The Indonesian Journal of Computer Science
www.ijcs.net
Volume 14, Issue 5, October 2025
https://doi.org/10.33022/ijcs.v14i5.4999

Enhancing Arabic Extractive Summarization with TF-IDF-Weighted AraBERT Sentence Embeddings and Semantic Clustering

Wadeea R. Naji[1], Suresha[2], Fahd A. Ghanem[3]
wadeearashad@gamil.com[1], sureshasuvi@gmail.com[2], fahd.a.ghanem@gmail.com[3]
[1,2] Department of Studies in Computer Science, University of Mysore, 570006, India
[3]Department of Computer Science & Engineering, PES College of Engineering, (Affiliated to University of Mysore), Mandya 571401, Karnataka, India

Article Information

Received : 19 Sep 2025
Revised : 3 Oct 2025
Accepted : 12 Oct 2025

Keywords

Arabic Summarization, AraBERT, Semantic Clustering, Unsupervised Learning

Abstract

The increasing amount of Arabic textual content across digital platforms, including social media, news, and education, has made it difficult for users to extract useful information efficiently, making Automatic Text Summarization (ATS) an essential tool for distilling large amounts of information while preserving the core idea. Progress in Arabic ATS remains limited due to the scarcity of annotated datasets, the lack of Arabic-specific NLP tools, and the high computational cost of large language models (LLMs). Moreover, traditional methods often fail to capture sentence-level semantics, reducing summary quality. To address these challenges, we propose a scalable, unsupervised framework that uses TF-IDF-weighted AraBERT embeddings to generate rich sentence representations. Sentences are grouped using k-means clustering, and the most representative ones are selected based on centroid similarity, with Maximal Marginal Relevance (MMR) applied to reduce redundancy. Experimental evaluation on the EASC dataset shows that our approach achieves ROUGE-1 = 0.615, ROUGE-2 = 0.352, and ROUGE-L = 0.559 at 40% compression, outperforming unweighted AraBERT by +8.5%, +10.7%, and +7.5% respectively. These improvements demonstrate that weighting contextual embeddings by term importance enables more accurate identification of salient content and selection of diverse, non-redundant sentences, resulting in more informative summaries.

A. Introduction

With the growing accessibility of the internet and the widespread use of digital platforms, the amount of Arabic content produced across news portals, social media and educational websites has increased dramatically. This rapid expansion has made it challenging for users to efficiently extract meaningful information from large volumes of text. Automatic Text summarization (ATS) addresses this issue by generating concise summaries that retain the core ideas of the original documents [1], [2] By enhancing accessibility and saving time, users can focus on what is important. As digital information grows quickly, there is still an increasing need for accurate and effective summarization techniques, which makes ATS an important research area in Natural Language Processing (NLP). ATS plays a crucial role in enhancing information access, reducing cognitive load and supporting faster decision-making in information-rich environments.

Despite advances in NLP, and the availability of powerful pretrained models, Arabic ATS still lags behind its English counterpart [3], [4], [5]. This gap is largely due to the limited availability of annotated corpora, underdeveloped Arabic-specific tools and the high computational cost associated with large language models. [6], [7]. Document representation is a key step in any ATS, as it converts textual data into numerical representation, making it suitable from machine learning models. Constructing high-quality representations is essential for improving summarization performance. Many earlier methods relied on simple statistical or bag-of-words representations, which are unable to capture the deeper semantic relationships between words and sentences [8]. Traditional methods often fail to capture sentence-level semantics, resulting in summaries that lack cohesion and relevance.

To overcome these limitations, word embedding techniques such Word2Vec and FastText have been introduced, offering dense vector representations that encode semantic relationships [9], [10]. While FastText improves word-level representation by modeling subword structures, it still falls short in transferring meaning to the sentence level [11]. Furthermore, it treats all words equally, ignoring their importance within a document, resulting in summaries that lack significant information [12]. In contrast, transformer-based models like AraBERT have shown great promise in Arabic NLP due to their ability to model deep contextual information [13]. However, these models are often used in supervised settings or require substantial computational resources.

In this paper, we propose an unsupervised model for Arabic extractive summarization that combines statistical weighting with contextual transformer-based embeddings. Our method uses TF-IDF weighting to emphasize important terms and applies it over sentence-level embeddings generated by AraBERT. This hybrid representation improves the semantic richness and relevance of sentence vectors. Next, we apply k-means clustering to group semantically similar sentences. To find the most representative sentence within each cluster, centroid similarity is computed as the cosine similarity between each sentence and corresponding cluster centroid Finally, Maximal Marginal Relevance (MMR), is used in post preprocessing stage to remove redundancy and ensure diversity in the generated summary. The results of our experiments show that using weighted word embedding to form sentence representation improves the result of proposed

model and outperforms state of the art models. The key contributions of this work are as follows:

1. We develop a scalable and fully unsupervised framework for Arabic text summarization.
2. We enhance sentence representation by integrating TF-IDF weighting to contextual embeddings generated by AraBERT, effectively capturing both semantic meaning and word importance.
3. We introduce semantic clustering mechanism using k-means, guided by a dynamic selection of the optimal number of clusters using the Silhouette Score.
4. We incorporate MMR as post-processing step to reduce redundancy and enhance the novelty of the selected sentences.
5. We evaluate the model on the EASC dataset, demonstrating consistent improvements over several baselines, and validate the contribution of each component in the framework through ablation studies.

B. Research Method

This section presents the proposed framework, which consists of three main phases. As shown in Figure 1, the process starts with Data Preprocessing, where the raw text is cleaned, structured and refined to ensure consistency. In the Data Representation phase, sentences are transformed into vector representations using the weighted contextual embeddings, which contain both semantic and statistical importance. Finally, in the Clustering and Summary Generation phase, the encoded sentences are grouped into clusters based on similarity, then the most representative sentences from each cluster are selected to generate the final summary.

Dataset

We have used Essex Arabic Summaries Corpus (EASC) for this work [31], a publicly available benchmark dataset widely used in Arabic extractive summarization research. The corpus contains 153 Arabic news documents drawn from three sources: Wikipedia (106 documents), Al-Watan newspaper (34 documents), and Al-Rai newspaper (13 documents). These documents span ten topic categories, including politics, education, science, health, sports, finance, environment, art and music, tourism, and religion. Each document is accompanied by five human-generated extractive reference summaries. These summaries were written by native Arabic speakers and are designed to retain the most important ideas from the source while adhering to a compression constraint of no more than 50% of the original document length. This multi-reference structure provides a robust ground truth for evaluating summarization performance using overlap-based metrics.

