Journal of Al-Qadisiyah For Computer Science And Mathematics
JQCM
University Of AL-Qadisiyah

Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

A Systematic Review and Experimental Evaluation of Classical and Transformer-Based Models for Arabic Extractive Text Summarization

***Hind R. Almayyali*[1]*, Ahmed H. Aliwy*[2]**

[1] University of Kufa, Faculty of computer science and mathematics, Al-Najaf, Iraq, hindr.almayyali@student.uokufa.edu.iq:

[2]University of Kufa, Faculty of computer science and mathematics, Al-Najaf, Iraq,.ahmedh.almajidy@uokufa.edu.iq:

ARTICLE INFO

Article history:
Received: 31/10/2025
Rrevised form: 27/11/2025
Accepted : 30 /11/2025
Available online: 30 /12/2025

Keywords:

Arabic text summarization
Extractive text summarization
Abstractive summarization
Natural language processing,
Transformer models, Arabic NLP.

ABSTRACT

Arabic text summarization has become an active research area due to the rapid growth of Arabic digital content. developing effective summarization models has many challenges result from the linguistic richness, the complex morphology, flexible syntax, and diverse writing styles. This study looks at Arabic extractive summarization, where the goal is to select the most relevant sentences from a text to create a concise version that still captures the original meaning. Many techniques were analyzed and synthesized with the existing data sets in addition to identifying the problems and gaps to understand the history of the given area and determine the direction of the research. Also, the metrics that used to evaluate the outcomes of the Arabic text summarization are mentioned. The review pointed out an evident development of classical statistical and graph-based extractive procedures to current transformer-based procedures.

Transformer architectures have been quickly embraced by the field and Arabic-specific pre-trained models have proven to perform better than multilingual counterparts. Nevertheless, there are still great gaps in multi-document summarization, dialect management and formalized evaluation systems. The research work that is to be done in the future is the creation of bigger Arabic corpora, better dialectal coverage, and creation of full-fledged evaluation standards.

MSC..

https://doi.org/ 10.29304/jqcsm.2025.17.42562

1. Introduction

Automatic summarization, the task of summarizing long pieces of texts into concise and detailed summaries that embrace the key details, has a long history as a NLP task but it is still has serious challenges in computational linguistics for long documents even in recent technological developments. Many methods, approaches and algorithms were used for this task but all of them can be one of two categories; extractive or abstractive. Extractive text summarization approaches, which have yielded most successful applications, works by choosing actual salient sentences explicitly. Abstractive text summarization tries to produce summary based on human intuitive shortening

∗Corresponding author Hind R. Almayyali

Email addresses: hindr.almayyali@student.uokufa.edu.iq

Communicated by ‘sub etitor’

mechanisms. In the two types of approaches, the situation becomes more difficult when it comes to a language has complex grammatical construction such as Arabic. These languages have many challenges such as a rich morphological structure based on the complex structures of roots, the variety of regional forms which bring with them their own vocabulary and their own syntactic and morphological principles, and the flexibility of the sentence structure. Linguistic complexity like this makes automated summarization especially challenging since the summary should be extracted/produced through these complex structures without taking away the main content of the original message[1]. This complexity in language requires the creation of more advanced and language-aware methods since the current condensation systems do not perform well in capturing, processing, and representing the differences and subtleties of Arabic textual contents[2].

This scientific paper presents a quantitative methodological study of traditional methods and a transformer for extractive text in Arabic. It includes a comprehensive survey of recent studies, as well as the datasets used in Arabic summarization. The next sections of this paper discusses related works, the evaluation methods, the datasets, research gaps and limitations, discussion, conclusions, and the future works.

2. Illustrations

The history of Arabic Text Summarization (ATS) has long history since 2000s as part of Natural Language Processing tasks. The attempts at ATS were based on rule-based and statistical methods. These methods used linguistic rules to identify salient sentences based on the properties of sentence position, cue phrases, and word frequency to quarry the meaning information[3]. This period created the principles of extractive summarization where the summaries are made by choosing the sentences or phrases of the original text without creating any new information. Although they were initially useful, they had certain limitations in the face of the Arabic rich morphology and orthographic complexities which did not always submit to simple rules.

With the development of the statistical and machine learning methods, the ATS landscape changed, and the Classical Extractive Paradigm appeared. Scholars varied the general rules of summarization, shifting the focus on explicit linguistic rules to the data-oriented techniques. Some methods used Term Frequency-Inverse Document Frequency (TF-IDF) weighting as data representation which were popular and important to rank the words in a document[4] [5]. Graph-based ranking algorithms, such as TextRank, also became popular, with documents represented as a graph of sentences where the presence of a connection was considered a semantic similarity between the word and its neighbor, and hence the centrality of a sentence is used to determine the most representative sentence[4] [6]. These approaches were a major breakthrough, providing stronger and more scalable solutions than their predecessors, which were based on rules. Recently, deep learning and transformer-based models like AraBERT and AraBART, are used where it learn from large datasets.

