

Meta-Heuristic Approaches with LM-based Selectors for Extractive Summarization

Shih-Wei Yang**, Bo-Yu Chen*, Shao-Chi Kuan*, Hau-Ching Chen*, Sy-Yen Kuo**, Jiann-Liang Chen*

*Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan

**Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan

d10921008@ntu.edu.tw, M11407505@mail.ntust.edu.tw, M11407504@mail.ntust.edu.tw, M11207522@mail.ntust.edu.tw, sykuo@ntu.edu.tw, Lchen@mail.ntust.edu.tw

Abstract- This work demonstrates the cost-effectiveness and suitability of a meta-heuristic algorithm as a selector for the extractive summarization task compared to the LM-based selector. Meta-heuristic-based selector achieves 0.05 higher ROUGE-1 score at 0.320 and 0.054 higher ROUGE-2 score than LM-based baseline at 0.116 while using at least one order of magnitude less computation time under the same condition. This work also demonstrate that the characteristics of LM-based selectors and meta-heuristic methods are complementary, which can be combined to achieve even better results of 0.351 ROUGE-1, 0.141 ROUGE-2, and 0.218 ROUGE-L.

Keywords- Extractive Summarization, News Summarization, NSGA-II, CNN/DailyMail, BERT, RoBERTa, XLNet

I. INTRODUCTION

The current trajectory in text summarization trends toward ever larger pretrained language models (PLMs) and embedding-based scoring pipelines. This direction improves average accuracy but increases computational cost, latency, and energy consumption. As the field's "scale-to-win" impulse grows, many extractive summarizers allocate most of their budget to dense PLM forward passes, which are used only as similarity heuristics at selection time. This is not sustainable for many practical deployments, and relying solely on semantic features might lead to unsatisfactory results.

Proposed method first revisited extractive summarization from an efficiency-first perspective. Using interpretable signals: term frequency with inverse sentence frequency (TF-ISF), sentence position, and target length adherence, combined with formulating sentence subset selection as a multi-objective search and using NSGA-II to explore Pareto-optimal trade-offs between salience, redundancy, and length, the result can match or exceed PLM-similarity heuristics while consuming far fewer compute resources.

Efficiency need not preclude accuracy. While our meta-heuristic alone is competitive, the later part of this work further demonstrates that ensembling it with PLM-based scoring, by adding a lightweight similarity signal from BERT/RoBERTa/XLNet to the objectives or as an auxiliary term in MMR, provides consistent gains. The ensemble leverages the complementary strengths of symbolic/statistical features, as well as contextual semantics, without relying heavily on PLM usage as the sole driver.

Our tests on CNN/DailyMail use 3-sentence outputs and report ROUGE-1, ROUGE-2, and ROUGE-L scores. In this

setting, our NSGA-II pipelines achieve higher scores on ROUGE-1 and ROUGE-2 than PLM-only similarity baselines, while also running faster. ROUGE-L is comparable but can be slightly higher for PLM-only. When candidates are combined with PLM-based scores (BERT, RoBERTa, XLNet) and rerank with MMR, the ensemble gives the best overall ROUGE in our runs.

II. RELATED WORKS

Extractive summarization commonly selects sentences using a relevance score and a redundancy penalty. Maximal Marginal Relevance (MMR) formalizes this trade-off and remains a standard diversification heuristic in both document reranking and summarization [1]. Beyond classical TF-IDF/graph centrality, recent works pair statistical features with search or learning to optimize the relevance-redundancy frontier. For example, a 2025 hybrid system integrates graph entropy with large language models to improve accuracy and time efficiency on extractive summarization benchmarks [2]. These lines of work motivate combining simple, interpretable signals with principled selection objectives.

Pretrained language models are frequently used to compute sentence salience via embedding similarity. BERT [3], RoBERTa [4], and XLNet [5] supply strong contextual embeddings that correlate with human relevance judgments. In many extractive systems, an LM-based similarity between candidate sentences and the document (or query) serves as the primary scoring heuristic, with MMR or cosine-based diversity to reduce redundancy. While effective, these pipelines can be computationally intensive and may over-favour semantically smooth but verbose sentences, motivating alternatives that emphasize efficiency and controllable bias toward concision.