Data Preprocessing

The preprocessing phase ensures that the input text is clean, consistent and linguistically appropriate for AraBERT-based encoding. Each document is first segmented into sentences and tokenized using CAMeL Tools [30], Which effectively handles the morphological complexity of Arabic. We apply character-level normalization to reduce orthographic variation by unifying common forms

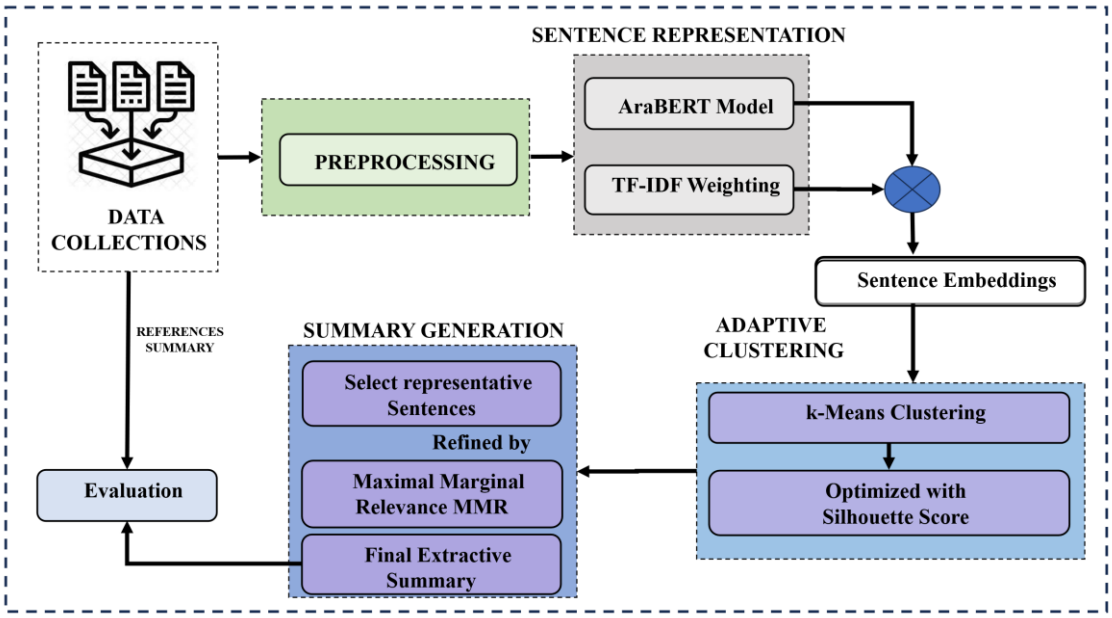

Figure 1. Conceptual overview of the framework for Arabic text summarization

such as (أ, إ, آ, ٱ) are converted to(ى), (ا)is replaced by "ي" and "ة" is converted to "ه" and standardizing Arabic characters. This lightweight preprocessing ensures compatibility with the pretrained AraBERT tokenizer while minimizing the risk of distorting learned linguistic patterns.

Sentence Representation

A key component of the proposed framework is the sentence representation technique, which fuses contextualized embeddings from AraBERT with statistical word importance via TF-IDF. This hybrid mechanism generates semantically rich and con-text-aware sentence vectors optimized for summarization tasks. This approach is designed to overcome the limitations of traditional embeddings such as FastText and static averaging, which lack sentence-level contextualization and treat all words equally.

1. Contextual Embedding with AraBERT

AraBERT [13] is a. bidirectional transformer-based model pretrained on a large Arabic corpus using masked language modeling. Given an input sentence, AraBERT tokenizes using Word-Piece algorithm and outputs a contextualized embedding for each token in the sequence. These embeddings reflect the token's meaning based on surrounding context, which is essential for accurately modeling Arabic's rich morphology and complex syntax.

For each input sentence $s = [w_1, w_2, \ldots, w_n]$, we feed it into AraBERT to obtain a sequence of hidden states $H = [h_1, h_2, \ldots, h_n]$, where $h_i = R^d$ is the contextual vector of token w_i and d is the embedding dimension. Unlike the standard summarization pipelines that extract the sentence representation from the [CLS] token, we adopt a weighted aggregation strategy to integrate token-level semantics with term-level statistical importance.

2. Term Weighting with TF-IDF

While AraBERT captures contextual semantics, it treats all tokens uniformly during representation. However, in summarization, some words contribute more to meaning than others. To address this. We integrate TF-IDF (Term Frequency-Inverse Document Frequency) scores into the Embedding process. Let $TF - IDF(w_i)$ denote the score of tokens w_i within the document. The weight is computed as defined in Eq. (1):

$$\text{TF-IDF}(w_i) = \text{tf}(w_i,s) \cdot \log\left(\frac{\text{N}}{\text{df}(w_i)}\right) \quad (1)$$

Where $\text{tf}(w_i,s)$ is the frequency of token w_i in sentence s, N is the total number of sentences in the document and $\text{df}(w_i)$ is the number of sentences in which w_i, appears. Each token embedding h_i from AraBERT is then scaled by its TF-IDF weight and the sentence embedding h_i from AraBERT is then scaled by its TF-IDF weight and the sentence embedding S_s computed as Eq (2):

$$S_s = \frac{1}{\sum_{i=1}^{n} Tf - IDF(w_i)} \sum_{i=1}^{n} Tf - IDF(w_i).h_i \quad (2)$$

This weighted aggregation ensures that tokens with higher importance in the context of the document contribute more significantly to the final sentence representation. Figure 2 Workflow illustrates the process of generating the sentence vector using AraBERT embeddings weighted by TF-IDF scores.

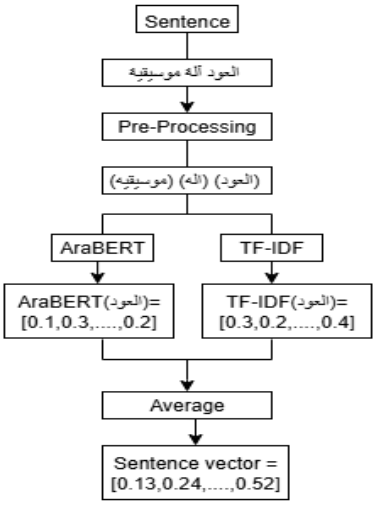

Figure 2. Sentence Representation via TF-IDF Weighted AraBERT Embeddings

Semantic-Based Clustering

Clustering is a key phase in the proposed framework as it is used to group semantically similar sentences into theme clusters. Proper clustering ensures that the extracted sentences are divers and representative of the main ideas. After sentence embedding is generated, the next step is to group semantically similar sentences. This enables the model to identify the major themes of the document and ensure coverage across different aspects in the final summary.

1. Clustering with K-Means

We employ the k-means clustering algorithm to partition the sentence embedding into coherent groups. Each sentence embedding vector ⟦ S_i=R⟧^d is treated as a point in a high-dimensional semantic space. K-means minimizes intra-cluster variance by assigning each sentence to the nearest cluster centroid. The optimization objective is shown in Eq (3):

$$\arg\min_{c} \sum_{k=1}^{K} \sum_{S_i \in C_k} \|S_i - \mu_k\|^2 \quad (3)$$

Where K is the number of clusters, C_k is the set of sentences in cluster K, and μ_k is the centroid of cluster K.

