Bridging NLP and Machine Learning: A Study of Supervised and Unsupervised Methods

1st Kiran B V
Computer Science and Engineering
Alva's Institute of Engineering and Technology
Moodbidri, Mangalore, India
Email: kiranbv@aiet.org.in

2nd K Jeevan Kumar
Computer Science and Engineering
Alva's Institute of Engineering and Technology
Moodbidri, Mangalore, India
Email: jeevanjeevan63643@gmail.com

3rd Kavyashree
Computer Science and Engineering
Alva's Institute of Engineering and Technology
Moodbidri, Mangalore, India
Email: kavyashreenayak8861@gmail.com

4th Kiran Kumar
Computer Science and Engineering
Alva's Institute of Engineering and Technology
Moodbidri, Mangalore, India
Email: kirankumark1707@gmail.com

5th Mahima Nayak
Computer Science and Engineering
Alva's Institute of Engineering and Technology
Moodbidri, Mangalore, India
Email: nayakmahima632@gmail.com

***Abstract*—Machine Learning (ML) and Natural Language Processing (NLP) have experienced significant advancements in recent years, driven by increased data availability, computational power, and enhanced algorithms. While ML concepts have been studied since the 1950s, their modern applications have transformed fields such as pharmacometrics, clinical pharmacology, and molecular biology. This paper introduces the foundational principles of ML and explores its relevance in analyzing scientific publications within these domains.**

NLP, a subfield of Artificial Intelligence, enables computers to understand, interpret, and generate human language. It facilitates various language-related tasks, including speech recognition, sentiment analysis, and text summarization. By leveraging ML, NLP models can process vast amounts of unstructured data, extract meaningful insights, and automate decision-making. NLP-powered tools have been widely adopted in applications such as social media sentiment classification, automated document processing, and entity recognition in business communications.

This paper highlights the synergy between ML and NLP, emphasizing their role in transforming data into actionable knowledge. Additionally, it examines key ML techniques used in NLP, discussing their impact on real-world applications.Through this conversation, we hope to offer a thorough grasp of the ways in which these technologies support developments across a range of fields.

***Index Terms*—Machine Learning, Natural Language Processing, Artificial Intelligence,Supervised and Unsupervised learning, Applications.**

I. Introduction

A subset of artificial intelligence (AI), machine learning (ML) allows systems to learn from data, spot trends, and make judgements with little assistance from humans. [6]. It involves algorithms and statistical models that allow computers to perform specific tasks without explicit programming [5].

The field has seen rapid advancements and widespread applications, particularly in areas like healthcare, finance, and autonomous systems [10]. Though ML has gained significant recognition in recent years, its foundations date back to Alan Turing's work in the 1950s and John McCarthy's introduction of AI in 1955 [1].

ML models are typically developed through the following steps:

1) Data collection and preprocessing.
2) Training a model using algorithms and hyperparameter tuning.
3) Using the trained model for predictions [1].

Alongside machine learning (ML), natural language processing (NLP) is a crucial area of artificial intelligence that enables machines to comprehend and react to human language. [3]. NLP applications such as translation, sentiment analysis, and text classification enhance human-computer interactions. [23] ML-driven NLP systems improve over time by learning from data, making them crucial for intelligent automation in various industries [3].

By leveraging ML and NLP, modern AI systems can efficiently process vast amounts of data, making them indispensable in today's digital world.

II. Natural Language Processing

A subfield of artificial intelligence called natural language processing (NLP) gives computers the ability to comprehend,

interpret, and process human language [1]. NLP makes jobs like speech recognition easier. Sentiment analysis, and text summarization. However, human language, often ambiguous and complex, poses challenges for machines, necessitating structured representations like the Object Constraint Language for better comprehension [7].

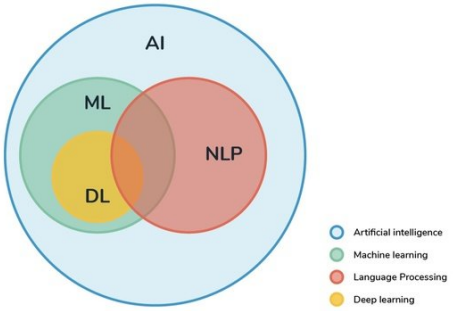

Fig. 1. Key components of the machine learning model architecture.

