



Article

Improved Graph-Based Arabic Hotel Review Summarization Using Polarity Classification

Ghada Amoudi *, Amal Almansour and Hanan Saleh Alghamdi *

Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

* Correspondence: gaamoudi@kau.edu.sa (G.A.); hsaalghamdi@kau.edu.sa (H.S.A.)

Abstract: The increasing number of online product and service reviews has created a substantial information resource for individuals and businesses. Automatic review summarization helps overcome information overload. Research in automatic text summarization shows remarkable advancement. However, research on Arabic text summarization has not been sufficiently conducted. This study proposes an extractive Arabic review summarization approach that incorporates the reviews' polarity and sentiment aspects and employs a graph-based ranking algorithm, TextRank. We demonstrate the advantages of the proposed methods through a set of experiments using hotel reviews from Booking.com. Reviews were grouped based on their polarity, and then TextRank was applied to produce the summary. Results were evaluated using two primary measures, *BLEU* and *ROUGE*. Further, two Arabic native speakers' summaries were used for evaluation purposes. The results showed that this approach improved the summarization scores in most experiments, reaching an F1 score of 0.6294. Contributions of this work include applying a graph-based approach to a new domain, Arabic hotel reviews, adding sentiment dimension to summarization, analyzing the algorithms of the two primary summarization metrics showing the working of these measures and how they could be used to give accurate results, and finally, providing four human summaries for two hotels which could be utilized for another research.

Keywords: *BLEU*; classification; deep learning; extractive; natural language processing; *ROUGE*; sentiment analysis



Citation: Amoudi, G.; Almansour, A.; Alghamdi, H.S. Improved Graph-Based Arabic Hotel Review Summarization Using Polarity Classification. *Appl. Sci.* **2022**, *12*, 10980. <https://doi.org/10.3390/app122110980>

Academic Editor:
Rafael Valencia-Garcia

Received: 9 September 2022

Accepted: 27 October 2022

Published: 29 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the rapid expansion of Web 2.0 applications, more and more people are using blogs, forums, online reviews, and bulletin boards to comment on their personal experiences. Online review platforms provide opportunities to share customer viewpoints, preferences, and experiences on a broad selection of products and services. The resulting agglomeration of online reviews is a valuable information source for consumers. While this comprehensive source of information can help individuals and businesses make better decisions, readers are faced with the daunting task of reading hundreds or thousands of reviews, which can overwhelm them. The explosive amounts of user-generated content enabled by Web 2.0 tools drive the initiation of Web 3.0, the semantic web [1]. While Web 3.0 technologies, such ontologies differ substantially from AI approaches used in this work—both share the same goal, managing information overload. For instance, reviews usually share similar opinions; while this redundancy is indeed important and confirms a particular feature of the product, eliminating repeated reviews allows the reader to get information about the other aspects of the product faster. Summarizing reviews for a specific product or service would save consumers time and provide precise details.

Automatic text summarization (ATS) gained more attention recently due to the increased use of online shopping and reservation services. ATS aims to produce a smooth and short summary that includes important information from the original text [2]. Several studies addressed ATS for English documents [3,4]; however, fewer studies are available for

the Arabic language due to its rich morphological structure, the range of dialects, and the scarcity of data and tools [5]. Arabic is the fifth most spoken language globally, with more than 400 million speakers worldwide [6]. The summarization of Arabic text still suffers from low performance and fewer research studies [7]. Moreover, most Arabic ATS targeted document summarization and very few focused on customer reviews.

ATS can be classified according to the number of simultaneously processed single- and multi-document documents [5]. Single-document summarization is an automatic procedure focusing on extracting useful information from a single text, such as an article, research, or webpage [8]. In contrast, multi-document summarization is an automated procedure that focuses on extracting critical information from multiple texts sharing the same topic, such as various types of research about the same issue [9]. Reviews can be viewed as multiple documents; however, to summarize reviews using TextRank, they should be combined as a single document, and then one final summary is produced.

Reviews usually include positive, negative, and neutral sentiments [3]; we hypothesize that involving sentiment in the summarization process could enhance the performance of the ATS system and improve the readability of the generated summary. This study proposes an approach for summarizing Arabic hotel reviews based on review polarity. TextRank extractive text summarization was applied in conjunction with sentiment analysis. Arabic natural language processing is an immature research area. Although there is a growing interest in the area, there is still a lack of specialized tools and lexicons. TextRank was chosen for this work for several reasons. First, it is language-independent. Second, the graph-based nature of this algorithm, that is, selecting a whole sentence instead of parts of the sentence, makes it a suitable option for reviews. This research does not focus on the summarization method per se; it focuses on the effect of combining the sentiment with the summarization. A comparison between the different summarization methods can be found in [9–12].

The proposed method comprises two main stages: labeling the reviews with the sentiment using supervised and unsupervised techniques; then summarizing the labeled reviews. Two human summaries were authored to evaluate the proposed ATS system. The BLEU and ROUGE scoring techniques were used to assess the performance of the proposed system. Thus, the research question of this study can be phrased as follows:

- Can the Arabic automatic review summarization improve by considering the reviews' sentiment factor?

