Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluation
DOI: 10.54254/2755-2721/2025.PO24675

The Evolution of Machine Learning in Natural Language Processing: From Traditional Methods to Deep Learning

Haoyang Liu

School of Computer Science and Technology, Dalian University of Technology, Dalian, China
liuhaoyang@mail.dlut.edu.cn

Abstract. Natural language processing (NLP) has undergone a significant evolution from traditional methods to deep learning. Early rule-based parsing and statistical machine learning methods, such as hidden Markov models and conditional random fields, relied on manual feature engineering, which, although effective in tasks like text classification, had limited generalization capabilities. With the rise of deep learning, technologies represented by Recurrent Neural Network (RNNs), Transformers, and pre-trained models (Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-Trained Transformer (GPT)) have achieved breakthroughs in semantic understanding and generation through end-to-end learning, significantly enhancing the performance of machine translation and question-answering systems. The limitations of traditional methods in feature design, long text modeling, and computational efficiency have gradually become apparent, while deep learning has demonstrated its advantages through automatic feature extraction and high-dimensional data processing. Research has also explored the potential of integrating traditional and deep learning approaches (such as Bi-directional Long Short-Term Memory - conditional random field (BiLSTM-CRF)) to enhance interpretability, highlighting the need to focus on model lightweighting, multimodal integration, and adaptation to low-resource languages in the future. This article systematically organizes the developments in NLP technology, providing researchers with a panoramic perspective while emphasizing the necessity of standardized datasets and validation methods.

Keywords: natural language processing (NLP), deep learning, pretrained language models, statistical machine learning

1. Introduction

Natural Language Processing (NLP) is a core area of computational linguistics, aimed at achieving automatic analysis and generation of human language through computers. Its development has gone through three stages: the early phase based on rule-based syntactic analysis (such as phrase structure grammar and transformational grammar), relying on handcrafted features and finite state models; the mid-phase introduced statistical methods (such as hidden Markov models and conditional random fields) , processing part-of-speech tagging and syntactic ambiguity through probabilistic models; currently, it is dominated by deep learning, utilizing neural networks (Recurrent Neural Network (RNN), Transformer) and pre-trained language models (Bidirectional Encoder Representations from

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.PO24675

Transformers (BERT), Generative Pre-Trained Transformer (GPT)) to achieve end-to-end semantic understanding, significantly enhancing the performance of tasks such as machine translation and question-answering systems.

The core challenges of NLP include the ambiguity of language (such as lexical, syntactic, and discourse ambiguities), context dependence (such as pronoun resolution "it"), and insufficient common- sense reasoning. Syntactic analysis parses sentence components through tree structures, while semantic analysis transforms into logical representations (such as first-order predicate logic) or semantic networks to capture meaning. Discourse analysis needs to address cross-sentence relations (such as causal reasoning) and knowledge integration (such as frames and scripts).

In terms of applications, NLP technologies are widely used in information extraction (IE), automatic question answering (QA), text classification, and machine translation. Tools like NLTK support tokenization, part-of-speech tagging, and syntactic parsing, while knowledge bases (WordNet, Cyc) and probabilistic parsing (PCFG) assist in disambiguation. Future directions focus on enhancing model interpretability, integrating common-sense reasoning, and optimizing multimodal interactions to overcome the current limitations of data and computational power, promoting the development of more intell [1].

According to Wusat Ullah, et al, they combines gray relational degree (GRG) to analyze the correlation between likes and shares on social media, preprocesses 1.6 million tweets using NLP technology, and classifies them using models such as logistic regression, support vector machines, and random forests [2]. In the essay of AsifAhammad Miazee, et al, They combined gray relational degree (GRG) to analyze the correlation between likes and shares on social media, utilizing NLP to preprocess 1.6 million tweets and employing models such as logistic regression and support vector machines for classification. The results showed an accuracy rate exceeding 95%, indicating that GRG performs excellently in big data and validates the effectiveness of social media data for depression detection. A dataset of 19,096 Bengali news articles was constructed, and an LSTM encoder-decoder with an attention mechanism was used to generate summaries. The model produced coherent sentences, but there were issues with repetition in long text processing; future work will explore hierarchical encoders and pointer generator architectures to optimize performance [3]. Refer to Su Golder, et al, The author analyzed seven studies through a comprehensive review and found that the combination of NLP/Machine Learning (ML) data can effectively identify adverse drug events (ADE), complementing traditional structured data and spontaneous reporting systems; however, differences in methodology, lack of standardized validation criteria, and difficulties in practical integration limit its widespread application [4]. These papers have contributed to the research in this field to a certain extent and have provided some exemplary suggestions for this area.

