Dataset-Aware Automated Model Selection for Abstractive Summarization: A Meta-Learning Approach

[1]Dhivya Bino*, [2]Manish Shrivastava

[1]Department of Computer Science and Engineering, Vivekananda Global University, Jaipur, India
[2]Middle East College, Muscat, Oman
[1]dhivya@mec.edu.com
[2]manish.shrivastava@vgu.ac.in

Abstract. Pretrained transformer-based models have significantly advanced abstractive summarization. However, selecting the right model for a new dataset remains challenging because exhaustively fine-tuning multiple candidate models for each dataset is computationally expensive and limits the scalability of summarization pipelines. This paper presents a dataset-aware meta learning framework for automated model selection in this context. We first construct an empirical oracle by fine-tuning key transformer-based models under a standardized training budget across heterogeneous summarization datasets. Model performance is evaluated using both lexical (ROUGE-L) and semantic (SBERT cosine similarity) metrics, producing objective-aware oracle labels. Dataset-level meta-features are then used to train supervised selectors that predict suitable models for previously unseen datasets using the LODO strategy. Experimental results reveal substantial dataset-dependent variability and a 71.4% disagreement between lexical and semantic oracle models, highlighting the need for automatic model selection. The proposed meta-learning selector achieves up to 85.7% accuracy in predicting lexical oracle models and 71% accuracy for semantic objectives, while significantly reducing the need for exhaustive fine-tuning.

Keywords: Automated Model Selection, Meta Learning, Abstractive Summarization, AutoML.

1 Introduction

Recent advances in deep learning have led to the rapid development of powerful pre-trained language models for natural language processing (NLP) tasks such as summarization, translation, and dialogue generation. Models including BART, PEGASUS, and LED have demonstrated strong performance across a variety of benchmarks. However, selecting the most suitable model for a new data set remains a challenging problem. Model performance frequently and significantly changes depending on dataset characteristics such as document length, domain, and summarization style, leading to a time consuming and computationally expensive manual model selection process. In practice, researchers and practitioners typically evaluate multiple candidate models independently and select the best-performing model through exhaustive experimentation,

B. Singh et al. (eds.), *Proceedings of the International Conference on Advances in Computing Technology and Artificial Intelligence (COMPUTATIA 2026)*, Atlantis Highlights in Intelligent Systems 18,
https://doi.org/10.2991/978-94-6239-713-2_43

which can be inefficient and resource-intensive. Automated machine learning (AutoML) aims to reduce the need for manual experimentation by automatically identifying suitable models and configurations [1][2]. While existing AutoML approaches have achieved notable success in domains such as computer vision and text classification, relatively very little work has been done for generative NLP tasks such as abstractive summarization.

To address this challenge, we present the first objective-aware meta-learning framework for automated selection of pretrained transformer models in abstractive summarization. The proposed approach learns from prior experiments conducted across multiple summarization datasets and uses dataset-level meta-features to predict which summarization model is most likely to perform well on a new dataset. Instead of retraining all candidate models for every new dataset, the meta-learning selector recommends a suitable model by leveraging patterns observed across previously evaluated datasets. The main contributions of this work are the following: 1. A dataset-aware meta-learning framework for automated model selection in abstractive summarization. 2. A full-fledged experimental study across fourteen diverse summarization datasets and four transformer models. 3. An empirical analysis of dataset characteristics influencing model suitability.

Experiments conducted on diverse summarization datasets using widely used transformer models demonstrate that model performance is substantially different across datasets when different evaluation objectives are considered. The results further show that meta-learning-based supervised selectors can effectively predict suitable models for unseen datasets while reducing the regret (defined later) associated with incorrect model selection. These findings highlight the potential of dataset-aware automated model selection as an efficient approach for improving the scalability of summarization pipelines.

