MACHINE LEARNING ROLE IN LANGUAGE PROCESSING: PRACTICAL PERSPECTIVE

M Mary Sujatha Assistant Professor, Dept of Computer Science, National Sanskrit University, Tirupati, India.
AC Priya Ranjini, Associate Professor, Dept of Computer Science and Applications, KLEF, India

ABSTRACT:
Machine Learning is a branch of artificial intelligence that gives systems the ability to learn automatically and improve themselves from the experience without being explicitly programmed or without the intervention of human. Language Processing is an application of machine learning in which computational techniques are used to understand and analyze human languages in smart and useful way. This article mainly focuses on machine learning techniques impact on language processing. Language processing flow graph explores step by step implementation details which include data acquisition, data extraction and cleaning, data preprocessing, feature engineering and model building. It also highlights open source libraries available in machine learning based algorithms to use in different stages of language processing.

Keywords: Machine Learning (ML) Artificial Intelligence (AI), Natural Language Processing (NLP).

I. INTRODUCTION

Machine learning and natural language processing are important subfields of artificial intelligence. In recent days their interrelation grown prominently. The combination of machine learning with natural language processing increased the importance of artificial intelligence. Artificial intelligent system turned to user friendly system by the advancement of language processing and artificial intelligent system can do better performance by the adoption of machine learning techniques. An ML technique gives accurate predictions to AI system by the ability to learn from past experience and examples. Traditional algorithms are less effective to handle unknown problems and they suit to fixed set of problems. The generalization feature of machine learning algorithms can solve unknown or new problems which is more suitable to real world problems having many unknown variables. Deep learning is a sub set of machine learning algorithms which is based on artificial neural network [1]. In recent days deep learning is widely adopting and producing accurate results due to its flexibility upon its architecture. Deep learning techniques made machine leaning techniques as most advanced and widely used in natural language processing research.

On the other hand Natural language processing gives ability to a system in understanding and processing human languages. Traditional computers can understand binary language in form of 1's and 0's and it does not understand human languages like English or Hindi. By embedding machine learning techniques in language processing system, they get the ability to understand human languages. It became much familiar in recent times due its user friendly nature. Machine learning played a vital role in processing of natural language to control electronic appliances with voice commands to other important applications of NLP such as chatbot application, Machine Translation system and email classification etc., Natural language processing involve different set of operations and machine learning played constructive role for better working of NLP based applications [2] .

II. NATURAL LANGUAGE PROCESSING BASED APPLICATIONS

Applications based on natural language processing include machine translation, human voice recognition, sentiment analysis, chatboat assistance, data summarization, data classification, and auto correction[3][4]. These applications are useful for both individual and organizations in analysis of data for decision making. It automates many operations and retrieves efficiency in lesser time.

2.1 Machine Translation:

It laid distinguished milestone in the era of natural language processing based applications. In earlier days dictionary based word to word translation techniques were used having language specific grammar rules and limitations. The underlying challenge with early translation is the input text need to get translated without changing meaning. As understanding grammar is difficult task for computer systems, new algorithms were practiced with the help of machine learning techniques to dismantle a sentence and reassemble it in target language without losing its meaning. The best machine translation technique include Google translate service. It uses machine learning algorithms to understand language patterns to perform natural language processing.

2.2 Human Voice Recognition:

Artificial Intelligence introduced smart machines to understand human voice by translating spoken language into machine understandable format. It uses machine learning algorithms to process natural languages through systems, which imitates human interactions and mimic human responses. Siri, and Alexa are popular applications using voice recognition technology. Smart phones are updated enough to understand voice instructions to make calls based on contact list.

2.3 Sentiment Analysis:

It is used to analyze user’s opinion and comments on a particular product. Sentiment analysis plays vital role in customer relationship management process. Every negative report of a product causes damage to the product reputation and can lead to stock price dropping. Machine learning and deep learning techniques made tremendous effect on decoding and analyzing human emotions towards a product or a service. Automatic sentiment analysis is used to measure public opinion and in turn effects product reputation. Financial domain is heavily influenced by human sentiment and emotion.

2.4 Chatbot Assistance:

Online chatbots are computer based applications used to generate machine based answers to common queries of consumers with the help of automated pattern recognition system with a rule of thumb response mechanism. Unlike help window option chatbots are user friendly and provide solutions based on customer request. Initially chatbots were designed to answer fundamental questions to minimize call center assistance to customer support service. By the existence of natural language processing and machine learning algorithms chatbot application is updated to manage complicated consumer requests in both voice and written format, called flow based chatbots whose response is based on previous dialogue. Open ended chatbots are another type of application used to make regular conversation.