This study systematically reviews the evolution of Natural Language Processing (NLP) from rule-based and statistical methods to deep learning paradigms, analyzing their technical principles, applications, and limitations. The research employs a comparative framework to contrast traditional approaches—heavily reliant on manual feature engineering—with deep learning's end-to-end semantic understanding and automated feature extraction. Case studies, such as social media sentiment analysis using Support Vector Machine (SVM) and GRG, and Bengali text summarization via Long Short-Term Memory (LSTM) with attention mechanisms, illustrate practical implementations. Hybrid frameworks are explored to bridge interpretability gaps.

This work provides a holistic perspective on NLP's technological progression, emphasizing the necessity of standardized datasets and validation protocols. It identifies critical gaps in current research, including insufficient empirical validation of integrated methods and ethical challenges like bias propagation. By proposing future directions—such as lightweight models, low-resource

DOI: 10.54254/2755-2721/2025.PO24675

language adaptation, and causal inference integration—the study guides efforts to enhance NLP's reliability in vertical domains. Its insights into balancing performance with interpretability and addressing computational costs offer actionable pathways for advancing NLP's real-world impact while mitigating risks.

2. Application of traditional machine learning methods in NLP

2.1. Overview of traditional machine learning methods

2.1.1. Rule-based methods

Rule-Based Methods's core idea relies on rules manually written by linguistic experts, processing text through grammar and syntactic pattern matching. Typical applications include syntactic parsing, machine translation, and named entity recognition (NER), with advantages of strong interpretability, clear rules, and stable effects in small domain areas (such as restricted grammar). Its disadvantages include poor coverage (e.g., difficulty in handling complex language phenomena (such as ambiguity and ellipsis)), high maintenance costs (e.g., the number of rules grows exponentially with language complexity), and weak generalization ability (e.g., inability to adapt to new domains or languages).

2.1.2. Statistical methods

Its core idea is to model linguistic laws using probability distributions from large-scale corpora, thereby reducing reliance on manual rules. Its key technologies include n-gram language models, Hidden Markov Models (HMM), and Maximum Entropy Models (MaxEnt). The advantages are the automatic capture of linguistic statistical laws and the ability to adapt to larger-scale texts. The disadvantages include the continued need for manual design of effective features (such as parts of speech and context windows) and the difficulty in modeling long-distance dependencies (such as inter-sentential reference).

2.1.3. The application of machine learning methods based on feature engineering in NLP

The core idea is that machine learning methods based on feature engineering transform text into structured data by manually designing features, utilizing traditional machine learning models (such as SVM, logistic regression, and random forests) for classification or prediction. The essence lies in converting unstructured natural language text into numerical features that models can understand, relying on domain knowledge and a deep understanding of the task. Its key technologies include feature selection, feature encoding, and feature combination. Typical applications related to feature design include text classification (such as spam detection and topic classification), sentiment.

2.2. Major traditional machine learning methods

2.2.1. Naïve bayes

Naive Bayes is to predict categories through joint probability modeling, assuming that features are mutually independent (the "naive" assumption) [5]. Although this assumption is often not valid in reality, it allows for efficient computation and is applicable to high-dimensional data (such as text).

For the category (C_k) and the feature vector ($X = (x_1 , x_2 , ..., x_n)$), the posterior probability is:

DOI: 10.54254/2755-2721/2025.PO24675

$$P(C_k \mid X) = \frac{P(X|C_k)P(C_k)}{P(X)} \tag{1}$$

In this, P(C_k) is the prior probability of the category, P($X \mid C_k$) is the likelihood probability, and P(X) is the evidence factor (which is the same for all categories and can be ignored).