NLP relies on techniques like sentence segmentation, tokenization, and semantic analysis to enable effective human-computer interaction [6]. The Semantic Web enhances NLP by linking data for improved contextual understanding, benefiting search engines and information retrieval [13].

Deep learning has significantly advanced NLP, allowing models to perform complex tasks like language modeling and sentiment analysis with high accuracy. However, challenges such as processing vast datasets and interpreting complex language structures persist [13].

NLP applications span multiple industries, including healthcare for medical record analysis, social media for trend detection, and customer service for chatbots [6]. Recent advancements focus on optimizing query processing and integrating NLP with emerging technologies like machine learning and big data for enhanced efficiency [6].

III. The History of NLP

The emergence of early AI research in the 1950s marked the beginning of NLP's history. It has gone through several stages of evolution over the years: Rule-Based Systems in the 1950s: Computers were able to read phrases, translate text, and respond to queries using the first NLP systems, which were built on sets of manually created rules. The Turing Test, which Alan Turing developed to assess machine intelligence, is the most well-known example from this era.

1960s– 1970s:Symbolic Approaches and ELIZA During this period, symbolic AI dominated NLP research. The famous chatbot ELIZA (created in 66 by Joseph Weizenbaum) simulated conversation using basic pattern matching but was unable to understand the meaning of the input.

1980s–Probabilistic Models: The 1980s saw the introduction of probabilistic models in NLP, moving from rule-based approaches to statistical methods. This shift occurred because of the limita tions of hand-coded rules when dealing with real-world language complexity. [26] Hidden Markov Models (HMMs)and early machine learning algorithms were applied to tasks like speech recognition and part of-speech tagging.

1990s–Riseof Machine Learning: Machine learning techniques, particularly supervised learn ing algorithms, becamewidely adoptedin the 1990s. [22] This allowed for more flexible models and better handling of language variations.Machine learning models were trained using corpora, which are big text collections.

2010s– Deep Learning and Neural Networks: The introduction of deep learning techniques, specifically recurrent neural networks (RNNs) and transformers, revolutionized NLP. NLP's skills have been greatly enhanced by models like word vec, BERT, and GPT, which let machines to produce and comprehend text with fluency comparable to that of humans. [22]

2020s–LargeLanguage Models(LLMs): LargeLanguage-Models (LLMs) have significantly advanced natural language processing (NLP) in recent years. Models like GPT-, GPT-, and OpenAI's ChatGPT are trained on large data sets. LLMs leverage transformer architectures, which allow them to process and generate human-like text at an unprecedented scale.
They can do many different things, such as answering questions, summarising texts, translating, and writing creatively. Their ability to generate contextually relevant and coherent text with out explicit rule-based systems marks a significant shift in NLP. With techniques like transfer learning andself-supervised learning, LLMs require less task-specific training data, making them highly adapt able across domains. [4]

IV. Equations

A. Bayes' Theorem in Text Classification

Bayes' Theorem is widely applied in text classification problems such as spam filtering and sentiment analysis. It allows us to compute the probability of a document belonging to a specific category based on prior knowledge of word distributions. The theorem is given by:

$$P(C|W) = \frac{P(W|C)P(C)}{P(W)} \tag{1}$$

where:

- $P(C|W)$ represents the probability of a document belonging to class C given a set of words W.
- $P(W|C)$ denotes the likelihood of observing the words W given class C.
- $P(C)$ is the prior probability of class C in the dataset.
- $P(W)$ is the overall probability of encountering words W in any document.

Using Bayes' Theorem, machine learning models like the **Naïve Bayes classifier** can efficiently categorize text by assuming that individual words contribute independently to the overall probability, simplifying computation for large datasets.

B. n-gram Language Model

In Natural Language Processing (NLP), an **n-gram model** is used to predict the probability of a sequence of words occurring in a sentence. According to this model, a word's likelihood is solely determined by its preceding $n-1$ words. The general form of the n-gram model is:

$$P(w_1, w_2, ..., w_n) \approx \prod_{i=1}^{n} P(w_i | w_{i-(n-1)}, ..., w_{i-1}) \quad (2)$$

where:

- $P(w_1, w_2, ..., w_n)$ is the probability of a sequence of words appearing in a sentence.
- $P(w_i | w_{i-(n-1)}, ..., w_{i-1})$ represents the conditional probability of a word w_i based on the previous $n-1$ words.

