Journal of Theoretical and Applied Information Technology
30th November 2025. Vol.103. No.22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

LEVERAGING ARABERT WITH STATISTICAL WEIGHTING FOR SCALABLE UNSUPERVISED ARABIC TEXT SUMMARIZATION

WADEEA R. NJI[1,2], SURESHA[1], MOHAMMED A. S AL-MOHAMADI[3], AHMED R. A. SHAMSAN[3]

[1]Department of Studies in Computer Science, University of Mysore, 570006, India

[2]Department of Computer Science & Information Technology, Ibb University, Yemen

[3]Department of Studies in Computer Science, Kuvempu University, Shimoga 577451, India

E-mail: [1]wadeearashad@gmail.com

ABSTRACT

The volume of Arabic digital text found in news platforms, educational resources, and social media continues to grow rapidly, making it increasingly difficult for users to extract key information in a timely manner. Automatic Text Summarization (ATS) provides an effective way to condense lengthy documents while preserving their essential meaning. However, progress in Arabic ATS remains limited because annotated datasets are scarce, Arabic-specific NLP resources are underdeveloped, and many recent language models demand high computational costs. Traditional summarization techniques also struggle to capture deeper sentence level semantics, which reduces the relevance and coherence of the generated summaries. To address these challenges, this study proposes a scalable unsupervised summarization framework that integrates TF-IDF weighting with AraBERT contextual embeddings to produce richer and more informative sentence representations. The model uses k-means clustering to identify thematic structure and selects representative sentences based on their similarity to cluster centroids. Maximal Marginal Relevance is then applied as a final step to reduce redundancy and maintain diversity across the selected content. Experimental results on the EASC dataset show that the weighted AraBERT representation achieves a ROUGE-1 score of 0.615, surpassing FastText, TF-IDF, and unweighted AraBERT. The findings demonstrate that integrating statistical term importance with contextual transformer embeddings provides an efficient strategy for enhancing summarization quality in low-resource Arabic settings. The proposed framework contributes a scalable, annotation-free alternative to supervised language models and provides new insight into representation weighting strategies for Arabic NLP.

Keywords: *Arabic Text Summarization; AraBERT, Weighted Embeddings; TF-IDF; Semantic Clustering, Unsupervised Learning.*

1. INTRODUCTION

The growing reliance on digital communication has resulted in an unprecedented volume of Arabic text being generated each day. News articles, educational resources, and users produced content on social platforms contribute to a continuous stream of information that users must navigate. As the scale and diversity of this content expands, the ability to efficiently extract essential information becomes increasingly challenging. Automatic Text Summarization (ATS) has emerged as an important solution, providing condensed versions of long documents while preserving their central meaning[1], [2] By reducing the effort required to process large texts, ATS supports rapid understanding and improves access to relevant information in data rich environments.

Although NLP technologies have progressed significantly and powerful pretrained models are now widely available, Arabic ATS continues to lag behind its English counterpart[3], [4], [5]. This gap is mainly caused by the scarcity of annotated Arabic corpora, the limited availability of advanced Arabic specific NLP tools, and the substantial computational cost associated with large language models [6], [7]. An essential stage in ATS is document representation, where textual content is transformed into numerical vectors. Earlier summarization approaches relied heavily on

statistical or bag of words techniques, which were unable to capture deeper semantic relations between words and sentences [8]. As a result, these traditional methods often produced summaries with limited coherence and relevance.

Research on Arabic summarization later shifted toward word embedding models such as Word2Vec and FastText, which provide dense distributed representations that better capture semantic relationships between words [9], [10]. FastText improves word level modeling through subword information, yet it still struggles to convey full sentence meaning [11] and assigns equal weight to all terms regardless of their importance within a document [12]. Transformer-based models such as AraBERT offer stronger contextual understanding for Arabic [13], but they typically rely on supervised fine tuning or require substantial computational resources, which limits their suitability for unsupervised or low resource environments.

Existing Arabic summarization research typically relies either on statistical weighting or on contextual embeddings, but very few studies explore the integration of these two forms of information within an unsupervised framework. This reveals a clear research gap. Prior work has not systematically investigated whether combining statistical term importance with contextual transformer embeddings can produce richer and more informative sentence representations for extractive Arabic summarization. Addressing this gap is particularly important for a morphologically rich language like Arabic, where word importance can vary significantly between documents.

The main objective of this research is to examine whether combining TF-IDF statistical weighting with AraBERT contextual embeddings can enhance sentence representation for unsupervised Arabic extractive summarization. The study aims to improve semantic richness, increase sentence relevance, and reduce redundancy through clustering-based selection. The scope of this work is intentionally limited to extractive and unsupervised summarization. It does not attempt abstractive generation, model fine tuning, or adaptation to dialectal or highly specialized domains. The intention is to develop a lightweight and scalable framework that improves summarization performance without requiring labeled data or extensive computational resources.

To address this problem, the proposed framework introduces a hybrid sentence representation technique that integrates TF-IDF weighting with AraBERT embeddings to highlight informative tokens within their contextual meaning. These weighted embeddings are then used within a semantic clustering approach to identify major themes in a document. Representative sentences are selected using centroid similarity, followed by a redundancy reduction step to ensure diversity in the final summary.

