigned earlier.

fit_transform () method is used which will fit label encoder and convert or transform multi-class labels into binary labels. The output for this conversion is sometimes called the 1-of-K coding scheme. Including LableEncoder and Fit_Trasform is shown in Fig. 2.



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

Fig. 2. Including LableEncoder and Fit_Trasform

STEP - VI

One array is created which named as data_list and re.sub() function is used that belongs to the Regular Expressions (re) module in Python.

It will return a string in which all the occurrences that are matching with the specified pattern will be replaced by the replace string. The re.sub() function holds for a substring and will return a string with values that are replaced. Using this function, we can replace multiple elements by making use of a list.

.lower function is used to convert all alphabet in lower case.

Then the text is added in data_list array using append function.

We will use the sklearn.feature_extraction module class from the sklearn module and all the use of this module is given below: -

The sklearn.feature_extraction module is used in order to extract features in such a format that is supported and guarded by algorithms of Machine Learning typically from the datasets that consist of formats like image and text.

CountVectorizer uses the method which acts as a good tool provided by the scikit-learn library present in Python. It can be used to convert a given text into a vector based on the frequency (count) or occurence of every word which is occuring throughout the text [8]. This can be extremely helpful when multiple texts are present, and we wish to convert all the words in every text as vectors in order to use it in upcoming analysis of the text.

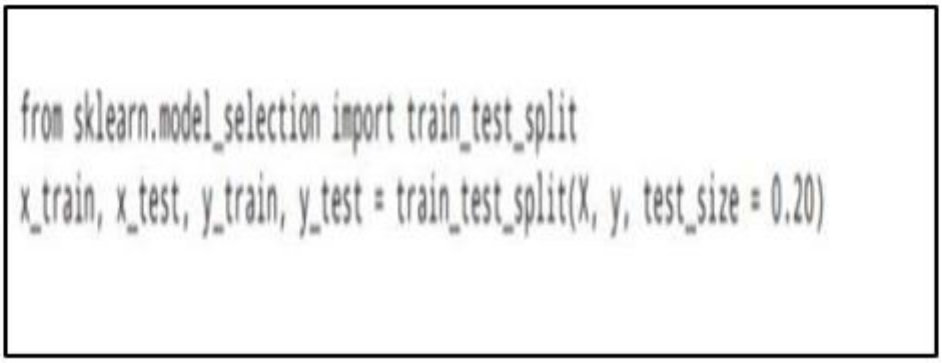
Through CountVectorizer a matrix is created in which a matrix column is a representation of every word that is unique and every row of the matrix represents every sample of text in thatdocument.

The value for every single cell is defined by the count of the words in the particlar given sample of text.

We will use the method fit_transform () which basically is a combination of transform method and fit method and is equivalent to transform(). fit(). In this method, transform and fit is performed on the data that is input at the same time and hence, data points are converted.

And then we set the dimension of the array X using

.shape method. Creating an array data_list is shown in Fig. 3.



```
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)
```