For example, in a **bigram model** ($n = 2$), a word depends only on the previous word:

$$P(w_n | w_{n-1}) = \frac{P(w_{n-1}, w_n)}{P(w_{n-1})} \quad (3)$$

Similarly, in a **trigram model** ($n = 3$), a word depends on the two preceding words:

$$P(w_n | w_{n-2}, w_{n-1}) = \frac{P(w_{n-2}, w_{n-1}, w_n)}{P(w_{n-2}, w_{n-1})} \quad (4)$$

These models are widely used in **speech recognition, text prediction, and machine translation**. By analyzing patterns in word sequences

V. Supervised Learning

Labelled datasets are used to train algorithms in supervised learning, a subset of machine learning. To help the model understand the connection between inputs and outputs, each training sample includes a set of input features and a matching output label. [24] The main goal is to predict outcomes for new, unseen data based on this learned relationship. [5]

Supervised machine learning algorithms require external assistance. [21] The input dataset is divided into training and testing datasets. [14] The output variable in the training dataset requires classification or prediction. Each algorithm identifies patterns from the training dataset and uses them to classify or predict the test dataset. [10]

As the name suggests, this type of algorithm requires supervision for prediction or decision-making. The input dataset is partitioned into training and test data. In supervised learning, target or output values are already assigned to training data for training the model. Supervised learning is most popularly used to find solutions for classification and regression problems. [11]

Training and testing are the two stages of the learning process in a basic machine learning model. Samples of training data are used as input throughout the training process, and the learner or learning algorithm uses these features to construct the learning model. The learning model makes predictions for the test or production data during the testing phase by using the execution engine. The learning model's output, known as "tagged data," provides the final forecast or classified data. [12] The probability for input is typically left unspecified in supervised learning, such as when the predicted result is known. This procedure yields a dataset with labels and features. [14] The primary objective is to build an estimator that can forecast an object's label based on the feature set. The learning algorithm then learns by comparing its actual output with corrected outputs to identify errors after receiving a set of features as inputs together with the correct outputs. [12] The most popular method for neural network and decision tree training is supervised learning. The information provided by the predetermined classification is necessary for both of these. [12] Applications where past data forecasts probable future events also make use of this learning.

There are numerous real-world applications of supervised learning, such as an application that can identify the species of iris based on a collection of flower measurements. [12]

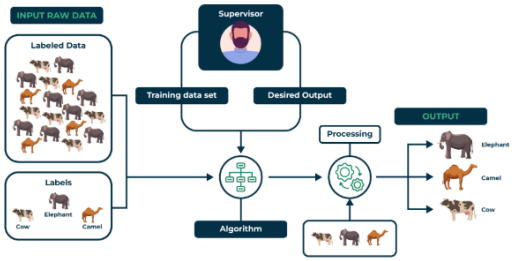

Fig. 2. Supervised Learning.

A. Key Characteristics

Labeled Data: Requires a dataset in which an output label is assigned to each occurrence.
Predictive Modeling: Focuses on predicting outcomes, which can involve classification (categorizing data into classes) or regression (predicting continuous values).
Training and Testing: Involves splitting the data into training and testing sets to evaluate the model's performance. [5]

B. Applications

Spam Detection: Using characteristics taken from the email content to determine whether or not an email is spam. [19]
Disease Diagnosis: Predicting the presence of diseases based on patient data and medical history.
Image Recognition: Identifying objects or features in images, commonly used in facial recognition and autonomous vehicles. [8]

C. Advantages and Disadvantages

1) Advantages: High accuracy and performance when sufficient labeled data is available.
Clear interpretability of results, especially with simpler

models.
Well-established techniques and frameworks for implementation.

2) Disadvantages: Requires extensive labeled datasets, which can be time-consuming and costly to create. [21]
If the training set is not representative, it might not generalize well to new data. It also has the potential to overfit, in which the model picks up noise rather than the underlying pattern.