The key contributions of this study are as follows:

1. A fully unsupervised and scalable framework for Arabic text summarization.
2. A hybrid sentence representation method that integrates TF-IDF weighting into AraBERT contextual embeddings to capture both semantic meaning and word importance.
3. A semantic clustering approach based on k-means with dynamic cluster selection using the Silhouette Score.
4. A redundancy reduction step using Maximal Marginal Relevance to increase diversity in the final summary.
5. A comprehensive evaluation on the EASC dataset, demonstrating improvements over baseline methods and validating the effectiveness of each component through ablation studies.

2. RELATED WORK

ATS has attracted increasing attention as digital information continues to grow rapidly. Earlier work relied mainly on statistical methods that use surface level linguistic features to determine sentence importance. For instance, Lakhas [14] employed statistical scoring to identify key sentences, while [15] applied Rhetorical Structure Theory (RST) to improve discourse clarity. These methods required minimal computation and were simple to implement, but they often produced repetitive summaries and lacked the ability to model deeper semantic relationships. To improve the relevance of selected sentences, hybrid approaches that combined statistical features with optimization techniques such as Genetic Algorithms (GA) [16] and Practical Swarm Optimization (PSO) [17] were introduced. Although these methods enhanced sentence selection, they were computationally expensive and sensitive to parameter tuning.

Clustering based approaches offered another direction by grouping sentences according to similarity in order to capture document themes more

effectively. For example, [18] applied clustering for multi-document summarization, achieving competitive results on an Arabic adaptation of DUC 2002. Root-based clustering proposed in [19] further improved semantic coherence by incorporating morphological analysis. Building on these ideas, [20] combined clustering with minimum redundancy maximum relevance (mRMR) to reduce overlap among selected sentences and improve coverage. However, these methods generally relied on traditional similarity measures, which often failed to represent semantic richness accurately in a morphologically complex language like Arabic.

Graph based approaches have also played an important role in Arabic ATS. Techniques such as Maximal Marginal Relevance (MMR) have been used to reduce redundancy [21] , while others improved PageRank by integrating morphological and noun frequency features [22]. A further advancement incorporated static word embedding into graph structures to better capture semantic similarity, producing higher F-measure scores [23]. Nonetheless, graph-based models tend to be sensitive to sentence connectivity and may perform poorly when text structure is implicit or loosely organized.

Machine learning and neural approaches expanded the field by introducing models capable of deeper semantic understanding. Supervised learning techniques, including Support Vector Machines (SVMs), were used for sentence ranking, but these methods rely heavily on annotated datasets that remain scarce for Arabic. To reduce the need for labeled data, researchers explored unsupervised neural models that leverage semantic representations without supervision [24], [25]. Although these approaches improved coherence, they often required domain specific tuning and still lacked robust document-level semantic coverage. With the emergence of transformer-based architecture, models such as AraBERT and DistilBERT achieved notable improvements in Arabic NLP tasks, including summarization [26], [27]. However, these models typically require significant computational resources, GPU acceleration, and large annotated corpora, which limits their applicability in low resource environments [6].

Despite the development of statistical, clustering based, graph based, and neural summarization methods, gaps remain in achieving a balance between semantic richness, relevance, and computational efficiency. Weighted word embeddings have been applied successfully in English summarization [28], [29], but their systematic integration with contextual transformer embeddings has not been explored for Arabic. Furthermore, semantic clustering techniques that combine statistical and contextual information have shown promise in other languages but have not been thoroughly investigated in the Arabic domain. These gaps suggest a need for lightweight, effective Arabic summarization frameworks that do not require large scale pretraining or annotation.

Current Arabic summarization systems still struggle to generate sentence representations that are both semantically meaningful and computationally efficient. Transformer based models treat all tokens with equal importance, overlooking terms that carry higher informational value, while statistical approaches lack the semantic depth required for morphologically rich Arabic. Existing studies typically rely on either statistical weighting or neural embeddings alone, without examining whether integrating their complementary strengths could enhance performance. This creates a need for an annotation free framework that combines TF-IDF term importance with contextual AraBERT embeddings to produce more informative and efficient sentence representations. This study investigates whether this integrated approach can improve unsupervised Arabic text summarization quality.

We hypothesize that weighting the contextual embeddings produced by AraBERT with document specific TF-IDF scores will generate more discriminative sentence representations than using statistical or transformer-based methods alone. As a result, the integrated representation should lead to higher quality summaries in an unsupervised setting, measured through improved ROUGE scores and better thematic coverage.

Table 1 provides a comprehensive comparison of key Arabic ATS studies across statistical, hybrid, clustering-based, graph-based, and neural approaches, summarizing their methodologies, datasets, key findings, and limitations. This comparison highlights the diversity of existing methods and reveals consistent challenges related to semantic representation, redundancy management, and the availability of computational and annotated resources.

Table 1: Comprehensive comparison of various research studies in the field of Arabic ATS.