Meta-heuristic optimization provides another route. NSGA-II [6] performs fast non-dominated sorting with elitism to approximate Pareto-optimal sets under multiple objectives. In the summarization task, multi-objective formulations typically maximize relevance while minimizing redundancy and length deviation, letting the search expose trade-offs that a single scalar heuristic can obscure. Prior extractive systems have used genetic algorithms and NSGA-II variants to tune sentence subsets or weights for sentence-level features [7]. Our approach follows this line but uses TF-ISF (a document-internal TF-IDF variant), sentence length, and position as explicit objectives combined with MMR at selection

time. This keeps features cheap, interpretable, and dataset-robust while letting NSGA-II explore diverse candidate subsets before MMR refines the final ordering.

CNN/DailyMail is a standard benchmark for news summarization and appears in two common forms: the early entity-anonymized [8], cloze-style variant and the later non-anonymized summarization variant [9]. The corpus consists of news articles paired with multi-sentence highlights. It exhibits a strong lead bias-key information tends to appear in the first few sentences, which favours extractive methods that include position features. Typical setups fix the output to a small number of sentences to compare systems under equal length budgets. This same practice is followed in our experiments by reporting ROUGE on fixed 3-sentence outputs, which aligns with prior work on this dataset.

Evaluation in extractive summarization standardly reports ROUGE-1/2/L, which measures n-gram overlap and longest common subsequence with reference summaries [10]. ROUGE-1, ROUGE-2, and ROUGE-L are adopted to enable direct comparison with embedding-centric baselines and prior NSGA-II-based systems.

III. METHOD

In extractive summarization, final quality strongly depends on the sentence pool assembled up front. A two-stage pipeline is proposed that couples a meta-heuristic search with lightweight language-model scoring, then finalizes the extract with Maximum Marginal Relevance (MMR). The workflow has five steps:

1. **Preprocessing.** Clean the input document by normalizing all characters into lower-cases and splitting it into sentences to form the candidate set.
2. **Feature computation and similarity.** For each sentence, compute three features: TF-ISF weight, a length score, a position score and build an inter-sentence cosine-similarity matrix using TF-ISF vectors.
3. **Meta-heuristic candidate generation.** Use NSGA-II to search under three objectives: maximize coverage, minimize redundancy, then take the Pareto-front solutions and select the 10 most frequently chosen sentences as the **NSGA-II candidate set**.
4. **LM-based candidate generation.** Independently, obtain sentence embeddings with three pretrained LMs (BERT, RoBERTa, XLNet), compute a document centroid, rank sentences by cosine similarity to the centroid, and collect the top 10 as the **LM candidate set**.
5. **Final extract with MMR.** Take the union of both candidate sets and apply MMR to select a 3-sentence summary by balancing relevance to the centroid and novelty with respect to already selected sentences.

A. Sentence features and similarity matrix

1) TF-ISF weighting: Term Frequency-Inverse Sentence Frequency is used to emphasize terms that are frequent in a few

sentences but down-weight corpus-level stop terms. The calculation of TF-ISF can be seen formulated in (1)(2)(3).

$$TF(s, t) = \frac{f_t}{|s|} \quad (1)$$

$$ISF(s, t) = \log \frac{S}{s_t} \quad (2)$$

$$TF - ISF(s, t) = TF(s, t) \times ISF(s, t) \quad (3)$$

Where t denotes a term, s denotes a sentence, f_t denotes the frequency of the term t , $|s|$ denotes the length of sentences s in quantity of terms, S denotes the total quantity of sentences in target article and s_t denotes the total quantity of sentences in target article containing at least one instance of t .

2) Length score: Normalize each sentence length against a target length to avoid favouring over-short or over-long sentences as formulated in (4).

$$Len(s) = \min\left(\frac{|s|}{L_{ref}}, 1\right) \quad (4)$$

Where L_{ref} denotes reference length of sentences, which is empirically determined to be 40.

