Ensemble Strategies for Turkish Legal Extractive Summarization Using Semantic Voting and ROUGE-Weighted Averaging

Maha ALBAYATI[1][0000−0002−4259−9168] and Oğuz FINDIK[2][0000−0001−5069−6470]

[1] Karabuk University, Karabuk 78000, Turkey
[2] Karabuk University, Karabuk 78000, Turkey
mahaa6022@gmail.com
oguzfindik@karabuk.edu.tr
https://www.karabuk.edu.tr/en

Abstract. The growing volume of Turkish legal documents poses an important challenge for legal professionals who must efficiently review complex, domain-specific texts. Although extractive summarization has shown promise across many languages, Turkish remains difficult due to its agglutinative morphology and the rigid structure of legal discourse. Existing approaches mainly rely on single methods and often fail to capture both semantic relevance and contextual structure. This study examines whether ensemble strategies can improve the extractive summarization of Turkish legal documents beyond traditional baseline methods. TF-IDF, TextRank, and a BERTurk-Legal-based extractive ranker serve as baseline methods, whereas Semantic Voting and ROUGE-guided weighting are evaluated as ensemble extensions built on these baselines. A curated dataset of 100 Turkish court decisions was compiled and annotated with manually prepared extractive reference summaries. Performance was assessed with ROUGE metrics. Semantic Voting obtained ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.4681, 0.3632, and 0.3968, respectively. ROUGE-Weighted Averaging obtained the highest scores, reaching 0.6423, 0.6377, and 0.6423 on ROUGE-1, ROUGE-2, and ROUGE-L. Overall, the findings suggest that the ensemble configuration, particularly ROUGE-Weighted Averaging, improves summary quality relative to the traditional baselines in this experimental setting.

Keywords: Ensemble methods · Extractive summarization · ROUGE evaluation · Semantic similarity · Turkish legal summarization

1 Introduction

Natural Language Processing (NLP) is a branch of artificial intelligence that enables machines to process and analyze human language. It is used in many tasks, including machine translation, sentiment analysis, information retrieval, question answering, and summarization. Text summarization condenses a document while

preserving its essential information [1]. Approaches are generally divided into extractive summarization, which selects original sentences [3, 13], and abstractive summarization, which paraphrases ideas in new wording [2, 7]. Because extractive methods preserve the original legal wording, they are particularly suitable for domains that require fidelity to the source language, such as Turkish court decisions. Classic extractive approaches rely mainly on statistical and graph-based evidence. The canonical traditional baselines are TF-IDF, which ranks sentences by the importance of their constituent terms, and TextRank, which propagates importance through a sentence-similarity graph [16]. To strengthen the comparison setting, this study also includes a modern neural extractive baseline based on the legal-domain transformer checkpoint BERTurk-Legal [17]. Against this background, the paper examines ensemble-based extensions of traditional methods. TF-IDF, TextRank, and BERTurk-Legal are treated as the baseline systems, while Semantic Voting and ROUGE-Weighted Averaging are evaluated as two ensemble strategies that combine or re-rank their outputs. On a corpus of 100 judgments, Semantic Voting reaches ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.4681, 0.3632, and 0.3968, whereas ROUGE-Weighted Averaging attains 0.6423, 0.6377, and 0.6423. These findings suggest that ensemble-based score fusion can improve on the underlying baselines in this experimental setting. The contribution of this study is primarily domain-oriented and empirical, focusing on controlled comparisons and evaluations of Turkish legal extractive summarization rather than on proposing a fundamentally new summarization architecture. Future work may extend the dataset beyond the present 100-case corpus and examine broader cross-validation settings or additional fusion strategies.

2 Problem Statement

Turkish summarization research has evolved along two broad tracks: general-domain work, mainly based on news corpora, and a smaller body of studies in specialized domains such as academic writing, biomedicine, social media, and, more recently, law. This development helps explain why Turkish legal summarization remains underexplored [11, 12]. Early work on Turkish summarization relied largely on surface-level statistics. TF-IDF weighting, Latent Semantic Analysis, and PageRank-style methods were applied to Turkish texts and news-style documents [4, 3]. As larger Turkish summarization resources became available, research gradually moved from lexical heuristics toward neural and sequence-to-sequence approaches [5, 7]. VBART, a Turkish sequence-to-sequence language model trained from scratch on a large Turkish corpus, reported competitive results for Turkish generation and summarization tasks [6]. Research outside mainstream news has also explored several specialized settings. In academic writing, researchers have examined deep learning approaches for summarizing Turkish scientific papers [8]. Biomedical work has investigated extractive summarization for Turkish medical papers [9]. Social-media studies remain exploratory, with recent work examining summary informativeness in multilingual