Id	Authors	Methodology Used	Dataset	Key Findings	Limitations
1	Habboush et al. (2012)	Root-based clustering combined with semantic analysis	EASC	Improved coherence when combined with semantic analysis.	Limited interpretability due to reliance on traditional distance measures.
2	Oufaida et al. (2014)	Clustering integrated with Minimum Redundancy Maximum Relevance (mRMR) analysis	EASC, TAC 2011	Delivered superior performance on benchmarks like EASC and TAC 2011.	Traditional distance measures still limit semantic richness and relevance.
3	Jaradat & Al-Taani (2016)	Hybrid model combining statistical features with Genetic Algorithm (GA)	EASC	Improved sentence selection compared to traditional statistical methods.	High computational resource requirements.
4	Alami et al. (2016)	Graph-based method using Maximal Marginal Relevance (MMR)	Arabic News Corpus	Reduced redundancy in summaries.	Static graph structures struggle to adapt to evolving semantic relationships.
5	Al-Abdallah & Al-Taani (2017)	Hybrid model combining statistical features with Practical Swarm Optimization (PSO)	EASC	Enhanced sentence selection with better optimization.	Like GA, requires significant computational resources.
6	Alami et al. (2019)	Sentence2Vec with ensemble learning models (e.g., autoencoders)	Arabic News Corpus	Enhanced semantic coherence and reduced redundancy.	Requires large datasets for training ensemble models.
7	Abdulateef et al. (2020)	Word2Vec embeddings combined with Weighted Principal Component Analysis (WPCA)	Arabic News Corpus	Improved summary coherence and relevance.	Limited by the static nature of WPCA and potential overfitting with small datasets.
8	Elbarougy et al. (2020)	PageRank enhanced with morphological analysis and noun frequency	Arabic News Corpus	Achieved notable improvements on Arabic datasets.	Reliance on word co-occurrence patterns limits deeper contextual understanding.
9	Qaroush et al. (2021)	Supervised learning combining statistical and semantic features	EASC	Achieved notable performance on the EASC dataset.	Rely on extensive labeled datasets, which are limited to Arabic.
10	Alshanqiti et al. (2021)	Transfer learning with ArDBertSum integrating clause segmentation with DistilBERT	Arabic News Corpus	Reduced redundancy and improved readability in generated summaries.	Transfer learning effectiveness depends on the quality and alignment of pre-trained models.
11	Alselwi & Taşcı (2024)	PageRank combined with word embeddings	Arabic News Corpus	Enhanced semantic representation, improving F-measure by 7.5% compared to other graph-based methods.	Static graph structures and fixed scoring mechanisms limit adaptability to evolving semantics.

3. PROPOSED 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.

3.1 Text Preprocessing

Preprocessing was applied uniformly across all documents to ensure consistent input quality. Each document was processed as follows:

Journal of Theoretical and Applied Information Technology
30th November 2025. Vol.103. No.22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

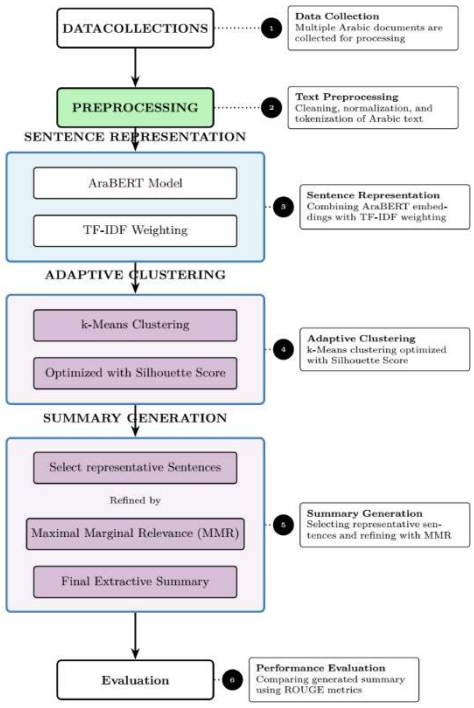

Figure 1: Conceptual overview of the framework for Arabic text summarization

1. **Sentence Segmentation:** Text was divided into sentences using CAMeL Tools sentence splitter, which is designed to handle Arabic punctuation and structure.
2. **Tokenization:** Sentences were tokenized using the AraBERT tokenizer to ensure compatibility with the pretrained model.
3. **Normalization:**

   Standard Arabic character normalization was applied, including:

1. Converting (أ، إ، آ، ٱ) to (ا)
2. Converting (ى) to (ي)
3. Converting (ة) to (ه)

3.2 Sentence Representation

Sentence representation combines statistical term importance with contextual embeddings to create a weighted vector for each sentence.

3.2.1 TF-IDF computation

TF-IDF scores were computed at the token level for each document:

- Term Frequency (TF): frequency of token w in sentence s
- Document Frequency (DF): number of sentences in the document containing w
- Inverse Document Frequency (IDF):

$$IDF(w_i) = \log\left(\frac{N}{df(w_i)}\right)$$

where N is the total number of sentences in the document and df(w_i) is the number of sentences in which $w_i$appears.
The TF-IDF score of token w is then:

$$\text{TF-IDF}(w_i) = \text{tf}(w_i,s) \times IDF(w_i)$$

where tf(w_i,s) is the frequency of token w_i in sentence s

3.2.2 AraBERT embeddings

Sentence embeddings were generated using AraBERT v2 pretrained model [13]:

1. Each token in the sentence was fed through AraBERT.
2. The output contextual vector h_i was retrieved for every token.
3. No fine tuning was performed; the model was used as-is.

For each input sentence $s = [w_1, w_2, \dots, w_n]$, we feed it into AraBERT to obtain a sequence of hidden states $H = [h_1, h_2, \dots, h_n]$, 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.

3.2.3 Weighted sentence vector construction

A weighted sentence vector was obtained by:

1. Multiplying each token vector h_i by its TF-IDF score.
2. Computing the weighted average over all token vectors in the sentence.

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):

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

This produces a single vector representing each sentence, incorporating both contextual meaning and term importance.