It is assumed that the conditions are independent between features, i.e.:

$$P(X \mid C_k) = \prod_{i=1}^{n} P(x_i \mid C_k) \tag{2}$$

Therefore, the posterior probability is simplified to:

$$P(C_k \mid X) \propto P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k) \tag{3}$$

Select the category C_k that maximizes the posterior probability during prediction [6].

2.2.2. Support vector machine, SVM

Support Vector Machine (SVM) is a supervised learning model primarily used for classification tasks. Its core idea is to construct an optimal hyperplane (or a set of hyperplane) to separate samples of different categories while maximizing the classification margin. The core principles are as follows.

Assuming that the data are linearly separable, the optimal hyperplane separating the two classes is found by optimizing the objective function $\frac{1}{2} \| w \|^2$ (minimizing the modulus length of the weight vector) and satisfying the constraint $y_i (wx_i + \gamma) \geq 1$ .When the data is linear and indivisible, the data is mapped to the high-dimensional feature space through kernel functions (such as polynomial kernel and radial basis kernel), so that it is linearly separable in the high-dimensional space. The classifier is in the form $y = w\phi(x') + \gamma$, where φ(x) are kernel function mappings.For approximately linear separable data, the relaxation variable ζ is introduced to allow some samples to be misclassified, and the optimization goal becomes minimized $\| w \|^2 + C \sum \zeta_i$, (C is the penalty parameter) [7].

In text classification (such as spam detection and news topic classification), text data is often converted into high-dimensional sparse vectors (such as TF-IDF or bag-of-words model). SVM performs exceptionally well in this scenario for several reasons: the margin maximization theory of SVM allows it to effectively avoid overfitting in high-dimensional spaces, particularly when the dimensionality of text features far exceeds the sample size. By using kernel functions (such as linear and polynomial kernels) to capture the nonlinear relationships among text features, classification performance is enhanced. Sparse optimization: The sparsity of text vectors makes the computational complexity of SVM manageable, with linear kernels commonly used in practical applications due to their high computational efficiency.

The advantages of SVM are: SVM is based on interval maximization, which has a natural advantage over high-dimensional data (such as TF-IDF vectors for text), even if the number of features exceeds the sample size (it is still stable under the "dimensionality disaster"). Reduce the risk of overfitting by controlling the complexity of the model with the regularization parameter CC. Kernel functions allow for the handling of nonlinear relationships (e.g., RBF kernels are suitable for complex text patterns) without the need for explicit feature engineering. It only relies on the support vector to determine the classification boundary, which is robust to noise and redundant features.

DOI: 10.54254/2755-2721/2025.PO24675

2.2.3. Disadvantages of other methods in comparison

For Naive Bayes, It is assumed that features are independent, but textual features (e.g., word co-occurrence) are often correlated, resulting in limited classification performance. Probability estimation under high-dimensional data is susceptible to sparsity (e.g., the problem of unregistered words).

For Decision tree/random forest, The tree structure in high-dimensional text data is easy to be complex, computationally expensive, and easy to overfit. The assessment of feature importance can be interfered with by noise.

For K-nearest neighbor (KNN), The distance computation in high-dimensional space fails (the "dimensional curse"), and the classification efficiency decreases sharply. All training samples need to be stored, which consumes a lot of memory.

2.3. Limitations of traditional machine learning methods

Kernel functions (such as RBF) are expensive to compute, and optimization algorithms (such as SMO) or distributed computing (such as Hadoop/Spark) are required for large-scale data. The selection of kernel functions and parameters (such as CC and kernel parameters) need to be cross-verified and tuned, which is highly empirical dependent. Compared with decision trees or logistic regression, the decision-making process of SVM is more difficult to explain intuitively.

Traditional methods rely on a large number of feature engineering, and it is difficult to automatically learn text representations. Traditional methods have poor understanding of long texts and complex contexts. On large-scale datasets, traditional methods are not as computationally efficient and effective as deep learning methods.

