JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Machine Learning in Natural Language Processing: A Review

[1] Ms. Urooj Sultana, [2] Ms. Rama Bhardwaj

[1,2] Assistant professor

[1,2] Vivekananda Institute of Technology, Jaipur

Abstract: Machine learning is a big part of artificial intelligence. Artificial intelligence systems are broad and complex, and they're programmed to solve complicated problems the same way humans would. Machine Learning is an emerging technology that is rapidly increasing in popularity. It is currently being used in a variety of different industries and is popular among many different types of businesses. It's a field of study where computer systems, software programs, virtual assistants, etc. become capable of automatically learning without explicitly programmed instructions. Natural Language Processing (NLP) is a powerful form of AI that gives machines the ability to read, understand, and interpret human language. With NLP, machines can complete tasks such as speech recognition, sentiment analysis, and automatic summarization. Machine learning for NLP and text analytics refers to a set of statistical techniques that identify parts of speech, entities, sentiment, and other aspects of text. The techniques are expressed as a model that is then applied to other pieces of text. This type of machine learning is called supervised machine learning. This paper reviews the concept of Machine Learning, its role in Natural Language Processing as well as also reviews the works in NLP.

Index Terms **– Natural Language Processing, Sentiment Analysis, Machine Learning.**

I. INTRODUCTION

Machine learning is a branch of artificial intelligence. It starts with facts and assumptions, and then improves depending on the accuracy of its predictions. Machine learning is an important part of data science. For example, through statistical methods and algorithms, machine learning enables computers to classify or predict, which in turn has a big impact on data mining projects. [1]

Insights from Big Data drives decision-making within applications and businesses, and ideally impacts key metrics of growth. As the demand for big data continues to grow, the market demand for data scientists will increase as well, requiring them to assist in identifying the right questions to be answered. [1] Artificial intelligence has grown in importance and prominence over the past few years. Which is why it's one of the key skills in data science. Machine Learning, a term coined by British computer scientist Arthur Samuel, remains an essential skill for any aspiring data analyst or data scientist. [1]

Machine Learning is defined as - the "Field of study that gives computers the capability to learn without being explicitly programmed." To put it in layman's terms, Machine Learning (ML) can be explained as automating and improving the computer's learning process based on their experiences without any human assistance. There are three main steps to our data-driven machine learning service: feeding it quality data, training the machine with different algorithms and picking the right algorithm. Naturally, this is a decision that needs to be made based on your particular needs and the type of data you're using. [2]

Example: Training of students during exams. When students prepare for an exam, they don't just cram the information. Instead, they try to understand the topic in-depth. This way, when it comes time for their exam, they have a good understanding of what will be on the test. Before the exam though, they will do some last minute studying using high quality content like questions and answers from different books or videos. [2]

They are training their brain by inputting as well as outputting. For example, what logic and approach do they adopt in solving a different type of problem? As they solve practice papers and see the performance (accuracy/score) they keep increasing their performance with the adopted approach. [2] To build a machine learning model, you first need to train it with both input data and outputs. Once the time comes, it can then be tested on data without an input and give you the model's predicted output by comparing it to what actually happened. Researchers are working diligently to improve algorithms and techniques so these models perform even better in the future. The basic difference between ML modelling and traditional programming : -

- Traditional Programming: We input DATA (Input) and use a PROGRAM (logic) to get the output on a machine.