2 Related Work

2.1 AutoML in NLP

Automated Machine Learning (AutoML) is an evolving field that aims to reduce human effort in building machine learning pipelines by automating different key stages like algorithm selection, hyperparameter optimization, and architecture design [1] [2]. Neural Architecture Search (NAS), one of its core branches, has achieved remarkable success in computer vision tasks such as in image classification and object detection [3]. However, NAS approaches focus on the design of the search space; search strategy and performance evaluation [4] are typically computationally intensive and require extensive training cycles [5]. In the context of NLP, the progress of AutoML and NAS has been comparatively slower, with the early works largely centered on classification tasks such as sentiment analysis or natural language inference [3]. Differentiable NAS has been integrated into dialogue summarization systems to refine BART-based architectures, yet such methods optimize structural configurations for a fixed dataset rather than address cross-dataset model selection [8]. Recent advances exploring transformer-based attention head search and depth scaling [6] remain computationally demanding and often tailored to specific tasks [7]. Moreover, most of such AutoML systems aim to generate new architecture within the same model family rather than select among existing pretrained models, which may be unnecessary when many strong pretrained candidates already exist. Furthermore, autogenerated frameworks using NAS approaches are often not transparent, with limited interpretability in how final architectures are selected or limited transferability to other domains [9] [10]. Parallel to NAS, frameworks such as AutoNLP [11], TextBrew [4], and stacked meta-learning pipelines with adversarial augmentation [13] demonstrate automated model selection and tuning for text classification. These systems predominantly target discriminative tasks and optimize single metrics (accuracy or F1). Evaluations typically rely on random splits rather than cross-dataset validation, and generative tasks such as abstractive summarization remain largely unexplored.

2.2 Meta-Learning for Algorithm Selection

Meta learning extends the foundational algorithm selection problem, which frames model choice as a mapping from problem characteristics to the most suitable algorithm [18] and seeks to apply prior experience across tasks to improve performance on new tasks [15]. A commonly seen approach, meta-feature-based learning, uses statistical and structural descriptors as inputs to train a meta-learner. Meta-learning has also been applied to recommend feature selection methods, beyond selecting model-predicting algorithms, demonstrating that the effectiveness of feature selection algorithms varies significantly across datasets and can be learned via dataset characteristics and ranking-based targets [16] [17]. Auto-sklearn style systems employ statistical meta-features to guide algorithm choice, while subsequent studies refine feature selection to improve predictive accuracy [18]. Embedding-based methods such as Auto-CASH learn latent dataset representations from historical performance matrices, enabling similarity-based recommendation [19]. TextBrew combines meta-learning, hyperparameter search, and synthetic meta-data augmentation for transformer-based text classification [12]. Similarly, stacked meta-learning with adversarial augmentation ensembles multiple pretrained transformers for improved robustness [13]. Within NLP, meta-learning has primarily been applied to classification or ensemble optimization mostly confined to tabular data or text classification under single-objective metrics. While effective, these systems operate within a single dataset and focus on ensemble construction rather than cross-dataset model recommendation.

2.3 Model Selection in Abstractive Summarization

Text Summarization primarily is done using two approaches, abstractive and extractive. While extractive summarization involves copying key phrases and sentences from the original document, abstractive summarization generates new concise content from the original text while preserving semantic similarity, coherency and fluency [20]. Abstractive summarization has advanced significantly with the introduction of pretrained encoder-decoder transformer models. BART, PEGASUS, and LED [21] [22] are widely used baselines across news, dialogue, and multi-document summarization tasks evaluating these models across datasets such as CNN/DailyMail, XSum, and SAMSum [23]. These studies typically analyse performance differences under fine-tuning or zero-shot settings. Meta-learning has been utilized for domain adaptation rather than model selection in abstractive summarization where lightweight MAML-based frameworks insert trainable modules into a fixed backbone such as BART to facilitate low-resource adaptation across summarization datasets [8]. Research showing that certain models perform better on news-style datasets while others are more robust for long-document summarization [21], consider model comparison as an empirical benchmarking exercise rather than a predictive selection problem. Transfer learning studies show that performance varies significantly across domains depending on dataset length, compression ratio, and discourse structure [14]. These approaches demonstrate that cross-dataset knowledge transfer can significantly improve summarization performance. However, such methods typically assume a fixed base architecture and focus on improving adaptation rather than selecting alternative pretrained models.