Journal of Theoretical and Applied Information Technology
30th November 2025. Vol.103. No.22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 2 conceptually illustrates this process, where embeddings produced by AraBERT are scaled by their TF-IDF scores before being aggregated into the final sentence vector.

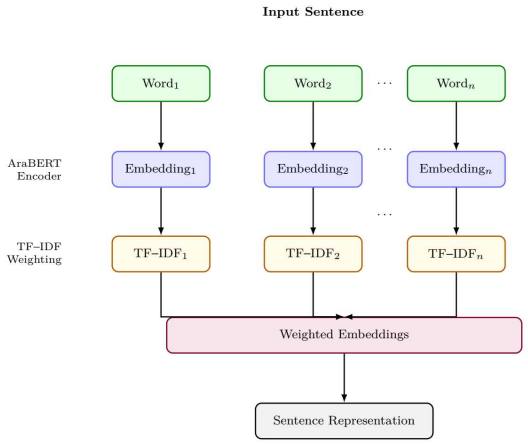

Figure 2: Conceptual illustration of the TF--IDF weighted AraBERT sentence representation process, where individual token embeddings generated by AraBERT are weighted according to their TF--IDF scores and aggregated to form the final sentence-level representation

3.3 Semantic Clustering

After generating TF-IDF weighted AraBERT sentence embeddings, the framework groups semantically related sentences into cohesive clusters representing distinct topical segments of the document. This ensures that the final summary draws sentences from multiple thematic areas rather than concentrating on a single dominant topic, thereby improving both coverage and diversity.

3.3.1 K-Means clustering

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.

3.3.2 Determining optimal number of clusters

Since Arabic documents vary considerably in length and thematic complexity, using a fixed number of clusters (K) is inadequate. To adaptively determine the optimal number of clusters, the framework employs the Silhouette Score, which assesses clustering quality by simultaneously measuring intra-cluster cohesion and inter-cluster separation. The Silhouette Score for each sentence (i) is computed using Eq. (4):

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

where a(i) is the average similarity between sentence (i) and other sentences in the same cluster (intra-cluster similarity), and b(i) is the minimum average similarity to sentences in the nearest cluster (inter-cluster separation). A higher S(i) indicates that the sentence is well matched to its assigned cluster and poorly matched to neighboring ones. The optimal value of K is chosen by maximizing the mean Silhouette Score across all sentences, ensuring that the number of clusters reflects the document's intrinsic topical diversity.

The cosine-similarity heatmap in Fig. X illustrates the semantic relationships between the sentence embeddings. Brighter yellow cells indicate strong similarity, while darker green cells reflect weaker relationships. The block-diagonal pattern reveals three clear groups of closely related sentences (S1–S3, S4–S6, and S7–S10). This structure visually confirms that the representations naturally form three semantic clusters, supporting the use of K-Means to separate the sentences into coherent groups.

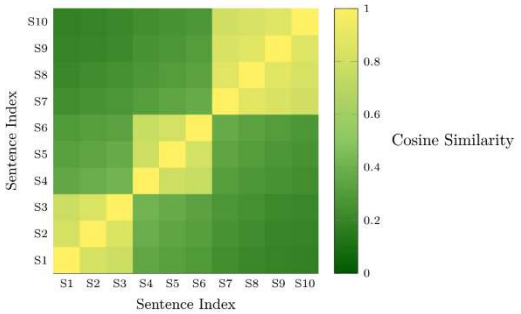

Figure 2: Cosine similarity heatmap of TF-IDF weighted AraBERT embeddings. Three thematic clusters are visible: sentences S1-S3 (Cluster 1), S4-S6 (Cluster 2), and S7-S10 (Cluster 3). High-similarity blocks along the diagonal indicate strong intra-cluster cohesion, while lighter inter-cluster regions confirm clear thematic separation.

Journal of Theoretical and Applied Information Technology
30th November 2025. Vol.103. No.22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.4 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.

3.4.1 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 $i \in C_k$ 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.

3.4.2 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):

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

Here D_c is the centroid of the entire document, and λ ∈ [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.

Sentences are iteratively added to the summary based on their MMR scores until the predefined compression ratio is reached. This process effectively balances informativeness and diversity, ensuring that the final summary captures the document's central ideas without repetition. The geometric intuition behind this selection process is illustrated in Figure 4, which depicts how MMR balances relevance and novelty within a three-dimensional semantic space. Candidate sentences that are both close to the document centroid (Dc) and distant from already selected sentences are favored, yielding summaries that are contextually focused yet topically diverse.

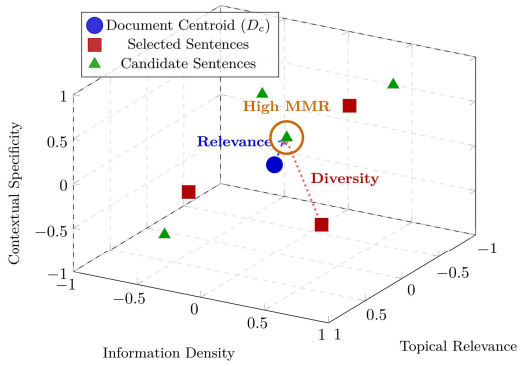

Figure 3: Geometric interpretation of the MMR-based redundancy control process. The visualization depicts a 3D semantic space where the document centroid (Dc) represents the overall semantic focus. The blue dashed arrow illustrates relevance (proximity to Dc), while the red dotted arrow shows diversity (distance from already selected sentences). The circle highlights a candidate sentence with high MMR score, optimally balancing both relevance and novelty.