- Machine learning is the process of teaching a computer to do something. Input + Output creates an algorithm, which can be evaluated while testing. [2]

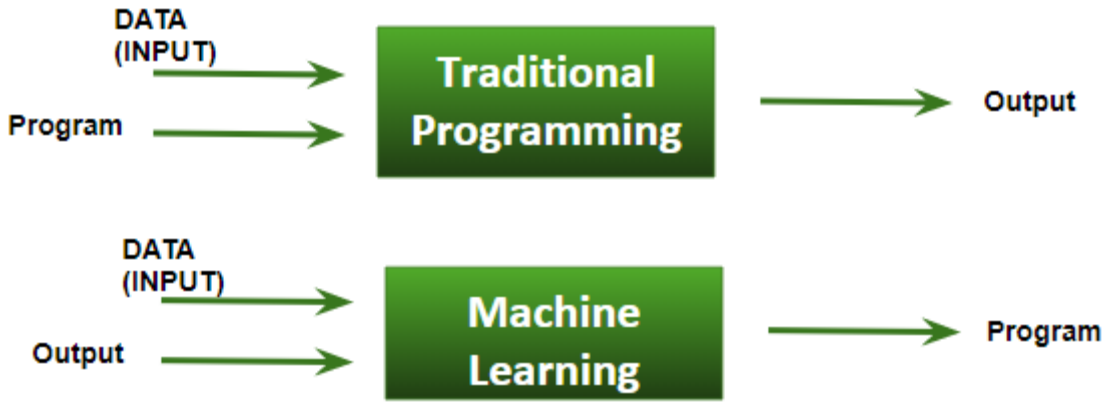

Fig 1. Traditional Programming V/S Machine Learning

A computer is said to be learning from experience if it performs better in a given task because of that experience. If a computer program learns to complete tasks as measured by some performance measure and task class, then it is said to be learning with respect to the tasks in that task class. There are two broad categories of machine learning: Types of Machine Learning. Supervised Learning involves showing a program many examples so that it can learn. Unsupervised Learning, on the other hand, involve using a program to find patterns in data without having any previous training. [3]

In the last few years, there has been a significant increase in online shopping. With an unlimited range of interests and the rapidity of innovation, it can be difficult to know exactly what is trending. But while online shopping, buyers often find themselves researching a number of products at once. Now, when a buyer searches for a product, social media sites will start recommending or showing offers on that particular product. Whether they use Facebook, check out the web pages in the Google search engine, or shop online using an e-commerce website, the advertisement will be directed at that type of marketplace. There's no need to code such a task for every user; it all happens automatically. [3]

Machine Learning is needed for researchers and data scientists to build the models on the machine. They use masses of quality data, but now their machines carry out tasks without human involvement and are even becoming more efficient with experience. Previously, an advertisement could only be done through newspapers, magazines, or radio. However, these outlets were limited in their ability to reach the most receptive audience. With the introduction of new technologies, a more efficient method of advertising (targeted online ads) has been made available. [3]

- Machine learning continues to help the health care industry. A current example of how ML is making a difference is in detecting cancer based on slide-cell images. It would've taken a lot of time for humans to perform this task - but not machine learning. But now, it's even easier. Doctors just need to give a call and the machine predicts If your patient has or will have cancer with some accuracy. Machines require high computation and good quality image data to get the best results. And doctors are also using ML for patients with different parameters in consideration. [4]

- You might have to use tools like IMDB ratings, Google Photos' face detection feature and other ML-powered ones like Lens. These can help you determine whether an E-mail is classified as social, promotion, updates or forums by using Text classifications. [4]

- Gathering past data in any form that's suited to processing. The better the quality of data, the more suitable it'll be for modeling. Sometimes, the data collected by the system is in raw form and needs to be processed before it can be used. For instance, there may be cases where certain fields are left blank in a particular tuple. In order for the machine learning technique to work properly, missing values must be filled with suitable values. [4]

- If a price of a house is missing, we will replace it with the mean price of houses in that area. If a house color is missing, we will use the most popular house color or assign it to all houses. Depending on the type of filters we want to use, data will need to first be in a form that the machine can understand. This may involve numerical conversion, text and image conversion, or any other type of conversion required by the task at hand. Simply put: Data needs to be made relevant and consistent before it is admitted into a form understandable by machines [4]