Three patterns emerge from AutoML, meta-learning, and summarization literature: a) NAS optimizes architecture rather than selects from pretrained models focusing on specific model families; b) AutoNLP systems work primarily on classification and single-objective metrics; and c) summarization research emphasizes benchmarking instead of predictive selection. No prior work, to our knowledge, constructs a meta-dataset of summarization tasks labelled with oracle models derived from standardized fine-tuning experiments and trains a cross-dataset supervised selector under multi-objective evaluation. The present study addresses this gap by formalizing pretrained transformer selection for abstractive summarization as an objective-aware meta-learning problem evaluated under controlled fine-tuning budgets and cross-dataset validation.

3 Methodology

3.1 Problem Definition

Let $D = \{D_1, D_2, \ldots, D_n\}$ denote a set of summarization datasets and $M = \{m_1, m_2, \ldots m_k\}$ represent a pool of candidate summarization models. For each dataset D_i, model performance is evaluated using a scoring function $s_o(m, D_i)$ where $o \in \{lex, sem\}$. The optimal model for dataset D_i is defined as the oracle model:

$$m_i^* = \arg\max_{m \in M} s_o(m, D_i) \tag{1}$$

The objective of the framework is to learn a selector function g(·) that predicts the most suitable summarization model for an unseen dataset:

$$g : \phi(D_i) \to m_i^* \tag{2}$$

where $\phi\ (D_i)$ denotes a vector of dataset-level meta-features describing dataset D_i.

3.2 Data Set Aware Meta Learning Framework

The proposed framework constructs a meta-dataset including dataset meta-features and oracle labels. The meta-features characterize key dataset properties, while oracle labels are obtained by fine-tuning and evaluating candidate models under a fixed training budget across multiple datasets and selecting the highest-performing model for each dataset using predefined evaluation metrics: ROUGE-L for lexical overlap and SBERT cosine similarity for semantic similarity.
The meta-dataset is defined as:

$$T = \{(x_i, m_i^*)\}_{\{i=1\}}^{n} \tag{3}$$

where x_i is for $\phi\ (D_i)$ and represents the dataset meta-features, and m_i^* denotes the corresponding oracle model. The meta-dataset T provides supervised training examples used to learn the selector function g(·) defined in Equation (2). Two oracle definitions considered based on the two evaluation objectives are:

$$m_{lex,i}^* = \arg\max_{m \in M} ROUGE - L(m, D_i) \tag{4}$$

$$m_{sem,i}^* = \arg\max_{m \in M} SBERT(m, D_i) \tag{5}$$

These definitions enable analysis of objective-dependent model behaviour. Disagreement between lexical and semantic oracle models is referred to as the oracle flip rate, defined as

$$FlipRate\ = \left|\left\{D_i \in D\ :\ m_{lex,i}^* \neq m_{sem,i}^*\right\}\right| / n \tag{6}$$

where n is the total number of datasets.
The overall procedure for constructing the meta-dataset and training the selector is summarized in Algorithm 1.

Algorithm 1: Dataset-aware meta-learning for model selection

Require: datasets D, candidate models M, evaluation metric $s_o(.)$, meta-feature extractor $\phi(.)$, selector learner A.

Ensure: Trained selector g(·)

1: $T \leftarrow \phi$ T
2: for all $D_i \in D$ do
3: for all $m \in M$ do
4: Fine-tune and evaluate m on D_i
5: Compute performance score $s_o(m, D_i)$

end for
Compute oracle model $m_i^* \leftarrow \max_{m \in M} s_o(m, D_i)$
Extract meta-features $x_i \leftarrow \phi(D_i)$
$T \leftarrow T \cup \{(x_i, m_i^*)\}$
end for
Train selector $g \leftarrow A(T)$
return g

* The procedure is executed independently for each evaluation objective $o \in \{lex, sem\}$.

3.3 Meta Features and Selector Evaluation

Dataset-level meta-features are extracted from the training split of each dataset to characterize structural properties influencing summarization performance. The features include average document length, standard deviation of document length, average summary length, compression ratio, vocabulary size, type-token ratio, and training dataset size.

Several classifiers are evaluated as meta-learning selectors, including Logistic Regression (LR), Random Forest (RF), k-Nearest Neighbors (KNN), and Gradient Boosting (GB). To assess generalization performance, a Leave-One-Dataset-Out (LODO) validation strategy is used in which the selector is trained on n − 1 meta-datasets and evaluated on the remaining dataset.