3) Position score: In news and similar domains, early sentences often carry essential topic and crucial background context, so a normalized position function, formulated as (5), is proposed to capture such feature:

$$Pos(s) = 1 - \frac{pos(s) - 1}{S - 1} \quad (5)$$

Where $pos(s)$ represent the 0-indexed position of sentence s in the article.

4) Similarity matrix: Represent each sentence s_i with its TF-ISF vector v_i and compute pairwise cosine similarity to estimate redundancy, as formulated in (6).

$$Sim(v_i, v_j) = \frac{v_i \cdot v_j}{|v_i| |v_j|} \quad (6)$$

B. NSGA-II sentence selection

By defining 3 object functions, selection process is casted as a multi-objective optimization, and utilizing NSGA-II algorithm to search for the best extracting solution in Pareto Front:

1) Decision Vector: Each candidate extract is represented as a binary vector X of length S as formulated in (7):

$$X \in \{0,1\}^S \quad (7)$$

Where if the i -th element in X is equal to 1, denoted as $x_i = 1$, it represents that the i -th sentence is included into the candidate set.

2) Object Function: By minimizing three conflicting objectives: negative importance $f_1(X)$ as formulated in (8), negative coverage $f_2(X)$ as formulated in (9), and intra-summary redundancy $f_3(X)$ as formulated in (10), where l_i denotes the saliency score of the i -th sentence, S_{ji} denotes the similarity between the i -th sentence and the j -th sentence:

$$f_1(X) = - \sum_{i=1}^S X_i \times l_i \quad (8)$$

$$f_2(X) = - \frac{1}{S} \sum_{j=1}^S \max_{\{i|x_i=1\}} S_{ji} \quad (9)$$

$$f_3(X) = \frac{\sum_{i<j, \{i,j|x_i=x_j=1\}} S_{ij}}{(\sum x_i)(\sum x_i - 1)/2} \quad (10)$$

NSGA-II performs fast non-dominated sorting with crowding-distance diversity and elitism.

After a fixed number of generations, a set of non-dominated solutions on the Pareto front were obtained to form the 10-sentences NSGA-II set $C_{NSGA-II}$.

C. LM scoring track

To compare with, and later complement with the meta-heuristic track, sentence embeddings are computed using pretrained LMs (BERT, RoBERTa, XLNet), the process is as such:

1) Generate Embedding: First obtain the embedding vector V_i of the i -th sentence in the article by performing inference on pretrained language model, such as BERT, RoBERTa and XLNet.

2) Calculate article centroid: Then derive the article context vector V_c by calculating a centroid over all V_i as formulated in (11)

$$V_c = \frac{1}{S} \sum_{i=1}^S V_i \quad (11)$$

3) Ranking by similarity and selection: Finally, rank each sentence vectors V_i by the similarity score obtained via calculating $Sim(V_i, V_c)$ and include top 10 scored sentences into candidate set C_{LLM}

D. Final MMR selection

Let the final candidate pool be the union of the two sets as formulated in (12):

$$C_{final} = C_{NSGA-II} \cup C_{LLM} \quad (12)$$

MMR, as formulated in (13), is then applied iteratively to pick 3 sentences, trading off sentence-centroid relevance against redundancy with already selected sentences via a balance parameter λ .

$$\arg \max_{\{V_i|V_i \in C_{final} \wedge V_i \notin S_{summary}\}} [\lambda \cdot Sim(V_i, V_c) - (1 - \lambda) \cdot \max_{V_j \in S_{summary}} Sim(V_i, V_j)] \quad (13)$$

Where $S_{summary}$ denotes the set of currently selected final candidates, with $\lambda \rightarrow 1$ reduces to relevance-only ranking and $\lambda \rightarrow 0$ enforces novelty.

IV. PERFORMANCE EVALUATION

The CNN/DailyMail corpus is used as evaluation dataset, and ROUGE-1, ROUGE-2, and ROUGE-L F1 scores are reported as a base of comparison. To ensure comparability, every system have exact same preprocess and outputs exactly 3 sentences.

