SAKARYA UNIVERSITY
INSTITUTE OF NATURAL SCIENCE

ARABIC TEXT SUMMARIZATION USING PAGERANK AND WORD EMBEDDING ALGORITHMS

M.Sc. THESIS

Ghadir Abdulhakim Abdo Abdullah ALSELWI

Department	:	INFORMATION SYSTEMS ENGINEERING
Supervisor	:	Dr. Tuğrul TAŞCI

July 2022

SAKARYA UNIVERSITY
INSTITUTE OF NATURAL SCIENCE

ARABIC TEXT SUMMARIZATION USING PAGERANK AND WORD EMBEDDING ALGORITHMS

M.Sc. THESIS

Ghadir Abdulhakim Abdo Abdullah ALSELWI

Department	**:**	**INFORMATION SYSTEMS ENGINEERING**
Supervisor	**:**	**Dr. Tuğrul TAŞCI**

This thesis has been accepted unanimously / ~~with majority of votes~~ by the examination committee on 29.07.2022

Head of Jury	**Jury Member**	**Jury Member**

DECLARATION

I declare that all the data in this thesis was obtained by myself in academic rules, all visual and written information and results were presented in accordance with academic and ethical rules, there is no distortion in the presented data, in case of utilizing other people's works they were refereed properly to scientific norms, the data presented in this thesis has not been used in any other thesis in this university or in any other university.

Ghadir ALSELWI

30.05.2022

ACKNOWLEDGMENT

I would like to thank the following people, without whom I would not have been able to finish my studies and earn my master's degree!

First and foremost, I would like to convey my sincere appreciation to my supervisor, Dr. Tuğrul TAŞCI, for his continuous support of my research and study and his patience, encouragement, enthusiasm, and thorough knowledge. His guidance was invaluable during the research and writing of this thesis. I could not have asked for a better adviser and supervisor for my master's program.

In addition to my supervisor, I would like to express my gratitude to my family: my parents, Abdulhakim Alsebaee and Kartalah Alselwi, for giving birth to me in the first place and for spiritually supporting me throughout my life. In addition, my heartfelt gratitude goes to my husband, Nashwan Alhakimi, and my brothers, sisters, and friends, for their unwavering support.

TABLE OF CONTENTS

CHAPTER 1.

CHAPTER 2.

CHAPTER 3.

CHAPTER 4.

CHAPTER 5.

LIST OF SYMBOLS AND ABBREVIATIONS

ArTS	: Arabic Text Summarization
ATS	: Automatic Text Summarization
BLEU	: Bilingual Evaluation Understudy
CNN	: Conolutional Neural Network
CR	: Compression Rate
DL	: Deep Learning
EASC	: Essex Arabic Summaries Corpus
GA	: Genetic Algorithm
GEATS	: Graph-based Extractive Arabic Text Summarization
IDF	: Inverse Term Frequenct
LSTM	: Long Short-Term Memory
MD	: Mutli-Document
ML	: Machine Learning
Mturk	: Mechanical Turk
NLP	: Natural Language Processing
NLTK	: The Natural Language Toolkit
OSSAD	: Analogy-based Summarization System for Arabic Documents
PR	: PageRank
PSO	: Particle Swarm Optimization
RNN	: Recurrent Neural Network
ROUGE	: Recall-Oriented Understudy for Gisting Evaluation
SD	: Single-Document
TF	: Term Frequency
TS	: Text Summarization
WE	: Word Embedding

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

SUMMARY

Keywords: Arabic text summarization, pagerank algorithm, word embedding, graph-based, word2vec, extractive Arabic text summarization, Farasa stemmer

Arabic is one of the world's most frequently spoken languages, with over 200 million people using it as their first language, and it is the official language of 26 nations. Although Arabic text summarization (ArTS) has increased in popularity in recent years, the quality of current ATS systems need improvement. Graph-based techniques on Arabic natural language processing have clearly gained popularity in recent years. Because of their ability to arrange large and difficult structures into standard and formal ways, graphs may be used and developed in a helpful way to assist in conquering and minimizing Arabic language challenges.

This study proposed a single-document Graph-based Extractive Arabic Text Summarization (GEATS). The PageRank method is used, along with word embedding. The similarity of any two sentences is calculated by ranking the sentences based on cosine similarity. The final score for each sentence is determined using PageRank scoring. Then, the summary includes the sentences with the highest ratings taking into account the compression ratio, which is 40% of the document's sentences.

The EASC Corpus is used as a standard corpus to test the performance of this technique. ROUGE-1, ROUGE-2, and BLUE metrics are also employed in the evaluation process. The findings demonstrated that the proposed strategy outperforms state-of-the-art approaches.

PAGERANK VE KELİME GÖMME ALGORİTMALARI KULLANARAK ARAPÇA METİN ÖZETLEME

ÖZET

Anahtar Kelimeler: Arapça metin özetleme, pageRank algoritması, kelime gömme, grafik tabanlı, word2vec, ekstraktif Arapça metin özetleme, Farasa stemmer

Arapça, 200 milyondan fazla insanın ilk dili olarak kullandığı, dünyanın en sık konuşulan dillerinden biridir ve 26 ülkenin resmi dilidir. Arapça metin özetleme (ArTS) son yıllarda popülaritesini artırmış olsa da, mevcut ATS sistemlerinin kalitesinin iyileştirilmesi gerekmektedir. Arapça doğal dil işlemede grafik tabanlı teknikler son yıllarda açıkça popülerlik kazanmıştır. Büyük ve zor yapıları standart ve biçimsel yollarla düzenleme yeteneklerinden dolayı, grafikler Arapça dil zorluklarını fethetmeye ve en aza indirmeye yardımcı olmak için yararlı bir şekilde kullanılabilir ve geliştirilebilir.

Bu çalışma, tek belgeli bir Grafik tabanlı Ekstraktif Arapça Metin Özetleme (GEATS) önerdi. PageRank yöntemi, kelime gömme ile birlikte kullanılır. Herhangi iki cümlenin benzerliği, cümlelerin kosinüs benzerliğine göre sıralanmasıyla hesaplanır. Her cümle için nihai puan PageRank puanlaması kullanılarak belirlenir ve yüksek puan alan cümleler, belgenin cümlelerinin %40'ı olan sıkıştırma oranı dikkate alınarak özete dahil edilir.

Bu tekniğin performansını test etmek için EASC Corpus kullanıldı. ROUGE-1, ROUGE-2 ve BLUE metrikleri de değerlendirme sürecinde kullanılmaktadır. Bulgular, önerilen yöntemin en gelişmiş yaklaşımlardan daha iyi performans gösterdiğini göstermiştir.

CHAPTER 1. INTRODUCTION

The Internet's online resources (for example, webpages, blogs, social media networks, user reviews, news, and so on) are massive sources of textual data. Furthermore, there is a wealth of textual information on the numerous archives of novels, news articles, medical documents, books, scientific papers, and so on. On a daily basis, the amount of textual information on the Internet and other archives grows tremendously.

Web-users frequently are reading enormous text pages due to the rapid increase in the volume and availability of online content, accessing and searching information has become challenging since it takes time to go through all of the textual data. In fact, researchers are struggling to keep up with all of the new publications to read. The widespread usage of the internet makes a high level of interest in Automatic Text Summarization (ATS) which is saving our time and effort, and addresses the issue of information overload that people encounter in the digital age. As a result, there is a growing demand for effective and sophisticated tools to automatically summarize texts [1], and the solution to this dilemma is found in using ATS. The primary purpose of ATS is to compress the original text while retaining the information content and overall meaning.

According to Radev et al. [2] a summary is “a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually, significantly less than that”. We may derive from Radev's definition of text summarizing that Text Summarization (TS) should contain the following features; summaries derived from one or more documents; only the most significant sentences should be extracted; summaries should be kept as brief as possible.

The TS process may be divided into three stages starting with the analysis step which examines the input text and picks a few key characteristics. Then, the transformation stage which converts the analytical results into a summary representation. Finally, the synthesis stage comes which takes the representation of the summary and generates a suitable summary based on the needs of the user.

There are several approaches to ATS that fall into one of the following categories: statistical, metaheuristic-based, hybrid, Machine Learning (ML), fuzzy-logic based, and graph-based approaches.

There are several issues with the supported languages in TS. The majority of ATS systems are focused on English language content and popular languages. Many other languages require improvements in terms of ATS systems. Despite the fact that ArTS has captured the interest of researchers in recent years, ArTS systems need to be improved especially considering F-measure, precision, and recall.

Because Arabic is one of the most frequently spoken languages, there is a tremendous need to summarize the enormous volume of textual data. Consequently, this work is focusing on using the graph-based approach. In fact, the PR algorithm was employed in combination with WE in this study. The focus of this study is on creating a model to generate automatic ArTS utilizing extraction techniques which are applicable to a wide range of domains and perform well.

1.1. Motivation and Problem Statement

Due to Arabic language complexity and the paucity of studies in this field, the Arabic summarization system continues to perform poorly. Many studies on ArTS employ graph-based algorithms such as the PR algorithm, as well as TF-IDF for text representation and feature extraction. Indeed, for feature extraction, this study used the PR method in conjunction with WE using Word2Vec. In addition, this study seeks to employ three algorithms, PR, LexRank, and TextRank, to get the best performance.

The subject addressed in this study is to create a model to generate automatic ArTS using extraction approaches that are applicable to a wide range of domains and have excellent performance. To solve this problem, some parameters must be determined, such as the appropriate list of stopwords, the appropriate data corpus for this system, the appropriate number of preprocessing steps, the best stemming technique, the best basic units to use, the most relevant features to be extracted, the best dataset to test the system, how to rebuild the summary, and how to evaluate the summary.

1.2. The Goal of the Study

This research examines a new method for extractive SD ArTS system that is based on Word Embedding (WE) and Google PageRank (PR) algorithm that uses graphs directly to build a summary for an Arabic document, highlighting uniqueness and guaranteeing that the final summary is both logical and thorough. This method begins with preprocessing approaches. By improving preprocessing approaches, we seek to enhance our suggested approach's performance. Then collecting the required features, forming the graph, then using the PR algorithm, and lastly extracting and evaluating the summary.

1.3. Significance of the Thesis

ArTS is crucial in the Arabic world for a variety of reasons, including encouraging the use of Arabic-language content on the internet, using ArTS in a variety of contexts and areas, helping Arabic readers by saving their time, money, and effort.

1.4. Limits and Scope

As previously stated, there are several types of text summarization. In this study, we focus on extractive single-document Arabic text summarization. The documents were evaluated in a specific domain described by the EASC corpus. The employed CR is 40% in the comparison stage.

1.5. Thesis Organization

This thesis is structured as follows: Chapter 2. will explore underlying methods. Chapter 3. is discussing the related work that has been done. Chapter 4. presents the theoretical background briefly. Chapter 5. explains Graph-based Extractive Arabic Text Summarization (GEATS) system’s methodology by using PR and WE methods and the stages that starts by collecting the dataset and ends by evaluating the system. In Chapter 6. experimentation and the results and GEATS approach evaluation are addressed and shown. Chapter 7. includes the conclusion as well as future works.

CHAPTER 2. UNDERLYING METHODS

2.1. Text Summarization (TS)

The basic purpose of TS is to reduce the amount of paragraphs and assertions in the document as much as possible in order to provide the readers with enough information to judge whether or not the document is beneficial. TS is a part of information retrieval, which is the act of locating and obtaining information resources appropriate to a certain information requirement from a collection of information resources.

2.2. Categories of TS

TS may be divided into several categories based on the factors we wish to consider. As a result, summaries can be classified based on several categories.

2.2.1. Type of the returned summary

Extractive TS takes essential paragraphs or sentences from a document and summarizes them without making any modifications. It extracts more relevant and useful data from the original document. It is less difficult and faster than abstractive TS. Many techniques are used, including hidden markov model, clustering, Deep Learning (DL) techniques, graphical methods, LexRank, bayesian method, support vector machine, logistic regression model, decision trees, binary classifier, Term Frequency-Inverse Document Frequency (TF-IDF), TextRank and maximal marginal relevance algorithm [60].

Abstractive TS provides meaningful summaries that may be available or may not be in the given source, Because it concentrates on making a novel summary. Because it generates a broad summary, abstractive TS outperforms extractive TS. Furthermore, it encounters greater challenges during computing than extractive TS. It employs a variety of approaches, including WordNet, support vector machine, Naive Bayes decision theory, K-means algorithm, Recurrent Neural Network (RNN), singular vector decomposition, CNN, neural network, sequence-to-sequence model, and so on [60].

2.2.2. Input factor or Size

Single-document TS takes the most significant information from one document and provides it to the system as one summary.

Multiple-documents creates one summary from various documents within the same topic, which is then sent to the system.

2.2.3. Summary’s nature or output form

Indicative summary gives the reader a concise summary of the content, and highlights the most essential parts in the document. The purpose of this summary is to assist the reader in determining whether or not the original document is useful to read. This is handy for creating a summary of URIs returned by the search engine. It employs a Compression Ratio or Rate (CR) of 5-10%.

Informative summary is more extensive than an indicative summary, which returns more specific information from the original text. This form of TS is important in the process of constructing a news feed summary. It employs a CR between 20% and 30%.

Critical or evaluative summary is a type of a summary that returns the writer's perspective on a specific subject [71].

2.2.4. The content of the summary

Generic summary provides generic facts from the document with general information, putting all main issues in the test to equal degree of relevance.

Query-based summary is created based on the needs of the reader or user. The user enters certain subjects or terms that he wishes to learn more about, and the summarizer provides a summary.

In *user-focused summary***,** the content of the produced summary will revolve around the user-centered demand.

Update summary responds to the inquiry "What's new?". It takes a document input stream and returns a substream of documents. This is accomplished by tracking the new information that constitutes a flow in the system. This type of system assumes that the reader has read the prior texts.

2.2.5. Input/output languages

In *monolingual* TS, there is just one language which means the input's and output's language is identical.

In *multilingual* TS, the summarizing system is capable of handling several languages. As a result, the input and output languages of the two documents are the same.

In *cross-lingual* TS, the input language differs from the output language.

2.2.6. Summary’s type

The length of the produced summary vary based on the ATS system's purpose.

Headline TS generally generates headlines that are less than a sentence long [77].

Highlights TS is generally determined by the desired summary length or a CR.

Sentence-Level TS creates a single sentence from the given document, often an abstractive sentence [77].

Full Summary TS generates a telegraphic style and extremely concise summary. It usually comes in a form of bullet points [78].

2.2.7. Summarization's domain

General domain TS summarizes documents which are from several domains.

Specific domain TS summarizes texts from a specific domain (e.g. technical or commercial documents).

2.3. TS Fields

Many disciplines in our everyday lives need TS such as the following subsections.

2.3.1. Commercial and advertising fields

The market has millions of items. Each item is well descriped. When a marketer wants to promote some items, he just needs a few words to explain them. In this case, ATS is required.

2.3.2. News area

Hundreds of economical, sports, political, and other news are posted every second. It's difficult, even it may be impossible to read them all or even a quarter of them. By using TS, the user may determine his own interesting news without having to read the entire text.

2.3.3. Legal area

There is unlimited number of lawful documents. The time of a legal professional is highly expensive. To ensure that legal experts perform properly, they must be presented with a summarized document. As a result, ATS systems will assist legal specialists in locating restated and concise material of important legal documents, such as proposed laws, applicable judicial decisions, or tribunal procedural summaries [72].