2.1. Traditional Extractive Method

The Classical Extractive Paradigm of text summarization try to select key sentences from the original text. There are many approaches, such as statistical (TF-IDF) based, cue words based, clustering based, similarity and Graph-based approaches. The Graph-based and statistical approaches are based on a set of explicit feature engineering and rules to extract salient sentences. They work well with languages that have simpler morphology but do not represent Arabic's rich semantics and contextualities.

One of the foundations of statistical extractive summarization is Term Frequency- Inverse Document Frequency (TF-IDF). It gives weight to the terms based on their frequency in the documents (Term Frequency), and how rare they are in the corpus (Inverse Document Frequency), considering the importance of the sentences with high scores. In the case of Arabic, though, TF-IDF makes use of the surface-level word forms, and this frequency makes it generate redundant and incoherent summary outcomes. TF-IDF does not find as much semantic relationship as syntactic similarity [7], and thus fails to find the contextual meaning of ambiguous undiacritized words.

TextRank is a graph ranking algorithm based on a document being represented by a graph of sentences (nodes) and semantic similarity (edges). It calculates the score of an individual sentence based on the scores of the neighboring sentences. The weights of the edges are usually based on cosine similarity of sentence vectors, usually TF-IDF representations. The quality of the graph representation is the determinant of the effectiveness of TextRank[8]. This is not easy in the case of Arabic where the graph can distort actual semantic relationships when sentence representations do not represent morphological and diacritization ambiguities. The implication of the algorithm being static is that it does not support dynamic and context-sensitive re-weighting which is a significant constraint to a highly contextual language.

In case of clustering based approaches, they try to find clusters of semantically similar sentences then choose a representative sentence from each cluster (e.g., closest to the centroid) and combine those to form the summary. As algorithms K-Means, Hierarchical Agglomerative Clustering (with Ward linkage), and Affinity Propagation are used. These methods are based on word dense representations (Embedding) of sentences which are trained on a set of Arabic corpora using FastText models. Embeddings contain more information than raw TF-IDF, but the clustering algorithms need a fixed distance measure. They group the sentences in a similarity representation in a fixed-space that does not allow them to analyse the compositionality of meaning or narrative flow across clusters[9].

2.2. Neural and Transformer Approaches for Arabic

All tables should be numbered with Arabic numerals. Every table should have a caption. Headings should be placed above tables, left justified. Only horizontal lines should be used within a table, to distinguish the column headings from the body of the table, and immediately above and below the table. Tables must be embedded into the text and not supplied separately. Below is an example which the authors may find useful.

Classical ATS approaches have constraints of handling the complexity of Arabic language which were solved using neural and transformer-based models. These neural networks based on self-attention and deep contextual embeddings have important improvements. They can find complicated patterns in language content without caring the explicit feature engineering that makes language understanding deeper. AraBERT, mBERT, RoBERTa, and CAMeLBert-MSA are examples of transformer models that have been modified to Arabic.

AraBERT is a transformer based on Arabic-specific textthat pre-trained on a large and diverse Arabic corpus of 77 GB[10]. It has a Byte-Pair Encoding (BPE), a special tokenizer, to splits the complex words into sub-word tokens. It successfully tackles the complexity of Arabic Linguistic such as rich morphology patterns and high out-of-vocabulary levels through learning representations of frequent prefixes, suffixes and root structures. AarBERT predicts masked tokens using context, which is facilitated by pre-training objectives, e.g., Masked Language Modeling (MLM), to disambiguate morphologically ambiguous words and support the Contextual Superiority Theory.

mBERT (Multilingual BERT) is pre-trained model in 104 languages, including Arabic, and it is able to use knowledge transfer across languages. It is not Arabic-centric, but due to the fine-tuning on the Arabic-specific data and target preprocessing, it performs better on Arabic tasks, such as summarization, and it also was proved to be adaptable. In the case of Arabic extractive summarization, the initial models on top of mBERT such as TRANS.ABS[11]have proven to be the first benchmarks. But their comparatively lower ROUGE scores suggested the infantile phase of extractive works by transformers[12]. Other models of transformers that are Arabic specific have also cropped up. ArabicBERT [13] is also based on BERT and is pre-trained over large amounts of 8.5 GB of Arabic data to understand the Arabic language in general.