tweet collections under noisy conditions [10]. However, these studies focus on domains with less rigid structure and less formal terminology than legal texts. Legal-domain studies remain limited. A hybrid NLP-based study on Turkish Constitutional Court decisions reported only modest gains over a TF-IDF baseline [11]. A later transformer-based pipeline generated multi-document civil-case summaries, but it relied on a single model and did not examine score-level ensemble fusion [12]. Published work, therefore, provides only limited evidence on how complementary extractive signals and sentence-level semantics can be combined and evaluated for Turkish legal summarization. Early ensemble work combined centroid-based extractors in English news domains [13], and later studies explored score aggregation across multiple summarizers [14]. More recent systems based on pretrained contextual encoders further improved the quality of extractive ranking [15]. Against this background, we examine whether ensemble strategies based on score fusion and semantic re-ranking can provide measurable improvements over TF-IDF and TextRank for Turkish court decisions.

3 Methodology

This section describes the methodology for improving extractive summaries of Turkish legal documents using ensemble strategies. The experimental framework consists of three baseline models, TF-IDF-based, TextRank-based, and BERTurk-Legal-based extractive summarization, together with two ensemble strategies, Semantic Voting and ROUGE-Weighted Averaging.

3.1 TF-IDF-Based Extractive Summarization

In the TF-IDF approach, the document is first preprocessed using the Zemberek NLP toolkit for Turkish. Preprocessing includes tokenization, lowercasing, stopword removal, and cleaning numerical and date-related information. The same preprocessed documents are used consistently throughout the experiments. For each sentence, we compute a score by summing the TF-IDF values of its words, where the TF-IDF score of a word w is calculated as:

$$\text{TF-IDF}(w) = \text{TF}(w) \times \log \frac{N}{\text{DF}(w)} \tag{1}$$

In this equation, $\text{TF}(w)$ is the term frequency of word w, representing how often it appears in the current document. $\text{DF}(w)$ is the document frequency, indicating how many documents in the corpus contain the word w, and N is the total number of documents. After computing the TF-IDF scores for all words, each sentence is assigned a cumulative score based on the sum of the TF-IDF values of its constituent tokens. The top-3 scoring sentences are then selected as the extractive summary. These summaries are evaluated under the shared ROUGE-based metrics and the common 5-fold cross-validation protocol described in the Experimental Setup.

3.2 TextRank-Based Extractive Summarization

TextRank generates a graph-based representation of the document using the same preprocessed sentence sequence, where nodes correspond to sentences and edges are weighted using cosine similarity between sentence TF-IDF vectors. The importance of each sentence S_i is calculated iteratively using the formula:

$$\text{Score}(S_i) = (1 - d) + d \sum_{S_j \in \text{In}(S_i)} \frac{\text{Score}(S_j)}{|\text{Out}(S_j)|} \tag{2}$$

In this formula, d is the damping factor, typically set to 0.85. The term $\text{In}(S_i)$ refers to the set of sentences that have an edge pointing to S_i (i.e., incoming links), and $\text{Out}(S_j)$ refers to the sentences that S_j links to (i.e., outgoing links). As illustrated in Figure 1, sentence S_1 links to both S_2 and S_3, so $S_1 \in \text{In}(S_2)$ and $S_1 \in \text{In}(S_3)$. In this case, $\text{Out}(S_1) = \{S_2, S_3\}$. The figure clarifies how incoming and outgoing links influence the flow of importance scores across the graph.

After convergence, the top-3 ranked sentences are selected and reordered to match their original positions in the text for improved coherence. These summaries are evaluated using the same ROUGE-based metrics and 5-fold cross-validation protocol as for the other methods.

In the sentence-similarity graph used by TextRank, the edge weight $\text{sim}(S_i, S_j)$ is symmetric, so every undirected edge appears in both directions. Thus, $\text{In}(S_i)$, the sentences that point to S_i, and $\text{Out}(S_i)$, those that S_i points to, contain the same neighbors. We keep the PageRank notation because the sum is taken over incoming neighbors, whereas the denominator $|\text{Out}(S_j)|$ normalizes the vote of each contributing node S_j by its identical out-degree.