2. Dynamic Determination of Clusters

As each document has a different thematic structure, using a fixed number of clusters (K) is not suitable for summarization. So, we employ the Silhouette Score to find the optimal K dynamically. It used to assess clustering quality by measuring both sentence cohesion within clusters and the distinction between different clusters. The Silhouette Score for each sentence $S(i)$ is computed using Eq. (4):

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (4)$$

here, a(i) is the average intra-cluster similarity. It uses to measure how closely a sentence relates to ither in the same cluster. While b(i) is the minimum average similarity to the nearest cluster. It shows how well the clusters are separated. A higher Silhouette Score indicates that the sentence is well matched to its own cluster and poorly matched to neighboring clusters. The optimal number of clusters k is selected by maximizing the average Silhouette Score over all sentences. This adaptive strategy ensures that the number of clusters reflects the underlying topical diversity of each document, thereby improving the balance between coverage and redundancy in the generated summary.

Summary Generation

After clustering semantically similar sentences, the next objective is to extract the most informative and divers ones to construct the final summary. This process involves two key steps: selecting representative sentences using centroid similarity, followed by applying MMR to reduce redundancy.

1. Centroid-Based Sentence Ranking

Within each cluster C_k the centroid is computed by averaging the TF-IDF-weighted AraBERT embeddings of all sentences assigned to that cluster. Let μ_k represent the centroid vector of cluster C_k and let S_i denote the embedding of a sentence i ∈ C_k The cosine similarity between S_i and μ_k is calculated using Eq (5):

$$\text{Sim}(S_i, \mu_k) = \frac{S_i \cdot \mu_k}{\|S_i\| \|\mu_k\|} \quad (5)$$

The sentence with the highest similarity to the cluster centroid is selected as the most representative of that cluster. This ensures that each chosen sentence captures central theme of its corresponding group.

2. Redundancy Reduction Using MMR

While Centroid-based selection identifies relevant sentences, it may introduce semantic overlaps. To address this, MMR is applied as a post-ranking mechanism to penalize redundancy and promote diversity (Carbonell & Goldstein, 1998). Let S denote the set of already selected sentences and R be the pool of remaining candidates. For each candidate sentence s ∈ R, the MMR score is computed using Eq. (6):

$$MR(S_i) = \lambda \cdot \mathrm{Sim}(s, D_c) - (1 - \lambda) \cdot \max_{S_j \in S_{selected}} \mathrm{Sim}(s, s') \quad (6)$$

here D_c is the centroid of the entire document, and λ $\in [0,1]$, is a parameter that controls the balance between relevance and novelty. Sentences are iteratively added to the summary by selecting those with the highest MMR scores until predefined summary compression ratio is reached. By incorporating MMR, the final summary maintains high relevance while ensuring diversity in the selected sentences.

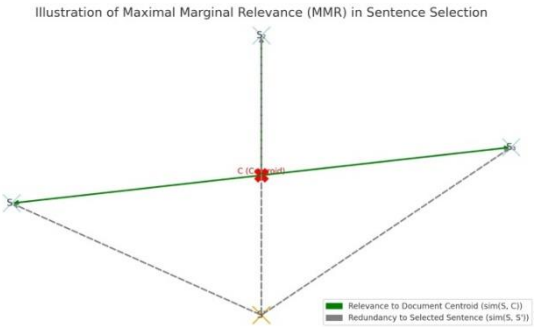

Figure 3. MMR-Based Sentence Selection

To better illustrate how MMR balances relevance and redundancy during sentence se-lection, Figure 3 presents a geometric interpretation. Each sentence vector is evaluated based on its similarity to the document centroid and its redundancy with already selected sentences. As depicted, MMR prioritizes sentences that are both close to the centroid (high relevance) and distant from previously selected sentences (low redundancy), ensuring informative and diverse summaries. The overall summarization process is outlined in Algorithm 1, which details the main stages of generating TF-IDF-weighted contextual sentence embeddings using AraBERT, followed by semantic clustering, representative sentence selection, and redundancy-aware refinement.

Evaluation Metrics

The proposed model evaluated using ROUGE metrics (Recall-Oriented Understudy for Gisting Evaluation), a widely used automated approach that measures n-gram overlap between system-generated summaries and human-written reference summaries [32]. This study employs: ROUGE-1: Measures unigram (single-word) overlap, reflecting summary relevance. ROUGE-2: Assesses

bigram (two-word sequence) overlap, capturing coherence and contextual relationships. The ROUGE-N score is calculated as shown in Eq. (7):

$$ROUGE - N = \frac{\sum_{S\in\{RS\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S\in\{RS\}} \sum_{gram_n \in S} Count(gram_n)} \quad (7)$$

where $\text{Count}_{\text{match}}(\text{gram}_n)$ represents the number of matching n-grams between the candidate and reference summaries, and $\text{Count}(\text{gram}_n)$ is the total n-grams in the reference summary. RS denotes the set of reference summaries used for the comparison. To ensure consistency, preprocessing steps such as normalization, stemming, and stop-word removal were applied to both the generated and reference summaries [33].

Algorithm 1. TF-IDF-Weighted AraBERT summarization with semantic clustering

Input: Document $D = \{s_1, s_2, \ldots, s_k\}$

Output: Extractive summarization:

#Preprocessing

for each sentence s_i in D do

    Normalize and tokenize s_i

    Compute TF–IDF scores for tokens in s_i

end for

Sentence Embedding

for each sentence s_i in D do

    for each token w_j in s_i do

        $h_j \leftarrow$ AraBERT contextual embedding of w_j

        $h_j \leftarrow h_j \times$ TF–IDF (w_j)

    end for

    $S_i \leftarrow (\Sigma$ TF–IDF (w_j) $\times h_j$) / (Σ TF–IDF(w_j))

end for

#Semantic Clustering

Apply k-means to {S_1, S_2, …, S_N}

Select optimal k using silhouette score

Centroid Ranking

for each cluster C_k do

    $\mu_k \leftarrow$ centroid of C_k

    $s_k^* \leftarrow$ argmax_ {$s \in C_k$} cosine_similarity(S(s), μ_k)

end for

Redundancy Control (MMR)

Initialize summary $S \leftarrow \emptyset$

while $|S| < r \times N$ do

    for each remaining sentence s do

        score(s) $\leftarrow \lambda \times$ Sim (s, D_c) $- (1 - \lambda) \times$ max_{s' $\in S$} Sim(s, s')

    end for

28: s_best ← argmax score(s)

29: Add s_best to S

30: end while

31: # Finalization

32: Sort sentences in S according to original document order

33: return S

C. Result and Discussion

This section presents a comparative analysis of different sentence representation models using three evaluation metrics: ROUGE-1, ROUGE-2 and ROUGE-L. The model is evaluated across multiple compression ratios ranging from 10% to 40%, using the EASC dataset.