The multilingual masked language **model XLM-R (Cross-lingual Language Model RoBERTa)**, which is based on 100 languages, such as Arabic, exploits this data to transfer training across languages.

Other models, like **ARBERT**[14]**MARBERT**[14]**and CAMeLBert-MSA** [15]**are** also specialize in Arabic which are trained on large and heterogeneous corpora to be able to represent morphological and syntactic variations of the Arabic language with high accuracy. These models rely on such objectives as Masked Language Modeling and Sentence Order Prediction to comprehend more complicated relations and discourse coherence, which is crucial in the case of effective extractive summarization.

Abstractive tasks demonstrate better results in models such as AraBERT, AraBART, mBART, and AraT5 even on a dialectal corpus. This means that these architectures can be applied in learning and exploiting complex linguistic patterns of Arabic. The experimental showed that transformers, having received representations, can offer a more sustainable solution to the language issues of Arabic than classical methods do and hence the substitution of the rule based by learned representation has a huge advancement. Many studies are summarized in Table 1 for Arabic Extractive Summarization.

Table 1 - Systematic Review of Arabic Extractive Summarization Studies.

Citation	*Title*	*Dataset*	*Used method*	*Accuracy*	*limitation*
[16]	Extractive Arabic Text	EASC	Modified	Precision=68.75	Complex

	Summarization Using Modified PageRank Algorithm		PageRank, Morphological Analysis	Recall= 72.94 F-measure= 67.99	morphology, noun extraction
[17]	Arabic Text Summarization Using AraBERT Model Using Extractive Text Summarization Approach	own dataset	AraBERT, BERT, XLNet, XLM	Precision=0.39 Recall =0.90 F-measure=0.54	Weak points identified
[18]	Unsupervised neural networks for automatic Arabic text summarization using document clustering and topic modeling	EASC	Clustering, Topic Modeling, Neural Networks	Improved ROUGE, F-measure	Context/domain not always considered
[19]	Extractive text summarization of arabic multi-document using fuzzy C-means and Latent Dirichlet Allocation	TAC-2011	Fuzzy C-means, LDA	Competitive vs. ant colony, DA	Topic clustering limits
[20]	Abstractive Arabic Text Summarization Based on Deep Learning	AHS, AMN	Seq2Seq, BiLSTM, GRU, LSTM	ROUGE-1= 51.49 best with BiLSTM	Few abstractive studies
[21]	Arabic Extractive Summarization Using Pre-Trained Models	KALIMA	QARiB, AraELECTRA, AraBERT	ROUGE-1:0.44 ROUGE-2:0.26, ROUGE-L: 0.44	Dataset scarcity
[22]	An Efficient Deep Learning Approach for Extractive Arabic Text Summarization Based on Multiple Encoders and a Single Decoder	HASD, EASC	Multi-encoder Seq2Seq	Compared by ROUGE, human eval	New eval measure, dataset
[23]	A Novel Gravity Optimization Algorithm for Extractive Arabic Text Summarization	EASC	Gravitational Optimization	ROUGE-1: Recall=68.04%	Metaheuristic complexity
[24]	Extractive Arabic Text Summarization Using PageRank and Word Embedding	Not specified	PageRank, Word Embedding	7.5% better F-measure than alternatives	Dataset not specified
[25]	Toward an efficient extractive Arabic text summarisation system based on Arabic large language models	EASC	AraT5, AraGPT2, AraBART, TextRank, KeyBERT	Precision: AraT5=53.5%	LLMs need more tuning
[26]	Leveraging Transformer Summarizer to Extract Sentences for Arabic Text Summarization	EASC	Transformer-based	Notable performance	Scenario-dependent strengths
[27]	AraTSum: Arabic Twitter Trend Summarization Using Topic Analysis and Extractive Algorithms	5 Twitter datasets	LDA, Pre-trained, Sentiment	Outperforms SOTA (ROUGE)	Topic polarity, aspect coverage

3. Evaluation metrics

All Extractive text summarization models require the use of standardized measures to conduct quantitative evaluation of the model as an objective and reproducible benchmark of performance. ROUGE[28], Precision-Recall-Fmeasure[29], and BLEU[30] are examples for metric evaluation of extractive text summarization. The research papers focused on the Recall-Oriented Understudy of Gisting Evaluation (ROUGE) suite, which is a lexical measure of the similarity between system-generated and human-generated summaries of reference lists. Evaluation is focused exclusively on the accuracy of the content selection, which is differentiated by the issues of generative quality of abstractive summing up. ROUGE-N is used to measure the similarity of n-grams of a candidate summary and a collection of summaries created by humans. ROUGE-1 is an unigram (single word) overlap measure, which evaluates inherent content preservation. ROUGE-2 compares the overlap of bigrams (two word sequence) and provides the information about the similarity at the phrase level. ROUGE-N has a mathematical formulation which is given as in equation 1.