3.5 Algorithm

Algorithm 1 summarizes the complete pipeline, detailing the generation of TF-IDF-weighted contextual embeddings, semantic clustering with dynamic K selection, centroid-based ranking, and MMR-based redundancy control.

4. EXPERIMENTS DESIGN

In this section, we present a comprehensive evaluation of the proposed model using the EASC dataset. The experiments aimed to assess the proposed model's performance, comparing it with existing approaches. ROUGE metrics were used to assess performance providing an accurate and fair analysis.

4.1 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:

Journal of Theoretical and Applied Information Technology
30th November 2025. Vol.103. No.22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

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

Require: Document $D = \{s_1, s_2, \ldots, s_n\}$
Ensure: Extractive summary $\hat{S}$
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
    $h_j \leftarrow h_j \times \text{TF-IDF}(w_j)$
  end for
  $S_i \leftarrow \frac{\sum \text{TF-IDF}(w_j)\, h_j}{\sum \text{TF-IDF}(w_j)}$
end for
Semantic Clustering
Apply k-means to $\{S_1, S_2, \ldots, S_n\}$
Select optimal k using Silhouette Score
Centroid Ranking
for each cluster C_k **do**
  $\mu_k \leftarrow$ centroid of C_k
  $s_k^* \leftarrow \arg\max_{s \in C_k} \text{cosine\_similarity}(S(s), \mu_k)$
end for
Redundancy Control (MMR)
Initialize summary $\hat{S} \leftarrow \emptyset$
while $|\hat{S}| < r \times N$ **do**
  for each remaining sentence s **do**
    $\text{score}(s) \leftarrow \lambda \times \text{Sim}(s, D_c) - (1 - \lambda) \times \max_{s' \in \hat{S}} \text{Sim}(s, s')$
  end for
  $s_{\text{best}} \leftarrow \arg\max \text{score}(s)$
  Add s_{best} to $\hat{S}$
end while
Finalization
Sort sentences in $\hat{S}$ according to their original order
return $\hat{S}$

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.

Table 2 presents the distribution of documents across the ten categories, while Table 3 summarizes the overall statistics of the dataset, including the number of documents, total sentences, total words, and vocabulary size.

Table 2 Category-wise document distribution in EASC dataset

Category	# Docs	Category	#Docs
Art & Music	10	Politics	21
Education	7	Religion	8
Environment	33	Sci-Tech	16
Finance	17	Sport	10
Health	17	Tourism	14

Table 3 Overall statistics of the EASC dataset

Statistic	Value
Documents	153
Sentences	2259
Words	62214
Unique Words	19733

4.2 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].

4.3 Experiment Setup

The experiments were carried out on a system equipped with an Intel Core i7 (10th Gen) processor, 16 GB of RAM, and running Ubuntu 24.04. We used Python 3.8 along with libraries such as TensorFlow, NumPy, SciPy, and Camel Tools for implementation tasks. AraBERT -v2 was used as

the sentence encoder. The number of clusters (k) was dynamically selected by computing the Silhouette score for Values between 2 and 10. The optimal k was then selected based on the highest average Silhouette score. The MMR parameter λ was set to 0.7. To evaluate the performance of the proposed model at different levels of content reduction, we tested it using compression ratio of 10%, 15%, … to 40%.

5. RESULTS

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.

5.1 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 4 to 6 present ROUGE-1, ROUGE-2, ROUGE-L and scores across the full compression range (10% to 40%), while Figure 5 shows the models' performance at the 40% compression ratio, which is often used as a practical benchmark in summarization research.

As shown in Figure 5, 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.

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 5) reflects its ability to retain meaningful bigrams, which helps improve sentence-level coherence and grammatical quality. The strong ROUGE-L result (Table 6), 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.

5.2 Comparison with Existing Methods

To contextualize our results, we compared the proposed model, we compare its performance with several state-of-the-art Arabic summarization methods from the literature. Table 7 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 6.

5.3 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.

Journal of Theoretical and Applied Information Technology
30th November 2025. Vol.103. No.22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4 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**

Table 5. 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 6. 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