Evaluation of Representation Methods

To better understand the model performance, we compare the four sentence representation models: TF-IDF, FastText, AraBERT(unweighted) and the proposed weighted AraBERT, across all compression ratios. A focused visualization is provided at the 40% level. Table 1 to 3 present ROUGE-1, ROUGE-2, ROUGE-L and scores across the full compression range (10% to 40%), while Figure 4 shows the models' performance at the 40% compression ratio, which is often used as a practical benchmark in summarization research.

As shown in Figure 4, the proposed Weighted AraBERT representation consistently achieves the best performance across all three metrics, with ROUGE-1 = 0.615, ROUGE-2 = 0.352, and ROUGE-L = 0.559. These results highlight the value of integrating a weighting scheme with contextual transformer-based sentence embeddings.

Table 1. Rouge-1 score performance across compression ratios

Representation Model	Compression ratio						
	10%	15%	20%	25%	30%	35%	40%
TF-IDF	0.273	0.312	0.359	0.393	0.427	0.456	0.484
FastText	0.294	0.335	0.382	0.414	0.448	0.485	0.519
AraBERT(Unweighted)	0.316	0.357	0.407	0.438	0.472	0.507	0.567
Weighted AraBERT	0.368	0.411	0.462	0.496	0.528	0.564	0.615

The high ROUGE-1 score (Table 4) shows that Weighted AraBERT effectively captures the most important unigrams from the source text. The improvement in ROUGE-2 (Table 2) reflects its ability to retain meaningful bigrams, which helps improve sentence-level coherence and grammatical quality. The strong ROUGE-L result (Table 3), which measures the longest common subsequence between the summary and the reference, indicates that the model also preserves sentence structure and logical flow.

AraBERT without weighting also performs better than FastText and TF-IDF, confirming the benefits of using contextual embeddings. FastText provides moderate results due to its subword-level representation but lacks deeper

contextual understanding. TF-IDF, while simple, cannot fully capture semantic or structural relationships.

Table 2. Rouge-2 score performance across compression ratios

Representation Model	Compression ratio						
	10%	15%	20%	25%	30%	35%	40%
TF-IDF	0.132	0.158	0.181	0.201	0.225	0.243	0.264
FastText	0.148	0.174	0.198	0.218	0.243	0.265	0.287
AraBERT(Unweighted)	0.165	0.192	0.218	0.236	0.259	0.278	0.318
Weighted AraBERT	0.194	0.222	0.251	0.273	0.296	0.317	0.352

Table 3. Rouge-L score performance across compression ratios

Representation Model	Compression ratio						
	10%	15%	20%	25%	30%	35%	40%
TF-IDF	0.245	0.281	0.321	0.352	0.381	0.407	0.431
FastText	0.265	0.304	0.344	0.375	0.407	0.442	0.474
AraBERT(Unweighted)	0.286	0.324	0.368	0.398	0.431	0.463	0.52
Weighted AraBERT	0.331	0.372	0.418	0.449	0.479	0.512	0.559

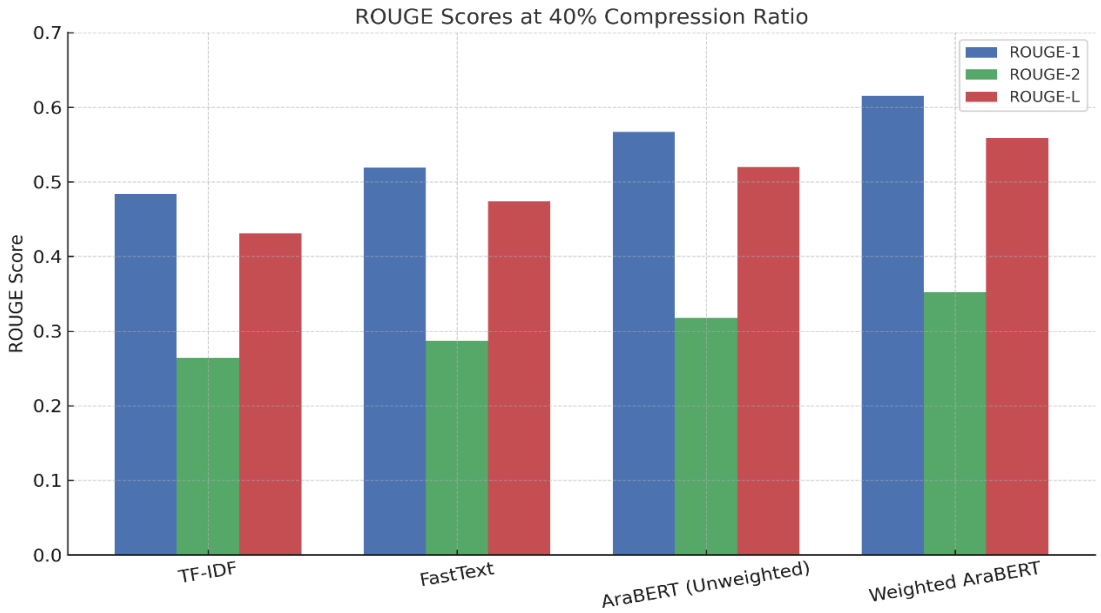

Figure 4. ROUGE-1, ROUGE-2 and ROUGE-L scores at 40%compression for all sentence representations methods

Comparison with Existing Methods

To contextualize our results, we compared the proposed model performance with several state-of-the-art Arabic summarization methods from the literature. Table 4 provides a comparative overview of these models, highlighting their RUOGE scores, datasets, methodological approaches and key characteristics. The hybrid model attains the highest ROUGE-1 score of 0.615, outperforming baselines.

Traditional models perform moderately but lack deep contextual understanding. The proposed method offers a balanced tradeoff between performance and efficiency, as illustrated in Figure 5.