$$ROUGE\text{-}N = \frac{\sum_{S\in References}\sum_{n\text{-}gram\in S} CountMatch\,(n\text{-}gram)}{\sum_{S\in References}\sum_{n\text{-}gram\in S} Count\,(n\text{-}gram)} \quad (1)$$

In the equation, CountMatch(n-gram) is the number of n-grams shared by the candidate and the reference summary with the longest length, and Count (n-gram) is the total number of n-grams in the reference summary[31]. ROUGE-L, which is defined as the Longest Common Subsequence (LCS) of words, finds the longest ordered sequence of words shared with the candidate and the reference summaries. This measure is less sensitive to the word order differences than ROUGE-N, thus it measures the structural regularity and the retention of essential phrases. ROUGE-L will be based on Precision (P_{LCS}), Recall (R_{LCS}), and F1-score (F_{LCS}) to offer a balanced content selection measure as in equation 2, 3, and 4 respectively.

Recall:

$$R_{LCS} = \frac{LCS(candidate,reference)}{length(reference)} \quad (2)$$

Precision:

$$P_{LCS} = \frac{LCS(candidate,reference)}{length(candidate)} \quad (3)$$

F-measure:

$$P_{LCS} = \frac{LCS(candidate,reference)}{length(candidate)} \quad (4)$$

Recall measures the size of the reference summary that is in the candidate; precision measures the content of the candidate that is available in the reference and F1-score is their harmonic mean and the balance. The aspects, provide all-encompassing evaluation of salient information capture and retention under the extractive summary. Though ROUGE metrics have a uniform structure, their use with morphologically rich languages such as Arabic have to be questioned. The partial dependence of ROUGE on lexical overlap is less effective in cases of semantic equivalence because of different morphological forms or where there are ambiguities caused by diacritic marks which makes the direct n-gram matching more ambiguous. Indicatively, in agglutinative languages, semantically identical sentences can score low ROUGE scores because of morphological differences, which prevent n-gram matches on the surface [32]. This is a linguistic difficulty in Arabic which states that human ratings of the quality of Arabic summaries are sensitive to linguistic quirks to a very high level. This in turn demands further advanced evaluation approaches that ROUGE scores cannot offer unaccompanied by additional qualitative measures.

figures should be numbered with Arabic numerals (1,2,3,….). Every figure should have a caption. All photographs, schemas, graphs and diagrams are to be referred to as figures. Line drawings should be good quality scans or true electronic output. Low-quality scans are not acceptable. Figures must be embedded into the text and not supplied separately. In MS word input the figures must be properly coded. Lettering and symbols should be clearly defined either in the caption or in a legend provided as part of the figure. Figures should be placed at the top or bottom of a page wherever possible, as close as possible to the first reference to them in the paper.

4. Datasets

Arabic extractive summarization has been greatly enhanced by the presence of various datasets that offer scale variation, different domains, and evaluation frameworks to further research activities. This section lists the Arabic summarization datasets that used in 2020 to 2024 with comparing their key attributes.

Large-scale Arabic News Summarization (LANS) dataset is the most extensive one and it includes 8.4 million articles that have a journalist-generated summaries of articles in 22 Arabic newspapers over a 20-year span (1999-2019). It provides a large-scale coverage of the subject and a high-quality set of human-rated summaries that can be used as gold standards in news summarization tasks [33]. To supplement this massive resource, Essex Arabic Summaries Corpus (EASC/ESAC) is a narrower corpus of extractive summarization evaluation, popular in graph-based and PageRank models of sentence extraction and ranking [34], [35]. When multilingual aspect is considered, the cross-language datasets like XL Sum and WikiLingua, including Arabic language in addition to others, are offering both highlight and full summary options, making it possible to conduct comparative research across linguistic borders [36]. Recent efforts expanded the Arabic summarization benchmark landscape that offer dual extractive and abstractive summary types such as HASD (≈43k articles) and AASD (≈150k articles) datasets providing researchers flexible evaluation options for different summarization approaches [37]. Additionally, the emerging AGS corpus represents a novel direction in the field by focusing on GPT-style summarization with automated generation and ROUGE-L evaluation metrics, reflecting the growing integration of large language models in Arabic text summarization research[38]. Collectively, these datasets provide a comprehensive foundation for Arabic extractive summarization research, spanning from traditional rule-based approaches to modern neural and transformer-based methodologies. Table 2 presents summary of these datasets size, type, domain, and notable notes reported in the source papers.