Selector performance is measured using accuracy, macro-F1 score, and average regret. Regret quantifies the performance loss caused by selecting a suboptimal model and is defined as:

$$Regret(D_i) = s(m_i^*, D_i) - s(\hat{m}_i, D_i) \tag{7}$$

where m_i^* denotes the oracle model for dataset D_i and $\hat{m}_i$ represents the model predicted by the selector.

4 Experimental Design

4.1 Meta Features and Selector Evaluation

Experiments were conducted on fourteen publicly available abstractive summarization datasets covering diverse domains and input characteristics, including news, legislative text, dialogue, scientific papers, patents, and user-generated content. Table 1 and Table 2 list the datasets, candidate models, and their typical properties, enabling analysis of model behavior under varied domains, document lengths, and summarization styles (single/multi-document and conversational/narrative). Transformer-based summarization models were selected as candidate models, including both standard encoder-decoder architectures and long-context models, representing commonly used pretrained summarization backbones.

Table 1 Datasets used in the experiments

Dataset	Domain	Length
BillSum(D1)	Legislative	Long
CNN/DailyMail (D2)	News	Med–Long
XSum(D3)	News	Medium
SAMSum(D4)	Dialogue	Short–Med
MultiNews(D5)	News	Long
WikiHow(D6)	Procedural	Medium
Reddit-TIFU(D7)	User stories	Med–Long
Newsroom(D8)	Journalism	Mixed
Gigaword(D9)	Headlines	Short
AmazonOpinions(D10)	Reviews	Short–Med
AESLC(D11)	Emails	Short
ArXiv(D12)	Scientific	Long
BigPatent(D13)	Patents	Long
GovReport(D14)	Government	Long

Table 2 Candidate models used in the experiments

Model	Architecture	Max Tokens
BART-large (BART)	Encoder–Decoder Transformer	1024
PEGASUS-large (PEG)	Encoder–Decoder Transformer	1024
LED-base (LED)	Long-context Transformer	16384
FLAN-T5-large (FT5)	Encoder–Decoder Transformer	1024

4.2 Training Configuration

To ensure fair comparison across datasets and models, fixed-size subsets were used for all experiments: 3000 training instances, 500 validation instances, and 500 test instances sampled without replacement from the available splits. All runs used a fixed seed (42) for dataset sampling and model training. All models were fine-tuned under the same training budget using HuggingFace Trainer: maximum steps = 1500, batch size = 1, gradient accumulation = 8, and decoding with beam search (4 beams), maximum new tokens = 256, and no-repeat n-gram size = 3. Inputs were truncated to 1024 tokens for standard models; LED used its long-context capability where applicable. All experiments were implemented in Python using HuggingFace Transformers and PyTorch. Meta-learning selectors were trained using scikit-learn. Experiments were executed in the Kaggle environment with GPU acceleration for model fine-tuning and CPU execution for selector training and analysis.

5 Result and Discussion

5.1 Dataset Wise Model Performance

Fig. 1 (a and b) shows that FLAN-T5-large often achieves the highest ROUGE-L scores, while BART-large consistently performs best under semantic evaluation. The results also indicate substantial performance variation across datasets, reinforcing the idea that summarization effectiveness is strongly dataset dependent. This divergence suggests that lexical overlap does not necessarily imply superior semantic quality. In particular, long-document datasets tend to benefit from models capable of capturing extended context like LED, whereas conversational and short-text datasets show different model preferences. Overall, the findings indicate that no single model consistently dominates all datasets and objectives.