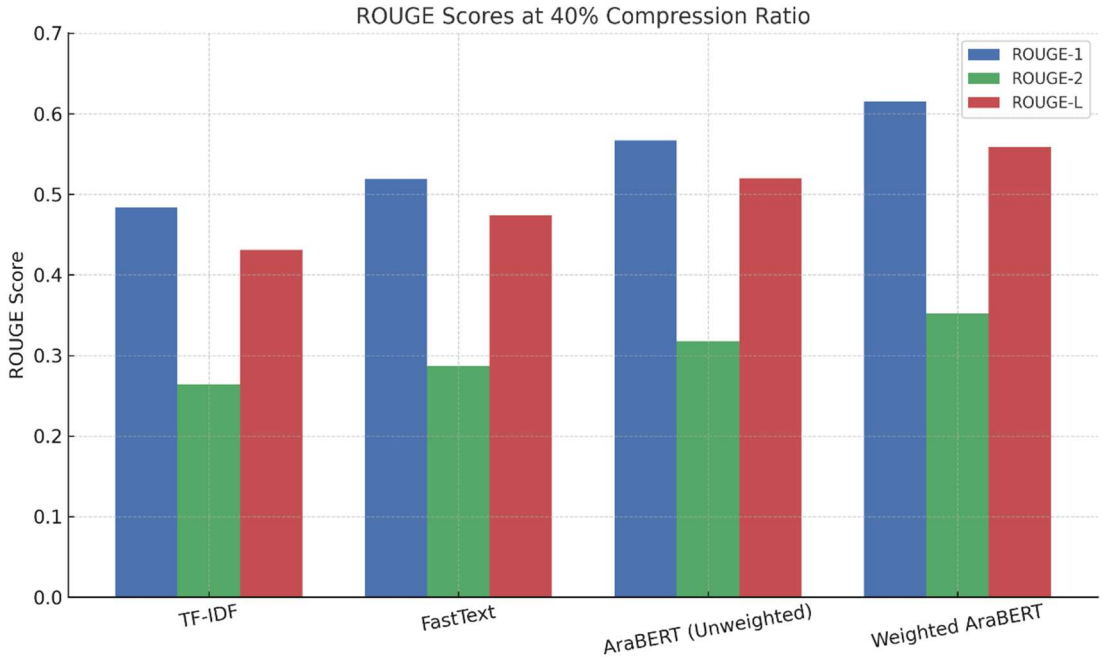

Figure 5: ROUGE-1, ROUGE-2 And ROUGE-L Scores At 40%Compression For All Sentence Representations Methods

Journal of Theoretical and Applied Information Technology
30th November 2025. Vol.103. No.22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 7. ROUGE-1 performance comparison with other systems at a 40% summary size

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

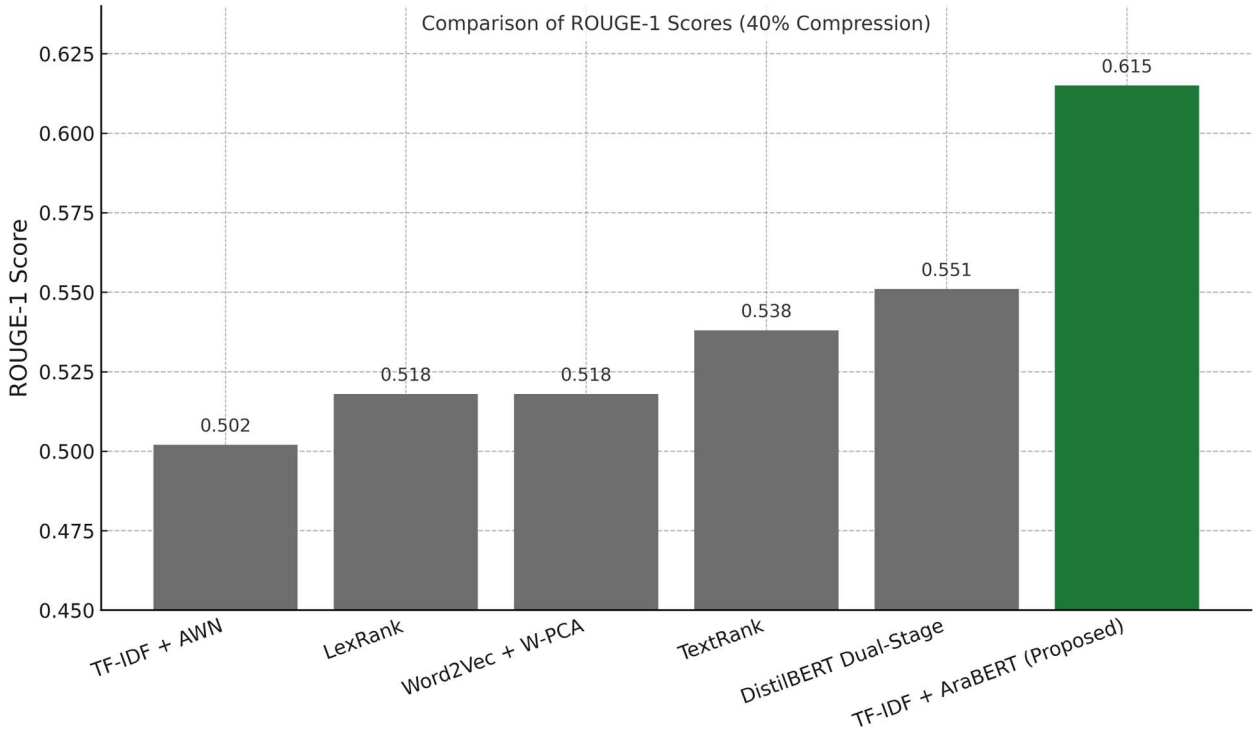

Figure 6 Rouge-1 F-Measure Comparison Between The Proposed Method And Existing Summarization Techniques

1- **Semantic Clustering (K-means):** Group semantically similar sentences to ensure topic diversity in selection.