2.3.4. Medical field

Every day, medical advances in the discovery of new diseases, surgical equipment, and methods of curing patients. As a result, researchers publish hundreds of documents each year to discuss their findings. Doctors and medical specialists must quickly locate pertinent information regarding their patients' illnesses. As a result, TS here preserves time and optimizes the accessibility of medical professionals.

2.3.5. Work and technical reports

Thousands of reports are created every day in the technical sector. Time of technical employees is not enough to enable them study these reports and make decisions based on them. As a result, TS is a useful procedure for assisting technicians in determining whether or not this report is significant.

2.4. ATS Approaches

ATS is challenging because when humans try to summarize a piece of document, they often read it thoroughly to improve their understanding, and then write a summary underlining its major ideas. Because computers lack human understanding and linguistic ability, ATS is a tough and time-consuming operation. ATS gained popularity as early as the 1950s. A great deal of study has been done on ArTS utilizing a variety of methodologies. Among these techniques are: statistical and semantics methods [10], [11], ML methodologies [12], [13], approaches based on meta-heuristics [14], [15], [16], [17], [18], hybrid-approaches [19], [20], DL approaches [21], [22], graph-based approaches [23], [24], etc.

2.5. TS and Languages

There are several languages spoken around the world. Language families and groups also differ. Some languages are more widely spoken and known more than others.

Furthermore, the six UN languages[1] – Arabic, English, French, Spanish (Castilian), Russian, and Chinese (Mandarin) – comprise either the first or second language of around 45 percent of the population[2] of the world. In the literature, many TS algorithms for Arabic and other languages have been presented. Many works have been produced for TS in Arabic [3], English [4],[5], French [59], Spanish [6], Russian [7], and Chinese [8].

2.5.1. Arabic language

One of the world's most frequently spoken languages[3] is Arabic. Arabic is the first laguage for more than 200 million people [9], and it is the official language of 26 nations[4]. It is also the Islamic liturgical language. It is not only the language that preserves the vast cultural heritage of the Arab world, but it is also a crucial instrument for doing business in this region of the world. Although ArTS has grown in popularity in recent years, the present ATS systems' quality must be enhanced.

2.5.2. Arabic Natural Language Processing (NLP)

NLP is a branch of computer science that studies human–computer interaction. Many NLP difficulties revolve around natural language comprehension, or enabling computers to derive meaning from human or natural language input, while others revolve around natural language production [70].

The Arabic language has an inherent structure, as well as a strong association with identity, Islam, and culture throughout history. Arabic NLP systems that do not take the characteristics of the Arabic language into account would be ineffective [73], [74]. Arabic NLP systems and applications must address a number of complicated issues related to the structure and the nature of Arabic.

[1] https://www.un.org/en/our-work/official-languages

[2] https://www.cia.gov/the-world-factbook/countries/world/#people-and-society

[3] https://www.cia.gov/the-world-factbook/countries/world/#people-and-society

[4] https://www.berlitz.com/en-uy/blog/most-spoken-languages-world

Some Arabic language characteristics are written here. Arabic language is written from right to left like some languages suhc as Hebrew, Korean, and Persian. Small letters and capital letters are not found in Arabic like the English language. The letters' shapes are changing in Arabic according to their position in a word for example it may take three positions: at the beginning, in the middle, or at the end of the word. Arabic language's morphology is complex. There are some diacritics in Arabic texts which may or may not be there. There is no Orthographic representation of short letters in modern standard Arabic. Arabic is a pro-drop language, which permits the dropping of subject pronouns [75] and deletion retrieval to subject [76].

2.5.3. ArTS

In recent years many works have been done in ArTS. Imam et al. used the Analogy-based Summarization System for Arabic Documents (OSSAD) which is a user-focused TS system [25]. Al-Taani and Al-Omour approach with the : short-path algorithm which is a graph-based approach that primarily focused on the semantic relationships between the sentences [11]. Al-Taani and Jaradat applied a hybrid-based approach and explored the impact of employing a scoring system that unites semantic and informative scoring strategies to address the issue of accuracy as well as the inattention to semantic connections within sentences [18]. Al-Abdullah and Al-Taani tried to obtain the best summary of a document, by combining informative and semantic scoring and using Particle Swarm Optimization (PSO) algorithm [15].

Al-Radaideh, and Bataineh employed a hybrid mix of statistical characteristics, semantic similarity, and a Genetic Algorithm (GA) which is beneficial since it produced improved summaries regarding precision, recall, and F-measure [16]. Al-Abdullah and Al-Taani utilized the firefly algorithm in which the informative and semantic scores are combined to obtain higher results [17]. Qaroush et al. deployed a generic extractive Single-Document (SD) summarizing technique. They used two strategies, the first strategy is score-based, while the second is based on ML [10]. These techniques are beneficial, but we still need to improve the ArTS system's precision, recall, and F-measure.

It is obvious that graph-based approaches for Arabic NLP have gained interest in recent years. Graphs might be utilized and built in a beneficial way to assist in conquering and reducing Arabic language issues because of their capacity to organize enormous and complicated structures into standard and formal ways. It is apparent that research in Arabic NLP is still in its early phases and requires further work and examination. When it comes to improving Arabic NLP applications, a fundamental difficulty in this subject is a shortage of Arabic NLP resources [24]. Consequently, the proposed model in this study is GEATS approach which is a graph-based approach.

CHAPTER 3. RELATED WORK

The Internet has provided and overwhelmed us with massive amounts of textual data. Furthermore, the increasing use of the internet raises interest in ATS, while addressing the issue of information overload that individuals face in the digital era. However, there are a number of difficulties with the supported languages in TS. The bulk of ATS systems is centered on English and other popular languages. As one of the most commonly spoken languages worldwide is Arabic, there is a huge demand to summarize the massive volume of Arabic textual data. The Arabic summarization system continues to perform poorly due to Arabic language complexity and scarcity of studies on this topic. Although there have been some good efforts in the field of ArTS, they are insufficient to cover this huge domain.

Various strategies, notably in the Arabic language, have been given during the development of ATS. ArTS is important in the Arabic world for a number of reasons, including boosting the use of Arabic-language content on the internet, employing ArTS in a range of situations and locations, and assisting Arabic readers by saving their time, money, and effort. Many researches on ArTS have been conducted between the years 2012 and 2021, as shown in Table 3.1. the methodologies, CR, a multi or single-document, and datasets all differ and provide distinct F-measures. This study intends to develop a new method for extracting SD ArTS systems based on WE and PR algorithms that work with graphs directly to construct a summary for Arabic documents while finding originality and guaranteeing that the final summary is both thorough and consistent and complete by using certain parameters.

Table 3.1. Some ArTS studies according to different years from 2012 until 2021

Reference	Year	Approach	Method	F-measure %	Corpus	SD or MD	CR %
[61]	2012	Statistical	Clustering Techniques	77.1	-	SD	-
[25]	2013	Ontology	OSSAD	49.8	EASC	SD	40
[62]	2014	Statistical	Clustering (K-Means Algorithm)	60	EASC	MD	-
[11]	2014	Graph-based	Short-Path Algorithm	48.6	EASC	SD	40
[63]	2015	Lexical-based	Lexical Cohesion And Text Entailment Relation	69.98	Arabic textual entailment	SD	-
[18]	2016	Hybrid	Hybrid	54.76	EASC	SD	40
[15]	2017	Metaheuristic-based	PSO	55.32	EASC	SD	40
[16]	2018	Metaheuristic-based	GA	60.5	EASC	SD	40
[17]	2019	Metaheuristic-based	Firefly	57.52	EASC	SD	40
[28]	2020	Hybrid	Unsupervised Score-Based (Clustering, Word2vec)	64.4	EASC	MD	30

Table 3.1. Some ArTS studies according to different years from 2012 until 2021. (Continued)

[23]	2020	Graph-based	Graph-based	76.37	EASC	SD	-
[40]	2020	Graph-based	Modified PR Algorithm	67.99	EASC	SD	-
[10]	2021	Hybrid	Statistical and Semantic Features	64.3	EASC	SD	50 [5]
[64]	2021	Hybrid	Documents Clustering, Topic Modeling, And Unsupervised Neural Networks	20.68 [6]	EASC	SD	40
[65]	2021	Metaheuristic-based	GA	41	EASC	SD	-
[66]	2021	Machine Learning	Knapsack Balancing Of Effective Retention	56.14	EASC	SD	-
[67]	2021	Machine Learning	ArDBertSum, DistilBERT model	49	EASC	SD	-

[5] Of document's word count

[6] With Ensemble NN_Sent2Topic_prob

3.1. Arabic Datasets

High-quality datasets are essential for successful NLP studies. There are a variety of Arabic datasets that are used for a variety of applications. Some of these are discussed further below in Table 3.2.

Table 3.2. Arabic datasets' groups, names and description.

Group	Dataset Name	Description
Text Classification	ANT corpus	The ANT dataset [94] is an Arabic corpus which is accessed online. It consists of news articles gathered from RSS feeds by Chouigui et al. in 2017. It is used for text classification. Each document is a represented in XML TREC format.
News Articles	1.5 billion Words Arabic Corpus Dataset	El-Khair et al. published the this Corpus in 2016 [92]. Over fourteen-year period, the data were gathered from newspaper articles in 10 major news sources from eight Arabic nations.
	Khaleej-2004 Corpus Dataset	Abbas[7] et al. created the Khaleej-2004 Corpus [89]. It has about 5690[8] Arabic articles, totaling almost 3 million words, divided into four categories: sports, local news, international news, and economy.
	Watan-2004 Corpus	Abbas et al. constructed the Watan-2004[9] Corpus [88]. It has 20291[10] documents divided into six categories: religion, sports, local news, economics, international news, and culture.
	Saudi Newspapers Corpus	Al-Hagri created the Saudi Newspapers Corpus[11] (SaudiNewsNet). There are 31030[12] Arabic newspaper articles in it.

[7] https://sites.google.com/site/mouradabbas9/arabic-corpora/text-corpora?authuser=0
[8] https://sourceforge.net/projects/arabiccorpus/
[9] https://sourceforge.net/projects/arabiccorpus/files/
[10] https://sourceforge.net/projects/arabiccorpus/
[11] https://github.com/inparallel/SaudiNewsNet/tree/master/dataset
[12] https://github.com/inparallel/SaudiNewsNet

Table 3.2. Arabic datasets' groups, names and description. (Continued)

Gender Bias	The Arabic Parallel Gender Corpus Dataset	Habash et al. built the Arabic Parallel Gender Corpus in 2019 [91], [90]. It is designed to aid studies on gender bias in Arabic NLP applications.
Text Summarization	Essex Arabic Summaries Corpus (EASC) Dataset	EASC dataset[13], created by El-Haj in 2013, comprises 153 Arabic articles and 765 human-generated extracted summaries. The extracted summaries were produced in Arabic by making use of Mechanical Turk.
	KALIMAT	KALIMAT [95] was created by Koulali, and El-Haj in 2013. It is a multiuse Arabic dataset. It includes 20291 Arabic articles, 20,291 extractive SD system summaries, 2057 Multi-Document (MD) system summaries, 20291 named entity recognized articles, 20291 part of speech tagged articles, and 20291 morphologically analyzed articles.
Text Localization	ASAYAR Dataset	ASAYAR is a dataset for Arabic-Latin scene text localization in highway traffic panels. Akallouch et al. [96] created this multilingual and multipurpose dataset in 2020. It is divided into three sub-datasets. The collection comprises 1763 photos taken along various Moroccan highways.
Text Commonsense Validation	Arabic Dataset for Commonsense Validation	Tawalbeh and Al-Smadi introduced a standard Arabic corpus for commonsense understanding and validation[14] in 2020. This is the first dataset in the field of Arabic text commonsense validation [97].
Songs	Habibi	Habibi was created by El-Haj in 2020 [99]. It is the first multi-dialect, multinational Arabic song lyrics dataset. The corpus contains about 30000 song lyrics in six Arabic dialects for singers from 18 various Arabic nations. There are about 500,000 sentences and over 3.5 million words in the lyrics.
Twitter Sentiment	Twitter Dataset for Arabic Sentiment Analysis	Abdulla built the Twitter dataset for Arabic sentiment analysis[15] which includes Arabic tweets.

[13] https://sourceforge.net/projects/easc-corpus/

[14] https://github.com/msmadi/Arabic-Dataset-for-Commonsense-Validationion

[15] https://archive.ics.uci.edu/ml/machine-learning-databases/00293/

Table 3.2. Arabic datasets' groups, names and description. (Continued)

Twitter Sentiment	Arabic Jordanian General Tweets Dataset	The Arabic Jordanian general tweets[16] dataset is created by Alomari in 2017. It consists of 1800 tweets that have been labeled as good or negative. Modern Standard Arabic or Jordanian dialect is used.
	Arabic Sentiment Tweets Dataset	This dataset was built in 2015 by Nabil et al. [101]. It comprises approximately 10K Arabic sentiment tweets. These twets are divided into four categories: objective, subjective negative, subjective positive, and subjective mixed.
Other	ArabicWeb16 Dataset	Suwaileh et al. created ArabicWeb16 in 2016 [100]. It has 150,211,934 Arabic Web pages with excellent coverage of dialectal Arabic and Modern Standard Arabic.
	CC100-Arabic Dataset	Conneau et al. [98] created CC100-Arabic in 2019. It is one of the 100 monolingual corpora. It is processed from the CC-Net source. This corpus is 5.4G in size.
	Arabic in Business and Management Corpora	El-Haj et al. created this corpora[17] in 2016. It comprises 400 Arab firms chairman and chief executive manager statements, 400 Arabic stock market news items, and 400 Arabic economic news pieces, all in Arabic Language.

3.2. ArTS Approaches

Some ArTS researchers employed a variety of methodologies and algorithms, which are detailed in this section. According to the methodologies employed for Arabic ATS, the literature is divided into six major categories: statistical, metaheuristic-based, ML, hybrid, fuzzy-logic based, and graph-based approaches.

[16] https://github.com/komari6/Arabic-twitter-corpus-AJGT
[17] https://sourceforge.net/projects/arabic-business-copora/

3.2.1. Statistical approaches

Statistical approaches need less processing and memory resources [79]. They do not need any additional linguistic understanding or extensive linguistic processing [93].

In 2020, Bialy et al. [26] propose a statistical-based extraction strategy for Arabic SD summarization. The suggested method was divided into three stages: pre-processing, sentence scoring, and summary production. In order to evaluate their system, the authors gathered thirtythree short articles from Wikipedia and had two human specialists describe them. The outcomes are being compared to the summaries of two human specialists. Bialy et al. stated that the outcomes were superior to those of human specialists.

By describing the relationship between documents and their related summaries, Elayeb et al. [27] established an extractive technique for Arabic SD TS by the use of analogical proportions. Two methods are applied: the first examines the document or summary for the existence of keywords, while the second evaluates the frequency of the keywords. ANT corpus and a short EASC test set are used to compare the two methods with several summarizers like Luhn, TextRank, LexRank, and LSA. The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and Bilingual Evaluation Understudy (BLEU) measures are used to compare the two methods. When compared against three other methodologies utilizing the same datasets, promising results are obtained.

Abdulateef et al. [28] proposed an unsupervised score-based method using clustering and Word2Vec to MD ArTS. To eliminate data redundancy in Arabic MDs, bag-of-words and vector space models are utilized with the k-means clustering technique. Preprocessing is used initially to reduce noise from the input. The Word2Vec model is used to convert words into vectors, and the semantic relationships among vectors are represented. Clustering is performed after preprocessing to extract significant sentences from each document. ROUGE measure and EASC dataset are employed throughout the examination procedure. When compared to earlier work in MD ArTS, promising results are obtained.