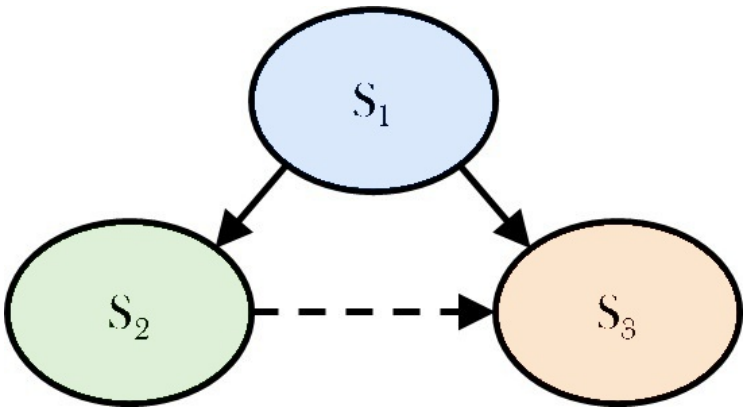

Fig. 1. TextRank sentence scoring formula with inbound and outbound links.

3.3 BERTurk-Legal-Based Extractive Summarization

To provide a stronger modern neural baseline for Turkish legal summarization, we also evaluated an extractive ranking method built on the domain-adapted transformer checkpoint `KocLab-Bilkent/BERTurk-Legal` [17]. The same preprocessed documents and sentence boundaries were used as in the other methods. For each document, sentence embeddings and a document embedding were obtained from the transformer encoder through mean pooling over the final hidden states. Sentence relevance was then estimated by cosine similarity between each sentence embedding and the corresponding document embedding. The top-3 highest-scoring sentences were selected and reordered according to their original document positions to preserve coherence. BERTurk-Legal was evaluated only as an independent baseline and was not used as an input source in the Semantic Voting or ROUGE-Weighted Averaging ensembles. This baseline keeps the extractive setting explicit while testing whether a legal-domain-pretrained encoder offers an advantage over traditional lexical and graph-based baselines. The resulting summaries were evaluated using the same ROUGE metrics and 5-fold cross-validation protocol used throughout the study.

3.4 Semantic Voting Ensemble

The semantic voting strategy merges and re-ranks sentences based on their semantic closeness to the original document. It operates on the same preprocessed documents used by the baseline methods. First, candidate sentences are collected from TF-IDF and TextRank summaries, and duplicates are removed. The full document and each candidate sentence are encoded using the SentenceTransformer model `paraphrase-multilingual-MiniLM-L12-v2`. Then, cosine similarity is computed between each document embedding and the corresponding sentence embedding. Sentences are ranked based on these similarity scores, and the top-3 most semantically relevant ones are selected and ordered according to their original position in the document. This ensemble construction step keeps sentence selection explicit while producing summaries that are evaluated under the same ROUGE metrics and 5-fold cross-validation protocol.

3.5 ROUGE-Weighted Averaging Ensemble

In the weighted averaging strategy, the contributions of TF-IDF and TextRank models are dynamically balanced based on validation-set ROUGE scores rather than test-set results. The method uses the same preprocessed documents and preserves explicit sentence selection through score fusion. To avoid evaluation leakage, the data used for weight tuning are explicitly separated from the data used for final reporting. Within each outer cross-validation run, TF-IDF and TextRank are first executed on an internal validation subset, and their average ROUGE-1, ROUGE-2, and ROUGE-L F1-scores on that subset are computed. These validation scores are then normalized to obtain fixed weights w_1 and w_2 such that:

$$w_1 + w_2 = 1 \tag{3}$$

For each document, candidate sentences are defined as the union of the top-ranked sentences returned by the TF-IDF and TextRank base summarizers. Each candidate sentence then keeps its original sentence-level TF-IDF score and its original sentence-level TextRank score. The TextRank component is therefore taken directly from the graph-based ranking scores described in Eq. (2), not from sentence length or any other proxy. Because the two scoring functions operate on different numeric scales, both score vectors are min-max normalized at the document level before fusion. After the weights are fixed on the internal validation subset, the same weights are transferred unchanged to the held-out test subset of that fold. A final score for each candidate sentence is calculated as:

$$\text{FinalScore}(S_i) = w_1 \times \widehat{\text{TF-IDF}}(S_i) + w_2 \times \widehat{\text{TextRank}}(S_i) \tag{4}$$