VI. Unsupervised Learning

The process of training algorithms on datasets without labelled outputs is known as unsupervised learning. [17]Without knowing the results beforehand, the objective is to find patterns, structures, or relationships in the data.This approach is particularly useful for exploratory data analysis and understanding the underlying structure of the data. [5]
To enable the model to discover hidden structures in the data, unsupervised learning involves training a data model on datasets without labels. [25] The two most important unsupervised learning techniques used commonly in NLP are clustering and topic modeling. [3]
Clustering entails forming a cluster where the pieces of text should be similar in some ways. For example, the article can be grouped into categories such as political, sport or technical articles in news based on the word frequencies and documents of the articles. These clusters help organizations to organize and search through the higher volumes of unstructured text. [3]
Another fundamental application is topic modeling that identifies topics hidden in a set of documents. Tools for this intention comprise Latent Dirichlet Allocation (LDA) and Non Negative Matrix Factorization (NMF). [13] For example, topic modeling can uncover latent topics when customers provide feedback about the product, for example, they complain that its quality is low, they like customer support, etc. This is particularly helpful for big data analysis, where the input figures are text in social media feeds or other sources that would take a long time to label by hand. [3]
The unsupervised learning algorithms learns few features from the data. When new data is introduced, It recognises the data's class using the previously learnt features. It is mainly used for clustering and feature reduction. [10]

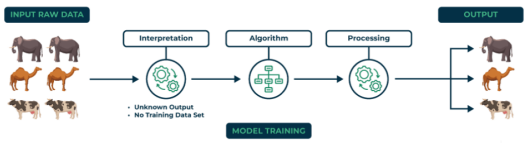

Fig. 3. Unsupervised Learning.

A. Key Characteristics

Unlabeled Data: Operates on datasets that lack explicit labels, relying solely on the inherent structure of the data. [13]
Pattern Recognition: Focuses on discovering hidden patterns or groupings in data, rather than predicting specific outcomes.
Flexibility: Can be applied to various data types, including numerical, categorical, and text data. [5]

B. Applications

Customer Segmentation: Identifying distinct groups within customer data to tailor marketing strategies or product offerings. [27]
Market Basket Analysis: Discovering associations between products purchased together, informing inventory and promotion strategies. [15]
Anomaly Detection: Identifying rare items or outliers in datasets, useful in fraud detection and network security. [8]

C. Advantages and Disadvantages

1) Advantages: No need for labeled data, making it easier to work with large datasets. [20]
Can reveal insights and structures that may not be immediately apparent, guiding further analysis. [15]
Useful for exploratory data analysis, helping to generate hypotheses and inform subsequent modeling efforts. [5]

2) Disadvantages: Compared to supervised learning, results may be more difficult to understand because output labels are unclear.
The choice of algorithm and parameters (e.g., number of clusters) can significantly impact results, requiring careful consideration.
May produce less accurate results when compared to supervised learning, as it lacks the guidance of labeled outputs. [5]

VII. Applications of NLP

Natural Language Processing (NLP) is transforming various industries, enhancing efficiency, personalizing experiences, and enabling data-driven decision-making [4].

A. Healthcare

NLP aids in processing unstructured medical data, supporting clinical decision-making by extracting patient information from electronic health records (EHRs) and matching it with medical literature [4].

B. Business and Finance

Sentiment analysis of customer feedback, social media, and financial news informs marketing and investment decisions. NLP also powers chatbots and virtual assistants like Siri and Alexa, improving customer service [9].

C. Education

NLP enhances learning with automated essay scoring and language learning platforms like Duolingo, which provide real-time corrections and personalized learning paths [9].

D. Media and Entertainment

NLP enables content recommendation systems on platforms like Netflix and Spotify by analyzing user preferences [4].

E. E-Commerce

NLP improves product search, personalized recommendations, and review analysis, enhancing customer experience and business insights [18].

F. Government and Public Policy

NLP streamlines governance by analyzing public sentiment from social media, news, and surveys, aiding policy-making [16].

G. Scientific Research

NLP facilitates literature mining, data extraction, and classification, helping researchers organize and analyze scientific findings more efficiently [4], [9].

VIII. Conclusion

In conclusion, the integration of Machine Learning (ML) and Natural Language Processing (NLP) has significantly advanced the field of artificial intelligence, enabling machines to process and understand human language with remarkable accuracy. The evolution of deep learning models and large language models has revolutionized various industries, from healthcare to e-commerce, improving applications such as translation, sentiment analysis, and content recommendation. Despite these advancements, challenges like data scalability, model interpretability, and domain adaptation remain. NLP will continue to be relevant and have an impact across a variety of fields if research and innovation in these areas continue.

IX. Acknowledgement

We would like to express our sincere gratitude to Mr. Kiran B.V., from the Department of Computer Science and Engineering at Alva's Institute of Engineering and Technology, for his continuous guidance, insightful feedback, and unwavering support throughout the development of this review paper. His expertise and dedication played a crucial role in refining and shaping this work to its success.