Table 2 - Summary of Arabic datasets for summarization.

Dataset	*Size or scale*	*type*	*Domain / source*	*Notes*
LANS	8.4 million articles	Journalist summaries (news)	22 Arabic newspapers (1999–2019)	Large-scale, diverse news summaries; human evaluation reports high quality for sampled items
EASC / ESAC	Corpus-scale (smaller than LANS)	Extractive summaries	News/articles (Essex Arabic Summaries Corpus)	Widely used for graph/PageRank and extractive evaluations
XL-Sum / WikiLingua	Cross-language summarization collections	Highlight / full summaries	Mixed domains; multilingual sources	Used as full/highlight summary benchmarks in Arabic experiments
HASD / AASD	HASD ≈43k, AASD ≈150k articles (reported)	Extractive + abstractive (HASD), abstractive (AASD)	Mixed (benchmark introduced by authors)	Proposed to expand Arabic benchmarks and evaluation options
AGS	Noted as a new corpus proposal	GPT-style summarization corpus	Not specified in summary	Recent corpus proposal focused on GPT-generated summarization labels and evaluation (Rouge-L used)

5. Research Gaps and Limitations

Although classical and transformer-based methods have been improved, extractive summarization in Arabic is still lacking in adequate research gaps and constraints. Such gaps impair the creation of powerful, generalizable systems, slowing down the Arabic Natural Language Processing (NLP) advancement. The major weakness here is that there is a dearth of extensive, multi-skilled, expert-curated extractive summarization data on Arabic to exploit the full potential of modern models.The limited scope and size of existing corpora do not support the training and stringent evaluation of modern models. One of the major gaps still exists in the form of datasets that have clear annotations on multi-document summarization labels, which are essential in synthesizing information across various sources. More to this, query-based salience, in which the summaries see the particular user information requirements and rhetorical roles, are not fully developed[39]. Such scarcity of quality data places a linguistic inconsistency upon Arabic NLP that makes the models use a small amount of data or data synthesized to fit, which are insufficient to capture the complexity and variety of the language. This lack of resources limit the generalizability of models and makes it difficult to fully assess these models. The second major gap is that there are no standardized and replicable pointers of comparison between models that used among the Arabic dialects. There are around 30 groups of Arabic dialects [40]. Although news, books and articles are written in Modern Standard Arabic (MSA), dialectal variations are introduced in quoted speech or informal sections and social media. The existing benchmarks are usually focused on MSA but not with real-world dialectal use and do not scale up to models with the capability to decode and summarize the whole spectrum of Arabic communication[41]. Such a lack of dialect-specific evaluation procedures makes the determination of model strength and transferability to the vast world of Arabic speakers more difficult. Also, it has been stated that the perceived quality of the summaries of Arabic is highly sensitive to these linguistic subtleties, and thus requires evaluation of higher complexity than the provided by current benchmarks.

Another problem that is confronting the research community lacked to quantify the precise influence of Arabic-specific preprocessing phases on transformer model performance. Although researchers acknowledge that preprocessing is essential, no steps, such as Alef normalization, Tashkeel removal, or custom-designed tokenization, have been used in a systematic way with transformer architectures. The optimization of preprocessing pipelines remains yet an art and not a data-driven based science, which can cause the use of suboptimal settings. These are the critical gaps in research that need to be filled by putting an effort where the emphasis should be made on the creation of High-Quality, diversified Arabic Datasets. It is connected with the designing of large, human-tagged datasets that consist of multi-domain content (e.g., legal, medical, social media) and have a variety of dialect. Community-based annotation projects, semi-automatic annotation tools, and crowdsource-based annotation tools are all possible methods to accelerate this process[42].

Simultaneously, language-specific measures of evaluation and the extra qualitative ones have to be constituted. This will entail the dislocation of the reliance upon ROUGE scores to the encompassment of human assessment facilities that are specifically intended at the coherence, factuality for ambiguity-solving of Arabic summaries. The second priority of direction is the creation of more linguistically sophisticated measures of evaluation that entail a mixture of morphological analysis or diacritization confidence scores. These measures will form the foundation of the future research as it will have a direct impact on the method of the experiment and analysis.