Divide the input data into training, cross-validation, and test sets. The ratio between each set should be 6:2:2. An algorithm is the set of rules a computer uses to perform a process. This includes identifying patterns and optimizing the efficiency of these rules when based on a training set. Testing our model with data that isn't in the range of the original training data and using metrics such as F1 score, precision, or recall. [4]

II. NATURAL LANGUAGE PROCESSING

The best way to explain Natural Language Processing is that it literally makes human language intelligible to machines and allows them to comprehend what we are saying. An example of NLP is Microsoft's Chinese-English language translation software, which uses computer programs based on Artificial Intelligence (AI) and machine learning. NLP relies heavily on the power of

linguistics and computer science in order to process human language and create intelligent systems capable of understanding, analyzing, and extracting meaning from text or speech. NLP is a subset of artificial intelligence that focuses on one thing in particular—understanding the structure and meaning of language. NLP analyzes syntax, semantics, pragmatics, and morphology to generate rules for machines to act upon. [5]

For example, Gmail is automatically sorted into categorizes like Promotions, Social, Primary and Spam. That's thanks to an NLP task known as keyword extraction. You just have to teach machines which words go with which categories by having them sort emails that contain those words. [5]

There are many benefits to NLP, but here are just a few top-level benefits that will help you achieve your goals:

- Natural Language Processing technology allows machines to automatically analyze huge amounts of unstructured text data and understand what it means. For example, you can use NLP when inputting text data such as social media comments, customer support tickets, online reviews, or news reports.

- Processes can be automated in real-time with the help of natural language processing tools. Machines, with little to no human interaction, are able to quickly and accurately sort information on their own, 24 hours a day.

- Tailor your NLP tool to your industry. Our natural language processing algorithms can be tailored to your needs, like complex, industry-specific language - even sarcasm and misused words. [5]

Text vectorization is a natural language processing tool that transforms written text into mathematical vectors representing the words and phrases present in the text. Machine learning algorithms are fed training data and expected outputs to train machines to make associations between a particular input and its corresponding output. Statistical analysis tools are employed to help machines build their own knowledge bank and learn which features represent the text best, before making predictions for new texts. If given the chance to train on more data, then these NLP algorithms will be able to produce more accurate text analysis models. [6]

Machine learning models can be used to analyze sentiment by assigning a polarity rank to text. You see this in the above chart and map. Machine learning models are a big help when it comes to working with increasing amounts of data. With them, you don't have to include strict rules about what's relevant or not. All you need is a set of training data that has several examples of the tags you want to analyze. For example, if you're trying to find information on sentiment analysis, keyword extraction, topic classification and intent detection in your sentence, they would work simultaneously and give you an even more detailed result! Syntax and semantics are two tasks that involve natural language processing. These help break human language into machine-readable pieces. [6]

Syntactic analysis is the process of analyzing and organizing text into a graded dependency structure diagram. The process identifies the syntactic structure of sentences and their dependencies on one another. Semantic analysis focusses on the meaning of language. But since language is polysemic and ambiguous, semantic analysis is one of the most difficult areas of NLP. [6]

III. ML AND NLP

Natural Language Processing is a part of daily life, whether we notice it or not. Even just using voice commands for things like our virtual assistants, smartphones, and vehicles requires processing language. Apps with voice-triggered commands such as Alexa, Siri, and Google Assistant use NLP and ML to answer our questions and do the tasks we ask them to. Not only does NLP make life easier, but it also has revolutionary implications on how we work, live, and play. There are many definitions of AI. Natural Language Processing, Machine Learning, and Artificial Intelligence are three subsets that fall under the umbrella. AI is an acronym for machines that can emulate human intelligence. ML and NLP are subsets of AI. [7]

Artificial Intelligence (AI) is the process of solving problems with computer systems. It has many applications in today's society, such as NLP and ML. Natural Language Processing is the newest form of artificial intelligence. With NLP machines have the ability to not only read, but also understand and interpret human language. This allows them to perform tasks including speech recognition, sentiment analysis, and automatic text summarization. [7]