The candidate sentences are then ranked by their final scores, and the top-3 highest-scoring sentences are selected as the final extractive summary. The selected sentences are reordered according to their original document positions to preserve coherence. This formulation keeps the weighted ensemble faithful to the actual TF-IDF and TextRank sentence rankings while preserving a leakage-free protocol, since the ROUGE-derived weights are tuned only on the internal validation subset of each fold and never updated with information from the held-out test documents. The resulting summaries are evaluated using the same ROUGE metrics and 5-fold cross-validation protocol used throughout the study.

4 Experimental Setup

4.1 Dataset

We constructed a custom dataset of 100 Turkish legal documents, each representing an actual court decision or judgment. The documents were manually cleaned to remove metadata, formatting inconsistencies, dates, and case numbers. The cleaned text consists primarily of the main legal reasoning and verdict sections, making it suitable for extractive summarization. Each document was paired with a manually prepared extractive reference summary composed of three source sentences. The annotation procedure followed a fixed guideline: one sentence to represent the court decision, one to capture the main legal issue or reasoning, and one to state the final outcome. After the initial sentence selection, each reference summary was checked for fidelity to the source wording, redundancy, and coverage of the core legal content before being included in the benchmark. Formal inter-annotator agreement was not measured in the current study. However, the reference summaries were prepared under a fixed annotation guideline and manually verified for fidelity, redundancy, and role coverage. Future work should incorporate multi-annotator validation to further strengthen annotation reliability.

The corpus size was kept moderate for practical reasons. Each document required manual cleaning, document-level review, and manual preparation of an extractive reference summary, and these steps were applied to legally sensitive

texts. Accordingly, the dataset should be regarded as a curated benchmark for controlled methodological comparison rather than as a large-scale corpus intended for broad statistical coverage.

4.2 Experimental Protocol

To ensure a leakage-free evaluation, we adopted document-level 5-fold cross-validation. In each outer run, one fold (20% of the corpus) was reserved as a held-out test fold, while the remaining four folds (80%) formed the development portion for that run. The development portion was then split internally into a development subset used for pipeline preparation and an internal validation subset. Because TF-IDF, TextRank, and the BERTurk-Legal baseline are unsupervised extractive methods, the development subset was used only to execute the base pipelines under the current fold configuration, whereas the validation subset was used exclusively to tune the ROUGE-based weights of the weighted ensemble. Semantic Voting was applied within the same outer folds without additional parameter tuning beyond the fixed embedding model. Once these weights were fixed for a given run, no further tuning was performed. The learned weights were then transferred unchanged to the held-out test fold, and the process was repeated until every fold had served as the outer test fold once. All five methods generated three-sentence summaries from the same preprocessed documents and were evaluated with the same ROUGE metrics on the held-out folds. Final ROUGE values reported in this paper correspond to the mean scores across the five held-out test folds. This nested protocol clearly separates development, validation, and final evaluation and prevents evaluation leakage.

4.3 Preprocessing

The preprocessing pipeline was implemented using the Zemberek NLP library, which supports Turkish morphology and sentence extraction. For each document, tokenization was applied using Zemberek's `TurkishTokenizer`, and sentence boundaries were detected using the `TurkishSentenceExtractor`. The text was further normalized by converting all characters to lowercase and removing numeric expressions, legal article numbers, and common functional patterns that act as legal-domain stopwords. The resulting standardized text served as input to the TF-IDF and TextRank summarization pipelines.

4.4 Summarization Pipeline

Each sentence's importance in the TF-IDF summarization method was calculated as the sum of the TF-IDF scores of its tokens, considering only terms present in the current document. The top-scoring sentences were then selected to form the extractive summary. The TextRank summarization method constructed a similarity graph in which each node represented a sentence, and edges were weighted by cosine similarity between the TF-IDF vectors. Sentence importance

was computed using the TextRank algorithm with a damping factor of 0.85, and the highest-ranking sentences were selected to build the summary. The BERTurk-Legal baseline encoded both full documents and their candidate sentences with a legal-domain transformer model, ranked sentences by document–sentence cosine similarity, and selected the top-3 sentences in their original document order.