The summary is created in these ways based on sentence ranking by picking relevant sentences from the source document(s) and applying a suitable CR [29]. Significant variables such as sentence title, sentence length, sentence location, keywords, TF-IDF, etc are used to choose relevant sentences.

3.2.2. Metaheuristic-based approaches

Metaheuristic-based techniques are evolutionary-based ways to find the solution to some difficulties such as complex issues that cannot be resolved in polynomial time. The firefly algorithm is one example of an evolutionary technique.

Baraka and Al Breem [30] presented a method for ArTS of large-scale MDs that makes use of GA and the MapReduce parallel programming methodology. The technique ensures summary generating scalability, speed, and correctness. It reduces sentence duplication while increasing readability and cohesiveness between summary phrases. The trials yielded satisfactory precision and recall results.

PSO was presented by Al-Abdallah and Al-Taani [15] for SD ArTS. The suggested model is assessed by the use of EASC dataset and ROUGE measure. When compared to several current techniques that employed GAs [18] and the harmony search algorithm [14], the obtained results were encouraging.

Based on the EASC corpus, Al-Radaideh et al. [16] suggested SD ArTS technique that combines GAs, statistical characteristics, and domain expertise to pick the final summary. ROUGE was employed as a framework for evaluation.

Al-Abdallah and Al-Taani [17] proposed For extractive SDs ArTS using firefly algorithm. A collection of semantic and informative scores are employed. Since each path in the graph presents a candidate summary, the proposed firefly algorithm is utilized to discover the optimum sub-path from graph’s candidate paths. As a result of the usage of "Term TF-IDF weight" as a novel heuristic characteristic in the computaion of informative ratings, firefly algorithm proposed approach outperformed the evolutionary-based approaches, harmony search and GA regarding F-measure, recall, and precision results. This feature enhances the summary's coherence and cohesion.

Alqaisi et al. [31] proposed an extractive technique for MD ArTS. This method employs multi-objective optimization and clustering methods. DUC-2002 dataset and TAC-2011 dataset are utilized for assessment. The outcomes demonstrated the suggested approach's effectiveness in comparison to other current techniques. Using the ROUGE metrics, the technique achieved good F-measure scores for both datasets.

3.2.3. Hybrid approaches

More than one method is employed as a combination in the hybrid-based strategies to improve the summarization process. When compared to existing summary systems, the hybrid mix of statistical characteristics, semantic similarity, and GAs produced better summaries regarding F-measure, precision, and recall [16].

Ibrahim et al. [32] presented a unique hybrid model for ArTS that combines vector space model and rhetorical structure theory. The suggested approach used rhetorical structure theory to identify the most crusial paragraphs on the basis of functional and semantic characteristics. The proposed model ranks important paragraphs using vector space model according to cosine similarity feature and its output summary is evaluated on three categories of 212 news articles of varying sizes. The statistical findings suggest that the proposed model enhances the average precision of the output text summary over rhetorical structure theory alone while retaining the benefits of rhetorical structure theory summarization.

Ibrahim, and Elghazaly [87] presented a new ArTS hybrid model that combines two sub-models: The first sub-model generates a primary summary by recognizing the most important parts of the text using rhetorical structure theory. Then, the second one ranks the primary important parts in the rhetorical-summary by using the cosine similarity. To examine the suggested approach, a prototype was created using a variety of articles that were divided into three groups. The experiment demonstrates that the offered method improves rhetorical-summary precision.

Cheragui and Lakhdar [33] proposed the SumSAT tool, which is an extraction-based ArTS system. The proposed work is distinguished by the use of a hybrid technique that integrates three methods: contextual investigation, indicative expression, and graph method. The suggested technique is tested by utilizing recall and precision measures to compare the acquired results with human summaries.

Fadel and Esmer [34] presented a hybrid strategy that combines abstractive and extractive techniques to provide an informative and cohesive summary from a lengthy text. The extraction technique offers a unique extraction formulation for a collection of semantic and statistical data from a sentence, taking into account its semantic, significance, and location. Only relevant sentences identified by the extractive technique will be trained with encoder-decoder bidirectional Long Short-Term Memory (LSTM) in the abstractive approach to construct a new summary. They demonstrate that the suggested hybrid technique outperforms and outperforms certain current Arabic summarizing systems.

Qaroush et al. [10] introduced an extractive Arabic SD summarizing method that makes use of supervised ML techniques and a mixture of semantic and statistical data. For evaluation, EASC dataset and ROUGE tool are employed. When compared to earlier techniques, good outcomes are obtained.

3.2.4. Machine learning approaches

Summarization is treated as a classification issue in ML and clustering algorithms, with summary sentences chosen depending on specific attributes. There are two types of ML approaches: supervised and unsupervised. Hidden markov models and the bayesian method are two examples of these techniques. Before extracting sentences from sentence clusters, clustering algorithms are utilized to represent them.

Ellouze et al. [35] developed a novel strategy for automatically assessing the overall responsiveness of ArTs. This approach is based on ML, which works by constructing a model with a range of linguistic features, such as syntactic, lexical, and named entity-based features, as well as a combination of content scores, such as ROUGE, AutoSummENG, MeMoG, NPowER, and SIMetrix scores. They used a regression approach to incorporate the aforementioned characteristics to create the prediction model. The outcomes demonstrate that the suggested technique outperforms the baselines.

Using language embedding models, Lamsiyah et al. [36] offer a DL strategy for SD ArTS. The DUC-2002 corpus is utilized for assessment, and three phrase embedding models are employed. The obtained findings demonstrated the efficacy of the three-sentence embedding models for ArTS when compared to other eight techniques.

Molham and Said [37] have generated an Arabic dataset of Arabic summaries. The collection contains 300,000 items, each of which contains an article introduction as well as the headlines for that introduction. For the summary of Arabic literature using DL algorithms, two abstractive models are presented. The experimental findings demonstrated that the proposed models produced satisfactory results when applied to the provided dataset. There are no comparisons with other methodologies.

Suleiman and Awajan [38] proposed using RNN to abstractly summarize ArTS. The model employs two levels of hidden states at the encoder and one layer of hidden states at the decoder. In the encoder and decoder layers, LSTM is employed. Using ROUGE tool, an artificial dataset is constructed and used to evaluate the summarization model. The experimental findings revealed that the suggested model performed well for ROUGE1-NOORDER and ROGUE-1. A differentiation is also done among the dependency-parsing-based word2Vec model and the original Word2Vec model, demonstrating that the dependency-parsing-based Word2Vec model is superior.

3.2.5. Fuzzy-logic based approaches

Fuzzy-based approach deals with input uncertainties because fuzzy inference systems may give logical assessments in an uncertain and ambiguous context [85]. Some fuzzy ontology-based processes are also being developed for Arabic document summarizing.

Atlam and El-Barbary [84] demonstrated a novel ArTS approach based on the field association fuzzy ontology technique. Initially, a domain ontology containing numerous events in Arabic is defined. The document preparation technique provides relevant terms based on the domain expert's Arabic corpus and Arabic language dictionary. After that, the relevant terms were categorized using a field association term classifier method. The results suggest that the Arabic document based on field association fuzzy ontology terms may be used for summarizing effectively.

Al Qassem et al. [85] developed a new ArTS strategy based on a novel noun extraction method and fuzzy logic. The suggested summarizer is tested against popular modern ArTS techniques using EASC corpus. According to the results, the fuzzy logic technique with noun extraction surpasses previous systems.

3.2.6. Graph-based approaches

The semantic connections among document sentences are the focus of these techniques. The document's sentences are represented as a graph, with nodes representing sentences as well as edges representing connections among sentences.

Al-Omour and Al-Taani [11] evaluated the impact of several fundamental units (word stems, words themselves, and n-grams) on the effictiveness of an extractive graph-based technique for ArTS. When tested on EASC corpus and ROUGE tool, the new strategy outperforms certain earlier approaches. When n-grams are employed in the summarizing process, the best results are obtained.

Elbarougy et al. [39] studied the effect of stopword removal as a preprocessing step on the effectiveness of an ArTS graph-based technique. Two tests are carried out: the first involves creating the summary with stopwords, and the second involves eliminating stopwords. Experiment findings on EASC corpus revealed that removing stopwords improved the performance of the summarization procedure.

Elbarougy et al. [40] described a graph-based solution for ArTS based on a modified PR algorithm. The suggested method is divided into three phases: preprocessing, the extraction of features and graph creation, and the extraction of summary utilizing the adjusted PR algorithm. When compared to previous techniques, the suggested methodology produced the best results on the EASC dataset after 10,000 iterations.

The best scores obtained for recall, F-measure, and precision are 0.729, 0.679, and 0.687, respectively.

Elbarougy et al. [23] examined the use of morphological analysis on the performance of the graph-based technique for ArTS using three morphological analyzers which are BAMA, Stanford NLP, and Safar Alkhalil. EASC corpus is utilized to assess the effectiveness of certain morphological analyzers. This research showed that the Safar Alkhalil analyzer outperformed the other analyzers.

3.3. Comparative Evaluation of ArTS Approaches

Statistical-based strategies demand less CPU capacity and memory [79], however certain key sentences may be excluded from the summary since they have a lower rank than others, yet alike sentences may be included since they have a higher score. To enhance sentence selection for summary in ML based approaches, a big quantity of training data is essential [80], which is manually generated extractive summaries in which each sentence in the initial training examples may be classified as "summary" or "non-summary" [80].

Although employing the strength of GAs in metaheuristic-based algorithms to determine the ideal weights is advantageous [81], it involves a significant amount of computational time, cost, and a fixed number of iterations or loops. In graph-based technologies, graphs that represent sentences as a bag of words and employ a similarity metric which possibly will miss terms that are semantically equivalent or identical [102]. The precision of similarity computation has an impact on the chosen sentences [103]. Consequently, WE has been used in GEATS.

Graph-based technologies improve coherence and detect duplicate information [80], are language-independent [82], and domain-independent [83]. Because of their ability to arrange vast and sophisticated structures into standard and formal ways, graphs may be used and developed in a useful method to assist in overcoming and minimizing Arabic language challenges.

NLP is a branch of computer science that studies human–computer interaction. NLP has many branches and one of them is TS. There are several approaches in TS, including working with various languages. This section discusses some of the most common approaches that deal with the Arabic language. As we can see in Table 3.3. the advantages and disadvantages of ATS approaches are shown. The employed method in this study is a graph-based strategy named GEATS. Because of their capacity to organize huge and challenging structures in standard and formal ways, graph-based techniques have obviously gained appeal in recent years. Furthermore, graphs may be utilized and developed to assist in overcoming and reducing Arabic language problems, such as complex morphological relations.

This section describes and categorizes various Arabic datasets, including text classification, news articles, text localization, text summarization, etc. Six primary approaches to text summarizing challenges are discussed and evaluated addressing advantages and disadvantages. Each strategy is best suited to specific types of issues and languages. The graph-based method is more convenient for Arabic language since graphs can be used and developed in a beneficial way to aid in conquering and minimizing Arabic language challenges due to their ability to arrange vast and intricate structures into standard and formal ways. Our technique is a graph-based approach with certain additional qualities that intends to establish a new way for extracting single-document ArTS systems using specified parameters based on word embedding and PR algorithms. The GEATS system was evaluated by the use of EASC dataset.

Table 3.3. Comparative Evaluation of ArTS Approaches.

Approach	Advantages	Disadvantages	Examples
Statistical	It needs minimal processing and memory resources [79]. There is no need for any additional linguistic understanding or extensive linguistic processing in statistical approach [93].	In a statistical approach, certain key sentences may be excluded from the summary since their rank is lower than others, yet alike sentences may be included since their score is higher.	[26], [27], [28], [29]
Metaheuristic-based	It is used to find a solution to some problems such as complex issues that their solution cannot be found in polynomial time. Employing the strength of GAs to determine the ideal weights is advantageous [81].	It involves a significant amount of computational time, cost, and a fixed number of iterations or loops.	[14], [15], [16], [17], [18], [30], [31]
Hybrid	It improves the summarization process. When compared to existing summary systems, the hybrid mix of statistical characteristics, semantic similarity, and GAs produced better summaries concerning precision, F-measure, and recall [16].	The system could be complex when using more than one approach and require computational time, more memory and CPU capacity.	[10], [32], [87], [33], [34]
Machine Learning	A significant volume of training data is necessary to enhance sentence selection for the summary [80]. Relatively simple regression models can outperform other classifiers [79].	It requires a big quantity of training data [80], which is manually generated extractive summaries in which each sentence in the initial training examples may be classified as "summary" or "non-summary" [80].	[35], [36], [37], [38]
Fuzzy-logic based	Fuzzy-based approach deals with input uncertainties because fuzzy inference systems may give logical assessments in an uncertain and ambiguous context [85].	In the summary, the duplication of chosen sentences is a negative issue that may arise and impact summary's quality [105]. To increase the quality of the final summary, a redundancy elimination strategy is necessary in the post-processing step.	[84], [85]
Graph-based	It improves coherence and finds out duplicate information [80], is language-independent [82], and domain-independent [83].	It supposes that the weights of the words are identical, hence it disregards terms' significance in the document [104]. It does not address difficulties such as the dangling anaphora [80].	[11], [34], [40], [23]

CHAPTER 4. THEORETICAL BACKGROUND

4.1. Graphs in TS

A graph G represents the Arabic text input. A directed graph G = (V, E) is a graph G of document D, where V is a collection of nodes and E is a set of edges [41]. Indeed, V and E are the two primary components of the graph. In other terms, G is a weighted directed network whose nodes represent D sentences and whose edge weights indicate sentence similarity.

G(V, E) is a mathematical structure that represents the pairwise relationship among items. Edges reflect the nature of the relationship between two vertices, whereas vertices represent the fundamental component of the depicted system. To use a graph model to design a solution, you must address three primary issues: (1) What are your application's fundamental components, in which text summary might be words, phrases, sentences, or even paragraphs? (2) The sort of relationship among nodes used to determine the weight of the edges in TS, such as cosine similarity or overlapping phrases, and etc. (3) the graph's vertices were ranked using a ranking algorithm [40]. Many techniques for text summarization exist, including LexRank [42], TextRank [43], and the PR algorithm [44].

TextRank is a graph-based, SD ranking model adapted on Google PR algorithm [44]. The similarity between phrases is represented as an edge weight in TextRank, which is an undirected linked graph. Both sentences and keywords are extracted using TextRank. Following the use of TextRank, sentences are ordered according to their score, with the highest-scored sentences being chosen as a summary.

LexRank is a graph-based MD summarizing model in which all sentences are represented as a graph. Two sentences are linked if their similarity exceeds a certain threshold. Following the construction of the graph, the most centric sentences are chosen as a summary.

4.2. PR Algorithm

An essential property of a node in compound networks such as the world wide web is its in-degree (out-degree), which is the number of inbound (outbound) connections on the node [45]. The in-degree of a specific page can be thought of as an esstimation of the significance or quality of that page [46]. The PR algorithm [46] has expanded on this concept by not counting incoming connections from all pages equivalently but instead normalizing based on the momentousness and quantity of outbound connections from nearby pages. In this regard, the PR value may be a superior measure of relevance because it integrates both the paper's visibility and authority by calculating the total number of citations and reputation of the citing publications into consideration [46]. $PR(A)$, is determined using a simple iterative method that corresponds to the primary eigenvector of the web's normalized link matrix [46]. The PR of Web page A, indicated by PR, as defined by (Equation 3.1).