AI is different from what you're used to. It's not the same as a human would do. To maintain a competitive edge, your AI needs to be capable of learning and improving through experience, without any explicit programming. Machine Learning can be used in many different applications of AI, like problem-solving and NLP tasks. [7]

3.1 ML in NLP

- **Translation:** NLP-based solutions have a wide range of uses, including translation, speech recognition and processing sentiment or emotion. These technologies are used by organizations to automate processes, gain competitive advantages and/or extract vital insights from data. Translating languages is a more complicated task than simply swapping words. Different languages have different grammar rules, so the challenge of translation is to preserve the meaning and style of the original text. Since computers don't understand grammar, they need a process which can deconstruct a sentence before reconstructing it in another language - keeping its sense and meaning intact. Google Translate is an extremely popular tool for translating texts from one language to another. Google Translate once used Phrase-Based Machine Translation (PBMT) but now uses Google Neural Machine Translation (GNMT). The difference is that GNMT relies on ML with NLP in order to identify patterns in languages. [8]

- **Speech recognition:** The speech recognition technology can detect word patterns and decipher even the most unclear voice with up to 97% accuracy. Speech recognition is a machine's ability to identify and interpret phrases and words from spoken language and convert them into a machine-readable format. It uses NLP to allow computers to situate computers as though they are interacting with humans, and ML to simulate human responses. Some of the most popular examples of speech recognition are Google Now, Alexa, and Siri. All you have to do is say "call Jane", and the given device will know what that means and will now make a call to the contact saved as Jane. [8]

- **Sentiment Analysis:** Sentiment analysis is a form of natural language processing and machine learning. It's used to identify the emotions in subjective data like news articles and tweets. Later, it can be used by companies to improve customer satisfaction, monitor their brand reputation, and gain deeper insight into their customers' interests. The stock market is a sensitive field that can be heavily influenced by human emotion. Negative sentiment may cause stock prices to drop, while positive sentiment may increase the price of the company's shares - and vice versa. [9]

- **Chats bots:** Chatbots are a type of program designed to interact and provide automated answers to common customer queries. They contain pattern recognition systems with heuristic responses, which they use to have conversations with humans. Initially, chatbots were used in call centers to answer simple questions and alleviate heavy workloads. But with AI-powered chatbots, things are getting more and more intuitive. Healthcare chatbots, for example, can collect intake data and help patients assess their symptoms. They can even recommend a treatment plan. [9]

- **Q&A Systems:** Question-answer systems are intelligent systems used to respond to customer queries. With their versatility and ability to answer complex questions with ease, these robust AI programs are a good alternative if you can't find an available chatbot that meets your needs. A question-answer system could hypothetically respond to a customer's query with "How do I get to the airport?" or "What is the capital of Denmark?" In 2011, IBM's Watson computer competed on Jeopardy against some of the show's biggest all-time champions and stunned everyone in the tech industry when it won first place. What was shocking about this stunt is that during the game, questions are given first, and the contestants have to supply their own responses. [10]

- **Automatic text summarization:** Automatic text summarization is a technique applied in many ways. This method of natural language processing is used in many different circumstances, including headlines, snippets in search engine results, and bulletins of market reports. [10]

- **Market Intelligence:** Gathering information on trends, consumers, products, and competitors through market intelligence helps give you a powerful competitive edge. Market Intelligence can analyze topics, sentiment, keywords, and intent in unstructured data and is less time-consuming than traditional desk research. Market Intelligence gathers all the latest search queries and knows what synonyms to select in order to make them diverse and to be sure organizations are targeting customers. It can also help organizations decide which products or services to discontinue and whether they should target new customers. [11]

- **Automatic text classification:** Automatic classification is also another core function of NLP. This process assigns tags to text, based on its contents and semantics, that help with rapid and easy retrieval of information during the search phase. This NLP application can differentiate spam from non-spam based on their content. [12]

Related Study in the NLP

J. Liu, et al. 2022 [13] This paper presents a system to identify social engineering attacks by analyzing textual input. Researchers can use this system in different environments with text input, such as SMS, chats, emails, etc. The system uses Natural Language Processing to extract features from a dialog that it can analyze. This includes URL extraction and counting, checking for spelling mistakes, counting blacklisted words, or other features. Researchers can then train Machine Learning algorithms (such as Neural Networks, Random Forests, and Support Vector Machines) on the data gathered to identify social engineering attacks. They found that their three classification algorithms could accurately detect at least 80% of these cyberattacks.