Our heartfelt thanks also go to Alva's Institute of Engineering and Technology for providing the necessary research facilities and resources to carry out this review.
In particular, we would like to acknowledge the Department of Library and Information Centre, IEEE Xplore,Google Scholar and SpringerLink for granting access to academic literature that was invaluable in gathering relevant material for this review.

We are also deeply appreciative of the Computer Science and Engineering Department's peers and colleagues for their valuable discussions and constructive suggestions. The collaborative spirit fostered within the department greatly contributed to creating an enriching research environment.

References

[1] V. Nilsen, "An Introduction to Machine Learning," May 22, 2018. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/Introduction-to-machime-learning/.

[2] T. P. Nagarhalli, "Role of Machine Learning in Natural Language Processing," 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/04/role-of-machine-learning-in-natural-language-processing/.

[3] L. Harris, "The Role of Machine Learning in Natural Language Processing: Bridging Communication Gaps," Aug. 19, 2021. [Online]. Available: https://www.researchgate.net/publication/386423203_The_Role_of_Machine_Learning_in_Natural_Language_Processing_Bridging_Communication_Gaps.

[4] K. Chen and C. Fei, "Deep Learning and Machine Learning – Natural Language Processing: From Theory to Application," Oct. 30, 2024. [Online]. Available: https://www.researchgate.net/publication/385700674_Deep_Learning_and_Machine_Learning_--_Natural_Language_Processing_From_Theory_to_Application.

[5] E. Oye, F. Edwin, and J. Owen, "Unsupervised and Supervised Models in Machine Learning," Dec. 17, 2024. [Online]. Available: https://www.researchgate.net/publication/387136555_Unsupervised_and_Supervised_Models_in_Machine_Learning.

[6] C. Isonkobong, "Machine Learning: A Review," Nov. 18, 2020. [Online]. Available: https://www.researchgate.net/publication/347059772_Machine_LearningA_Review.

[7] L. Zhao and F. Li, "Statistical Machine Learning in Natural Language Understanding: Object Constraint Language Translator for Business Process," 2008.

[8] P. Dayan, M. Sahani, and G. Deback, "Unsupervised learning," Pages 857-859 in The MIT Encyclopaedia of the Cognitive Sciences.

[9] D. M. Dutton and G. V. Conroy, "A review of machine learning," The Knowledge Engineering Review, vol. 12, no. 4, pp. 341-367, 1997.

[10] B. Mahesh, "Machine Learning Algorithms - A Review," International Journal of Science and Research (IJSR), vol. 9, pp. 381-386, 2020.

[11] D. Pandey, K. Niwaria, and B. Chourasia, "Machine Learning Algorithms: A Review," Machine Learning, vol. 6, no. 2, 2019.

[12] V. Nasteski, "An Overview of the Supervised Machine Learning Methods," Horizons, vol. 4, pp. 51-62, 2017.

[13] D. H. Maulud, S. Y. Ameen, N. Omar, S. F. Kak, Z. N. Rashid, and H. M. Yasin, "Review on Natural Language Processing Based on Different Techniques," Asian Journal of Research in Computer Science, vol. 10, no. 1, pp. 1-17, 2021.

[14] T. Brown, B. Mann, N. Ryder, et al., "Language Models are Few-Shot Learners," Advances in Neural Information Processing Systems, vol. 33, 2020.

[15] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need," Advances in Neural Information Processing Systems, vol. 30, 2017.

[16] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning," MIT Press, 2016.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, pp. 436-444, 2015.

[18] X. Zhang, Z. Yang, and Y. Yu, "Transformers in Natural Language Processing," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 7, pp. 1-14, 2021.

[19] K. Cho, B. Van Merriënboer, C. Gulcehre, et al., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in Proc. EMNLP, 2014.

[20] T. Mikolov, K. Chen, G. Corrado, et al., "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781, 2013.

[21] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in Proc. EMNLP, 2014.

[22] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[23] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019.

[24] A. Radford, J. Wu, R. Child, et al., "Language Models are Unsupervised Multitask Learners," OpenAI, 2018.

[25] G. Hinton, L. Deng, D. Yu, et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, 2012.

[26] Y. Zhang, X. Wu, and C. Xu, "Advancements in Transfer Learning for NLP," ACM Computing Surveys, vol. 55, no. 3, 2022.

[27] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A Neural Probabilistic Language Model," Journal of Machine Learning Research, vol. 3, pp. 1137-1155, 2003.