PR(A) is the primary eigenvector of the web's normalized link matrix and may be determined using a simple iterative approach.

$$PR(A) = (1 - d) + d * \sum_i \frac{PR(T_i)}{C(T_i)} \tag{3.1}$$

Where $PR(T_i)$ is the PR of page T_i that is linked to page A, $C(T_i)$ is the quantity of outbound connections on page T_i , and d is a damping factor that can adjusted between the range of [0,1].

The PR of A is recursively determined by PR algorithm of those pages that connect to page A, as shown in (Equation 3.1). The PR of pages T_i is always weighted by the quantity of outbound connections $C(T_i)$ inside the algorithm, resulting in a lower PR value transmitted from pages T_i to the receiving page A. It is also anticipated that each new inbound connection to a recipient page A would always boost A's PR.

CHAPTER 5. GEATS SYSTEM USING PR AND WE

Text summarization has recently captured the interest of researchers. The primary goal of TS is to limit the amount of paragraphs and statements in the text in order to offer readers with adequate information to determine whether the content is valuable or not. TS is used in several languages, including Arabic. ArTS has been used in various approaches such as statistical, hybrid, metahuristic-based, graph-based, machine learning, and so on, as well as various methods have been used such as clustering techniques, short-path algorithm, particle swarm optimization, GA, statistical and semantic features, knapsack balancing of effective retention, modified PR algorithm, etc. Various datasets have also been used to evaluate the proposed systems.

As previously stated, there are several types of text summarization processes and specifically of ArTS. Although there has been several methods to solve ArTS problem, there are some issues such as the significant computational time, cost, low performance and accuracy. In this study, Arabic text summarization problem is addressed by using PageRank and word embedding with a graph-based approach named GEATS.

This study concentrates on extractive TS using Arabic SD. Our proposed model is a graph-based model and is tested using EASC corpus with a 40% CR. This research aims at finding a way for extracting SD ArTS systems that is built on the WE and PR algorithms and works straight with graphs to build an Arabic document’s summary while identifying originality. This procedure has different processes. We want to enhance the effectiveness of our suggested model (GEATS) by better preprocessing procedures. Then comes gathering the necessary characteristics, constructing the graph, applying the PR algorithm, and finally extracting and assessing the summary.

The proposed approach is discussed in this section. Figure 5.1. depicts the GEATS's flow chart, which includes three primary steps. The first stage begins with text extraction from a document, followed by preprocessing operations like normalization, stop words removal, and stemming. The desired features are retrieved in the second stage, and the document is then represented as a graph. Eventually, in the third stage, the PR method is used to provide a summary, after which the performance is reviewed and the results are displayed. The stages of the GEATS system are explained in the following stages.

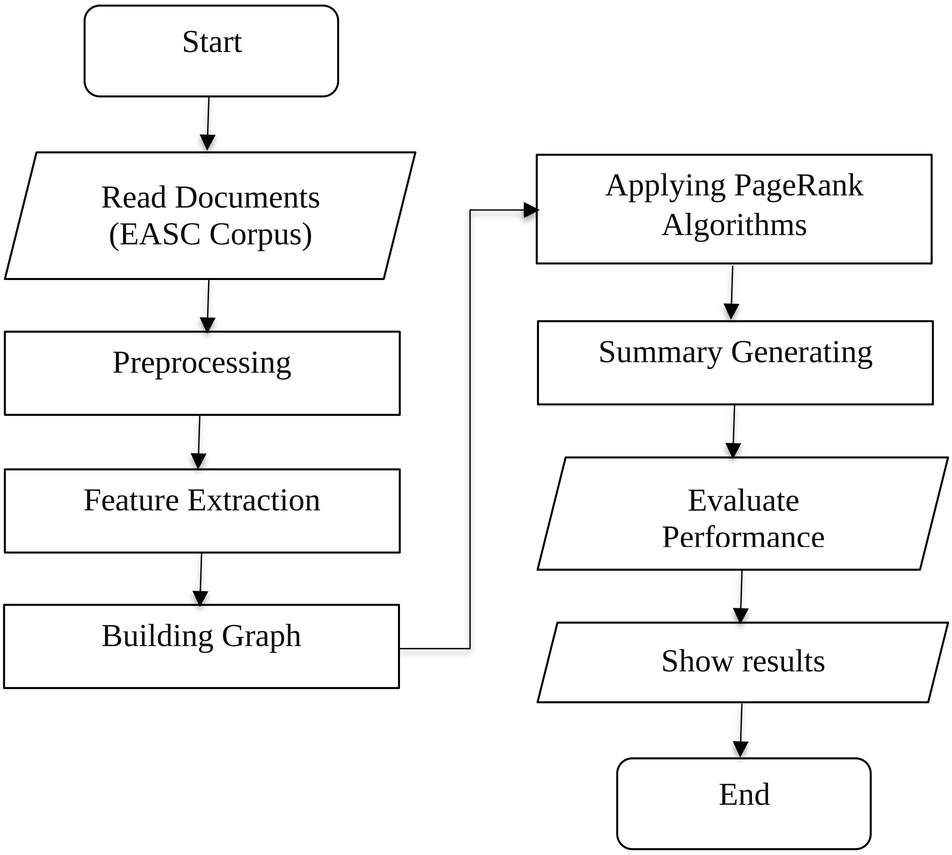

Figure 5.1. The flow chart of the GEATS approach

5.1. Stage 1: Preprocessing

The Arabic language is classified as having a wealthy and complicated morphological and syntactic flexibility [47]. Consequently, dealing with Arabic documents straight in information retrieval in the absence of any preprocessing phases will make dealing with the text more challenging and give us or the user inaccurate findings. As a result, some languages processing must occur before summarization phase and after entering the documents step, as we will see in the following steps. Consequently, at his point, the document is entered and processed in preparation for feature extraction.

5.1.1. Importing Documents

The used dataset is EASC dataset that contains 153 documents from different topics. This stage also involves loading and importing the Arabic document that will be summarized from the EASC dataset. Then, extracting text from each document that is written in the Arabic language. After importing all the documents normaliation will take place in this system.

5.1.2. Normalization

After importing all the documents normalization process is so essential. Normalization is the process of transforming text into another format in order to improve consistency utilizing various processing mechanisms. Normalization has a significant impact on the extracted summary's goodness since it removes repetitive phrases, duplicated white spaces, and so on.

Normalization entails the following steps: removing diacritics or Tashkeel such as " َ " Fatha, " ً " Tanwin Fath, " ُ " Damma, etc, removing punctuation and links, dealing with duplicate white spaces, and dealing with numerous full stops. This will enhance the system's performance positively.

In Arabic, there are distinctive notations known as diacritics. It is used to assist Arabic readers in correctly pronouncing Arabic words. Diacritics are assigned based on the Arabic grammar rules. When the word's location in the sentence changes, it results in various diacritics and a distinct meaning. Although diacritics are very important in Arabic texts, when creating the summary for each document it is better to omit them.

The available diacritics in Arabic are shown in Table 5.1. These diacritics will be eliminated from the text. This list of 16 diacritics in Arabic includes Fatha, Dama, Kasra, and so on.

Table 5.1. Diacritics in Arabic language.

Diacritic’s Name	Diacritic’s Shape	Diacritic’s Name	Diacritic’s Shape
Fatha	ـَ	Tanwin Fath	ـً
Dama	ـُ	Tanwin Dam	ـٌ
Kasra	ـِ	Tanwin Kasr	ـٍ
Shadda Fath	ـَّ	Shadda and Tanwin Fath	ـًّ
Shadda Dam	ـُّ	Shadda and Tanwin Dam	ـٌّ
Shadda Kasr	ـِّ	Shadda and Tanwin Kasr	ـٍّ
Shadda	ـّ	Shadda and Sukun	ـّْ
Sukun	ـْ	Madah	_~

An example of deleting diacritics from sentences is shown in Table 5.2. In this case, the initial text includes several words with diacritics, and the result after removing diacritics is presented in the table which is highlighted with light green 3.

Table 5.2. Diacritics removing example.

Original Text	After Removing Diacritics
تتباين رؤية الإسلام عن الرؤية اليهودية في مسألة التوراة، إذ يتّفق الفريقان أن التوراة من عند الله أنزلها على موسى كما توضّح الآية 3 من سورة آل عمران "نَزَّلَ عَلَيْكَ الْكِتَابَ بِالْحَقِّ مُصَدِّقاً لِّمَا بَيْنَ يَدَيْهِ وَأَنزَلَ التَّوْرَاةَ وَالإِنجِيلَ.." والآية 53 من سورة البقرة "وَإِذْ آتَيْنَا مُوسَى الْكِتَابَ وَالْفُرْقَانَ لَعَلَّكُمْ تَهْتَدُونَ". فالتوراة من عند الله ولكن يعتقد المسلمون بأن توراة اليوم طرأ عليها زيادة ونقصان (تحريف بشكل عام) مقارنة بالتوراة المنزّلة على موسى.	تتباين رؤية الإسلام عن الرؤية اليهودية في مسألة التوراة، إذ يتفق الفريقان أن التوراة من عند الله أنزلها على موسى كما توضح الآية 3 من سورة آل عمران "نزل عليك الكتاب بالحق مصدقا لما بين يديه وأنزل التوراة والإنجيل.." والآية 53 من سورة البقرة "وإذ آتينا موسى الكتاب والفرقان لعلكم تهتدون". فالتوراة من عند الله ولكن يعتقد المسلمون بأن توراة اليوم طرأ عليها زيادة ونقصان (تحريف بشكل عام) مقارنة بالتوراة المنزلة على موسى.

5.1.3. Removing punctuation

Arabic, like other languages, requires several marks to arrange texts and provide readers with a proper meaning of sentences. These signs or marks are known as punctuation marks. Because punctuation in the text summary has no meaning, we eliminate all punctuation except the full stop. Table 5.3. lists the punctuation that should be eliminated when it appears in the text. These punctuation marks include commas, brackets, and so on. The example in Table 5.4. explains how to remove punctuations ,which are bold and highlighted with yellow 3 color, from a sentence.

Table 5.3. Punctuation's Symbols.

Punctuations					
<>	'	{ }	~	¦	+
`	|	…	“”	–	-
-	‘,	""	/	:	"
_	()	*	&	^	%
؛	×	÷	][	,	»«
~	@	\	^	=	;
<	?	$	!	#	[ ]

Table 5.4. Punctuation removing example.

Original Text	After Removing Punctuation
تتباين رؤية الإسلام عن الرؤية اليهودية في مسألة التوراة، إذ يتّفق الفريقان أن التوراة من عند الله أنزلها على موسى كما توضّح الآية 3 من سورة آل عمران "نَزَّلَ عَلَيْكَ الْكِتَابَ بِالْحَقِّ مُصَدِّقاً لِّمَا بَيْنَ يَدَيْهِ وَأَنزَلَ التَّوْرَاةَ وَالإِنجِيلَ.." والآية 53 من سورة البقرة "وَإِذْ آتَيْنَا مُوسَى الْكِتَابَ وَالْفُرْقَانَ لَعَلَّكُمْ تَهْتَدُونَ." فالتوراة من عند الله ولكن يعتقد المسلمون بأن توراة اليوم طرأ عليها زيادة ونقصان (تحريف بشكل عام) مقارنة بالتوراة المنزّلة على موسى.	تتباين رؤية الإسلام عن الرؤية اليهودية في مسألة التوراة إذ يتّفق الفريقان أن التوراة من عند الله أنزلها على موسى كما توضّح الآية 3 من سورة آل عمران نَزَّلَ عَلَيْكَ الْكِتَابَ بِالْحَقِّ مُصَدِّقاً لِّمَا بَيْنَ يَدَيْهِ وَأَنزَلَ التَّوْرَاةَ وَالإِنجِيلَ.. والآية 53 من سورة البقرة وَإِذْ آتَيْنَا مُوسَى الْكِتَابَ وَالْفُرْقَانَ لَعَلَّكُمْ تَهْتَدُونَ. فالتوراة من عند الله ولكن يعتقد المسلمون بأن توراة اليوم طرأ عليها زيادة ونقصان تحريف بشكل عام مقارنة بالتوراة المنزّلة على موسى.

5.1.4. Unifying ALEF's style

ALEF is the first character in the alphabet of Arabic language. ALEF may be written in multiple forms or shapes like (" إ أ ٱ آ ا ") depending on its location in the word. The system changes each occurrence of ALEF in the text to (" ا ") format to make the whole different forms of ALEF character in one same form, that aids in the stemming process positively. The ALEF style was changed for all possible locations of the ALEF character in each word, that contains it, to one format that is highlighted by light green 3 color as shown in Table 5.5. which illustrates how to deal with ALEF style in a sentence.

Table 5.5. Check and unifying ALEF style example.

Original Text	After Checking and Unifying ALEF Style
تتباين رؤية الإسلام عن الرؤية اليهودية في مسألة التوراة، إذ يتّفق الفريقان أن التوراة من عند الله أنزلها على موسى كما توضّح الآية 3 من سورة آل عمران "نَزَّلَ عَلَيْكَ الْكِتَابَ بِالْحَقِّ مُصَدِّقاً لِّمَا بَيْنَ يَدَيْهِ وَأَنزَلَ التَّوْرَاةَ وَالإِنجِيلَ.." والآية 53 من سورة البقرة "وَإِذْ آتَيْنَا مُوسَى الْكِتَابَ وَالْفُرْقَانَ لَعَلَّكُمْ تَهْتَدُونَ". فالتوراة من عند الله ولكن يعتقد المسلمون بأن توراة اليوم طرأ عليها زيادة ونقصان (تحريف بشكل عام) مقارنة بالتوراة المنزّلة على موسى.	تتباين رؤية الاسلام عن الرؤية اليهودية في مسالة التوراة، اذ يتّفق الفريقان ان التوراة من عند الله انزلها على موسى كما توضّح الاية 3 من سورة ال عمران "نَزَّلَ عَلَيْكَ الْكِتَابَ بِالْحَقِّ مُصَدِّقاً لِّمَا بَيْنَ يَدَيْهِ وَاَنزَلَ التَّوْرَاةَ وَالاِنجِيلَ.." والاية 53 من سورة البقرة "وَاِذْ اتَيْنَا مُوسَى الْكِتَابَ وَالْفُرْقَانَ لَعَلَّكُمْ تَهْتَدُونَ". فالتوراة من عند الله ولكن يعتقد المسلمون بان توراة اليوم طرا عليها زيادة ونقصان (تحريف بشكل عام) مقارنة بالتوراة المنزّلة على موسى.

5.1.5. Removing stopwords

This step is critical because we want to eliminate all stopwords ,which are shown in the annex in Table 0.1., from the material that will be summarized. Thus, what exactly are stopwords? Stopwords are a group of words that are widely employed in many languages to do various jobs such as a connector or a different task that gives your phrase a nice meaning. They are recurred throughout the text, such as (من ، على ،إلى، في ،...). The process of removing stopwords are shown in Algorithm (5.1). and the example of removing stopwords, which are highlighted in light yellow 3 color, from a sentence is shown in Table 5.6. In general, removing stopwords improves information retrieval efficiency since common words have a high tendency to diminish frequency differences, and reduce the length of the document, which has an impact on the weighing procedure [48].