Two ensemble strategies were applied to refine the TF-IDF and TextRank-based summaries. The semantic voting ensemble re-ranked candidate sentences, collected from both TF-IDF and TextRank outputs, based on their semantic similarity to the entire document using multilingual sentence embeddings. The ROUGE-weighted averaging ensemble computed dynamic weights using only the internal validation subset of each outer fold and then applied these fixed weights during testing so that held-out test documents did not contribute to weight estimation. In other words, BERTurk-Legal remained a standalone baseline for comparison, whereas the two ensemble models operated only on TF-IDF and TextRank candidate sentences and sentence-level scores. Across all five models, preprocessing, sentence selection, ensemble construction where applicable, evaluation metrics, and the 5-fold cross-validation protocol were kept explicit and consistent. Figure 2 summarizes the overall experimental pipeline, while Figure 3 shows the model-specific summarization workflow for TF-IDF, TextRank, BERTurk-Legal, Semantic Voting, and ROUGE-Weighted Averaging.

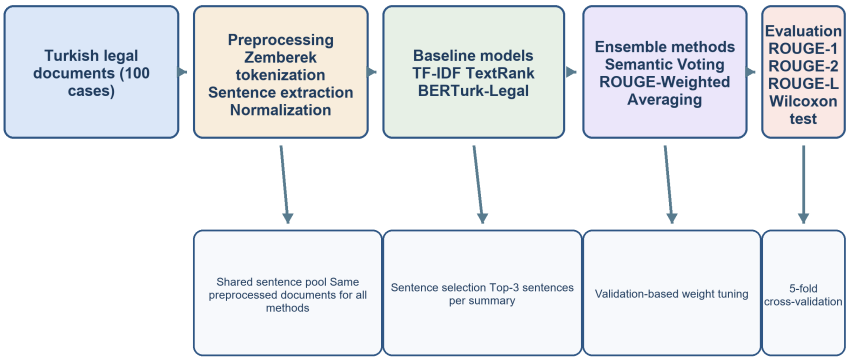

Fig. 2. Overall experimental pipeline. The figure shows the shared preprocessing stage, baseline generation with TF-IDF, TextRank, and BERTurk-Legal, ensemble construction, evaluation metrics, and the common 5-fold cross-validation protocol. BERTurk-Legal is evaluated only as a standalone baseline, whereas the ensemble models operate on TF-IDF and TextRank outputs.

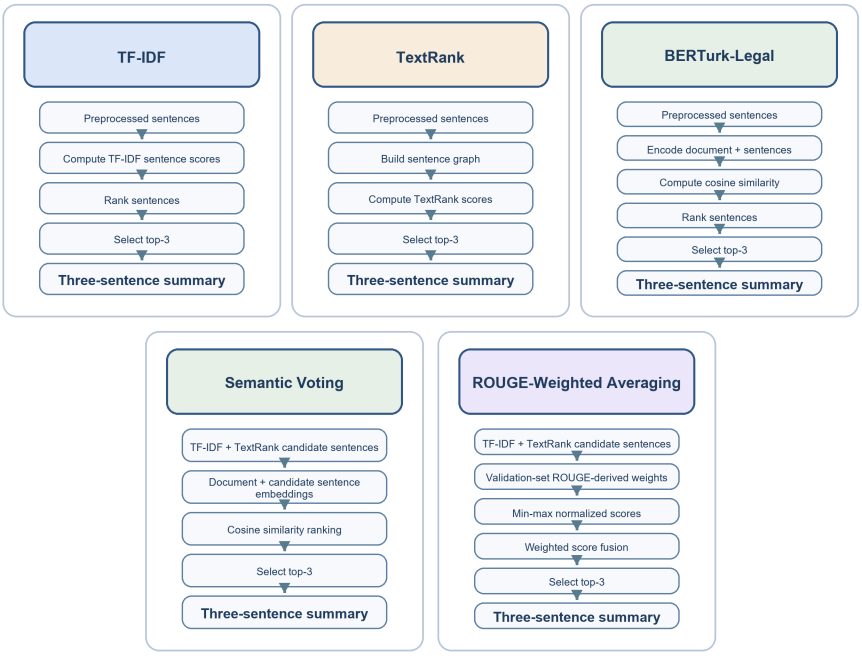

Fig. 3. Model-specific summarization workflows. TF-IDF, TextRank, and BERTurk-Legal produce baseline summaries directly, whereas Semantic Voting re-ranks candidate sentences through document–sentence cosine similarity, and ROUGE-Weighted Averaging combines validation-derived weights with min-max normalized baseline scores.