6. Coclusion

The modern outlook of the extractive summarization in the Arabic literature has been explained in this literature review stating that there is still a lot to be done for extractive summarization because there are challenges that persist. Transformer-based structures would have provided superior solutions to classical strategies, but the morphological complexity, dialectal diversity, and even orthographic variation of Arabic are key factors which require a special solutions. Three issues are critical, namely, the absence of high-quality datasets that can reflect the linguistic heterogeneity of Arabic, the poor alignment of the evaluation to the morphologically rich languages, and the insufficient systematic validation of the role played by preprocessing strategies in the performance of summarization. These gaps are not represented by academic research only but also of the practical application of Arabic summarization systems and hence the interventions must be specific to close the gap between the level of computation and the linguistic requirements.

7. Future Work

This review leads to some priority areas of research such as evaluation metric development, Cross-dialectal Transfer Learning, Preprocessing Impact Analysis, and establishing collaborative annotation projects to make gold standard corpora.

Multi-dimensional Evaluation Structures can be developed where all-encompassing evaluation principles that will incorporate morphological coherence, semantic consistency, and pragmatic suitability scores that will not rely on the typical ROUGE scores will be designed. It must also entail intrinsic quality measurement, and extrinsic task based measurement to provide complete performance measurement.

Cross-dialectal Transfer Learning can be achieved by capture the relationship between the modern standard Arabic and regional dialects without deteriorating the dialect. The systematic evaluation of the great dialect groups is the clue to the model generalizability and practical applicability to the Arabic-speaking world.

Preprocessing Impact Analysis can be done in order to measure the contribution of each step of preprocessing e.g. stemming, normalization, and tokenization strategies. This empirical finding is significant in transforming the heuristic-based pipeline optimization to the data-driven optimization.

Also, Domain Adaptation Strategies can be used where few-shot learning to special-purpose domain adaptations are necessary for Learning legal, medical, and financial text with little or no annotated data. Creation of field-sensitive attention processes may be used to develop better content recognition and propagation of field-specific information Resource Development Projects for establishing collaborative annotation projects on the basis of crowd forcing and semi-automated tools to generate various large-scale corpora. Quality control and standardization rules for the sake of reproducibility and consistency in datasets across research activities are needed. It takes concerted efforts to deal with these issues in a systematic way to advance the Arabic extractive summarization. The success will be reached with the creation of linguistically-aware methodologies that consider the specialties of Arabic, but that use the current computational techniques to address the varied demands of Arabic-speaking communities.

References

[1] M. Al-Maleh and S. Desouki, "Arabic text summarization using deep learning approach," *J Big Data*, vol. 7, no. 1, pp. 1–17, Dec. 2020, doi: 10.1186/S40537-020-00386-7/TABLES/6.

[2] A. M. Al-Numai and A. M. Azmi, "Arabic Abstractive Text Summarization Using an Ant Colony System," *Mathematics 2025, Vol. 13, Page 2613*, vol. 13, no. 16, p. 2613, Aug. 2025, doi: 10.3390/MATH13162613.

[3] H. Shakil, A. Farooq, and J. Kalita, "Abstractive text summarization: State of the art, challenges, and improvements," *Neurocomputing*, vol. 603, p. 128255, Oct. 2024, doi: 10.1016/J.NEUCOM.2024.128255.

[4] M. Gamal, M. A. Salam, H. F. A. Hamed, and S. Sweidan, "ACOSUM: Ant Colony Optimized Multi-Level Semantic Graph Summarization," vol. 1, no. 1, p. 16, 2025, doi: 10.21608/ijaici.2025.350645.1006.

[5] J. Zhou, Z. Ye, S. Zhang, Z. Geng, N. Han, and T. Yang, "Investigating response behavior through TF-IDF and Word2vec text analysis: A case study of PISA 2012 problem-solving process data," *Heliyon*, vol. 10, no. 16, p. e35945, Aug. 2024, doi: 10.1016/J.HELIYON.2024.E35945.

[6] A. A. Aladeemy *et al.*, "Advancements and challenges in Arabic sentiment analysis: A decade of methodologies, applications, and resource development," *Heliyon*, vol. 10, no. 21, p. e39786, Nov. 2024, doi: 10.1016/J.HELIYON.2024.E39786.

[7] S. Albitar, S. Fournier, and B. Espinasse, "An effective TF/IDF-based text-to-text semantic similarity measure for text classification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8786, pp. 105–114, 2014, doi: 10.1007/978-3-319-11749-2_8.

[8] M. Bugueño and G. de Melo, "Connecting the Dots: What Graph-Based Text Representations Work Best for Text Classification Using Graph Neural Networks?," *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8943–8960, Jan. 2024, doi: 10.18653/v1/2023.findings-emnlp.600.

[9] A. E. Martin, "A Compositional Neural Architecture for Language," *J Cogn Neurosci*, vol. 32, no. 8, pp. 1407–1427, 2020, doi: 10.1162/JOCN_A_01552.