A. Meta-heuristics vs LM-only baselines

Three selection strategies: NSGA-II, GRASP, and Greedy are compared against three LM-only baselines that rank by embedding similarity from BERT, RoBERTa, and XLNet. All systems share the same features (TF-ISF, length, position) where applicable.

TABLE 1. THREE SENTENCES ROUGE SCORE COMPARISON BETWEEN LM-BASED BASELINES AND META-HEURISTIC ALGORITHMS

Metric	Bert	RoBERTa	XLNet	NSGA2	GRASP	Greedy
ROUGE-1	0.238	0.249	0.268	0.320	0.317	0.317
ROUGE-2	0.053	0.062	0.062	0.116	0.113	0.113
ROUGE-L	0.206	0.216	0.219	0.192	0.189	0.190

TABLE 2. PER-ARTICLE TIME BREAKDOWN (IN MICROSECONDS)

Phase	Greedy	GRASP	NSGA2
preprocess	0.45 ms	0.45 ms	0.45 ms
select	2.29 ms	6.49 ms	9.20 ms
evaluate	2.78 ms	2.77 ms	2.73 ms

TABLE 3. AVERAGE RUNTIME PER ARTICLE (IN MICROSECONDS)

	Bert	RoBERTa	XLNet	NSGA2	GRASP	Greedy
Average time per article	950.6 ms	634.7 ms	382.0 ms	132.4 ms	9.7 ms	5.5 ms

As shown in TABLE 1, NSGA-II achieves the best performance on the 3-sentence task, with ROUGE-1/2/L of 0.3209/0.1162/0.1928, clearly higher than LM-only baselines (best 0.2386/0.0530/0.2069). Notably, even the simple Greedy algorithm reaches 0.3174 on ROUGE-1, surpassing all LM-only baselines, which further indicates the suitability of meta-heuristic methods for extractive summarization.

For time efficiency shown in TABLE 2 and TABLE 3, Greedy and GRASP require only 5.5 ms/article and 9.7 ms/article, respectively. NSGA-II needs 132.4 ms/article, which is still low compared with LM-only methods and acceptable in practice. Because LM-only inference averages 382–951 ms/article, the gap ranges from tens to over a hundred times, showing that meta-heuristics can maintain summary quality while substantially increasing throughput.

Looking at metric differences, ROUGE-2 shows the largest gains, suggesting that multi-objective algorithms handle coverage and redundancy more effectively and thus capture bigram content better. By contrast, LM-only baselines retain a slight advantage on ROUGE-L, likely because embedding-based similarity captures longer-range dependencies and semantic structure, which helps on longest-common-subsequence overlap. This highlights the limits of relying only on semantic features and the value of adding multi-objective optimization.

Overall, under the same feature conditions, meta-heuristic algorithms outperform LM-only baselines. This supports that multi-objective search-based methods can improve quality while keeping cost low, which is the main advantage of the proposed approach.

B. Two-stage pipeline (algorithmic candidates + LM assistance)

This experiment applies a two-stage pipeline for 3-sentence summaries. Stage 1 uses the same features as the former experiment and generates candidate sets via NSGA-II and via each LM (BERT, RoBERTa, XLNet). Stage 2 takes the union of these candidate sets, performs representation-based semantic re-ranking, and then produces the final summary:

TABLE 4 THREE SENTENCES ROUGE SCORE COMPARISON BETWEEN 3 DIFFERENT ENSEMBLED RESULTS.

Metric	ROUGE-1	ROUGE-2	ROUGE-L
NSGA2 + BERT	0.351331	0.141175	0.218475
NSGA2 + RoBERTa	0.350112	0.139600	0.217257
NSGA2 + XLNet	0.349913	0.139821	0.217609

As shown in TABLE 4, the mixed approach outperforms either component alone. NSGA-II + BERT gives the best results at 0.3513/0.1412/0.2185 for ROUGE-1/2/L. The combinations with RoBERTa and XLNet show similar levels and are better than standalone NSGA-II or any LM-only baseline. These

findings suggest that, although LM-only methods did not surpass meta-heuristics in the former experiment, within the two-stage framework their semantic coverage and scoring signals make a clear contribution to final quality.