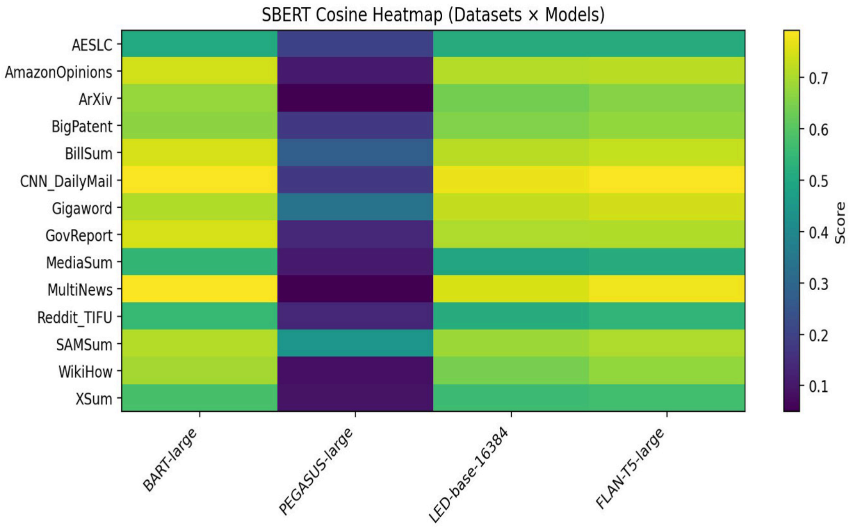

Fig. 1(a): Model performance across datasets under lexical and semantic objectives

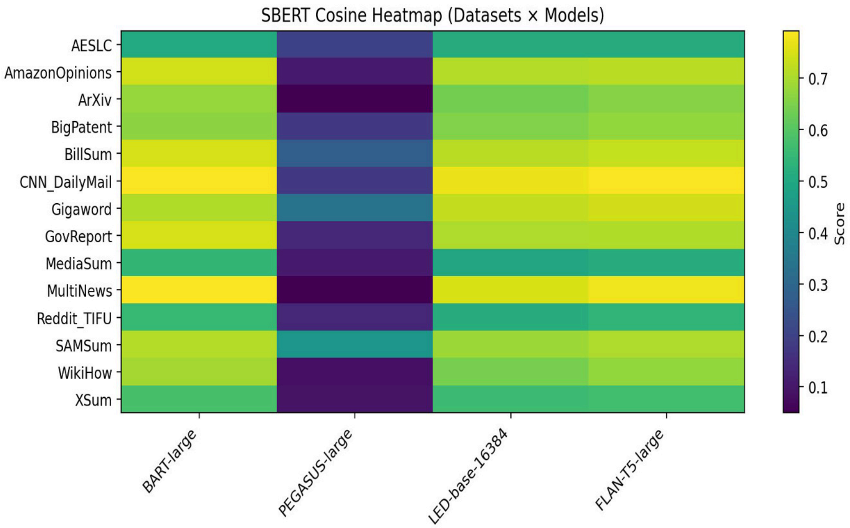

Fig. 1(b): Model performance across datasets under lexical and semantic objectives

5.2 Oracle Model Analysis

Oracle analysis reveals strong objective-dependent behaviour. FLAN-T5-large wins the model-wise distribution under lexical evaluation, with 64.3% of datasets, whereas BART-large emerges as the strongest semantic oracle with 57.1% of datasets, as shown in Table 3 and Table 4. The fourteen datasets exhibit different oracle models across the two objectives, yielding a high oracle flip rate of 71.4%, which is quite significant (Fig. 2). The dataset-wise oracle summary in Table 5 further highlights that oracle disagreement is widespread across domains rather than concentrated in a few datasets. This indicates that model suitability depends strongly on evaluation objectives and that single-metric model selection can lead to suboptimal choices when diverse datasets are under consideration.