[10] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," 2020. Accessed: Sep. 28, 2025. [Online]. Available: https://aclanthology.org/2020.osact-1.2/

[11] K. N. Elmadani, M. Elgezouli, and A. Showk, "BERT Fine-tuning For Arabic Text Summarization," *presented at AfricaNLP Workshop, ICLR 2020*, Mar. 2020, Accessed: Oct. 24, 2025. [Online]. Available: https://arxiv.org/pdf/2004.14135

[12] M. Kahla, Z. Gy″, O. Yang, and A. Novák, "Cross-lingual Fine-tuning for Abstractive Arabic Text Summarization," 2021. doi: 10.26615/978-954-452-072-4_074.

[13] A. Safaya, M. Abdullatif, and D. Yuret, "KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media," *14th International Workshops on Semantic Evaluation, SemEval 2020 - co-located 28th International Conference on Computational Linguistics, COLING 2020, Proceedings*, pp. 2054–2059, 2020, doi: 10.18653/V1/2020.SEMEVAL-1.271.

[14] M. Abdul-Mageed, A. R. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic," *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, vol. 1, pp. 7088–7105, 2021, doi: 10.18653/V1/2021.ACL-LONG.551.

[15] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models," 2021. Accessed: Oct. 24, 2025. [Online]. Available: https://aclanthology.org/2021.wanlp-1.10/

[16] R. Elbarougy, G. Behery, and A. El Khatib, "Extractive Arabic Text Summarization Using Modified PageRank Algorithm," *Egyptian Informatics Journal*, vol. 21, no. 2, pp. 73–81, Jul. 2020, doi: 10.1016/J.EIJ.2019.11.001.

[17] A. M. A. Nada, E. Alajrami, A. A. Al-Saqqa, and S. S. Abu-Naser, "Arabic Text Summarization Using AraBERT Model Using Extractive Text Summarization Approach," 2020.

[18] N. Alami, M. Meknassi, N. En-nahnahi, Y. El Adlouni, and O. Ammor, "Unsupervised neural networks for automatic Arabic text summarization using document clustering and topic modeling," *Expert Syst Appl*, vol. 172, Jun. 2021, doi: 10.1016/J.ESWA.2021.114652.

[19] M. Gouiouez, "A Fuzzy Near Neighbors Approach for Arabic Text Categorization Based on Web Mining Technique," *Lecture Notes in Networks and Systems*, vol. 211 LNNS, pp. 575–584, 2021, doi: 10.1007/978-3-030-73882-2_52.

[20] Y. M. Wazery, M. E. Saleh, A. Alharbi, and A. A. Ali, "Abstractive Arabic Text Summarization Based on Deep Learning," *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/1566890.

[21] Y. Einieh, A. AlMansour, and A. Jamal, "Arabic Extractive Summarization Using Pre-Trained Models," *Journal of King Abdulaziz University: Computing and Information Technology Sciences*, vol. 12, no. 1, pp. 63–73, Jul. 2023, doi: 10.4197/Comp.12-1.6.

[22] A. Elsaid, A. Mohammed, L. Fattouh, and M. Sakre, "An Efficient Deep Learning Approach for Extractive Arabic Text Summarization Based on Multiple Encoders and a Single Decoder," *1st International Conference of Intelligent Methods, Systems and Applications, IMSA 2023*, pp. 1–6, 2023, doi: 10.1109/IMSA58542.2023.10217361.

[23] M. J. Hadi, A. R. Abbas, and O. Y. Fadhil, "A Novel Gravity Optimization Algorithm for Extractive Arabic Text Summarization," *Baghdad Science Journal*, vol. 21, no. 2, pp. 537–547, 2024, doi: 10.21123/BSJ.2023.7731.

[24] 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, Sep. 2024, doi: 10.1007/S13369-024-08890-1/TABLES/9.

[25] G. Bourahouat, M. Abourezq, and N. Daoudi, "Toward an efficient extractive Arabic text summarisation system based on Arabic large language models," *Int J Data Sci Anal*, vol. 20, no. 3, pp. 2445–2457, Sep. 2024, doi: 10.1007/S41060-024-00618-6/METRICS.

[26] H. Zaiton, A. Fashwan, and S. Alansary, "Leveraging Transformer Summarizer to Extract Sentences for Arabic Text Summarization," *Procedia Comput Sci*, vol. 244, pp. 353–362, Jan. 2024, doi: 10.1016/J.PROCS.2024.10.209.