In summary, the “algorithmic candidates + LM assistance” design has a clear advantage: Stage 1’s multi-objective search secures coverage and diversity, and Stage 2’s LM signals add semantic strength. The final summaries improve in content completeness and semantic consistency, showing the complementarity between meta-heuristics and LMs.

C. Cross-Domain Generalization on SciTLDR-AIC

TABLE 5. ROUGE SCORES ON SCITLDR-AIC (ONE-SENTENCE SUMMARIES)

Model Variant	ROUGE-1
NSGA-2 + BERT	0.239148
NSGA-2 + RoBERTa	0.236803
NSGA-2 + XLNet	0.233836

TABLE 6. AVERAGE TIME PER ARTICLE ON SCITLDR-AIC (IN MS)

Model Variant	Stage1-Base	Stage1-LM	Stage2	Total Eval Time
NSGA-2 + BERT	194.36 ms	323.71 ms	4.02 ms	522.09 ms
NSGA-2 + RoBERTa	195.77ms	421.72 ms	4.98 ms	622.47 ms
NSGA-2 + XLNet	229.34 ms	770.49 ms	3.86 ms	1003.69 ms

To evaluate the cross-domain robustness of the proposed approach, we further tested the model on the SciTLDR-AIC dataset, which contains 618 scientific abstracts paired with single-sentence author-written highlights. Unlike CNN/DailyMail, SciTLDR-AIC represents an extremely compressed summarization task with long, technical sentences and inherently low lexical overlap. Moreover, the dataset provides only ROUGE-1 as the official evaluation metric; no ROUGE-2 or ROUGE-L references are available. Therefore, our evaluation focuses exclusively on ROUGE-1. To establish a meaningful baseline for interpreting performance, we analyzed the ROUGE-1 distribution across the entire validation set. For each instance, we computed the ROUGE-1 score between the human-written highlight and the oracle extraction consisting of all source sentences, following the same preprocessing used in our system. The resulting distribution shows a mean of 0.1357, a median of 0.1212, an interquartile range from 0.0625 (Q1) to 0.1892 (Q3), and values spanning from 0.0 to 0.999999995. This indicates that ROUGE-1 scores above 0.20 already reflect strong performance under the extreme compression of this task. As shown in Table 5, the proposed two-stage framework achieves ROUGE-1 scores of 0.2338–0.2391 across BERT, RoBERTa, and XLNet variants,

placing the method well above typical quality levels within the dataset. Inference efficiency is also preserved, with total evaluation times of approximately 5 ms per article (Table 6). These results demonstrate that the meta-heuristic and LM-assisted framework generalizes effectively to scientific summarization despite the limited evaluation metric and challenging task characteristics.

V. CONCLUSION

This work first demonstrated that under identical features, NSGA-II and other meta-heuristic algorithms outperform LM-only baselines on ROUGE-1/2. At the same time, meta-heuristics achieve such a strong ROUGE score with far lower inference cost. Even the costliest NSGA-II at 0.1324 s/article is still much faster than LM-only pipelines. Later, our proposed method of combining NSGA-II with PLM scoring is shown to further enhances ROUGE-2/L scores, highlighting their complementary strengths: meta-heuristics ensure coverage and diversity, while LMs reinforce semantic coherence. Our work provided robust evidence of the effectiveness of meta-heuristic algorithms as suitable methods for the extractive summary task. With its ensemble version incorporating LM-based methods, a more effective, higher-quality, and semantically precise solution provides an empirical foundation for further applications and system design.