Algorithm 5.1. The pseudo code of stopwords removal.

```
For each document
    For each sentence
        For each word
            If (word is a stopword ):
                Continue;
            Else:
                Add_To_Text;
            End IF
        End For
    End For
End For
```

Stopwords can be grouped into diverse groups [48] such as adverbs, prepositions, pronouns, coin names, relative pronouns, conditional pronouns, verbal pronouns, interrogatice pronouns, measurement units, referral names or determiners, transformers (verbs, letters), etc.

Table 5.6. Stopwords removing example.

Original Text	After Removing Stopwords
الفضة من المعادن الكريمة ابيض اللون، وهو معدن ثمين معروف منذ القدم حيث عرفه قدماء المصريين والعرب والصينيون واستخدموه في صناعة الحلي وفي الطب والوقاية من الامراض. تستخدم في النقود والحلي تماماً كالذهب الا أنها اقل قيمة.	الفضة المعادن الكريمة أبيض اللون، معدن ثمين معروف القدم عرفه قدماء المصريين والعرب والصينيون واستخدموه صناعة الحلي وفي الطب والوقاية الأمراض . تستخدم النقود والحلي تماماً كالذهب أنها قيمة .

In fact, there is no particular or fixed stopword list in Arabic language. It differs from one researcher to researcher and from one topic to a different one, however in this thesis, the Arabic stopwords list is utilized from The Natural Language Toolkit[18] (NLTK).

5.1.6. Stemming

The technique of reducing words to their origins or roots is known as stemming. Roots or fundamental forms of words are stemmed by removing any affixes that have been applied to them. The goal of this stage is to obtain the origin or the root of the term, which will enhance the relationship weighing process among sentences, resulting in an improvement in the goodness of the summarization process. For instance, stemming an Arabic word ‘writing’ "كتابة" gives the origin or source word ‘write’ "كتب". The word ‘writer’ "كاتب" can also be used to form this root. Following the reduction of words to their origins, the resulting origins could be utilized for a variety of implementations or applications such as compression, spell checking, and text searching. To this end, in this process, Farasa stemmer[19] [49] is utilized to extract the origin or root of every word in the sentence as illustrated in Table 5.7. This process is used to minimize the quantity of distinguished words in the document to make a finer term frequency computations when using word2Vec representation. Gensim[20] [69], a Python library, has been used to assist in implementing Word2Vec.

Table 5.7. Stemming example.

Original Text	After Stemming
الفضة من المعادن الكريمة ابيض اللون، وهو معدن ثمين معروف منذ القدم حيث عرفه قدماء المصريين والعرب والصينيون واستخدموه في صناعة الحلي وفي الطب والوقاية من الامراض. تستخدم في النقود والحلي تماماً كالذهب الا انها اقل قيمة.	فضة من معدن كريم أبيض لون ، هو معدن ثمين معروف منذ قدم حيث عرف قديم مصري عرب صيني استخدم في صناعة حلي في طب وقاية من مرض . استخدم في نقد حلي تمام ذهب إلا أن أقل قيمة .

[18] https://www.nltk.org/

[19] https://github.com/MagedSaeed/farasapy

[20] https://pypi.org/project/gensim/

5.1.7. Tokenization

Tokenization is known as segmentation. It is the process of reducing any text to a lower number of units. These parts or units can be words, sentences, and so on. During this process, each document is broken into paragraphs, after that into sentences, and eventually into words. Words have been used in the stemming process. After that words were represented by word embedding using Word2Vec representation.

5.2. Stage 2: Feature Extraction and Graph Construction

In feature extraction and graph construction the required features are retrieved at this stage, and each document is then shaped as a graph.

5.2.1. Features extraction

At this stage, two sorts of characteristics are extracted. Here, the term is equal to word's origin or root.

5.2.2. Word2Vec and cosine similarity amongst two sentences

Word2Vec is a helpful way for creating WE [50]. This is equivalent to representing a term as a vector [51]. During this method, we first create a vocabulary set from the full training data set after preprocessing the input corpus. Word2Vec is used to generate an embedding vector for each term in the documents and to derive the semantic relationship between the word lists. Following training, each word attaches a vector with a dimension of 100.

Cosine similarity is a metric that computes the cosine angle between two vectors [52]. This technique determines the degree of similarity between sentences in documents that are represented by vectors. If the two vectors are equal, the similarity is strong, and we get a value of one. For each document, sentences’ vectors are compared and the highest value will be chosen to produce sentence representation for the summary.

Each term is represented as a vector in this context, with each vector including the Word2Vec representation, and the cosine formula is then used for two sentences as follows.

To calculate the cosine similarity amongst two sentences (S_i, S_j) in the same document, we follow the following procedures: (1) Compute Word2Vec for each term in the two sentences. (2) find the term representation in vectors for the two sentences and see the similar or close words, the length of this list is “n”. (3) we perform repeatedly on the similar or close in the meaning list and apply (Equation 4.1).

$$cosine - similarity(S_i, S_j) = \frac{\sum_l^n Word2Vec(S_i) * Word2Vec(S_j)}{\sqrt{\sum_l^n Word2Vec(S_i)^2 * \sum_l^n Word2Vec(S_i)^2}} \quad (4.1)$$

5.2.3. Graph construction and weighing

As seen in Figure 5.2, one of the documents is represented as a graph. The representation of one document as a graph that has 11 nodes, and 66 edges. Sentences are the nodes or the vertices. Each pair of vertices is connected by an edge with an equal weight to cosine similarity, as shown in (Equation 4.1).

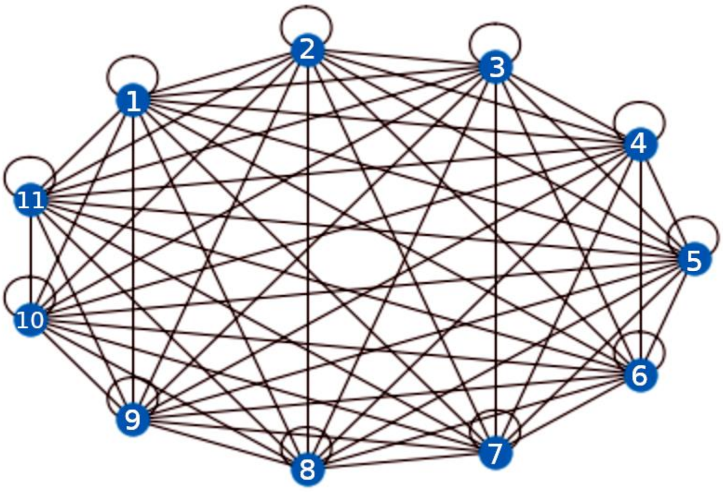

Figure 5.2. The representation of one document as a graph that has 11 nodes, and 66 edges, before ranking

5.3. Stage 3: Using the PR Algorithm and the Extraction of the Summary

The PR method is used at this stage, and the summary is extracted.

5.3.1. Using the PR algorithm

In this phase, the PR algorithm is implemented to all documents, yielding a rating for each sentence. The PR algorithm is used with a maximum of 10,000 iterations to achieve significant results [23].

5.3.2. The extraction of the summary

Nodes are sorted in this phase based on their final ranks. The best 'n' sentences are picked based on the greatest number of summary’s cut-off sentences. The value of 'n' is determined by the CR. This rate represents a specific proportion of the sentence’s quantity in the initial text. Consequently, Sentences are retrieved one at a time and added to the summary until the CR of 40% is met.

5.3.3. Rearranging the sentences

After the summary is generated, the sentences are reordered to provide the best representation of the summaries' sentences and to enhance precision.

5.3.4. Selecting and comparing with ground-truth summary files

In this phase, a ground-truth summary is picked out from the datsaset to assess the summary's goodness. EASC corpus contains five ground-truth summaries. They are compared to the extracted or generated summary. Algorithm (5.2) depicts the pseudo-code for the suggested methodology, which begins with document reading and progresses through preprocessing, feature extraction and graph formation, applying PR algorithm, summary extraction, and evaluation stages.

Algorithm 5.2. GEATS approach.

```
Input: Entire EASC corpus's documents
Output: All documents
For each Document:
    For each sentence:
      Normalization()
      RemovingStopWords()
      For each word:
          Stemming_FarasaStemmer()
          Tokenization()
          WordEmbedding_Word2Vec()
      GenerateSentenceVector()
      UseCosineSimilarity()
    MakeSimilarityMartix()
    GraphBuilding()
    UsePageRankAlgorithm()
    Generate_Summary(CR)
Output: Generated Summary
```

5.4. Implementation

Python is free and open source, general purpose programming language[21]. Many programming tools, applications, and libraries which are crucial in this implementation, therefore Python is employed in the GEATS system programming process.

5.4.1. Used tools and programs

Particular tools and programs are utilized to complete the implementation of automatic ArTS and documentation process of the stعdy. The main ones are specified here.

[21] https://en.wikipedia.org/wiki/Python_(programming_language)

Farasa and Arabic light stemmer have been used in the stemming process. Google Sheets have been used to make the tables and compute evaluation results and calculate the final precision, recall, F-measure. Colaboratory[22], or “*Colab*”, is used for the execution and writing of the Python codes. Google Docs sare used fundamently in documentation’s writing for both the application and thesis final report.

[22] https://research.google.com/colaboratory/faq.html

CHAPTER 6. EXPERIMENTATION AND RESULTS

In TS, there is no summary that could be considered as a gold or a standard one. In other terms, there are several summaries that may be created for each document, depending on an expert human who develops the summary as well as his educational and technological background. Surfing through research publications on TS, we discover that the expert human evaluator does not concur on a single summary for each paragraph. As a result, evaluating TS is a challenging procedure.

6.1. Dataset (Corpus)

For the evaluation of the proposed methodology, the EASC[23] corpus generated by Mechanical Turk (Mturk) [53] is used. EASC is utilized as a standard dataset. The corpus is containing 153 documents, each with five summaries which are written by humans. The articles in this corpus were gathered from three different sources: Wikipedia, Alwatan newspaper, and Alrai newspaper; with 106 articles, 34 articles, and 13 articles, respectively. The subjects or general topics in the EASC corpus are: religion, education, science and technology, environment, finance, tourism, health, politics, sports, art, and music. For each document, the system generates one summary. The corpus's details are shown in Table 6.1.

Table 6.1. EASC corpus’s details.

Corpus	Documents’ Number	Summaries’ Number	Topics’ Number	Average Number of Sentences
EASC	153	765	10	17

[23] https://sourceforge.net/projects/easc-corpus/

6.2. Evaluation Metrics

An evaluation process is performed to evaluate the quality of GEATS strategy. There are two forms of evaluation. The first is manual evaluation, which requires humans to determine the technique's quality, but this is time-consuming and expensive. The second way is known as automatic evaluation, and it is both faster and less expensive than the manual one. We employed ROUGE metric [54] which counts the quantity of overlapping units, e.g. word sequences, amongst the computer-generated summary to be assessed and the ground-truth summaries made by humans when evaluating the summary. ROUGE measurements are classified as ROUGE-N, ROUGE-W, ROUGE-L, and ROUGE-S [54]. Considering F-score, precision, and recall, we employed ROUGE-N (ROUGE-1, ROUGE-2) and this allows us to assess the overlap between the machine or generated summary and the reference or golden-truth summary by counting the similarity units between each of them as unigrams or bigrams, as specified in (Equation 6.1-6.4), respectively.

The BLEU [5] score is also used. BLEU is based on precision. It is used for summarization evaluation, but it is most frequently employed for automatic machine translation evaluation. In (Equation 6.5), we represent the modified precision score p_n. Recall: is calculated by dividing right sentences' number by outcomes' number that should be returned. Recall is referred to as sensitivity in binary classification. As a result, it can be viewed as the likelihood that the query will return a relevant document [55].

$$Recall = \frac{gram_{ref} \cap grams_{gen}}{grams_{ref}} \tag{6.1}$$

Precision: is defined as right sentences' number divided by the total quntity of returned outcomes. Precision in binary classification is equivalent to positive predictive value. Precision takes into account all retrieved documents [56].

$$Precision = \frac{gram_{ref} \cap grams_{gen}}{grams_{gen}} \tag{6.2}$$

Where the ground-truth (reference) summary grams are represented by gramsref and the generated summary grams are represented by gramsgen.

F-measure: using precision alone is insufficient. Furthermore, employing recall alone is inaccurate, thus we utilize the F-measure to balance these two metrics.

$$F - measure = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{6.3}$$

ROUGE-N: is a recall related metric that is calculated by counting the quantity of overlaps amongst the generated summary and the reference (ground-truth) summary.

$$ROUGE - N = \frac{\sum_{s \in ReferenceSummaries} \sum_{N-gram \in S} Count_{match(N-gram)}}{\sum_{s \in ReferenceSummaries} \sum_{N-gram \in S} Count(N-gram)} \tag{6.4}$$

Where N is the total size of the n-gram, count match N-gram is the highest quantity of grams found in both the system and human (ground-truth) summaries, and count over N-gram is the total quantity of n-grams that are in the human summary.

BLEU: is an adjusted form of precision that evaluates the degree of similarity amongst the reference summary and the generated summary.

$$P_n = \frac{\sum_{C \in \{Candidates)} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in \{Candidates)} \sum_{n-gram \in C'} Count(n-gram)} \tag{6.5}$$

Where $Count(n - gram)$ indicated the n-grams quantity that candidate in the test set, $Count_{clip}(n - gram)$ represents the clipped n-grams quantity for the candidate sentences.

6.3. Experiment Setup

This section demonstrates how to graph formation and the PR algorithm function. As we can see in Table 6.2. An example of Arabic one document which contains six sentences. For each document, the graph will be built. Figure 5.2. depicts one document as a graph before ranking. Figure 6.1. depicts the graph after the PR algorithm has been applied; every node in the graph has its own rank or weight.

After using the PR algorithm. Figure 6.1. depicts the scores or ranks of network nodes; because the CR is 40%, the number of sentences in the output summary will be four. Furthermore, according to the diagram, nodes representing sentences 5, 7, 9, and 11 return the greatest rank, hence they are included in the summary.

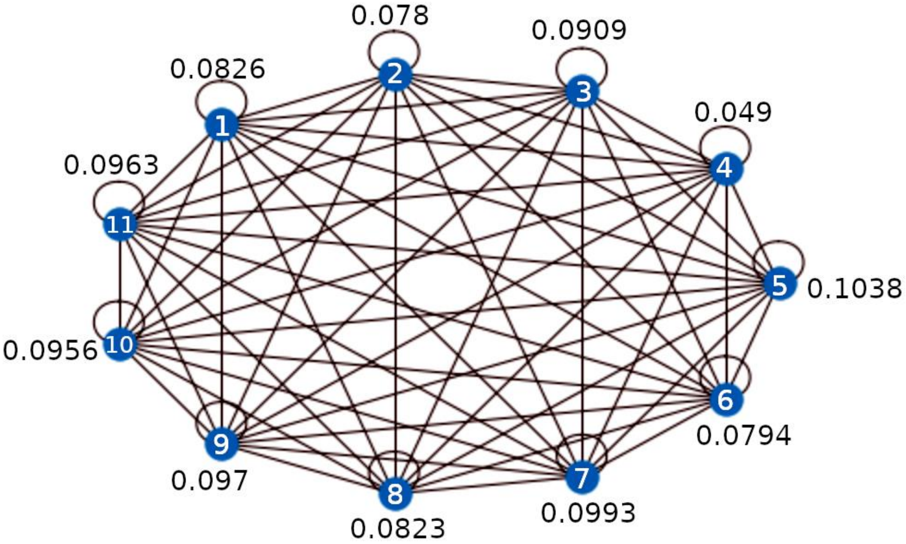

Figure 6.1. The representation of one document as a graph after ranking

Table 6.2. Example of an Arabic SD sentences.