[27] E. Monir and A. Salah, "AraTSum: Arabic Twitter Trend Summarization Using Topic Analysis and Extractive Algorithms," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, pp. 1–18, Dec. 2024, doi: 10.1007/S44196-024-00546-0/FIGURES/8.

[28] "ROUGE: A Package for Automatic Evaluation of Summaries - ACL Anthology." Accessed: Oct. 24, 2025. [Online]. Available: https://aclanthology.org/W04-1013/

[29] A. Nenkova, S. Maskey, I. Research, and Y. Liu, "Automatic Summarization Why summarize?," *Information Retrieval*, vol. 5, pp. 103–233, 2011.

[30] K. Owczarzak, J. M. Conroy, H. T. Dang, and A. Nenkova, "An Assessment of the Accuracy of Automatic Evaluation in Summarization," 2012. Accessed: Oct. 24, 2025. [Online]. Available: https://aclanthology.org/W12-2601/

[31] D. Yadav, J. Desai, and A. K. Yadav, "Automatic Text Summarization Methods: A Comprehensive Review," Mar. 2022, Accessed: Sep. 26, 2025. [Online]. Available: http://arxiv.org/abs/2204.01849

[32] F. B. Fikri, K. Oflazer, and B. Yanıkoğlu, "Semantic Similarity Based Evaluation for Abstractive News Summarization," *GEM 2021 - 1st Workshop on Natural Language Generation, Evaluation, and Metrics, Proceedings*, pp. 24–33, 2021, doi: 10.18653/V1/2021.GEM-1.3.

[33] A. Alhamadani, X. Zhang, J. He, A. Khatri, and C. T. Lu, "LANS: Large-scale Arabic News Summarization Corpus," *ArabicNLP 2023 - 1st Arabic Natural Language Processing Conference, Proceedings*, pp. 89–100, 2023, doi: 10.18653/v1/2023.arabicnlp-1.8.

[34] Y. A. AL-Khassawneh and E. S. Hanandeh, "Extractive Arabic Text Summarization-Graph-Based Approach," *Electronics (Switzerland)*, vol. 12, no. 2, Jan. 2023, doi: 10.3390/ELECTRONICS12020437.

[35] N. Burmani, H. Alami, S. Lafkiar, M. Zouitni, M. Taleb, and N. E. Nahnahi, "Graph based method for Arabic text summarization," *2022 International Conference on Intelligent Systems and Computer Vision, ISCV 2022*, 2022, doi: 10.1109/ISCV54655.2022.9806127.

[36] T. Hasan *et al.*, "XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4693–4703, Jun. 2021, doi: 10.18653/v1/2021.findings-acl.413.

[37] A. Elsaid, A. Mohammed, L. Fattouh, and M. Sakre, "Abstractive Arabic Text Summarization Based on MT5 and AraBart Transformers," *1st International Conference of Intelligent Methods, Systems and Applications, IMSA 2023*, pp. 7–12, 2023, doi: 10.1109/IMSA58542.2023.10217539.

[38] A. Atef, F. Seddik, and A. Elbedewy, "AGS: Arabic GPT Summarization Corpus," *4th International Conference on Electrical, Communication and Computer Engineering, ICECCE 2023*, 2023, doi: 10.1109/ICECCE61019.2023.10441794.

[39] H. Rhel and D. Roussinov, "Large Language Models and Arabic Content: A Review," May 2025, Accessed: Sep. 26, 2025. [Online]. Available: https://arxiv.org/pdf/2505.08004

[40] M. Kurt Pehlivanoğlu, R. T. Gobosho, M. A. Syakura, V. Shanmuganathan, and L. de-la-Fuente-Valentín, "Comparative analysis of paraphrasing performance of ChatGPT, GPT-3, and T5 language models using a new ChatGPT generated dataset: ParaGPT," *Expert Syst*, vol. 41, no. 11, Nov. 2024, doi: 10.1111/EXSY.13699.

[41] B. Mousi *et al.*, "AraDiCE: Benchmarks for Dialectal and Cultural Capabilities in LLMs," *n Proceedings of the 31st International Conference on Computational Linguistics, pages 4186–4218, Abu Dhabi, UAE. Association for Computational Linguistics.*, pp. 4186–4218, 2025, Accessed: Sep. 29, 2025. [Online]. Available: https://aclanthology.org/2025.coling-main.283/

[42] A. Charfi, M. Bessghaier, A. Atalla, R. Akasheh, S. Al-Emadi, and W. Zaghouani, "Stance detection in Arabic with a multi-dialectal cross-domain stance corpus," *Soc Netw Anal Min*, vol. 14, no. 1, Dec. 2024, doi: 10.1007/S13278-024-01335-5.