4.5 Evaluation Metrics

All generated summaries were evaluated against reference summaries using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric [18]. The evaluation focused on three key variants: ROUGE-1, which measures the overlap of unigrams (individual words); ROUGE-2, which captures the overlap of bigrams (two-word sequences); and ROUGE-L, which identifies the longest common subsequence between generated and reference summaries. For each metric, F1 scores were reported. In addition to the individual scores, we also computed a ROUGE sum metric, defined as the cumulative sum of ROUGE-1, ROUGE-2, and ROUGE-L F1 scores, to provide a consolidated measure of overall summarization performance. The reported averages were computed on the five held-out test folds only, whereas the ROUGE-weighted ensemble parameters were estimated separately from the internal validation subset inside each fold. To complement the mean scores, we also applied paired Wilcoxon signed-rank testing to the document-level ROUGE outputs collected from the held-out test folds to assess whether the observed differences between the ensemble models and the strongest baseline were statistically reliable. Statistical significance was assessed at the $p < 0.05$ level.

4.6 Reproducibility Statement

The experimental pipeline is described in sufficient detail to support methodological transparency, including preprocessing, sentence selection, ensemble construction, evaluation metrics, and the 5-fold cross-validation protocol. The implementation details reported in the manuscript are sufficient to support methodological replication of the experimental design and its evaluation procedure. The original dataset is not publicly released due to legal and institutional constraints governing the source documents.

Table 1. Mean ROUGE scores for summarization models across 5-fold cross-validation.

Approach name	**Model name**	**R_1**	**R_2**	**R_L**	**R_SUM**
Baseline models	TF-IDF	0.4879	0.3940	0.4303	1.3122
	TextRank	0.4761	0.3817	0.4227	1.2805
	BERTurk-Legal	0.4410	0.3217	0.3796	1.1424
Ensemble models	Semantic Voting	0.4681	0.3632	0.3968	1.2281
	ROUGE-Weighted Averaging	**0.6423**	**0.6377**	**0.6423**	**1.9223**

5 Results and Discussion

We evaluated three base models (TF-IDF, TextRank, and BERTurk-Legal) together with two ensemble approaches (Semantic Voting and ROUGE-Weighted

Averaging) using ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), and the cumulative score (RSUM). The detailed results are reported in Table 1. The table presents mean scores over the five held-out test folds, while the ROUGE-based weights used by the weighted ensemble were tuned only on the internal validation subset of each outer fold. In addition to the mean scores, paired Wilcoxon signed-rank tests were applied to the held-out ROUGE values at the document level to reduce the risk of over-interpreting partition-specific variation. Among the single-model baselines, TF-IDF slightly outperformed TextRank, achieving ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.4879, 0.3940, and 0.4303, with a cumulative RSUM of 1.3122. TextRank obtained 0.4761, 0.3817, and 0.4227, with an RSUM of 1.2805. The BERTurk-Legal baseline reached 0.4410, 0.3217, and 0.3796, with a cumulative RSUM of 1.1424. The difference among the three baselines, therefore, remains clear, and TF-IDF remains the highest-scoring single-model reference under the present evaluation setup.

Figure 4 visualizes the same ROUGE comparison and clearly separates the baseline group from the weighted ensemble across all three reported metrics.