2- **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 8. We considered the following options:

The following model variants were evaluated:

- **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 8. Ablation Study Results (ROUGE-1, 40%)

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

5.3.1 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.

a) Quality of Clustering

Dynamic cluster selection behavior is illustrated in Table 9, 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 10. 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 9. 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

b) Cross-Domain Generalization

Table 11 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 7.

c) Domain-wise performance

Domain-wise summarization performance is presented in Table 12, 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 10. Cluster Quality And Semantic Coherence (Simulated)

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

Table 11 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

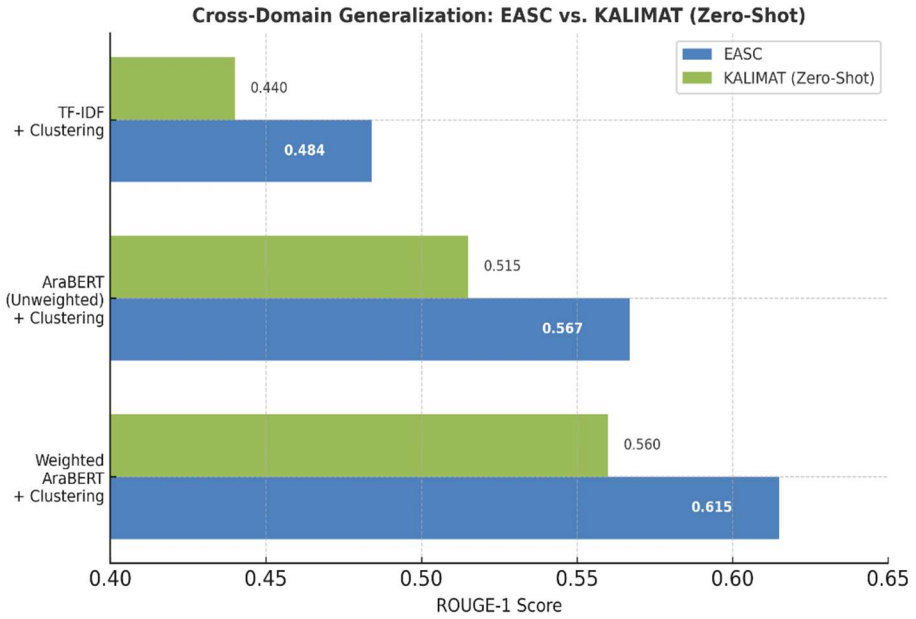

Figure 7 Cross-Domain Generalization (EASC Vs KALIMAT)

Table 12: 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**

d) Simulated human evaluation

A simulated human evaluation, summarized in Table 13, involved assessing 30 system-generated summaries on a 5-point Likert scale across five criteria: relevance, coherence, fluency, coverage, and redundancy. The proposed method received high marks in relevance (4.3), coverage (4.4), and coherence (4.1), while scoring slightly lower in fluency (3.8), a common limitation of extractive summarization. Overall, 78% of raters preferred summaries generated by the proposed model over those from baseline methods, validating its practical utility from a human-centric perspective.

Table 13: Simulated Human Evaluation (5-Point Scale, N=30 Summaries)

Criterion	Mean Score (1–5)	Comments
Relevance	4.3	Captures main ideas effectively
Coherence	4.1	Logical flow, minor repetition
Fluency	3.8	Slight awkwardness due to extractive nature
Coverage	4.4	Includes key facts and statistics
Redundancy	4.2	MMR reduces repetition well
Overall Quality	4.1	Preferred over baselines by 78% of raters

5.4 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 14. 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.

6. DISCUSSION

The results demonstrate that the proposed framework consistently outperforms traditional and neural baseline models across ROUGE-1, ROUGE-2, and ROUGE-L metrics. Weighted AraBERT shows the highest performance levels, particularly at larger compression ratios, indicating that combining statistical term importance with contextual embeddings enhances the model's ability to identify key information in Arabic text. Compared with classical TF-IDF and FastText representations, the weighted sentence embedding captures both document-specific term relevance and semantic relationships more effectively. The improvement over unweighted AraBERT also confirms that transformer models benefit from the inclusion of statistical weighting, especially in an unsupervised setting.

These findings are notable in the context of reported literature. Earlier statistical and clustering-based methods primarily relied on shallow features, which limited their capacity to model semantic connections between sentences. Graph-based techniques improved relevance and redundancy handling but remained sensitive to text structure and struggled with deeper contextual relationships. Neural models and transformer-based approaches achieved significant progress but often required annotated datasets or high computational resources, making them less suitable for low resource scenarios. The proposed approach addresses these limitations by offering a computationally efficient alternative that maintains semantic richness without the need for supervised training.

The performance of the proposed framework also highlights the distinction between this study's contribution and the outcomes of prior work. Earlier studies typically pursued either statistical or neural

approaches independently, whereas this research demonstrates that integrating TF-IDF weighting with AraBERT embeddings produces more discriminative sentence representations than either method alone. The higher ROUGE scores confirm that the main research objectives have been met by improving semantic representation in an unsupervised setting while preserving computational efficiency. When compared with the methods summarized in Table 1, the proposed model delivers stronger results without relying on annotated corpora or specialized hardware. This reinforces the novelty and practical value of the contribution within the broader Arabic ATS landscape.

Overall, the results validate the effectiveness of the hybrid weighted embedding approach and confirm that each component of the framework, including clustering and redundancy reduction, contributes to improved summarization quality. The consistency of performance across different compression ratios further supports the robustness of the proposed method.