2.5 Data Summarization:

It is a process of reducing a portion of text into a concise version without disturbing its meaning. It is possible by extracting main concept and preserving the meaning of the content. Natural language processing is used to understand key points of financial market reports, to extract sports results and used to create news headlines.

2.6 Data Classification:

Data can be in form of text or image. Text classification is a process of allocating digital tags to text based on content and semantics. It helps effortless and efficient retrieval of data in searching process. Text classification is made up of importance of input text such as high, medium and low. This machine learning application is helpful to classify spam and non spam mails based on subject over a period of time.

2.7 Auto Correction:

It is a trending feature in language modeling. Auto correction is possible by predicting possible word. It is also helpful in auto grammar checking and auto completion of a sentence. Auto grammar checking process will highlights error by underlying word in red.

III. MACHINE LEARNING ROLE IN PROCESSING NATURAL LANGUAGE BASED APPLICATIONS

Machine learning and deep learning algorithms made much influence on language processing applications. These algorithms attain more accuracy in natural language based applications by adopting flexibility feature by the availability of high performance computing devices. It is noted that millions and trillions of data been produced every day and machine learning and deep learning algorithms are capable of learning new things from available data. They train from the given examples of input and make its own conclusions without human assistance. The ability of machine learning and deep learning algorithms popular in different domains such as healthcare, agriculture, transportation, etc,. At this extent some of the techniques of deep learning extensively produced good result for these applications [5].

Natural language processing is a subfield of artificial intelligence which provides an ability to understand human spoken or written language to computer system. Ambiguity might cause in processing linguistic data as same words have different meaning at different contexts. There exist different stages of analysis [6] to make a better understanding of human language such as morphological analysis, syntactic analysis, semantic analysis, discourse analysis and pragmatic analysis. Some popular natural language processing applications include Apple's Siri, Amazon's Alexa and Google assistant. In the following grounds figure 1 shows flowchart for building a language processing based application using machine learning algorithms.

Step1: Prepare required data for modeling
Step2: Formatting data to make it available for built model
Step 3: Data aggregation and ungrouping of data based on requirement.
Step 4: working with data attributes to build better model using domain knowledge.
Step 5: Running the model with processed data
Step 6: Evaluating model performance using different metrics, based on results one can go back to step 3.
Step 7: Deployment and going back to step 1 based on requirement to work on new data items.