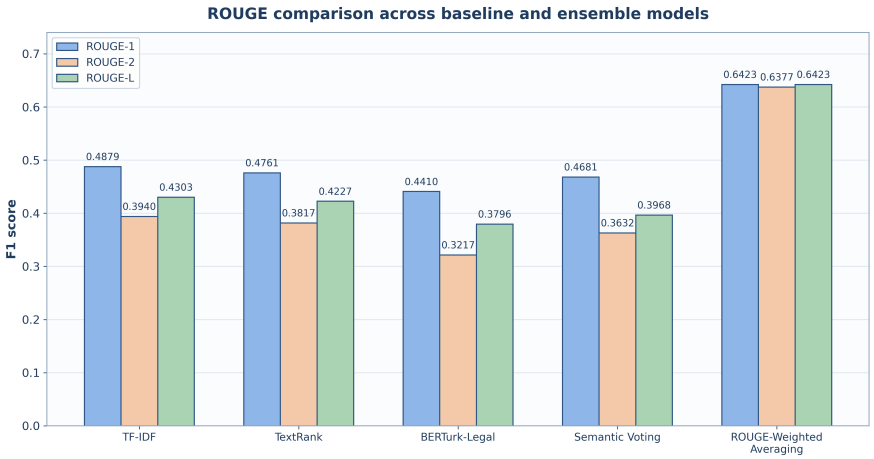

Fig. 4. Comparison of ROUGE-1, ROUGE-2, and ROUGE-L F1 scores across the baseline and ensemble models. The grouped bars visually reinforce the stronger overall performance of ROUGE-Weighted Averaging.

The Semantic Voting ensemble reached ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.4681, 0.3632, and 0.3968, respectively, with a cumulative RSUM of 1.2281. Under the present configuration, this semantic re-ranking strategy did not surpass the two base models. This suggests that semantic similarity alone was not sufficient to preserve the lexical and structural evidence most relevant to this legal-domain extractive task.

By contrast, the ROUGE-Weighted Averaging ensemble achieved the highest overall scores, with exact ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.6423,

0.6377, and 0.6423, respectively, and a cumulative RSUM of 1.9223. Relative to the strongest baseline, TF-IDF, this corresponds to gains of 31.6% in ROUGE-1, 61.9% in ROUGE-2, 49.3% in ROUGE-L, and 46.5% in RSUM. These findings suggest that score-level fusion of TF-IDF and TextRank yielded higher scores than direct semantic re-ranking in this setting and that the weighted ensemble combines complementary evidence from both base models.

The BERTurk-Legal baseline remains an informative methodologically because it introduces a stronger, modern neural comparator based on a legal-domain-pretrained transformer rather than a purely lexical or graph-based heuristic. However, under the present unsupervised document–sentence similarity formulation, it did not exceed TF-IDF or TextRank. A likely reason is that the checkpoint was originally designed for legal retrieval rather than sentence-level extractive summarization, while the current baseline uses document–sentence cosine similarity without any task-specific fine-tuning. Under these conditions, the model can capture broad legal-domain-relatedness, yet may remain less effective at identifying the most summary-worthy sentences in a short three-sentence extractive output. This pattern suggests that domain adaptation alone was not sufficient to outperform the stronger traditional baselines without additional task-specific training or role-aware supervision.

Overall, these results indicate that the ensemble methods, particularly ROUGE-based weighting, yielded higher scores than the traditional baselines. They also suggest that score-level fusion was more effective in this setting than semantic re-ranking alone.

Several limitations should be noted. The dataset is moderate in size and therefore constrains broad generalization beyond the document types represented in this study, even though it remains suitable for controlled methodological comparison. In addition, formal inter-annotator agreement was not measured in the current study. Future work should expand the corpus and incorporate multi-annotator validation to strengthen annotation reliability and improve the benchmark's generalizability.

6 Conclusion

This study examined improvements in Turkish legal extractive summarization using ensemble approaches built on traditional methods. In addition to TF-IDF and TextRank, the evaluation included a BERTurk-Legal-based extractive baseline as a stronger modern neural comparator. Semantic Voting obtained ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.4681, 0.3632, and 0.3968, with a cumulative RSUM of 1.2281. In contrast, ROUGE-Weighted Averaging obtained the highest scores, reaching 0.6423, 0.6377, and 0.6423 on ROUGE-1, ROUGE-2, and ROUGE-L, with a cumulative RSUM of 1.9223. Compared with the strongest baseline, TF-IDF, the weighted ensemble improved ROUGE-1 by 31.6%, ROUGE-2 by 61.9%, ROUGE-L by 49.3%, and RSUM by 46.5%. These findings suggest that the ensemble strategy provided the strongest overall performance among the methods considered in this setting. The proposed methods

are lightweight and may be suitable for integration into practical legal NLP workflows.

Future work may explore several extensions. First, incorporating deep learning models such as BERT- or LSTM-based classifiers into the ensemble may improve semantic understanding and contextual relevance. Second, experiments on larger datasets and additional legal document types, such as contracts, regulations, or administrative decisions, would help assess generalizability more fully. Human evaluation by legal experts would also provide useful qualitative evidence. Finally, combining extractive and abstractive components may support the development of hybrid summarization systems for the legal domain.