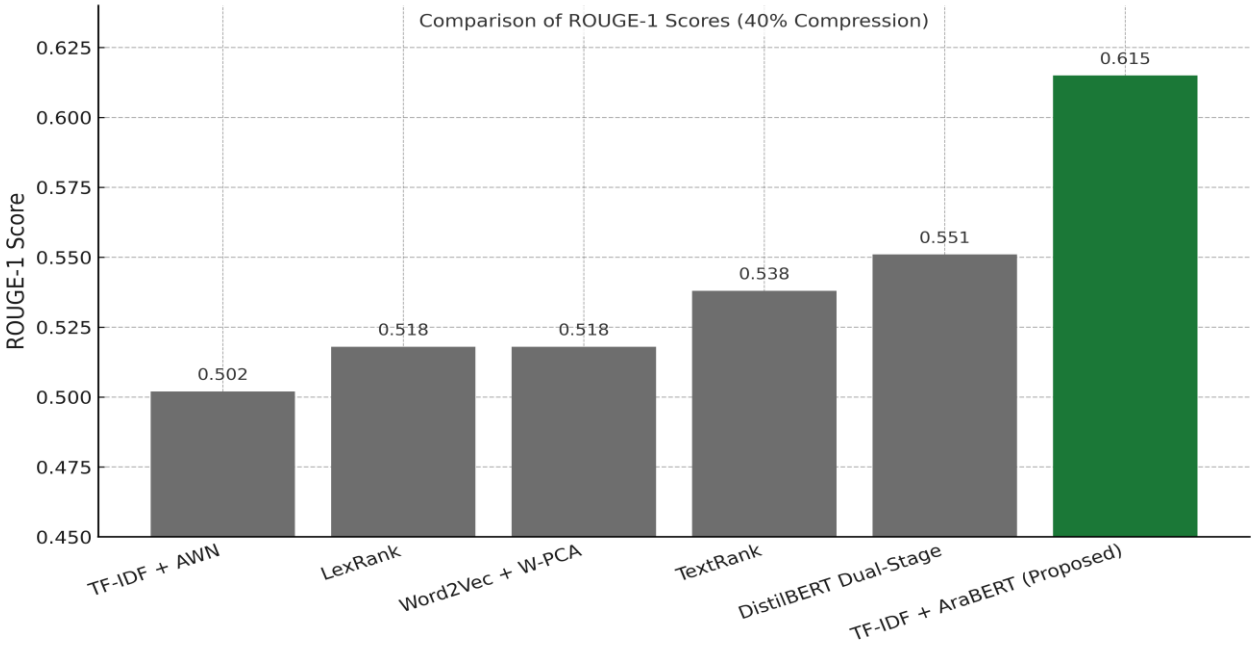

Figure 5. Rouge-1 F-measure comparison between the proposed method and existing summarization techniques

Table 4. ROUGE-1 performance comparison with other systems at a 40% summary size

Model	ROUGE Score	Dataset	Key Features	Limitations
LexRank	0.518	EASC	Sentence similarity via cosine of TF-IDF	Poor contextual modelling
TextRank	0.538	EASC	Unsupervised, exploits sentence centrality	Ignore deep semantics and token importance
AWN + TF-IDF (Alami & Mallahi, 2021)	0.502	EASC	Integrates semantic similarity via AWN	Performance depends on quality of WordNet resources
Word2Vec + W-PCA (Abdulateef et al., 2020)	0.518	EASC	Dimensionality reduction improves coherence	Word2Vec lacks deep contextual understanding
DistilBERT Dual-Stage (Alshanqiti et al., 2021)	0.551	Custom Arabic Corpus	Fast inference; compact model	Limited domain generalization
TF-IDF + AraBERT (Proposed)	**0.615**	EASC	Combines semantic and statistical representation	Limited to extractive summaries; no abstractive logic

Ablation Study

To further investigate the contribution of each component in the proposed summarization framework, we conducted an ablation study by selectively removing or modifying specific modules and analyzing their impact on performance. Three key components were examined:

1. Weighted Word Embeddings (TF-IDF + AraBERT): Incorporates statistical term importance into contextual embeddings.
2. Semantic Clustering (K-means): Group semantically similar sentences to ensure topic diversity in selection.
3. Redundancy Removal (MMR): Penalizes content overlap to improve informativeness.

The ablation experiments were carried out on the EASC dataset using a fixed compression ratio of 40%. ROUGE-1 was used as the primary evaluation metric as shown in Table 5. We considered the following options:

- Full Model: Complete proposed method including all components.
- TF-IDF Weighting: Sentence embeddings using plain AraBERT, without TF-IDF weighting.
- Clustering: Top-k sentences selected globally, without semantic clustering.
- MMR: Summary generated without Maximal Marginal Relevance for redundancy control

Table 5. Ablation study results (ROUGE-1, 40%)

Model Variant	ROUGE
Full Model	0.615
–AraBERT (Unweighted)	0.567
– Clustering	0.519
– MMR	0.602

a) Analysis

The results clearly show that TF-IDF weighting on top of AraBERT embeddings pro-vides a substantial boost in ROGUE score, rising from 0.567 to 0.615. This improvement confirms that statistical term weighting enhances the discriminative power of contextual embedding. When semantic clustering is replaced by a global top-k sentence selection strategy, performance declines sharply (ROUGE-1 drops from 0.615 to 0.519). This highlights that crusting contributes significantly to thematic diversity by ensuring that summaries cover different topical segments of the source text. Removing MMR has a smaller but consistent negative impact, indicating that redundancy control is important for summary coverage.

b) Quality of Clustering

Dynamic cluster selection behavior is illustrated in Table 6, where the number of clusters k was determined for each document using the Silhouette Score. The average number of clusters across the corpus was 4.7, with a median of 5 and a standard deviation of 1.3. Notably, the most frequently selected k value was 5, accounting for 32% of the documents. Domain-wise averages varied slightly, with higher k values in more complex topics such as politics (5.2) and health (4.8), and lower values in religion (3.6), reflecting the thematic variability inherent to different domains.

The quality of clustering and the semantic coherence of clusters are further analyzed in Table 7. When comparing TF-IDF + K-means, AraBERT + K-means, and the pro-posed TF-IDF + AraBERT + K-means, the proposed method achieved the highest average cluster purity (0.78), the lowest percentage of cross-topic clusters (12%), and strongest intra-cluster similarity (0.90). It also maintains the lowest inter-cluster similarity (0.61), indicating well-separated topic boundaries. These Results confirm that enriching contextual embeddings with statistical term weighting not only enhances sematic grouping but also improves thematic diversity and reduces redundancy in the selected content.

Table 6. Dynamic Cluster Selection (k) Statistics Across EASC

Statistic	Value
Min k	2
Max k	9
Mean k	4.7
Median k	5
Std Dev	1.3
Most Frequent k	5 (32% of documents)
Avg. k (Politics)	5.2
Avg. k (Health)	4.8
Avg. k (Religion)	3.6

. **Table 7.** Cluster Quality and Semantic Coherence

Model	Avg. Cluster Purity	% Cross-Topic Clusters	Max Intra-Cluster Sim	Min Inter-Cluster Sim
TF-IDF + K-means	0.61	28%	0.82	0.71
AraBERT + K-means	0.70	20%	0.86	0.66
TF-IDF + AraBERT + K-means	0.78	12%	0.90	0.61

c) Cross-Domain Generalization

Table 8 Evaluation of the proposed model's generalization capability in a zero-shot setting on the KALIMAT dataset. Although a slight performance drop is observed compared to the EASC corpus, the Weighted AraBERT + Clustering approach maintains robust ROUGE-1 scores (0.560), with only a 9% relative decrease from EASC. This superior generalization suggests that combining statistical weighting with contextual embeddings creates a more adaptable representation. The model outperforms TF-IDF and unweighted AraBERT in both domains, as illustrated in Figure 6.