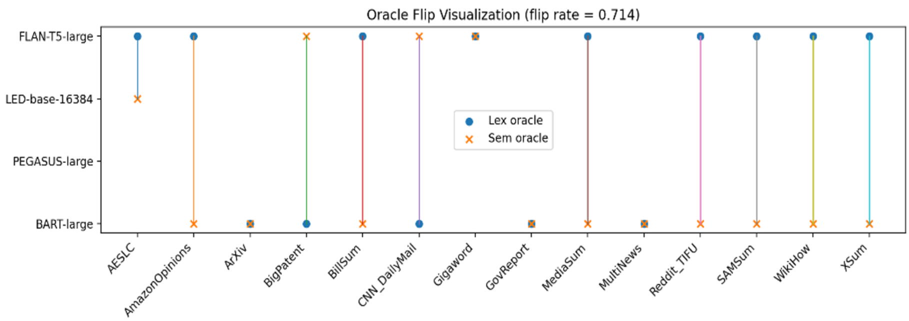

Fig. 2 Oracle disagreement between lexical and semantic objectives

5.3 Meta Learning Selector Performance

As indicated in Table 5, for lexical objectives, KNN performs the best, achieving the highest accuracy (85.7%) and lowest regret (0.006), reducing regret by more than 60% compared to the majority baseline (0.017). Random Forest and Logistic Regression also demonstrate strong performances with relatively low regret values, indicating that the proposed approach is effective in training meta-learning selectors for suitably predicting models for unseen datasets. For the semantic objectives, the differences between selector performances are smaller due to the dominance of BART-large as the oracle. Despite this, meta-learning selectors still consistently achieve lower regret compared to the majority baseline. Macro-F1 scores further suggest that selectors produce balanced predictions across oracle classes rather than being biased towards frequently selected models.

Table 3. Model-wise Oracle Distribution

Model	Lex (%)	Sem (%)
FT5	64.3	14.3
BART	14.3	57.1
LED	14.3	14.3
PEG	7.1	14.3

Table 4 Dataset-wise Oracle Summary

Dataset	**Lex**	**Sem**	**Dataset**	**Lex**	**Sem**
D1	*FT5*	BART	D8	*FT5*	BART
D2	*FT5*	BART	D9	PEG	PEG
D3	*FT5*	*FT5*	D10	FT5	BART
D4	BART	BART	D11	LED	LED
D5	*FT5*	BART	D12	*FT5*	*FT5*
D6	*FT5*	*FT5*	D13	*FT5*	BART
D7	*LED*	LED	D14	BART	BART

Table 5. Meta-learning selector performance

Task	**Selector**	**Accuracy**	**Macro F1**	**Avg Regret**
Lexical	Majority	0.64	-	0.017
	RF	0.79	0.61	0.008
	LR	0.79	0.61	0.009
	KNN	**0.86**	**0.66**	**0.006**
Semantic	Majority	0.57	-	0.010
	RF	0.71	0.52	0.006
	LR	0.71	0.49	0.007
	KNN	**0.71**	**0.53**	**0.006**

5.4 Meta-Feature Importance

Feature importance analysis was undertaken using Random Forest impurity-based importance scores, which indicated that vocabulary size, average document length, and type-token ratio are the most influential predictors. Vocabulary size reflects the lexical diversity of a dataset, document length captures its contextual complexity, and type-token ratio represents its lexical richness. Together, these dataset-level characteristics provide meaningful inputs that help in predicting the most suitable model for a given dataset.

5.5 Discussion and Limitations

The results highlight that summarization performance is highly dependent on the dataset, and the selection of the best models depends on the evaluation objectives, like lexical or semantic. The use of dataset-level meta-features enables effective prediction of suitable models, significantly reducing the computational cost of exhaustive model comparison. This study considered only fourteen datasets and four pretrained models; hence, expanding dataset diversity and model pools may further improve generalization. Additionally, richer embedding-based dataset representations could enhance predictive performance. Despite these limitations, the proposed framework and related findings suggest that lightweight dataset descriptors are sufficient for practical automated model selection.

6 Conclusion

This paper explored automated model selection for abstractive summarization using a meta-learning framework. Experiments conducted using diverse datasets showed that summarization performance varies significantly with dataset characteristics and evaluation objectives, with a high oracle flip rate observed between lexical and semantic metrics. The results of the proposed approach demonstrate that meta-feature-trained meta-selectors can effectively recommend models for unseen datasets while also reducing the regret associated with incorrect model selection. Dataset characteristics such as vocabulary size, document length, and lexical diversity acting as meta-features were found to play an important role in determining model suitability. Overall, these findings highlight the potential of dataset-aware meta-learning approaches for efficient and reliable model selection in summarization systems and other NLP pipelines involving generative tasks. Future work will focus on extending the approach to larger model pools, incorporating richer meta-feature representations, and considering additional natural language generation tasks.