J. C. Lopez and J. E. Camargo, 2022 [14] Computer-human language interactions are a computer science and artificial intelligence topic that involves designing computers to be able to process, explore, and understand a wide variety of natural language data. A non-expert users' use of the Structured Query Language to store their data in a database can be challenging. To improve this interaction, an intelligent interface is necessary. Utilizing a natural language instead of a structured query language has led to the creation of natural language interfaces in databases. This research aims to build an algorithm for machine learning to represent information with the user's demands for answesring the query and obtaining information. For conversion of Natural Language Query into Structured Query, lowercase conversion was utilized, words with special symbols were removed, words were tokenized; POS tags were assigned; word similarity was analyzed using the Jaro-Winkler Indexing Algorithm; and Naive Bayes method was used.

M. Arefin, et al. 2021 [15] social media seems to be getting a lot of attention in the last decade - both good and bad. To begin with, people are able to communicate directly without any cultural or economic barriers being in the way. However, many negatives have arise as well. One such problem has been hate speech - an offensive, hostile language taking place on social media. For instance, it might affect the person's caste and creed, causing them to not feel their race among others. Don't worry - authors believe that there's much that can be done to minimize hate speech all together! By diving deep into natural language processing and using methods like machine-learning models for the best results possible - Authors come up with no shortage of solutions for this problem!

D. Bhimani, et al. 2021 [16] From the past decade, social media has gained a lot of momentum, both in a positive and negative way. With this rapid increase in networking regions with no cultural or economic gap, people are able to communicate directly with one another without issue. While there are many benefits to social media, there are also just as many negatives. One such problem that has arisen over the past few years is hate speech. Hate speech happens when people use offensive language while communicating on social media. It can refer to any person or group of people with shared interests. In this paper, authors've introduced their way of dealing with this hateful communication and lessening it to the largest degree. People will post their hatred or anger about certain topics straight on social media which would hurt the feelings of other people and would negatively affect them based on their caste, creed, religion, race, and so forth. Some comments might not be intentional toward anyone but those comments will be classified as hate speech due to the foul language that was used. Authors have gone deep into natural language processing for the elimination of hate speech and authors've done various machine learning models for determining which one should be utilized as per its accuracy.

B. Kynabay,et al. 2021 [17] Automatic text summarization software has become a top priority as the internet generates more and more information on a daily basis. The primary goal of this work is to propose an efficient method for automatic text summarization by using natural language processing and machine learning techniques. This research introduces a simple, easily understandable, and uncomplicated method of implementing this technique through overuse of python programming language. Their efficiency is necessary in web search tasks where websites frequently provide barely readable data that needs to be summarized in seconds. Their novelty is that their work only focuses on extracting keywords from Kazakh texts. Their contribution is creating manually created stop words specific only to Kazakh language and scraping articles from Kazakhstan's largest international news portal, www.inform.kz

IV. CONCLUSION

There are a lot of existing social good applications for NLP, such as identifying hate speech or signs of depression. It can also be used for more proactive applications like increasing well-being or fostering constructive conversations. Natural language processing is important because it helps resolve ambiguity and adds numeric structure to text data. Using NLP, researchers can sort through unorganized information to improve patient care and research. This branch of artificial intelligence within computer science focuses on helping computers understand how humans write and speak.