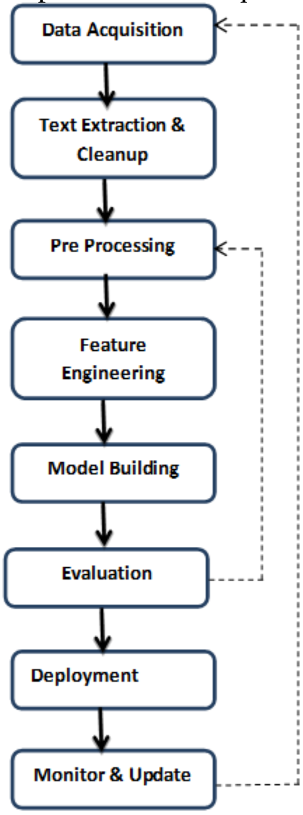

Figure 1: Natural Language Processing Architecture

1. Data Acquisition: It is a process of converting the data which is available in human understandable format to machine readable format to perform computational tasks. In other words it is a process of text extraction from physical format to digital format. The input can be written text or spoken words. There are different repositories for raw data accessibility from local disk files to public data sets, different approaches [7] will be followed based on its availability
2. Text extraction and cleanup: It is the process of fetching entities in the text, which highlights fundamental concepts and references of the text such as people, places, utilities etc,. Machine learning algorithms automatically extracts keywords and required phases from unstructured data [8]. Cleanup process makes the machine to produce error free analysis. Text cleaning is the process of eliminating stop words, normalizing Unicode characters such that machine can understand and converting complex words to their root level.
3. Preprocessing: Data preprocessing is one of the important step in building machine learning model for language processing. Data is aggregated from different sources using various data warehousing and data mining techniques consists of noise, incomplete, duplicate and missing values. Outliers and inconsistent data may cause false predictions and disturb data analysis. Data cleaning [9] involves dealing with missing values, removing noisy data by eliminating random error or variance in measured variable. Noisy data can also be handled by grouping similar data items together and treating an item as noise which does not fell into the group. These ungrouped items are also called as outliers or inconsistent data. Data integration is another part of pre processing where data present at multiple sources are merged together as single unit.
4. Feature Engineering: Feature selection and feature extraction are part of feature engineering. It is a process of retrieving essential data attributes from raw data and transforming into required attributes for improving model accuracy using domain knowledge. To build a better model it is necessary to train with better features. Feature engineering is a process of creating new features, removing unwanted variables, transformation of features from one representation to other, extraction of features from huge data to meaningful information and data analysis to find new patterns in the large quantity of data [10]. TF_IDF vectorizer and one hot encoding are the techniques used in feature engineering to convert raw data into system understandable format. These are applicable on both numerical and categorical data sets.
5. Model Building: A model would be built based on available data. There are different approaches in building a model such as heuristic approach, Machine learning approach and deep learning approach. A problem with fewer amounts of data can be solved with traditional heuristic approach. A problem with good amount of data will go with machine learning approaches and a problem with large amount of data will take the help of deep learning approaches. Type of problem also plays vital role in selection of model building approaches. GridSearchhCV [11] helps in identifying best model for representing data by finding optimal parameter values for a model from a given set of parameters grid.
6. Model Evaluation: It is a process of considering different set of metrics to check the performance of a model. A model can use different set of metrics such as accuracy, precision, recall, F1 score etc., to check out goodness of a model. Confusion metrics can be used to indicate truth and a prediction can be made to compare truth. Based on evaluation results the working process might go back to previous steps.
7. Deployment: In this step built model can deploy on to cloud for users for their usage. There are three stages of deployment such as deploying model into cloud for users, monitoring cloud continuously with the help of a dashboard to check evaluation metrics and finally retraining model with new data to redeploy.

IV. CONCLUSION:

Language is the most important communication tool to interact with each other. Because of its importance, computers are trained to understand human language. Traditionally a computer system understands binary language in form of 1's and 0's and it does not understand human languages like Hindi or English. Natural language processing helps the system to understand human languages efficiently. Its existence made electronic devices smart by making them to operate with the help of voice commands. On the other hand machine learning existence in NLP made an end to end automated implementation. It improves accuracy in processing by the ability of machine learning to learn from past experiences and examples. Given natural language processing framework elaborates an end to end implementation flow with supporting packages and built in methods. One can implement given work flow using any of open source based programming languages like python. Built-in packages and methods help to retrieve accurate and efficient results.

REFERENCES

[1] Dr.Tatwadarshi P. N. "Role of Machine Learning in Natural Language Processing" Natural Language Processing | Role of ML in Natural Language Processing (analyticsvidhya.com)

[2] Anuj Kumar Dwivedi, Mani Dwivedi "A Study on The Role of Machine Learning in Natural Language Processing", International Journal of Scientific Research in Computer Science, Engineering and Information Technology ISSN : 2456-3307 Volume 8, Issue 4, pp.192-198, July-August-2022. Available at doi : https://doi.org/10.32628/CSEIT228429

[3] "Natural Language Processing (NLP) For Machine Learning" https://kili-technology.com/data-labeling/nlp/natural-language-processing-machine-learning, KILI TECHNOLOGY © 2022

[4] XimenaBolaños " How Natural Language Processing with Machine Learning is driving innovation", https://www.encora.com/insights/natural-language-processing-and-machine-learningSeptember 29, 2021

[5] Tatwadarshi P. Nagarhalli , Dr. Vinod Vaze , Dr. N. K. Rana "Impact of Machine Learning in Natural Language Processing: A Review", IEEE Xplore Part Number: CFP21ONG-ART; 978-0-7381-1183-4

[6] Jurafsky, Daniel , Martin, James. (2008). "Speech and Language Processing: An Introduction to Natural Language Processing", Computational Linguistics, and Speech Recognition, : https://www.researchgate.net/publication/200111340.

[7] Shankar297 "An End to End Guide on NLP Pipeline", https://www.analyticsvidhya.com/blog/2022/06/an-end-to-end-guide-on-nlp-pipeline/

[8] Tiedemann, J. (2014). Improved Text Extraction from PDF Documents for Large-Scale Natural Language Processing. In: Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2014. Lecture Notes in Computer Science, vol 8403. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-54906-9_9

[9] Kadhim, Ammar. (2018). An Evaluation of Preprocessing Techniques for Text Classification. International Journal of Computer Science and Information Security,. 16. 22-32.

[10] Rawat, Tara & Khemchandani, Vineeta. (2019). Feature Engineering (FE) Tools and Techniques for Better Classification Performance. 10.21172/ijiet.82.024.

[11]Scott Okamura , "GridSearchCV for Beginners" https://towardsdatascience.com/gridsearchcv-for-beginners-db48a90114ee