REFERENCES

- [1] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," *SIGIR*, pp. 335-336, 1998.
- [2] T. Uçkan, "A hybrid model for extractive summarization: Leveraging graph entropy with LLMs," *Engineering Reports*, 2025.
- [3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019.
- [4] Z. Liu, W. Lin, Y. Shi, and J. Zhao, "A Robustly Optimized BERT Pre-training Approach with Post-training," in *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, 2021.
- [5] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le., "XLNet: generalized autoregressive pretraining for language understanding," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.
- [6] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 2, no. 6, pp. 182-197, april 2002.
- [7] A. R. P. Manju, D. Gupta, "Extractive Single Document Summarization Using NSGA-II," *Advances in Intelligent Systems and Computing*, no. 1148.
- [8] K. M. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Proceedings of the 29th International Conference on Neural Information Processing Systems*, 2015.
- [9] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gülçehre, and B. Xiang, "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond," in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, 2016.
- [10] C. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," *Text Summarization Branches Out*, pp. 74-81, 2004.
- [11] M. Mojriani, S. A. Mirroshandel, "A novel extractive multi-document text summarization system using quantum-inspired genetic algorithm: MTSQIGA," *Expert Systems with Applications*, vol. 171, 2021.



Shih-Wei Yang was born in Taiwan in 1992. He is currently working at the National Institutes of Applied Research (NIAR) and pursuing the Ph.D. degree in Electrical Engineering at National Taiwan University (NTU). His research interests include cybersecurity and AI-enabled computing systems.



Bo-Yu Chen was born in Taiwan in 2003. He received the B.S. degree in Computer Science from Tunghai University (THU), Taichung, in 2025. He is currently pursuing the M.S. degree in Electrical Engineering at the National Taiwan University of Science and Technology (NTUST), Taipei. His research interests include machine learning and large language models.



Shao-Chi Kuan was born in Taiwan in 2003. He received the B.S. degree in Computer Science from Tunghai University (THU), Taichung, in 2025. He is currently pursuing the M.S. degree in Electrical Engineering at the National Taiwan University of Science and Technology (NTUST), Taipei. His research interests include machine learning and large language models.



Hau-Ching Chen was born in Taiwan in 1999. He received a B.S. degree in Electronic Engineering in 2023. He is currently pursuing an M.S. degree in Electrical Engineering at the National Taiwan University of Science and Technology (NTUST), Taipei. His main research interests include Deep Learning and Computer Vision.

Sy-Yen Kuo (Fellow, IEEE) received the B.S. degree from National Taiwan University (NTU), Taipei, Taiwan, in 1979, the M.S. degree from the University of California at Santa Barbara, in 1982, and the Ph.D. degree in computer science from the University of Illinois at Urbana-Champaign, in 1987. From 1982 to 1984, he was an Engineer with Fairchild Semiconductor, California, and Silvar-Lisco, California. From 1988 to 1991, he was a Faculty Member with the Department of Electrical and Computer Engineering, University of Arizona. From 2001 to 2004, he was the Chairman of the Department of Electrical Engineering, NTU. From 2012 to 2015, he was the Dean of College of Electrical Engineering and Computer Science, NTU. He is currently the Pegatron Chair Professor with the Department of Electrical Engineering, NTU. He has published more than 450 articles in journals and conferences. He holds 22 U.S. patents, 23 Taiwan patents, and 15 patents from other countries. His current research interests include dependable and secure systems, edge and cloud computing, the Internet of Things, and quantum computing. He was a member of the IEEE Fellow Committee, from 2018 to 2020. He was a Core Member and a member of the Board of Governors of the IEEE Computer Society, from 2017 to 2020.



Jiann-Liang Chen (Senior Member, IEEE) was born in Taiwan on December 15, 1963. He received a Ph.D. in Electrical Engineering from the National Taiwan University, Taipei, Taiwan, in 1989. Since August 1997, he has been with the Department of Computer Science and Information Engineering of National Dong Hwa University, where he is a professor and Vice Dean of Science and Engineering College. Prof. Chen joins the Department of

Electrical Engineering, National Taiwan University of Science and Technology, as a Distinguished Professor and Dean. His research interests are cellular mobility management, cybersecurity, personal communication systems, and the Internet of Things (IoT)