REFERENCES

1. Dongbo Zhang "Vocabulary and Grammar Knowledge in Second Language Reading Comprehension: A Structural Equation Modeling Study" The Modern Language Journal vol. 96 no. 4 pp. 558-575 2012.
2. Jun Liu and Yuji Matsumoto "Sentence Complexity Estimation for Chinese-speaking Learners of Japanese" In Proceedings of the 31st Pacific Asia Conference on Language Information and Computation (PACLIC 31) pp. 296-302 November 2017.
3. Dongli Han and Xin Song "Japanese Sentence Pattern Learning with the Use of Illustrative Examples Extracted from the Web" IEEJ Transactions on Electrical and Electronic Engineering vol. 6 no. 5 pp. 490-496 2011.
4. Takahiro Ohno Zyunitiro Edani Ayato Inoue and Dongli Han "A Japanese Learning Support System Matching Individual Abilities" In Proceedings of the PACLIC 27 Workshop on Computer-Assisted Language Learning pp. 556-562 2013.
5. Keiko Hori Jae-Ho Lee and Yoichiro Hasebe "HAGOROMO: A Usage Database of Function Words in Japanese" Mathematical linguistics (in Japanese) vol. 30 no. 5 pp. 275-285 2016.
6. Kosho Shudo Toshifumi Tanabe Masahito Takahashi and Kenji Yoshimura "MWEs as non-propositional content indicators" In proceedings of the 2nd ACL workshop on Multiword Expressions: Integrating Processing (MWE-2004) pp. 32-39 2004.
7. B. Bogin M. Gardner and J. Berant "Representing schema structure with graph neural networks for text-to-sql parsing" arXiv preprint 2019.
8. M. Mony J. M. Rao and M. M. Potey "An overview of nlidb approaches and implementation for airline reservation system" International Jour nal of Computer Applications vol. 107 no. 5 2014.
9. H. Kim B.-H. So W.-S. Han and H. Lee "Natural language to sql: Where are we today?" Proceedings of the VLDB Endowment vol. 13 no. 10 pp. 1737-1750 2020.
10. M. Uma V. Sneha G. Sneha J. Bhuvana and B. Bharathi "Formation of sql from natural language query using nlp" 2019 International Conference on Computational Intelligence in Data Science (ICCIDS). IEEE pp. 1-5 2019.
11. A.Giordani and A. Moschitti "Generating sql queries using natural language syntactic dependencies and metadata" in International Con ference on Application of Natural Language to Information Systems Springer pp. 164-170 2012.
12. M. Norouzifard S. Davarpanah M. Shenassa et al. "Using natural language processing in order to create sql queries" Computer and Communication Engineering 2008. ICCCE 2008. International Confer ence on. IEEE pp. 600-604 2008.
13. J. Liu, Y. Fang, Z. Yu and T. Wu, "Design and Construction of a Knowledge Database for Learning Japanese Grammar Using Natural Language Processing and Machine Learning Techniques," 2022 4th International Conference on Natural Language Processing (ICNLP), Xi'an, China, 2022, pp. 371-375.
14. J. C. Lopez and J. E. Camargo, "Social Engineering Detection Using Natural Language Processing and Machine Learning," 2022 5th International Conference on Information and Computer Technologies (ICICT), New York, NY, USA, 2022, pp. 177-181.
15. M. Arefin, K. M. Hossen and M. N. Uddin, "Natural Language Query to SQL Conversion Using Machine Learning Approach," 2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 2021, pp. 1-6.
16. D. Bhimani, R. Bheda, F. Dharamshi, D. Nikumbh and P. Abhyankar, "Identification of Hate Speech using Natural Language Processing and Machine Learning," 2021 2nd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2021, pp. 1-4.
17. A. Kynabay, A. Aldabergen and A. Zhamanov, "Automatic Summarizing the News from Inform.kz by Using Natural Language Processing Tools," 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST), Nur-Sultan, Kazakhstan, 2021, pp. 1-4.