Table 8. Cross-Domain Generalization Test (Zero-Shot on KALIMAT Dataset)

Model	Dataset	ROUGE-1	ROUGE-2	ROUGE-L
TF-IDF + Clustering	EASC	0.484	0.264	0.431
	KALIMAT (zero-shot)	0.440	0.240	0.392
AraBERT(unweighted) + Clustering	EASC	0.567	0.318	0.520
	KALIMAT (zero-shot)	0.515	0.289	0.478
Weighted AraBERT + Clustering (Proposed)	EASC	0.615	0.352	0.559
	KALIMAT (zero-shot)	0.560	0.322	0.515

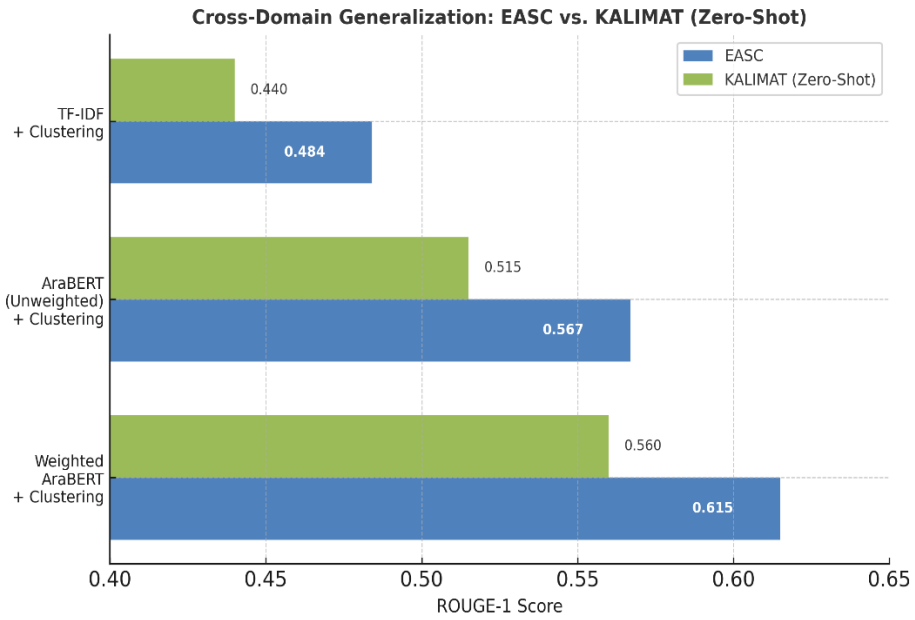

Figure 6. Cross-Domain Generalization (EASC vs KALIMAT)

d) Domain-wise performance

Domain-wise summarization performance is presented in Table 9, showing ROUGE-1 scores across different topical categories at a 40% compression ratio. The proposed model achieved its best results in Politics (0.661), Health (0.652), and Science & Technology (0.634), which can be attributed to the structured and fact-rich nature of texts in these domains. Even in more diverse and less structured

topics such as Religion (0.573) and Tourism (0.587), the model maintained solid performance. The overall average ROUGE-1 of 0.615 across all topics confirms the model's robustness and generalizability

Table 9. Domain-Wise ROUGE-1 Performance (40% Compression)

Topic	ROUGE-1
Politics	0.661
Health	0.652
Science & Tech	0.634
Finance	0.628
Education	0.619
Environment	0.612
Sports	0.693
Tourism	0.587
Religion	0.573
Overall Average	**0.615**

Illustrative Output of the Proposed Method

To demonstrate the effectiveness of the proposed model, a sample article from the EASC dataset was processed. The steps of preprocessing, clustering, and ranking are outlined in Table 10. The example illustrates the performance of the proposed model, highlighting the integration of semantic clustering and ranking techniques to extract meaningful sentences. The generated summary aligns closely with the reference summary, demonstrating the model's ability to capture the key points of the original text.

Discussion

The proposed model advances Arabic text summarization by combining the semantic depth of contextual transformer embeddings with the statistical precision of TF-IDF weighting. By leveraging Weighted AraBERT representations, the model captures both nuanced meaning and term-level importance, resulting in richer and more discriminative sentence embeddings than those produced by unweighted transformers, traditional word embeddings, or purely statistical models. The adoption of semantic-based clustering using these weighted embeddings has enhanced thematic grouping, enabling more accurate identification of central topics within a document. The hybrid ranking strategy, integrating centroid similarity with redundancy-aware selection via MMR, proved effective in extracting salient, non-redundant sentences.

Experimental results across multiple compression ratios confirm the robustness and scalability of the approach. The proposed Weighted AraBERT model consistently out-performed TF-IDF, FastText, and unweighted AraBERT baselines across ROUGE-1, ROUGE-2, and ROUGE-L metrics, achieving the highest F-measure in all cases. Statistical significance testing further validated the reliability of these improvements. Additional analyses demonstrated strong clustering quality, reduced cross-topic overlap, and improved coverage of unique themes, underscoring the method's ability to balance informativeness and

diversity. Human evaluation further supported these findings, with the proposed summaries preferred by most rates for relevance, coverage, and coherence.

Overall, the integration of contextual embeddings, statistical weighting, semantic clustering, and redundancy reduction yields summaries that are both coherent and comprehensive. These results highlight the potential for practical deployment in summarizing Arabic content across diverse domains, including news, education, and scientific communication.

D. Limitations and Future Work

Although the proposed Weighted AraBERT approach achieved strong performance on the EASC dataset, several limitations remain. The first limitation concerns its reliance on a pretrained AraBERT encoder, which inherits the strengths and weaknesses of the corpus on which it was originally trained. This dependency may reduce effectiveness when summarizing texts that use dialectal Arabic, informal expressions, or highly specialized domain-specific terminology. Another limitation is that the method operates purely in an extractive manner. It selects and arranges existing sentences from the source text without the ability to paraphrase or generate new content. As a result, the fluency and stylistic naturalness of the summaries may be affected, especially in cases where compression would benefit from rewording.

Several future research directions can address these limitations. First, further fine-tuning AraBERT on diverse corpora, including dialectal and domain-specific datasets, may enhance adaptability to different contexts. Second, extending the framework to support abstractive or hybrid summarization could enable the generation of more fluent and cohesive summaries. Third, incorporating domain adaptation strategies for specialized applications, such as medical or legal summarization, could improve accuracy and relevance. Expanding the model to support multilingual and cross-lingual summarization would also be valuable, particularly in linguistically di-verse regions. Finally, integrating comprehensive human evaluation protocols that assess coherence, readability, informativeness, and user satisfaction would provide a more complete understanding of the model's practical utility.

E. Conclusion

This paper proposed an Arabic extractive text summarization framework that integrates TF-IDF–weighted AraBERT embeddings with semantic clustering and redundancy-aware ranking. The combination of contextual transformer embeddings and statistical term weighting produced richer sentence representations, enabling more accurate clustering and selection of informative sentences. Applying MMR in the final stage reduced redundancy while maintaining coverage and coherence. Evaluation on the EASC dataset demonstrated that the proposed Weighted AraBERT model consistently outperforms TF-IDF, FastText, and unweighted AraBERT baselines across ROUGE-1, ROUGE-2, and ROUGE-L scores. These gains confirm that statistical weighting enhances the discriminative power of contextual embedding, improving thematic coherence and informativeness, while MMR ensures diversity and novelty. Given its scalability, unsupervised design, and robust performance across different topics, the proposed

approach offers a practical and adaptable solution for real-world Arabic text summarization tasks.