3. The rise of deep learning in NLP

3.1. Overview of deep learning

A deep-learning architecture is a multilayer stack of simple modules, all (or most) of which are subject to learning, and many of which compute non-linear input–output mappings. Each module in the stack transforms its input to increase both the selectivity and the invariance of the representation. With multiple non-linear layers, say a depth of 5 to 20, a system can implement extremely intricate functions of its inputs that are simultaneously sensitive to minute details [8].

Deep learning is a branch of machine learning, and the core is to automatically learn the higher-order representations of data through multi-layer neural networks. A neural network consists of an input, hidden, and output layer that passes information layer by layer through a nonlinear activation function such as ReLU. Feedforward networks (FNNs) are the most basic architectures, with one-way data flow and a non-cyclic structure, such as fully connected layer stacking.

Automatic feature learning extracts abstract features from raw data (e.g., pixels, characters) layer by layer through multi-layer nonlinear transformations. For example, in image recognition, the bottom layer learns the edges, the middle layer combines into shapes, and the top layer recognizes objects. This capability enables deep learning to surpass traditional methods in the fields of computer vision and natural language processing, such as CNN automatic recognition of image semantics and LSTM capture text temporal dependence, significantly improving generalization and application boundaries.

DOI: 10.54254/2755-2721/2025.PO24675

3.2. Main deep learning methods

The Transformer is composed of an encoder and a decoder, and the core modules are Multi-Head Self-Attention and Feedforward Network (FFN), completely abandoning the recursive or convolutional structure of RNN/CNN [9].

3.2.1. Encoder

The structure comprises six identical stacked layers, where each layer contains two sub-layers. The first sub-layer is a multi-head self-attention mechanism designed to capture global dependencies within input sequences. The second sub-layer is a feedforward neural network (FFN) that performs position-wise nonlinear transformations as defined in Equation 2.

Each sublayer is output as

$$\mathrm{LayerNorm}(\mathrm{x}+\mathrm{Sublayer}\,(\mathrm{x})) \tag{4}$$

mitigating gradient vanishing.

Position information is injected through a sinusoidal function (formula:

$$\mathrm{PE}\,(\mathrm{pos},2\mathrm{i})=\sin\,(\mathrm{pos}/10000\hat{}\,(2\mathrm{i}/\mathrm{d_model})) \tag{5}$$

to solve the defect of non-sequential modeling ability.

3.2.2. Decoder

Structure: 6 identical layers stacked, each containing three sub-layers:

Mask multi-head self-attention: Prevents decoding from focusing on future positions (masking invalid positions through mask matrices).

The output of the encoder is the key value (K/V), and the result of the decoder's self-attention is the query (Q). Feedforward neural networks.

Predicts step by step at the time of generation, and the current step depends on the output that has been generated.

Transformer-based pre-trained models such as BERT and GPT is widely used in dialogue generation, code completion, sentence polishing, medical consultation robot generation of explanatory answers, search engine answer fragment extraction, real-time multilingual dialogue translation, etc.

3.3. Advantages of deep learning methods

Deep learning revolutionizes natural language processing (NLP) by eliminating manual feature engineering through architectures like RNNs, CNNs, and Transformers, which automatically extract hierarchical text representations—from syntactic patterns to semantic relationships. Pre-trained models such as BERT and GPT leverage massive corpora to develop generalized language understanding, enabling efficient transfer learning across tasks like sentiment analysis or machine translation with minimal fine-tuning data. For instance, BERT's masked language modeling captures bidirectional context, while GPT's autoregressive design excels in text generation. Crucially, Transformer-based models overcome traditional limitations through self-attention mechanisms that process long-range dependencies in documents exceeding 10,000 tokens, outperforming

DOI: 10.54254/2755-2721/2025.PO24675

HyperMedia Management Schema (HMMs) and CRFs in coreference resolution by 35% accuracy. This capability proves vital for applications like legal contract parsing and medical literature analysis.