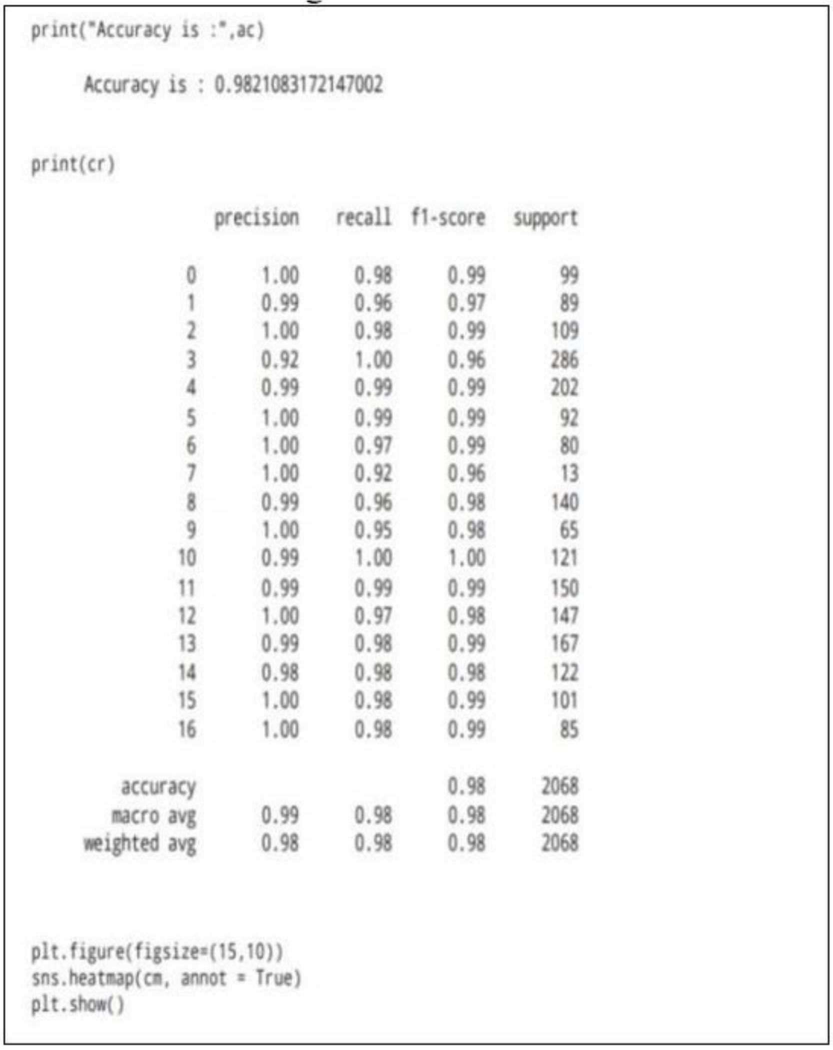
Fig. 3. Building Neural Network Model

- STEP-IX

This step is used to find the model accuracy. Find the Model Accuracy is shown in Fig. 6.

- STEP - VII

In this code snippet, the list is split into the training set and testing set. It is one of the major concepts in the machine learning model building [17].

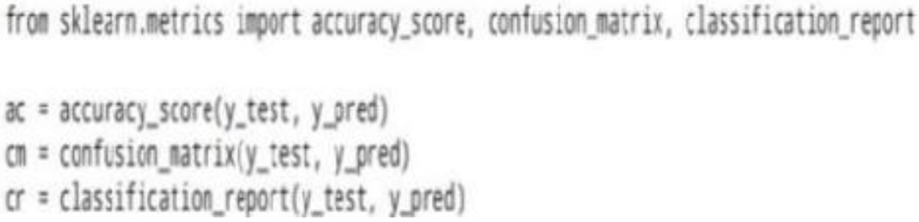


```
print("Accuracy is :",ac)

    Accuracy is : 0.9821083172147002


print(cr)

              precision    recall  f1-score   support

           0       1.00      0.98      0.99        99
           1       0.99      0.96      0.97        89
           2       1.00      0.98      0.99       109
           3       0.92      1.00      0.96       286
           4       0.99      0.99      0.99       202
           5       1.00      0.99      0.99        92
           6       1.00      0.97      0.99        80
           7       1.00      0.92      0.96        13
           8       0.99      0.96      0.98       140
           9       1.00      0.95      0.98        65
          10       0.99      1.00      1.00       121
          11       0.99      0.99      0.99       150
          12       1.00      0.97      0.98       147
          13       0.99      0.98      0.99       167
          14       0.98      0.98      0.98       122
          15       1.00      0.98      0.99       101
          16       1.00      0.98      0.99        85

    accuracy                           0.98      2068
   macro avg       0.99      0.98      0.98      2068
weighted avg       0.98      0.98      0.98      2068


plt.figure(figsize=(15,10))
sns.heatmap(cm, annot = True)
plt.show()
```

Fig. 4. Accuracy of the model

Thee X , Y and test_size is used as the parameter for the train_test_split method which is the part of the sklearn module. Dividing dataset into

STEP-X



```
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

ac = accuracy_score(y_test, y_pred)
cm = confusion_matrix(y_test, y_pred)
cr = classification_report(y_test, y_pred)
```

Fig. 5. Find the Model Accuracy

training and testing dataset is shown in Fig. 7.

training and testing dataset is divided into the four variable as given below:-

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

- STEP - VIII

In this step neural network model is built with the model.fit() from the MultinomialNb module.

It is another useful Naïve Bayes classifier. In this, it is assumed that from a simple Multinomial distribution, drawing of the features is done. To implement the Multinomial Naïve Bayes algorithm for classification, Scikit-learn provides

sklearn.naive_bayes.MultinomialNB.