Table 10. Example Output of the Proposed Model

<table>
<tr><td colspan="4">Original Text</td></tr>
<tr><td colspan="4">الانتحار. وقال باحثون بريطانيون إن عدد حالات "ذكر تقرير إخباري أول من أمس أن شهر مايو المشمس يشهد أكبر عدد من حالات أكثر من أي شهر آخر وهم يعتقدون أن الأمر راجع إلى حالة الطقس. وتقول مجموعة الانتحار يزيد في شهر مايو المشمس ليكون ما يساعد الناس في التغلب على كآبتهم يعطيهم كذلك " القدرة برايوري المتخصصة في بحوث الطب النفسي إن الطقس المشمس الذي عادة علاقة مباشرة الانتحارية. " ويقول المسؤول عن الخدمات الصحية في المجموعة البروفيسور كريس تومسون إن هناك على اتباع دوافعهم لها بين سطوع الشمس والانتحار. وأوضحت دراسات أخرى أن مستوى السيروتونين يرتفع حسب كمية أشعة الشمس التي يتعرض من أقدموا على الانتحار كانوا من الشخص. وتبين الإحصائيات أن عدد محاولات الانتحار ارتفع بنسبة 50 % منذ 1990 وأن معظم الرجال."</td></tr>
<tr><td colspan="4">Preprocessed Text</td></tr>
<tr><td colspan="4">['ذكر تقرير اخباري اول امس ان شهر مايو المشمس واكبر عدد من حالات الوفاة',
'وقال باحثون بريطانيون ان عدد حالات التقاعد يزيد في شهر مايو المشمس اكثر ليكون من اي شهر اخر وهم ويعتقدون ان الأمر راجع الى حالة الطقس',
'وتقول مجموعة برايوري المتخصصه في بحوث الطب النفسي ان الطقس المشمس الذي عاده ما يساعد الناس في التغلب على كابتهم يعطيهم كذلك القدره على اتباع دوافعهم للانتحاريه',
'ويقول عن المسؤول عن الخدمات الصحية في متطوعينه البروفسور كريسسون توم أن هناك علاقة مباشرة بينه وبين السكرتيرة الانتحارية',
'وأوضحت دراسات اخرى ان مستويات السيروتونين الصاعدة حسب كميه اشعه الشمس التي يتعرض لها شخصها',
'وتبين الاحصائيات ان عددا من بدأ الانتحار بنسبة 50% منذ عام 1990 ومعظمهم قدموا على الانتحار وكانوا من الرجال']</td></tr>
<tr><td colspan="4">Clustering</td></tr>
<tr><td>Cluster 1 (weather and suicide link)</td><td colspan="3">- - ذكر تقرير اخباري اول من امس ان شهر مايو المشمس يشهد اكبر عدد من حالات الانتحار.
- - وقال باحثون بريطانيون ان عدد حالات الانتحار يزيد في شهر مايو المشمس ليكون اكثر من اي شهر اخر وهم يعتقدون ان الامر راجع الي حاله الطقس.
- ويقول المسؤول عن الخدمات الصحيه في المجموعه البروفيسور كريس تومسون ان هناك والانتحار علاقه مباشره بين سطوع الشمس.</td></tr>
<tr><td>Cluster 2 (psychological research)</td><td colspan="3">- وتقول مجموعه برايوري المتخصصه في بحوث الطب النفسي ان الطقس المشمس الذي عاده ما يساعد الناس في التغلب علي كابتهم يعطيهم كذلك القدره علي اتباع دوافعهم الانتحاريه.
-- واوضحت دراسات اخري ان مستوي السيروتونين يرتفع حسب كميه اشعه الشمس التي يتعرض لها الشخص</td></tr>
<tr><td>Cluster 3 (statistics and trends)</td><td colspan="3">وتبين الاحصائيات ان عدد محاولات الانتحار ارتفع بنسبه منذ وان معظم من اقدموا علي الانتحار كانوا من الرجال</td></tr>
<tr><td>Ranking</td><td>Cluster 1:
Sentence 1: 0.89
Sentence 2: 0.94
Sentence 3: 0.87</td><td>Cluster 2:
Sentence 4: 0.95
Sentence 5: 0.91</td><td>Cluster 3:
Sentence 6: 1.00</td></tr>
<tr><td>Selected sentence</td><td colspan="3">The selected sentences were [2, 4, 6], corresponding to the highest scores of [0.94, 0.95, 1.00],</td></tr>
<tr><td>Generated summary</td><td colspan="3">"وقال باحثون بريطانيون ان عدد حالات الانتحار يزيد في شهر مايو المشمس ليكون اكثر من اي شهر اخر وهم يعتقدون ان الامر راجع الي حاله الطقس. وتقول مجموعه برايوري المتخصصه في بحوث الطب النفسي ان الطقس المشمس الذي عاده ما يساعد الناس في التغلب وتبين الاحصاءيات ان عدد .علي كابتهم يعطيهم كذلك القدره علي اتباع دوافعهم الانتحاريه ".محاولات الانتحار ارتفع بنسبه منذ وان معظم من اقدموا علي الانتحار كانوا من الرجال</td></tr>
<tr><td>Reference summary</td><td colspan="3">"وقال باحثون بريطانيون ان عدد حالات الانتحار يزيد في شهر مايو المشمس ليكون اكثر من اي شهر اخر وهم يعتقدون ان الامر راجع الي حاله الطقس. ويقول المسءول عن الخدمات الصحيه في المجموعه البروفيسور كريس تومسون ان هناك علاقه مباشره بين سطوع الشمس والانتحار. وتبين الاحصاءيات ان عدد محاولات الانتحار ارتفع بنسبه منذ وان معظممن اقدموا علي الانتحار كانوا من الرجال."</td></tr>
</table>

F. References

[1] L. Abualigah, M. Q. Bashabsheh, H. Alabool, and M. Shehab, "Text Summarization: A Brief Review," Studies in Computational Intelligence, vol. 874, no. December 2019, pp. 1–15, 2020, doi: 10.1007/978-3-030-34614-0_1.

[2] A. K. Yadav, Ranvijay, R. S. Yadav, and A. K. Maurya, "State-of-the-art approach to extractive text summarization: a comprehensive review," Multimed Tools Appl, vol. 82, no. 19, pp. 29135–29197, 2023, doi: 10.1007/s11042-023-14613-9.

[3] Y. Albalawi, J. Buckley, and N. S. Nikolov, "Investigating the impact of pre-processing techniques and pre-trained word embeddings in detecting Arabic health information on social media," J Big Data, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00488-w.