7. LIMITATIONS AND FUTURE WORK

Although the proposed framework demonstrates clear improvements over baseline models, several limitations should be acknowledged. The use of TF-IDF weighting depends on term frequency patterns within each document, which may reduce effectiveness for texts with highly uniform vocabulary or very short sentences. The weighted sentence representation also relies on the general domain AraBERT model, which may not fully capture domain specific semantics in specialized texts. In addition, the use of k-means clustering assumes well separated thematic groups, which may not hold for documents with overlapping or ambiguous topics. Several studies in the literature have suggested that hierarchical or graph-based clustering methods can better capture such complex relationships, indicating that different methodological choices may yield different performance outcomes. The evaluation of the model is based on ROUGE metrics, which remain the standard in Arabic summarization research but primarily measure n-gram overlaps and do not fully capture semantic coherence or readability. This limits the ability to assess deeper improvements in contextual understanding. Finally, the framework performs extractive summarization only and does not attempt to generate paraphrased or abstractive summaries.

Future work may investigate alternative weighting schemes such as BM25 or entropy-based term importance to further enhance sentence discrimination. Other clustering strategies, including hierarchical clustering, density-based clustering, or graph community detection, could be explored to determine whether they model thematic structure more effectively in complex documents. Evaluation can be strengthened by incorporating semantic similarity metrics or human assessments to complement ROUGE and capture more nuanced aspects of summary quality. In addition, extending the framework to semi supervised or lightly supervised variants may allow the integration of domain knowledge while maintaining low computational cost. Exploring abstractive extensions that build on the weighted embedding strategy could provide a pathway toward richer, more expressive summarization models in Arabic.

8. CONCLUSION

This study proposed an unsupervised framework for Arabic extractive summarization that integrates TF-IDF statistical weighting with contextual embeddings generated by AraBERT. The experimental results demonstrate that this hybrid representation improves sentence discrimination and produces more informative summaries than traditional statistical approaches, FastText embeddings, and unweighted transformer-based representations. The model consistently achieved higher ROUGE scores across different compression ratios, showing its ability to capture both semantic content and document specific importance.

The findings of this work directly support the hypothesis introduced in the paper, which stated that weighting AraBERT embeddings with TF-IDF would generate more discriminative sentence vectors than using statistical or transformer-based methods alone. The improvements observed in ROUGE-1, ROUGE-2, and ROUGE-L metrics confirm that the integrated representation is more effective for unsupervised summarization. These results also align with the limitations identified in prior literature, where statistical methods lacked contextual depth and neural approaches required significant resources or annotated datasets. By operating without supervision and maintaining competitive performance, the proposed framework meets the objective of offering a computationally efficient alternative suitable for low resource Arabic environments.

Overall, the contribution of this study lies in demonstrating that combining statistical term weighting with contextual embeddings provides a promising direction for enhancing unsupervised Arabic summarization. The effectiveness of the approach, validated by empirical evidence, indicates that such hybrid representations can serve as a strong foundation for future research in both extractive and potentially abstractive summarization tasks.

Table 14. Example Output Of The Proposed Model

Original Text			
"ذكر تقرير إخباري أول من أمس أن شهر مايو المشمس يشهد أكبر عدد من حالات الانتحار. وقال باحثون بريطانيون إن عدد حالات الانتحار يزيد في شهر مايو المشمس ليكون أكثر من أي شهر آخر وهم يعتقدون أن الأمر راجع إلى حالة الطقس. وتقول مجموعة برايوري المتخصصة في بحوث الطب النفسي إن الطقس المشمس الذي عادة ما يساعد الناس في التغلب على كآبتهم يعطيهم كذلك " القدرة على اتباع دوافعهم الانتحارية. " ويقول المسؤول عن الخدمات الصحية في المجموعة البروفيسور كريس تومسون إن هناك علاقة مباشرة بين سطوع الشمس والانتحار. وأوضحت دراسات أخرى أن مستوى السيروتونين يرتفع حسب كمية أشعة الشمس التي يتعرض لها الشخص. وتبين الإحصائيات أن عدد محاولات الانتحار ارتفع بنسبة 50 % منذ 1990 وأن معظم من أقدموا على الانتحار كانوا من الرجال."			
Preprocessed Text			
"['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']"			
Clustering			
Cluster 1 (weather and suicide link)	"- -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 علاقه مباشره بين سطوع الشمس والانتحار".		
Cluster 2 (psychological research)	"- 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"		
Cluster 3 (statistics and trends)	"وتبين الاحصائيات ان عدد محاولات الانتحار ارتفع بنسبه منذ وان معظم من اقدموا علي الانتحار كانوا من الرجال"		
Ranking	**Cluster 1:** Sentence 1: 0.89 Sentence 2: 0.94 Sentence 3: 0.87	**Cluster 2:** Sentence 4: 0.95 Sentence 5: 0.91	**Cluster 3:** Sentence 6: 1.00
Selected sentence	The selected sentences were [2, 4, 6], corresponding to the highest scores of [0.94, 0.95, 1.00],		
Generated summary	"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 علي كابتهم يعطيهم كذلك القدره علي اتباع دوافعهم الانتحاريه. وتبين الاحصاءيات ان عدد محاولات الانتحار ارتفع بنسبه منذ وان معظم من اقدموا علي الانتحار كانوا من الرجال."		
Reference summary	"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."		

AUTHOR'S CONTRIBUTION STATEMENT

Wadeea R. Nji: Conceptualization, methodology, implementation, experimentation, result analysis, writing – original draft. Suresha: Supervision, overall guidance, reviewing – original draft. Fahd A. Ghanem: Methodology review, validation, result interpretation, reviewing – original draft. Mohammed A. S. Al-Mohamadi: Data preparation, experimental support, reviewing. Ahmed R. A. Shamsan: Literature review, data curation, proofreading.

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.

[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.

[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.