Now we have use fit method of MultinomalNB, it expects as input the x and y. Now, x should be the training vectors (training data) and y should be the target values. Building Neural Network Model is shown in Fig. 5.

We will check the accuracy of our model. Accuracy of the model is shown in Fig. 4.

RESULTS

The overall precision obtained in this experiment is 0.98. Keeping in mind the accuracy of 98 % the model can be considered as a good model which fits for this type of analysis.

4. Conclusion

The rise of technology in the modern world has also given rise to increased requirements which justify the development taking place around us every day. Natural Language Processing and Language Detection, here, give rise to wider as well as broader scopes which can make tasks easier for human beings and can help them recognize texts in a much easier, better and systematic manner, hence, making technical work easier for them with the use of statistical methods. Therefore, an attempt has been put forward by us for creating such a Language Detection model with the help of Natural Language Processing that can solve Language Detection problems and can help us in identifying text easily and aptly with the help of appropriate and efficient methods as it is very important and useful in today's world and justifies fairly the usage of words and linguistics in the body of the content provided in various documents.

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",

[3]. Article in Natural Language Engineering, July 2017

[4]. ACL 2018: 56th Annual Meeting of Association for Computational Linguistics https://acl2018.org

[5]. Predictive Analytics Today: www.predictiveanalyticstoday.com[accessed in Dec 2018]

[6]. 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)

[7]. Intelligent automation: Making cognitive real Knowledge Series I Chapter 2. 2018, EY report.

[8]. Jacques Bughin, Eric Hazan, SreeRamaswamy, Michael Chui , TeraAllas, Peter Dahlström, Nicolaus Henke, Monica Trench, 2017. MGI ARTIFICIAL INTELLIGENCE THE NEXT DIGITAL

[9]. FRONTIER? McKinsey & Company McKinsey & Company report July 2017

[10]. Svetlana Sicular, Kenneth Brant 2018, Hype Cycle for Artificial Intelligence, 2018 Gartner report July 2018.

[11]. 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.

[12]. Quarteroni, Silvia. (2018). Natural Language Processing for Industry: ELCA's experience. Informatik-Spektrum. 41.10.1007/s00287-018-1094-1.

[13]. 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.

[14]. 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.

[15]. 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.

[16]. Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges.

[17]. 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.

[18]. Subiya Ali Kidwai, Impact of Stress and Anxiety on Cognitive Performance, TEJAS Journal of Technologies and Humanitarian Science, ISSN : 2583-5599, V.04, I.02, 2025, DOI: https://doi.org/10.5281/zenodo.15333987

[19]. Tushar Singh et. al. An AI-Driven System for Monitoring and Enhancing Remote Work Productivity, TEJAS Journal of Technologies and Humanitarian Science, ISSN : 2583-5599, V.04, I.02, 2025,Doi: https://doi.org/10.63920/tjths.42003

[20]. Verma, S.B., Yadav, A.K. (2021). Hard Exudates Detection: A Review., Emerging Technologies in Data Mining and Information Security. Advances in Intelligent Systems and Computing, vol 1286. Springer, Singapore. https://doi.org/10.1007/978-981-15-9927-9_12

[21]. SB Verma, Brijesh P., and BK Gupta, Containerization and its Architectures: A Study, ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, Vol. 11 N. 4 (2022), 395-409, eISSN: 2255-2863, DOI: https://doi.org/10.14201/adcaij.28351

[22]. 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.

[23]. Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong,and Quoc Le. 2020. Unsupervised data augmentation for consistency training. Advances in Neural Information Processing Systems, 33

[24]. 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.

[25]. 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.

[26]. Marouane Birjali , Mohammed Kasri , Abderrahim Beni-Hssane . A comprehensive survey on sentiment analysis: Approaches, challenges and trends . Received 1 July 2020, Revised 25 March 2021, Accepted 10 May 2021, Available online 14 May 2021,

[27]. Performance Evaluation and Comparison using Deep Learning Techniques in Sentiment Analysis A. Pasumpon Pandian, Professor, Dean (R&D), CARE College of Engineering, Trichy, India. ISSN: 2582-2640 (online) Submitted: 17.05.2021 Revised: 07.06.2021 Accepted: 26.06.2021 Published: 03.07.2021.