[4] A. Elsaid, A. Mohammed, L. Fattouh, and M. Sakre, "Hybrid Arabic text summarization Approach based on Seq-to-seq and Transformer A Hybrid Arabic text summarization Approach based on Seq-to-seq and Transformer," pp. 1–21, 2023, [Online]. Available: https://doi.org/10.21203/rs.3.rs-2856782/v1

[5] Y. Jaafar, D. Namly, K. Bouzoubaa, and A. Yousfi, "Enhancing Arabic stemming process using resources and benchmarking tools," Journal of King Saud University - Computer and Information Sciences, vol. 29, no. 2, pp. 164–170, Apr. 2017, doi: 10.1016/j.jksuci.2016.11.010.

[6] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, "Arabic natural language processing: An overview," Journal of King Saud University - Computer and Information Sciences, vol. 33, no. 5, pp. 497–507, Jun. 2021, doi: 10.1016/j.jksuci.2019.02.006.

[7] M. Zaytoon, M. Bashar, M. A. Khamis, and W. Gomaa, "Amina: an Arabic multi-purpose integral news articles dataset," Neural Comput Appl, vol. 7, no. 0123456789, 2024, doi: 10.1007/s00521-024-10277-0.

[8] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," Trans Assoc Comput Linguist, vol. 5, pp. 135–146, 2017, doi: 10.1162/tacl_a_00051.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[10] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[11] K. Al-Sabahi and Z. Zuping, "Document Summarization Using Sentence-Level Semantic Based on Word Embeddings," International Journal of Software Engineering and Knowledge Engineering, vol. 29, no. 2, pp. 177–196, 2019, doi: 10.1142/S0218194019500086.

[12] W. Zhao, L. Zhu, M. Wang, X. Zhang, and J. Zhang, "WTL-CNN: a news text classification method of convolutional neural network based on weighted word embedding," Conn Sci, vol. 34, no. 1, pp. 2291–2312, 2022, doi: 10.1080/09540091.2022.2117274.

[13] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," arXiv preprint arXiv:2003.00104, 2020.

[14] F. Douzidia and G. Lapalme, “Lakhas, an Arabic summarization system,” in Proceedings of DUC2004, 2004, pp. 128–135. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.125.7420&rep=rep1&type=pdf
[15] A. M. Azmi and S. Al-Thanyyan, “A text summarizer for Arabic,” Comput Speech Lang, vol. 26, no. 4, pp. 260–273, 2012, doi: 10.1016/j.csl.2012.01.002.
[16] Y. A. Jaradat and A. T. Al-Taani, “Hybrid-based Arabic single-document text summarization approach using genatic algorithm,” in 2016 7th International Conference on Information and Communication Systems, ICICS 2016, IEEE, 2016, pp. 85–91. doi: 10.1109/IACS.2016.7476091.
[17] R. Z. Al-Abdallah and A. T. Al-Taani, “Arabic Single-Document Text Summarization Using Particle Swarm Optimization Algorithm,” Procedia Comput Sci, vol. 117, pp. 30–37, 2017, doi: 10.1016/j.procs.2017.10.091.
[18] M. El-Haj, U. Kruschwitz, and C. Fox, “Exploring clustering for multi-document Arabic summarisation,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7097 LNCS, no. February, pp. 550–561, 2011, doi: 10.1007/978-3-642-25631-8_50.
[19] A. K. Habboush et al., “Arabic text summarization model using clustering techniques,” Researchgate.Net, vol. 2, no. 3, pp. 62–67, 2012, [Online]. Available: https://www.researchgate.net/profile/Ahmad-
[20] H. Oufaida, O. Nouali, and P. Blache, “Minimum redundancy and maximum relevance for single and multi-document Arabic text summarization,” Journal of King Saud University - Computer and Information Sciences, vol. 26, no. 4, pp. 450–461, 2014, doi: 10.1016/j.jksuci.2014.06.008.
[21] N. Alami, M. Meknassi, S. Alaoui Ouatik, and N. Ennahnahi, “Arabic text summarization based on graph theory,” in Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA, 2016, pp. 1–8. doi: 10.1109/AICCSA.2015.7507254.
[22] R. Elbarougy, G. Behery, and A. El Khatib, “Extractive Arabic Text Summarization Using Modified PageRank Algorithm,” 2020. doi: 10.1016/j.eij.2019.11.001.
[23] G. Alselwi and T. Taşcı, “Extractive Arabic Text Summarization Using PageRank and Word Embedding,” Arab J Sci Eng, vol. 49, no. 9, pp. 13115–13130, 2024, doi: 10.1007/s13369-024-08890-1.
[24] N. Alami, M. Meknassi, and N. En-nahnahi, “Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning,” Expert Syst Appl, vol. 123, pp. 195–211, 2019, doi: 10.1016/j.eswa.2019.01.037.
[25] S. Abdulateef, N. A. Khan, B. Chen, and X. Shang, “Multidocument Arabic text summarization based on clustering and word2vec to reduce redundancy,” Information (Switzerland), vol. 11, no. 2, p. 59, 2020, doi: 10.3390/info11020059.
[26] W. Antoun, F. Baly, and H. Hajj, “AraBERT: Transformer-based Model for Arabic Language Understanding,” 2020, [Online]. Available: http://arxiv.org/abs/2003.00104

[27] A. Alshanqiti, A. Namoun, A. Alsughayyir, A. M. Mashraqi, A. R. Gilal, and S. S. Albouq, "Leveraging DistilBERT for Summarizing Arabic Text: An Extractive Dual-Stage Approach," IEEE Access, vol. 9, pp. 135594–135607, 2021, doi: 10.1109/ACCESS.2021.3113256.
[28] R. Rani and D. K. Lobiyal, "A weighted word embedding based approach for extractive text summarization," Expert Syst Appl, vol. 186, no. June, p. 115867, 2021, doi: 10.1016/j.eswa.2021.115867.
[29] E. Yulianti, N. Pangestu, and M. A. Jiwanggi, "Enhanced TextRank using weighted word embedding for text summarization," International Journal of Electrical and Computer Engineering, vol. 13, no. 5, pp. 5472–5482, 2023, doi: 10.11591/ijece.v13i5.pp5472-5482.
[30] O. Obeid et al., "CAMeL tools: An open source python toolkit for arabic natural language processing," LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, pp. 7022–7032, 2020.
[31] M. El-Haj, U. Kruschwitz, and C. Fox, "Exploring clustering for multi-document Arabic summarisation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7097 LNCS, Dubai, United Arab Emirates: Springer-Verlag, 2011, pp. 550–561. doi: 10.1007/978-3-642-25631-8_50.
[32] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Proceedings of the workshop on text summarization branches out (WAS 2004), 2004, pp. 25–26. [Online]. Available: papers2://publication/uuid/5DDA0BB8-E59F-44C1-88E6-2AD316DAEF85
[33] E. Lloret and M. Palomar, "Text summarisation in progress: A literature review," Artif Intell Rev, vol. 37, no. 1, pp. 1–41, 2012, doi: 10.1007/s10462-011-9216-z.