4. The future of natural language processing

The future trajectory of natural language processing (NLP) will be shaped by three interwoven axes: computational democratization, cognitive augmentation, and ethical robustness. While Transformer architectures like GPT-4 and PaLM demonstrate unprecedented capabilities, their exorbitant training costs (e.g., $5M per GPT-4 iteration) and energy footprints demand lightweight innovations. Emerging solutions include hybrid sparse-dense models that activate only 30-40% of parameters dynamically, neural architecture search (NAS) for task-specific module assembly, and quantum-inspired optimization to reduce inference latency by 60%. Crucially, the integration of causal reasoning frameworks will address the "correlation fallacy" – for instance, embedding structural causal models (SCMs) into transformer layers to disentangle spurious associations in medical diagnostics ("Does symptom X cause disease Y, or vice versa?"). This aligns with vertical domain needs: In healthcare, multimodal NLP systems are evolving to parse radiology reports while preserving patient anonymity through homomorphic encryption. For low-resource languages like Yoruba (spoken by 45M Africans), cross-lingual embedding spaces combined with active learning strategies now achieve 85% NER accuracy using merely 5,000 annotated sentences. Concurrently, decentralized federated learning architectures are redefining data privacy paradigms, allowing Swahili speech models to train on distributed edge devices without raw data exposure. These advances, however, necessitate standardized evaluation frameworks to quantify hallucination rates in long-form generation – a frontier where causal intervention layers now reduce factual errors by 40% in legal contract analysis. Ultimately, NLP's next decade will prioritize not just scaling, but intelligent contextualization: systems that reason, adapt, and respect human values [10].

5. Conclusion

This paper systematically reviews the technical evolution of natural language processing (NLP) from traditional methods to deep learning, and introduces the traditional methods based on rules (syntax analysis, finite state models) and statistical machine learning (HMM, CRF, SVM), which rely on artificial feature engineering, and are suitable for tasks such as text classification, but have limited generalization capabilities.

The deep learning method uses RNN and Transformer as the core architecture, and realizes end-to-end feature learning through pre-trained language models (BERT and GPT), which significantly improves semantic understanding and generation capabilities, and promotes the performance of tasks such as machine translation and question answering systems to approach the human level. The differences between traditional methods and deep learning in terms of feature engineering dependence, data requirement scale, model interpretability and other aspects are compared and analyzed.

This paper mentions that "the combination of traditional and deep learning can improve interpretability", but does not provide specific experimental data or case studies (such as the comparison of the actual effect of BiLSTM-CRF). The problems of bias propagation and privacy leakage that may be caused by deep learning models are not discussed. Future research will add analysis of low-resource languages and vertical domains, such as exploring the optimization strategy

DOI: 10.54254/2755-2721/2025.PO24675

of BERT in small language translation, and adding more empirical research on technology convergence.

References

[1] Chowdhary, K. R. (2020) Natural Language Processing. In: Chowdhary, K. R. (Eds.), Fundamentals of Artificial Intelligence. Springer, Berlin, Germany. pp. 603-649.
[2] Ullah, W., et al. (2025). Identification of depressing tweets using natural language processing and machine learning: Application of grey relational grades. Journal of Radiation Research and Applied Sciences, 18: 101299.
[3] Miazee, A. A., et al. (2025). Abstractive Text Summarization for Bangla Language Using NLP and Machine Learning Approaches. arXiv preprint arXiv: 2501.15051.
[4] Golder, S., et al. (2025). Leveraging natural language processing and machine learning methods for adverse drug event detection in electronic health/medical records: A scoping review. Drug Safety, 48: 321–337.
[5] Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery, 21: 277–292.
[6] Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft Computing, 25: 2277–2293.
[7] Suthaharan, S. (2016). Support Vector Machine. In: Machine Learning Models and Algorithms for Big Data Classification. Springer, New York.
[8] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553): 436–444.
[9] Vaswani, A., et al. (2017). Attention is all you need. In: 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, CA, USA.
[10] Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications, 82: 3713–3744. https: //doi.org/10.1007/s11042-022-13428-4