Sentence Content	Number
الزلزال هو ظاهرة طبيعية عبارة عن اهتزاز أرضي سريع يعود إلى تكسر الصخور وإزاحتها بسبب تراكم إجهادات داخلية نتيجة لمؤثرات جيولوجية ينجم عنها تحرك الصفائح الأرضية.	1
قد ينشأ الزلزال كنتيجة لأنشطة البراكين أو نتيجة لوجود انزلاقات في طبقات الأرض.	2
تؤدي الزلازل إلى تشقق الأرض ونضوب الينابيع أو ظهور الينابيع الجديدة أو حدوث أمواج عالية إذا ما حصلت تحت سطح البحر (تسونامي) فضلا عن آثارها التخريبية للمباني والمواصلات والمنشآت وغالبا ينتج عن حركات الحمل الحراري في الأستينوسفير والتي تحرك الصفائح القارية متسببة في حدوث هزات هي الزلازل.	3

Table 6.2. Example of an Arabic SD sentences. (Continued)

كما أن الزلازل قد تحدث خرابا كبيرا و تحدد درجة الزلزال بمؤشر وتقيسه من 1 إلى 10: من 1 إلى4 زلازل قد لا تحدث اية اضرار أي يمكن الاحساس به فقط، من4 إلى 6 زلازل متوسطة الاضرار قد تحدث ضررا للمنازل والاقامات، اما الدرجة القصوى اي من 7الى10 فيستطيع الزلزال تدمير المدينة باكملها وحفرها تحت الأرض حتى تختفي مع اضرار لدى المدن المجاورة لها.	4
زلزال لشبونة 1755 قتل فيه ما بين ال60 إلى 100 ألف نسمة و كان من أشد الزلازل تدميرا على مر التاريخ.	5
من أشهر الزلازل : .زلزال سان فرانسيسكو 1906 قتل فيه ما يقارب ال3 الألاف شخص و بلغت خسائره حوالي 400 مليون دولار و كان من أشد الزلازل التي ضربت كاليفورنيا .زلزال غوجرات غرب الهند 26 يناير 2001 .زلزال بم في إيران حيث قتل حوالي 40 الف شخص فيه .زلزال المحيط الهندي 26 ديسمبر 2004 الذي أعقبه أشهر موجة تسونامي حيث ضربت سواحل العديد من الدول منها اندونيسيا، سريلانكا ، تايلاند ، الهند ، الصومال وغيرها حيث وصفت هذا الزلزال بأنه أحد أسوأ الكوارث الطبيعية التي ضربت الأرض على الإطلاق قتل فيه ما يقارب ال250000 .زلزال كشمير 2005 قتل فيه حوالي 79 ألف شخص	6

6.4. Results, Discussion, and Analysis

This subsection covers GEATS system's findings and compares them to the results of alternative approaches.

Figure 6.2. presents the precision, recall, and F-score of the two feature extraction approaches, TF-IDF and WE, demonstrating that the WE method is superior to the TF-IDF method. Although TF-IDF has a close value to WE for some criteria, WE has a higher average percent for all criteria considering ROUGE-1 and ROUGE-2.

In other words, Figure 6.2. compares the method outcomes when the TF-IDF is used versus when the WE is used to represent the words in documents. According to the results in the table, utilizing the WE improves algorithm performance according to the ROUGE-1 and ROUGE-2 metrics. As a result, in this work, we use WE as a fundamental building block to create text representations. In contrast to the term TF-IDF vectors, that represent a single word by a single-hot vector, such an embedding is a distributed vector representation of one word in a fixed-dimensional semantic space [57], [58].

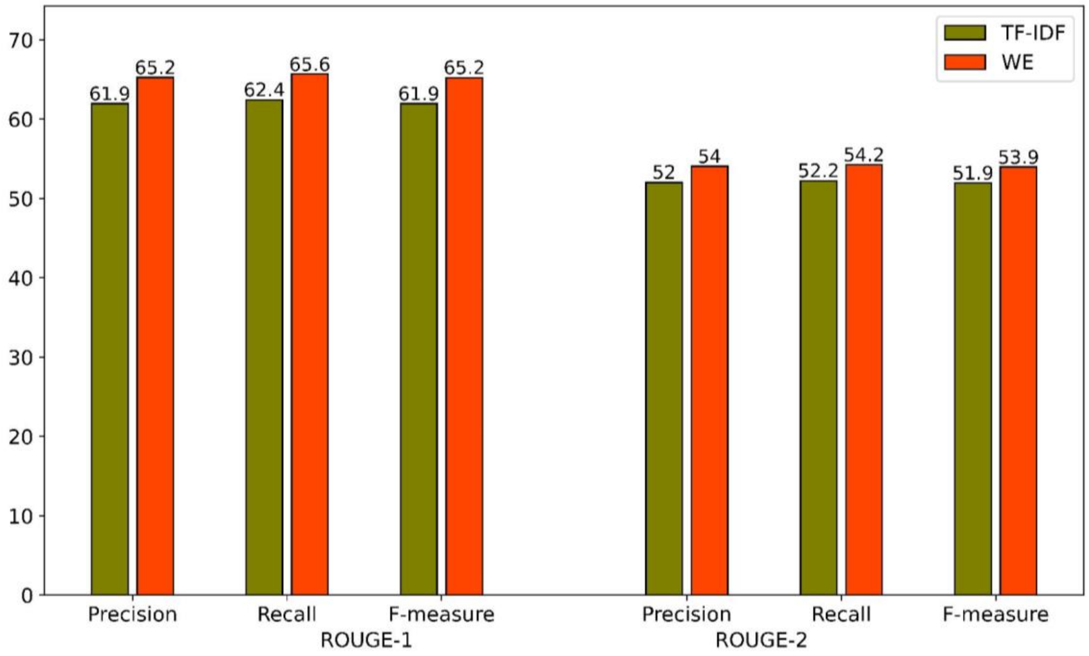

Figure 6.2. TF-IDF and WE using Word2Vec representation techniques' ROUGE-1 and ROUGE-2 values

In the stemming process one stemmer which is Farasa stemmer[24] has been used. In addition to Farasa, Arabic light stemmer [68] (Tashaphyne[25]) has been used to compare the two. Using the proposed method Table 6.3. indicates the results with CR = 40% when using Farasa and Arabic Light Stemmer for the five ground-truth summaries (S1-S5) regarding BLEU, ROUGE-1, and ROUGE-2 where Precision, Recall, and F-measure, are abbreviated as P, R and F respectively. Consequently, it is obvious that when using Farasa stemmer we obtained higher results as the F-measure in ROUGE-2 equals 53.973 which is higher than the Arabic light stemmer's value (53.180). On the other hand, when using ROUGE-1 and BLEU the light stemmer's F-measure (65.474, 51.404) respectively are larger but with a small difference between it and Farasa stemmer which is (0.25, 0.975), which indicates that in GEATS system Farasa stemmer is more suitable and provide us with better accuracy in ROUGE-2 but in terms of ROUGE-1 and BLEU Arabic light stemmer is more suitable.

[24] https://alt.qcri.org/farasa/

[25] https://pypi.org/project/Tashaphyne/

Figure 6.3. compares the current research results to the results of six state-of-the-art methodologies that utilized the same 40% CR and dataset as this work, which is the EASC corpus. The methods used in the comparable studies are as follows: (1) OSSAD [25], (2) graph-based approach with the Short-Path Algorithm (SPA) [11], (3) Hybrid-based approach [18], (4) PSO algorithm [15], (5) GA [16], and (6) Firefly (FF) algorithm [17]. The comparison demonstrates that the PR with WE results outperform the state-of-the-art approaches.

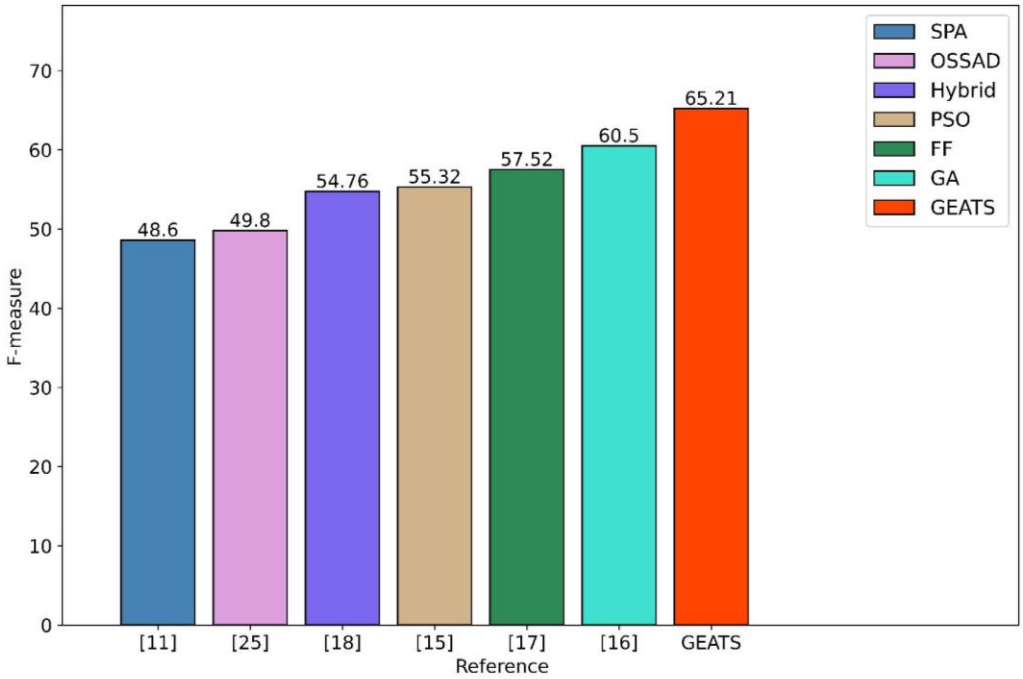

Figure 6.3. The F-measure values of the state-of-the-art methodologies with GEATS method

Table 6.3. The Results with CR = 40% when using Farasa Stemmer and Arabic light stemmer for the Five Ground-Truth Summaries by Using GEATS Method.

Stemmer		Farasa Stemmer							Arabic Light Stemmer						
ROUGE-N		ROUGE-1			ROUGE-2			BLEU	ROUGE-1			ROUGE-2			BLEU
Metric		P	R	F	P	R	F		P	R	F	P	R	F	
Number of Summary	S1	64.286	64.366	64.159	53.733	53.379	53.493	50.264	64.923	65.032	64.896	52.679	52.505	52.567	51.294
	S2	65.881	66.409	65.679	54.375	54.560	54.128	50.569	66.526	66.342	66.144	54.226	53.689	53.695	51.539
	S3	64.829	65.380	64.916	53.527	53.587	53.497	50.418	65.686	66.013	65.795	53.594	53.594	53.594	51.355
	S4	65.643	66.729	65.841	54.446	55.276	54.508	50.408	65.359	65.032	65.141	53.464	53.178	53.267	51.343
	S5	65.642	65.580	65.500	54.252	54.441	54.241	50.487	65.234	66.013	65.392	52.679	52.941	52.777	51.491
Mean		65.256	65.693	65.219	54.067	54.249	53.973	*50.429*	65.546	65.686	65.474	53.328	53.181	53.180	*51.404*

The CR affects the system's F-measure score respectively. As it is clear in Table 6.4. when the CR is 10% the number of sentences and the F-measure will be small because the summary will contain less information which means the summary will not be meaningful and coherent. On the other hand, when the CR is very big the F-measure will not be in its best cases because the summary's sentences will be a lot which means there will be redundant sentences. The change of the values is obvious in Figure 6.4. where F-measure values are represented as follows in terms of ROUGE-1 ,which starts by th white color and ends by the black color, and ROUGE-2 ,which starts by th floralwhite color and ends by the goldenrod color.

Table 6.4. The Obtained F-measure Scores with Different CRs by using ROUGE 1, 2 with the GEATS Method.

CR	10%	15%	20%	25%	30%	35%	40%	45%	50%
ROUGE -1	35.947	46.535	53.725	59.211	62.764	*65.992*	65.469	63.147	60.331
ROUGE -2	23.137	27.712	34.901	42.177	47.261	50.997	*53.560*	52.471	49.644

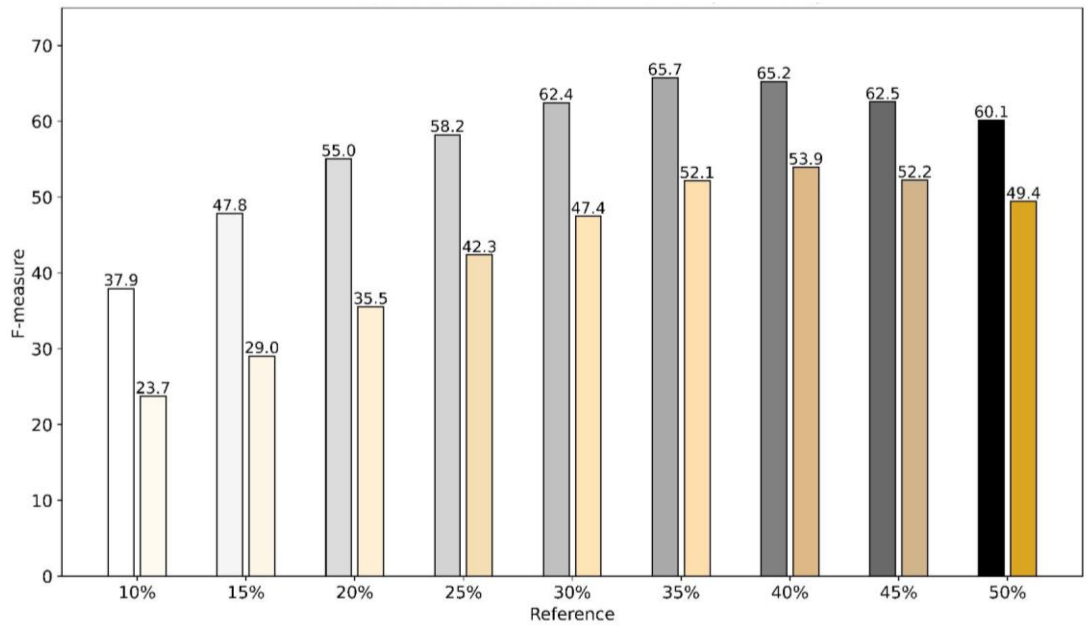

Figure 6.4

Figure 6.4. The ROUGE-1 and ROUGE-2 F-measure values by using (10-50)% CRs with the GEATS system

In Figure 6.4., the F-measure scores dramatically start to increase in both ROUGE-1 and ROUGE-2 by the increase of CR, then when the CR goes bigger it starts to decrease. Consequently, in the GEATS system the best obtained F-measure in terms of ROUGE-1 is 65.744 with a 35% CR. Moreover, in terms of ROUGE-2 the best F-measure is 53.973 with 40% CR. In addition, when the LexRank [42] and TextRank [43] algorithms were utilized under identical conditions, the suggested technique produced finer results than LexRank concerning ROUGE-1 and ROUGE-2. GEATS F-measure score in ROUGE-2 is also higher than TextRank's which gets the highest F-measure in ROUGE-1, as shown in Table 6.5.

Table 6.5. The obtainbed Precision, Recall, F-measure score with 40% CR by using ROUGE 1, 2 using LexRank, TextRank and GEATS (PageRank-based).