References

1. [R. Barbudo, S. Ventura, and J. R. Romero, "Eight years of automl: categorisation, review and trends," *Knowl. Inf. Syst.*, vol. 65, no. 12, pp. 5097–5149, 2023.
2. K. T. Chitty-Venkata, M. Emani, V. Vishwanath, and A. K. Somani, "Neural architecture search for transformers: A survey," *IEEE Access*, vol. 10, pp. 108374–108412, 2022.
3. I. Dagan, R. Vainshtein, G. Katz, and L. Rokach, "Automated algorithm selection using meta-learning and pre-trained deep convolution neural networks," *Inf. Fusion*, vol. 105, Art. no. 102210, 2024.
4. R. Desai, A. Shah, S. Kothari, A. Surve, and N. Shekokar, "Textbrew: Automated model selection and hyperparameter optimization for text classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, 2022.
5. T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search," in *Automated Machine Learning: Methods, Systems, Challenges*, F. Hutter, L. Kotthoff, and J. Vanschoren, Eds. Cham, Switzerland: Springer, 2019, pp. 63–77.
6. A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "Summeval: Re-evaluating summarization evaluation," *Trans. Assoc. Comput. Linguist.*, vol. 9, pp. 391–409, 2021.
7. X. He, K. Zhao, and X. Chu, "AutoML: A survey of the state-of-the-art," *Knowl.-Based Syst.*, vol. 212, Art. no. 106622, 2021.
8. T. Huh and Y. Ko, "Lightweight meta-learning for low-resource abstractive summarization," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (SIGIR)*, 2022, pp. 2629–2633.
9. L. Jiang, Y. Hao, and J. Lin, "Dialogue text summarization method combined self-supervised learning and neural architecture search," in *Proc. 2nd Int. Conf. Big Data, Artif. Intell. Risk Manage.*, 2022, pp. 65–69.
10. N. I. Kulin and S. B. Muravyov, "A meta-feature selection method based on the auto-sklearn framework," *Sci. Tech. J. Inf. Technol. Mech. Opt.*, vol. 21, no. 5, pp. 702–708, 2021.
11. M. Lewis et al., "Bart: Denoising sequence-to-sequence pre-training for natural language generation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguist. (ACL)*, 2020, pp. 7871–7880.
12. A. Mosquera, "Towards autonlp via meta-learning and adversarial data augmentation," in *Proc. 16th Int. Workshop Semantic Eval. (SemEval)*, 2022, pp. 485–489.
13. T. Mu, H. Wang, C. Wang, Z. Liang, and X. Shao, "Auto-cash: A meta-learning embedding approach for autonomous classification algorithm selection," *Inf. Sci.*, vol. 591, pp. 344–364, 2022.
14. M. Munaf, H. Afzal, K. Mahmood, and N. Iltaf, "Low resource summarization using pre-trained language models," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 23, no. 10, 2024.
15. A. R. S. Parmezan, H. D. Lee, N. Spolaôr, and F. C. Wu, "Automatic recommendation of feature selection algorithms based on dataset characteristics," *Expert Syst. Appl.*, vol. 185, Art. no. 115589, 2021.
16. A. Rao, S. Aithal, and S. Singh, "Single-document abstractive text summarization: A systematic literature review," *ACM Comput. Surv.*, vol. 57, no. 3, 2024.
17. P. Ren et al., "A comprehensive survey of neural architecture search: Challenges and solutions," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–34, 2022.
18. J. R. Rice, "The algorithm selection problem," *Adv. Comput.*, vol. 15, pp. 65–118, 1976.
19. S. Saleem and S. Kumarapathirage, "AutoNLP: A framework for automated model selection in natural language processing," in *Proc. 18th Iberian Conf. Inf. Syst. Technol. (CISTI)*, 2023, pp. 1–4.
20. S. Salmani Pour Avval, N. D. Eskue, R. M. Groves, and V. Yaghoubi, "Systematic review on neural architecture search," *Artif. Intell. Rev.*, vol. 58, no. 3, Art. no. 73, 2025.

21. R. Vilalta and Y. Drissi, "A perspective view and survey of meta-learning," *Artif. Intell. Rev.*, vol. 18, no. 2, pp. 77–95, 2002.
22. M. T. Wu and C. W. Tsai, "Training-free neural architecture search: A review," *ICT Express*, vol. 10, no. 1, pp. 213–231, 2024.
23. J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "Pegasus: Pre-training with extracted gap sentences for abstractive summarization," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, 2020.