References

1. Yang, B., Luo, X., Sun, K., Luo, M.Y.: Recent progress on text summarisation based on BERT and GPT. In: *Natural Language Processing and Information Systems*, LNAI, vol. 14120, pp. 225–241. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-40292-0_19
2. Alipour, N., Aydın, S.: Abstractive summarization using multilingual text-to-text transfer transformer for the Turkish text. *IAES International Journal of Artificial Intelligence (IJ-AI)* **14**(2), 1587–1596 (2025). https://doi.org/10.11591/IJAI.V14.I2.PP1587-1596
3. Akülker, E., Tatar, Ç.: Extractive text summarization for Turkish: Implementation of TF-IDF and PageRank algorithms. In: *Proceedings of the International Symposium on Intelligent Systems*, LNNS, vol. 544, pp. 688–704. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-16075-2_51
4. Ozsoy, M.G., Cicekli, I., Alpaslan, F.N.: Text summarization of Turkish texts using latent semantic analysis. In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pp. 869–876. Coling 2010 Organizing Committee, Beijing, China (2010). `https://aclanthology.org/C10-1098/`
5. Baykara, B., Güngör, T.: Abstractive text summarization and new large-scale datasets for agglutinative languages: Turkish and Hungarian. *Language Resources and Evaluation* **56**(3), 973–1007 (2022). https://doi.org/10.1007/S10579-021-09568-Y
6. Türker, M., Ari, M.E., Han, A.: VBART: The Turkish LLM. *arXiv preprint* arXiv:2403.01308 (2024). `https://arxiv.org/abs/2403.01308`
7. Baykara, B., Güngör, T.: Turkish abstractive text summarization using pretrained sequence-to-sequence models. *Natural Language Engineering* **29**, 1275–1304 (2023). https://doi.org/10.1017/S1351324922000195
8. Alagoz, N.K., Kucuksille, E.U.: Extractive summarization of scientific papers in Turkish with deep learning. *Research Square preprint* (2022). https://doi.org/10.21203/RS.3.RS-2292145/V1
9. Kuş, A., Acı, Ç.I.: An extractive text summarization model for generating extended abstracts of medical papers in Turkish. *Bilgisayar Bilimleri ve Teknolojileri Dergisi* **4**(1), 19–26 (2023). https://doi.org/10.54047/bibted.1260697
10. Mahindrakar, S.M., Mondal, T., Arosh, S., Dhakne, A., Chavan, D.D.: Exploration of summary informativeness and topic richness through mining multilingual tweets: A study on Turkey earthquake 2023. *Social Network Analysis and Mining* **14**(1), 117 (2024). https://doi.org/10.1007/S13278-024-01386-8

11. Turan, T., Kucuksille, E.U.: Summarization, prediction, and analysis of Turkish Constitutional Court decisions with explainable artificial intelligence and a hybrid natural language processing method. *IEEE Access* **13**, 59766–59779 (2025). https://doi.org/10.1109/ACCESS.2025.3556725
12. Albayati, M.A.A., Findik, O.: A hybrid transformer-based framework for multi-document summarization of Turkish legal documents. *IEEE Access* **13**, 37165–37181 (2025). https://doi.org/10.1109/ACCESS.2025.3545750
13. Radev, D.R., Jing, H., Budzikowska, M.: Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In: *Proceedings of the NAACL-ANLP Workshop on Automatic Summarization*, pp. 21–30, Seattle, WA, USA (2000)
14. Umejiaku, A.P., Zhou, F., Sheng, V.S.: Ensemble summarization models to leverage performance. In: *5th International Conference on Computing and Big Data (ICCBD)*, pp. 56–61. IEEE (2022). https://doi.org/10.1109/ICCBD56965.2022.10080715
15. Liu, Y., Lapata, M.: Text summarization with pretrained encoders. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3730–3740, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1387
16. Mihalcea, R., Tarau, P.: TextRank: Bringing order into text. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404–411. Association for Computational Linguistics, Barcelona, Spain (2004). `https://aclanthology.org/W04-3252/`
17. Öztürk, C.E., Özçelik, Ş.B., Koç, A.: A transformer-based prior legal case retrieval method. In: *2023 31st Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4. IEEE, Istanbul, Turkey (2023). https://doi.org/10.1109/SIU59756.2023.10223938
18. Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004). `https://aclanthology.org/W04-1013/`