Method	ROUGE-1			ROUGE-2		
	P	R	F	P	R	F
LexRank	63.788	64.353	63.723	52.240	52.344	52.170
TextRank	67.916	68.804	*68.011*	53.218	54.713	53.235
GEATS	65.256	65.693	65.219	54.067	54.249	*53.973*

CHAPTER 7. CONCLUSION AND FUTURE WORK

7.1. Conclusion

Since the amount of textual material on the internet is increasing all the time, there is a rising necessity for efficient and sophisticated methods for ATS. The original text will be compressed into a shorter version while keeping the information content and general meaning by applying ATS. There are two types of TS: extractive TS and abstractive TS. Although ArTS has grown in popularity in recent years, the quality of current ATS systems requires to be enhanced. Graph-based approaches for Arabic NLP have obviously gained favor in recent years. The process of TS is composed of three significant stages: pre-processing phase, features extraction and graph building phase, and eventually applying the PageRank algorithm and summary extraction and evaluation. This work proposes a SD extractive graph-based ArTS. The PageRank algorithm is employed, combined with WE using Word2Vec.

The similarity of any two sentences is measured by ranking them according to cosine similarity. The final score for each phrase is established using PageRank ranking, and the high-ranked sentences are included in the summary. To evaluate the effectiveness of this methodology, the EASC dataset is employed as a standard corpus. In addition, ROUGE-1, ROUGE-2, and BLUE measures are used in the review process. With a CR of 40% of the document's sentences using ROUGE-1 the obtained F-measure is 65.47%. The results showed that GEATS strategy is superior to state-of-the-art methdologies. Furthermore, employing the Farasa stemmer in the stemming process and word embeddings in the feature extraction phase generated better results than using the Arabic light stemmer and TF-IDF, respectively.

7.2. Future Work

Some recommendations for future work on GEATS system are listed here: more characteristics, such as sentence topic relevance, or other morphological qualities, might be used. Linguistic morphological tools may be used to address complicated morphological relationships in Arabic. To improve the system's output, specific stopwords can be used for each category of the EASC dataset. To improve the outcomes, apply modification with PR. Stemming techniques at the preprocessing step or employing the lemmatization approach Attempting to create a hybrid system for text processing can be improved by employing both graph-based and rhetorical methods. The system's preprocessing and normalization can be enhanced, too. Use additional tokenization approaches to enhance the tokenization process. Various assessment methodologies can be employed to have a deeper grasp of the nature of the ArTS challenge.

REFERENCES

[1] Lloret, E. and Palomar, M., Text summarisation in progress: a literature review. Artificial Intelligence Review, 37(1), 1-41, 2012.

[2] Radev, D., Hovy, E. and McKeown, K., Introduction to the special issue on summarization. Computational linguistics, 28(4), 399-408, 2002.

[3] El-Kassas, W.S., Salama, C.R., Rafea, A.A. and Mohamed, H.K., Automatic text summarization: A comprehensive survey. Expert Systems with Applications, 165, 113679, 2021.

[4] Dhankhar, S. and Gupta, M.K., Automatic Extractive Summarization for English Text: A Brief Survey. In Proceedings of Second Doctoral Symposium on Computational Intelligence. Springer, Singapore, 183-198, 2022.

[5] Papineni, K., Roukos, S., Ward, T. and Zhu, W.J., Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 311-318, 2002.

[6] Contreras, I.R.L., Carreón, A.M., Rodas-Osollo, J. and Varela, M.C., Extractive Text Summarization Methods in the Spanish Language. In Handbook of Research on Natural Language Processing and Smart Service Systems. IGI Global, 379-391, 2021.

[7] Chernyshev, D. and Dobrov, B., Abstractive Summarization of Russian News Learning on Quality Media. In International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham, 96-104, 2020.

[8] Deng, Z., Ma, F., Lan, R., Huang, W. and Luo, X., A two-stage Chinese text summarization algorithm using keyword information and adversarial learning. Neurocomputing, 425, 117-126, 2021.

[9] Versteegh, K., Arabic language. Edinburgh University Press, 2014.

[10] Qaroush, A., Farha, I.A., Ghanem, W., Washaha, M. and Maali, E., An efficient single document Arabic text summarization using a combination of statistical and

semantic features. Journal of King Saud University-Computer and Information Sciences, 33(6), 677-692, 2021.

[11] Al-Taani, A.T. and Al-Omour, M.M., An extractive graph-based Arabic text summarization approach. In The International Arab Conference on Information Technology, 2014.

[12] Touati, I., Ellouze, M., Graja, M. and Hadrich Belguith, L., Appraisal of two Arabic opinion summarization methods: statistical versus machine learning. The Computer Journal, 2020.

[13] El-Fishawy, N., Hamouda, A., Attiya, G.M. and Atef, M., Arabic summarization in twitter social network. Ain Shams Engineering Journal, 5(2), 411-420, 2014.

[14] Jaradat, Y.A., Arabic Single-Document Text Summarization Based on Harmony Search (Doctoral dissertation, Yarmouk University), 2015.

[15] Al-Abdallah, R.Z. and Al-Taani, A.T., Arabic single-document text summarization using particle swarm optimization algorithm. Procedia Computer Science, 117, 30-37, 2017.

[16] Al-Radaideh, Q.A. and Bataineh, D.Q., A hybrid approach for Arabic text summarization using domain knowledge and genetic algorithms. Cognitive Computation, 10(4), 651-669, 2018.

[17] Al-Abdallah, R.Z. and Al-Taani, A.T., Arabic text summarization using firefly algorithm. In 2019 amity international conference on artificial intelligence (AICAI). IEEE, 61-65, 2019.

[18] Jaradat, Y.A. and Al-Taani, A.T., Hybrid-based Arabic single-document text summarization approach using genatic algorithm. In 2016 7th International Conference on Information and Communication Systems (ICICS). IEEE, 85-91, 2016.

[19] Etaiwi, W. and Awajan, A., Graph-based Arabic text semantic representation. Information Processing & Management, 57(3), 102183, 2020.

[20] Moulay Lakhdar, S. and Chéragui, M.A., Building an extractive Arabic text summarization using a hybrid approach. In International Conference on Arabic Language Processing. Springer, Cham, 135-148, 2019.

[21] Al-Maleh, M. and Desouki, S., Arabic text summarization using deep learning approach. Journal of Big Data, 7(1), 1-17, 2020.

[22] Abu Nada, A.M., Alajrami, E., Al-Saqqa, A.A. and Abu-Naser, S.S., Arabic text summarization using arabert model using extractive text summarization approach, 2020.

[23] Elbarougy, R., Behery, G. and KHATIB, A.E., Graph-Based Extractive Arabic Text Summarization Using Multiple Morphological Analyzers. Journal of Information Science & Engineering, 36(2), 2020.

[24] Etaiwi, W. and Awajan, A., Graph-based Arabic NLP techniques: a survey. Procedia computer science, 142, 328-333, 2018.

[25] Imam, I., Nounou, N., Hamouda, A. and Khalek, H.A.A., An ontology-based summarization system for arabic documents (ossad). International Journal of Computer Applications, 74(17), 38-43, 2013.

[26] Bialy, A.A., Gaheen, M.A., ElEraky, R.M., ElGamal, A.F. and Ewees, A.A., Single arabic document summarization using natural language processing technique. In Recent Advances in NLP: The Case of Arabic Language. Springer, Cham, 17-37, 2020.

[27] Elayeb, B., Chouigui, A., Bounhas, M. and Khiroun, O.B., Automatic arabic text summarization using analogical proportions. Cognitive Computation, 12(5), 1043-1069, 2020.

[28] Abdulateef, S., Khan, N.A., Chen, B. and Shang, X., Multidocument Arabic text summarization based on clustering and Word2Vec to reduce redundancy. Information, 11(2), 59, 2020.

[29] Al-Taani, A.T., Recent Advances in Arabic Automatic Text Summarization. International Journal of Advances in Soft Computing & Its Applications, 13(3), 2021.

[30] Baraka, R.S. and Al Breem, S.N., Automatic arabic text summarization for large scale multiple documents using genetic algorithm and mapreduce. In 2017 Palestinian International Conference on Information and Communication Technology (PICICT). IEEE, 40-45, 2017

[31] Alqaisi, R., Ghanem, W. and Qaroush, A., Extractive multi-document Arabic text summarization using evolutionary multi-objective optimization with K-medoid clustering. IEEE Access, 8, 228206-228224, 2020.

[32] Ibrahim, A., Elghazaly, T. and Gheith, M., A novel Arabic text summarization model based on rhetorical structure theory and vector space model. International

Journal of Computational Linguistics and Natural Language Processing, 2(8), 480-485, 2013.

[33] Cheragui, M.A. and Lakhdar, S.M., 2019, SumSAT: Hybrid Arabic Text Summarization Based on Symbolic and Numerical Approaches. In Proceedings of the 3rd International Conference on Natural Language and Speech Processing, 120-127, 2019.

[34] Fadel, A. and Esmer, G.B., A Hybrid Long Arabic Text Summarization System Based on Integrated Approach Between Abstractive and Extractive. In Proceedings of the 2020 6th International Conference on Computer and Technology Applications, 109-114, 2020.

[35] Ellouze, S., Jaoua, M. and Belguith, L.H., Arabic Text Summary Evaluation Method, 2020.

[36] Lamsiyah, S., Mahdaouy, A.E., Alaoui, S.O.E. and Espinasse, B., A supervised method for extractive single document summarization based on sentence embeddings and neural networks. In International Conference on Advanced Intelligent Systems for Sustainable Development. Springer, Cham, 75-88, 2019.

[37] Molham, A.M. and Said, D., Arabic text summarization using deep learning approach. Journal of Big Data, 7(1), 2020.

[38] Suleiman, D. and Awajan, A., Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges. Mathematical problems in engineering, 2020.

[39] Elbarougy, R., Behery, G., and El Khatibm A., The Impact of Stop Words Processing for Improving Extractive Graph-Based Arabic Text Summarization, International Journal of Scientific & Technology Research, 8(11), 2134-2139, 2019.

[40] Elbarougy, R., Behery, G. and El Khatib, A., Extractive Arabic text summarization using modified PageRank algorithm. Egyptian informatics journal, 21(2), 73-81, 2020.

[41] Wills, R.S., Google's PageRank: The math behind the search engine. Math. Intelligencer, 28(4), 6-11, 2006.

[42] Erkan, G. and Radev, D.R., Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research, 22, 457-479, 2004.

[43] Mihalcea, R. and Tarau, P., Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing 404-411, 2004.

[44] Page, L., Brin, S., Motwani, R. and Winograd, T., The PageRank citation ranking: Bringing order to the web. Stanford InfoLab, 1999.

[45] Cohen, R., Havlin, S. and ben-Avraham, D., Structural properties of scale-free networks. Handbook of Graphs and Networks: From the Genome to the Internet, 85-110, 2002.

[46] Brin, S. and Page, L., The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems, 30(1-7), 107-117, 1998.

[47] Attia, M.A., Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation. The University of Manchester (United Kingdom), 2008.

[48] El-Khair, I.A., Effects of stop words elimination for Arabic information retrieval: a comparative study. International Journal of Computing & Information Sciences, 4(3), 119-133, 2006.

[49] Abdelali, A., Darwish, K., Durrani, N. and Mubarak, H., Farasa: A fast and furious segmenter for arabic. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations, 11-16, 2016.

[50] Ren, Y., Wang, R. and Ji, D., A topic-enhanced word embedding for Twitter sentiment classification. Information Sciences, 369, 188-198, 2016.

[51] Enríquez, F., Troyano, J.A. and López-Solaz, T., An approach to the use of word embeddings in an opinion classification task. Expert Systems with Applications, 66, 1-6, 2016.

[52] Xia, P., Zhang, L. and Li, F., Learning similarity with cosine similarity ensemble. Information Sciences, 307, 39-52, 2015.

[53] El-Haj, M., Kruschwitz, U. and Fox, C., Using mechanical turk to create a corpus of Arabic summaries, 2010.

[54] Lin, C.Y., Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, 74-81, 2004.

[55] Baxendale, P.B., Machine-made index for technical literature—an experiment. IBM Journal of research and development, 2(4), 354-361, 1958.

[56] Powers, D.M., Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061, 2020.

[57] Achananuparp, P., Hu, X. and Shen, X., The evaluation of sentence similarity measures. In International Conference on data warehousing and knowledge discovery. Springer, Berlin, Heidelberg, 305-316, 2008.

[58] Manning, C., Raghavan, P. and Schütze, H., Introduction to information retrieval. Natural Language Engineering, 16(1), 100-103, 2010.

[59] Fendji, J.L.E.K., Taira, D.M., Atemkeng, M. and Ali, A.M., WATS-SMS: A T5-Based French Wikipedia Abstractive Text Summarizer for SMS. Future Internet, 13(9), 238, 2021.

[60] Yadav, A.K., Maurya, A.K. and Yadav, R.S., Extractive Text Summarization Using Recent Approaches: A Survey. Ingénierie des Systèmes d'Information, 26(1), 2021.

[61] Haboush, A., Al-Zoubi, M., Momani, A. and Tarazi, M., Arabic text summarization model using clustering techniques. World of Computer Science and Information Technology Journal (WCSIT) ISSN, 2221-0741, 2012.

[62] Waheeb, S.A., Multi-document text summarization using text clustering for Arabic Language (Doctoral dissertation, Universiti Utara Malaysia), 2014.

[63] AL-Khawaldeh, F.T. and Samawi, V.W., Lexical cohesion and entailment based segmentation for arabic text summarization (lceas). World of Computer Science & Information Technology Journal, 5(3), 2015.

[64] Alami, N., Meknassi, M., En-nahnahi, N., El Adlouni, Y. and Ammor, O., Unsupervised neural networks for automatic Arabic text summarization using document clustering and topic modeling. Expert Systems with Applications, 172, 114652, 2021.

[65] Tanfouri, I., Tlik, G. and Jarray, F., An automatic arabic text summarization system based on genetic algorithms. Procedia Computer Science, 189, 195-202, 2021.

[66] Ayed, A.B., Biskri, I. and Meunier, J.G., Arabic text summarization via Knapsack balancing of effective retention. Procedia Computer Science, 189, 312-319, 2021.

[67] Alshanqiti, A., Namoun, A., Alsughayyir, A., Mashraqi, A.M., Gilal, A.R. and Albouq, S.S., Leveraging DistilBERT for Summarizing Arabic Text: An Extractive Dual-Stage Approach. IEEE Access, 9, 135594-135607, 2021.

[68] Zerrouki, T., Tashaphyne, Arabic light stemmer, 2010.

[69] Rehurek, R. and Sojka, P., Software framework for topic modelling with large corpora. In In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks, 2010.

[70] Sakai, T. and Sparck-Jones, K., Generic summaries for indexing in information retrieval. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 190-198, 2001.

[71] Delort, J.Y., Bouchon-Meunier, B. and Rifqi, M., Enhanced web document summarization using hyperlinks. In Proceedings of the fourteenth ACM conference on Hypertext and hypermedia, 208-215, 2003.

[72] Saravanan, M., Ravindran, B. and Raman, S., Improving legal document summarization using graphical models. Frontiers in Artificial Intelligence and Applications, 152, 51, 2006.

[73] Shaalan 1, K.F., An intelligent computer assisted language learning system for Arabic learners. Computer Assisted Language Learning, 18(1-2), 81-109, 2005.

[74] Shaalan, K.F., Arabic GramCheck: A grammar checker for Arabic. Software: Practice and Experience, 35(7), 643-665, 2005.

[75] Farghaly, A., Subject pronoun deletion rule in Egyptian Arabic. In *Discourse Analysis: Theory and Application Proceedings of the Second National Symposium on Linguistics and English Language Teaching*, 60-69, 1982.

[76] Frankel, D.S., *Model driven architecture applying MDA*. John Wiley & Sons, 2003.

[77] Dernoncourt, F., Ghassemi, M. and Chang, W., A repository of corpora for summarization. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[78] Woodsend, K. and Lapata, M., Automatic generation of story highlights. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 565-574, 2010.

[79] Gambhir, M. and Gupta, V., Recent automatic text summarization techniques: a survey. Artificial Intelligence Review, 47(1), 1-66, 2017.

[80] Moratanch, N. and Chitrakala, S., A survey on extractive text summarization. In 2017 international conference on computer, communication and signal processing (ICCCSP). IEEE, 1-6, 2017.

[81] Meena, Y.K. and Gopalani, D., Evolutionary algorithms for extractive automatic text summarization. Procedia Computer Science, 48, 244-249, 2015.

[82] Nasar, Z., Jaffry, S.W. and Malik, M.K., Textual keyword extraction and summarization: State-of-the-art. Information Processing & Management, 56(6), 102088, 2019.

[83] Nallapati, R., Zhai, F. and Zhou, B., Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In Thirty-first AAAI conference on artificial intelligence, 2017.

[84] Atlam, E.S. and El-Barbary, O., Arabic document summarization using FA fuzzy ontology. International Journal of Innovative Computing, Information and Control, 10(4), 1351-1367, 2014.

[85] Kumar, A.K.S.H.I. and Sharma, A.D.I.T.I., Systematic literature review of fuzzy logic based text summarization. Iranian journal of fuzzy systems, 16(5), pp.45-59, 2019.

[86] Al Qassem, L., Wang, D., Barada, H., Al-Rubaie, A. and Almoosa, N., Automatic Arabic text summarization based on fuzzy logic. In Proceedings of the 3rd international conference on natural language and speech processing, 42-48, 2019.

[87] Ibrahim, A. and Elghazaly, T., Improve the automatic summarization of Arabic text depending on Rhetorical Structure Theory. In 2013 12th Mexican International Conference on Artificial Intelligence. IEEE, 223-227, 2013.

[88] Abbas, M., Smaïli, K. and Berkani, D., Evaluation of topic identification methods on Arabic corpora. J. Digit. Inf. Manag., 9(5),185-192, 2011.

[89] Abbas, M. and Smaili, K., Comparison of topic identification methods for arabic language. In Proceedings of International Conference on Recent Advances in Natural Language Processing, RANLP, 14-17, 2005.

[90] Alhafni, B., Habash, N. and Bouamor, H., The Arabic Parallel Gender Corpus 2.0: Extensions and Analyses. arXiv preprint arXiv:2110.09216, 2021.

[91] Habash, N., Bouamor, H. and Chung, C., Automatic gender identification and reinflection in Arabic. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, 155-165, 2019.

[92] El-Khair, I.A., 1.5 billion words arabic corpus. arXiv preprint arXiv:1611.04033, 2016.

[93] Ko, Y. and Seo, J., An effective sentence-extraction technique using contextual information and statistical approaches for text summarization. Pattern Recognition Letters, 29(9), pp.1366-1371, 2008.

[94] Chouigui, A., Khiroun, O.B. and Elayeb, B., ANT corpus: an Arabic news text collection for textual classification. In 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA). IEEE, 135-142, 2017.

[95] El-Haj, M. and Koulali, R., KALIMAT a multipurpose Arabic Corpus. In Second workshop on Arabic corpus linguistics (WACL-2), 22-25, 2013.

[96] Akallouch, M., Boujemaa, K.S., Bouhoute, A., Fardousse, K. and Berrada, ASAYAR: A dataset for Arabic-Latin scene text localization in highway traffic panels. IEEE Transactions on Intelligent Transportation Systems, 2020.

[97] Tawalbeh, S. and Al-Smadi, M., Is this sentence valid? an arabic dataset for commonsense validation. arXiv preprint arXiv:2008.10873. 2020.

[98] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V., Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116. 2019.

[99] El-Haj, M., Habibi-a multi dialect multi national Arabic song lyrics corpus, 2020.

[100] Suwaileh, R., Kutlu, M., Fathima, N., Elsayed, T. and Lease, M., ArabicWeb16: A new crawl for today's Arabic Web. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, 673-676, 2016.

[101] Nabil, M., Aly, M. and Atiya, A., Astd: Arabic sentiment tweets dataset. In Proceedings of the 2015 conference on empirical methods in natural language processing, 2515-2519, 2015.

[102] Khan, A., Salim, N., Farman, H., Khan, M., Jan, B., Ahmad, A., Ahmed, I. and Paul, A., Abstractive text summarization based on improved semantic graph approach. International Journal of Parallel Programming, 46(5), 992-1016, 2018.

[103] Wang, S., Zhao, X., Li, B., Ge, B. and Tang, D., Integrating extractive and abstractive models for long text summarization. In 2017 IEEE International Congress on Big Data (BigData Congress). IEEE, 305-312, 2017.

[104] Fang, C., Mu, D., Deng, Z. and Wu, Z., Word-sentence co-ranking for automatic extractive text summarization. Expert Systems with Applications, 72, 189-195, 2017.

[105] Patel, D., Shah, S. and Chhinkaniwala, H., Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique. *Expert Systems with Applications*, *134*, 167-177, 2019.

ANNEX

Table 0.1. Arabic stopword list.

آض	ثلاثين	ّلكن	ّحي	هاء	ثمان	وإن	لستم	تلكما	إذ
أمسى	اربعين	َء	ّحي	واو	سبت	ولا	لستما	ته	إذا
انقلب	خمسين	أجل	دونك	ياء	أحد	ولكن	لستن	تي	إذما
بات	ستين	ًإذا	رويدك	همزة	اثنين	ولو	لسن	تين	إذن
تبدّل	سبعين	أمّا	سرعان	ي	ثلاثاء	وما	لسنا	تينك	أف
تحوّل	ثمانين	إمّا	َشتان	نا	أربعاء	ومن	لعل	ثم	أقل
حار	تسعين	ّإن	َشَتَّان	ك	خميس	وهو	لك	ثمة	أكثر
رجع	بضع	ّأن	ْصه	كن	جمعة	يا	لكم	حاشا	ألا
راح	نيف	أى	ٍصه	ه	أول	ٌأب	لكما	حبذا	إلا
صار	أجمع	إى	طاق	إياه	ثان	ٌأخ	لكن	حتى	التي
ّظل	جميع	أيا	طَق	إياها	ثاني	ٌحم	لكنما	حيث	الذي
عاد	عامة	ب	ْعَدَس	إياهما	ثالث	فو	لكي	حيثما	الذين
غدا	عين	ّثم	كِخ	إياهم	رابع	ِأنت	لكيلا	حين	اللاتي
كان	نفس	جلل	مكانَك	إياهن	خامس	يناير	لم	خلا	اللائي
ما انفك	لا سيما	جير	مكانَك	إياك	سادس	فبراير	لما	دون	اللتان
ما برح	أصلا	ّرُب	مكانَك	إياكما	سابع	مارس	لن	ذا	اللتيا

Table0.1. Arabic stopword list. (Continued)

مادام	أهلا	س	مكانكم	إياكم	ثامن	أبريل	لنا	ذات	اللتين
مازال	أيضا	ّعل	مكانكما	إياك	تاسع	مايو	له	ذاك	اللذان
مافتئ	بؤسا	ف	ّمكانكن	إياكن	عاشر	يونيو	لها	ذان	اللذين
ابتدأ	بعدا	ّكأن	ْنَخ	إياي	حادي	يوليو	لهم	ذانك	اللواتي
أخذ	بغتة	كلَّا	َهاك	إيانا	أ	أغسطس	لهما	ذلك	إلى
اخلولق	تعسا	كى	ْهَج	أولالك	ب	سبتمبر	لهن	ذلكم	إليك
أقبل	حقا	ل	هلم	ِتان	ت	أكتوبر	لو	ذلكما	إليكم
انبرى	حمدا	لات	هيّا	تانِك	ث	نوفمبر	لولا	ذلكن	إليكما
أنشأ	خلافا	ّلعل	هَيْهات	تِه	ج	ديسمبر	لوما	ذه	إليكن
أوشك	خاصة	ّلكن	وا	تِي	ح	جانفي	لي	ذو	أم
جعل	دواليك	ّلكن	ًواها	ِتَيْن	خ	فيفري	لئن	ذوا	أما
حرى	سحقا	م	وراءَك	ّثم	د	مارس	ليت	ذواتا	أما
شرع	سرا	ّن	َوُشْكَان	ثمّة	ذ	أفريل	ليس	ذواتي	إما
طفق	سمعا	هلّا	ْوَي	ِذان	ر	ماي	ليسا	ذي	أن
علق	صبرا	وا	يفعلان	ذِه	ز	جوان	ليست	ذين	إن
قام	صدقا	أل	تفعلان	ذِي	س	جويلية	ليستا	ذينك	إنا
كرب	صراحة	إلّا	يفعلون	ِذَيْن	ش	أوت	ليسوا	ريث	أنا
كاد	طرا	ت	تفعلون	هَؤلاء	ص	كانون	ما	سوف	أنت
ّهب	عجبا	ك	تفعلين	ِهَاتان	ض	شباط	ماذا	سوى	أنتم
تلكم	عيانا	لمّا	بين	هَاتِه	ط	آذار	متى	شتان	أنتما

Table0.1. Arabic stopword list. (Continued)

لست	غالبا	ن	اتخذ	هَاتِي	ظ	نيسان	مذ	عدا	أنتن
وإذا	فرادى	ه	ألفى	ِهَاتَيْن	ع	أيار	مع	عسى	إنما
عشر	فضلا	و	تخذ	هَذا	غ	حزيران	مما	عل	إنه
نون	قاطبة	ا	ترك	ِهَذان	ف	تموز	ممن	على	أنى
ِحَذار	كثيرا	ي	تعلَّم	هَذِه	ق	آب	من	عليك	أنى
لات	لبيك	تجاه	جعل	هَذِي	ك	أيلول	منه	عليه	آه
عشرين	معاذ	تلقاء	حجا	ِهَذَيْن	ل	تشرين	منها	عما	آها
أضحى	أبدا	جميع	حبيب	الألى	م	دولار	منذ	عن	أو
تلك	إزاء	حسب	خال	الألاء	ن	دينار	مه	عند	أولاء
لدى	أصلا	سبحان	حسب	أل	ه	ريال	مهما	غير	أولئك
وإذ	الآن	شبه	خال	أنّى	و	درهم	نحن	فإذا	أوه
تسع	أمد	لعمر	درى	ّأي	ي	ليرة	نحو	فإن	آي
ميم	أمس	مثل	رأى	أيّانّ	ء	جنيه	نعم	فلا	أي
حاي	آنفا	معاذ	زعم	أنّى	ى	قرش	ها	فمن	أيها
قلما	آناء	أبو	صبر	ّأي	آ	مليم	هاتان	في	إي
تسعون	أنّى	أخو	ّظن	أيّانّ	ؤ	فلس	هاته	فيم	أين
أصبح	أول	حمو	ّعد	ذيت	ئ	هللة	هاتي	فيما	أين
بيد	أيّان	فو	علم	ّكأي	أ	سنتيم	هاتين	فيه	أينما
لاسيما	تارة	مئة	غادر	فلان	باء	يوان	هذا	قد	إيه
والذين	ّثم	مئتان	ذهب	وا	تاء	شيكل	هذان	كأن	بخ

Table0.1. Arabic stopword list. (Continued)

استحال	ثمّة	ثلاثمئة	وجد	َآمين	ثاء	واحد	هذه	كأنما	بس
ثمانون	حقا	أربعمئة	ورد	ِآه	جيم	اثنان	هذي	كأي	بعد
طالما	صباح	خمسمئة	وهب	ٍآه	حاء	ثلاثة	هذين	كأين	بعض
َبَلْه	مساء	ستمئة	أسكن	ًآها	خاء	أربعة	هكذا	كذا	بك
لام	ضحوة	سبعمئة	أطعم	ٍّأُف	دال	خمسة	هل	كذلك	بكم
ثماني	عوض	ثمنمئة	أعطى	ٍّأُف	ذال	ستة	هلا	كل	بكما
ّارتد	غدا	تسعمئة	رزق	ٍّأف	راء	سبعة	هم	كلا	بكن
سبعون	غداة	مائة	زود	أمامك	زاي	ثمانية	هما	كلاهما	بل
ساء	ّقط	ثلاثمائة	سقى	َأمامك	سين	تسعة	هن	كلتا	بلى
بطآن	كلّما	أربعمائة	كسا	ْأوّه	شين	عشرة	هنا	كلما	بما
كاف	لدن	خمسمائة	أخبر	َإلَيْك	صاد	أحد	هناك	كليكما	بماذا
	لمّا	ستمائة	أرى	َإليك	ضاد	اثنا	هنالك	كليهما	بمن
	مرّة	سبعمائة	أعلم	ّإليكن	طاء	اثني	هو	كم	بنا
	قبل	ثمانمئة	أنبأ	ٍإيه	ظاء	إحدى	هؤلاء	كما	به
	خلف	تسعمائة	حدَث	ٍبخ	عين	ثلاث	هي	كي	بها
	أمام	عشرون	خبَّر	ّبس	غين	أربع	هيا	كيت	بهم
	فوق	ثلاثون	نبَّا	ْبَس	فاء	خمس	هيت	كيف	بهما
	تحت	اربعون	أفعل به	سبع	قاف	ست	هيهات	كيفما	بهن
	يمين	خمسون	ما أفعله	والذي	كأيّن	ة	يورو	هاك	بي
	شمال	ستون	بئس	لا	بضع	ألف	ين	هاهنا	فيها

RESUME

Name Surname **:** Ghadir Abdulhakim Abdo Abdullah ALSELWI

EDUCATION

Degree	School	Graduation Year
Master	Sakarya University / Graduate School of Natural and Applied Sciences / Information System Engineering	Continue
Degree	Taiz University / Information and Engineering Faculty / Information Technology Engineering	2017
High School	Zaid Almushki School	2011

JOB EXPERIENCE

Year	Place	Position
2018-2019	Sheba Youth foundation for development and peace – Yemen	Resposible for the community media and awareness unit
2018-2019	ASCEND institute, American French institute, and ALTEC institute – Yemen	Instructor

FOREIGN LANGUAGE

English, Turkish

PRODUCTS (article, paper, project, ets.)

1. (Alselwi, and Taşci, 2020)[26].

2. (Alselwi, and Kaynak, 2021)[27].

3. (Ghaleb, Bin-Thalab, and Alselwi, 2021)[28].

4. (Alselwi, and Taşci, 2021)[29].

[26] https://dergipark.org.tr/en/pub/bufbd/issue/63046/878481

[27] https://ieeexplore.ieee.org/abstract/document/9664773

[28] https://peerj.com/articles/cs-776/

[29] https://ieeexplore.ieee.org/abstract/document/9664773/?casa_token=y5OUmMMLINkAAAAA:uAWfxKIWF5Sp2A_NGGMtTffqlaqGO6wMa_G7RZ4gTghckd2HLHHZ7FdffD6GTxs80KxWoIzWZaL2