The rest of this study is organized as follows: Section 2 reviews the literature, and three discusses some related works. The proposed methodology is presented in Section 4. The experimental results are shown in Section 5. Section 6 provides a discussion, and Section 7 concludes this study with final remarks and future directions.

2. Background

ATS refers to one of the natural language processing (NLP) applications that aim to reduce the amount of original text and retrieve important and essential information from the original text [7]. ATS produces a new version of the text, preserving the original text's context and original meaning [13]. ATS procedure can be classified according to the following criteria:

- The number of documents simultaneously processed—single- and multi-document [9].
- The type of retrieved sentences, extractive abstractive, and hybrid summarization [5]. Extractive summarization selectively includes in the summary some sentences from the original text without modifying the sentence structure, while abstractive summarization maintains the text meaning without adhering to its structure.
- General-based and query-based summary; the general-based summary retrieves the sentence regardless of any consideration for question or relations with the title, whereas the query-based summary is returned due to the summary and questions between the title and sentence [2].

Summarizing reviews can be viewed as a multi-document summarization process. Extractive summarization comprises four main methods, graph-based, statistical-based, fuzzy logic, and semantic-based [5]. In this study, we consider graph-based extractive summarizations as extractive methods, in general, yield better outcomes compared to abstractive methods, which still face many challenges, such as the difficulty of semantic representation, inference, and natural language generation [14]. Graph-based approaches are entirely unsupervised, so no training data are required. A document is reconstructed as a graph, where nodes are sentences and edges are similarities between nodes. The fundamental graph-based methods are described next.

- **PageRank:** PageRank is the backbone of the infamous search engine Google, created by Brin and Page in 1998. Google revolutionized web search by introducing PageRank, which finds web pages not only by considering keywords and indexing but also according to the page importance compared to other web pages. In this algorithm, the web pages are represented by nodes, and edges represent the links between pages. A page with a high rank is a page with a high in-degree value, i.e., many other pages are pointing to it. Additionally, this is done recursively, so if the edge comes from a significant page, it will have extra importance.
- **TextRank:** PageRank is the heart of TextRank. TextRank is a graph-based ranking model developed by Mihalcea and Tarau [15] for extracting important keywords and sentences from the text. TextRank is considered an extractive text summarization method. It generates a fully connected, undirected graph of sentences or words of a single document. In this graph, each sentence within the document is represented by a node. In contrast, the similarity between sentences measured as a function of joint concepts is defined by an edge. A weight value marks the importance of each edge. The sentences are ordered based on their scores, thus, creating a list of ranked sentences. For text summarization, usually, the top-ranked sentences are the most representative.
- **LexRank:** a graph-based extractive text summarization technique [14], similar to TextRank, where a node represents each sentence; however, the edge weights are defined by computing the cosine similarities. LexRank uses graph centrality measures such as degree and eigenvector centrality to find important sentences. During the summarization process, the relevance between the sentences is considered by calculating their importance for neighboring sentences. The positive contribution enhances the importance value of the neighbor sentence, and the negative contribution minimizes the importance value of neighbor sentences. Then, the sentences are ordered based on rank [16].

3. Related Work

Text summarization research has been around since 1958 [17], where the problem was approached considering the frequency of terms in a document. As tools and techniques developed, recent studies applied various methods, from word and phrase counts to deep learning architectures and graph-based methods. In [18], the authors proposed a multi-document summarization using TextRank, and to reduce redundancy in the output summary, they used Maximal Marginal Relevance (MMR). The evaluation metrics *ROUGE-1* and *ROUGE-2* were used, and the results achieved were F scores of 0.5103 and 0.4257, respectively. In [19], a system for summarizing hotel reviews collected from the Tripadvisor website is proposed. The method is based on textual features achieving a *ROUGE-1* score of 0.2101 and a *BLEU* score of 0.7820.

Several studies investigated Arabic single-document summarization [20–25] and multi-document summarization [23]. The approaches taken in prior work to build Arabic summarization systems include deep learning models [26–30], the Particle Swarm Optimization (PSO) method [20], clustering techniques [23,24], a fuzzy logic approach to select important sentences [25], and graph-based approaches [7,11].

Abstractive summarizing models for the Arabic text have been explored in [28–32]. Etaiwi and Awajan [31] proposed a graph embedding-based abstractive text summarization technique for the Arabic language. The model used a deep neural network to generate the

summary. The study evaluated the model against the baseline word embedding technique word2vec. The study employed the standard *ROUGE* evaluation metric and subjective evaluation. The model achieved 21.4% better than word2vec in the F1 measure. In [32], the authors developed an abstractive Arabic text summarization system using deep neural networks, including Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), and Bidirectional Long Short-Term Memory (BiLSTM). The study also employed AraBERT to preprocess text. Additionally, the study compared different word embedding mechanisms. The result showed that BiLSTM achieved the best performance, and regarding word embedding, models using skip-gram outperformed the models that used the continuous bag of words (CBOW).

Graph-based approaches were applied in many Arabic document summarization tasks. Elbarougy et al. [7] proposed a system for Arabic text summarization by using PageRank with an additional feature: each node (sentence) within the graph has an initial value representing the number of nouns in the sentence. They created the EASC dataset, Essex Arabic Summaries Corpus, a collection of 153 articles divided into ten categories: art, music, tourism, technology, and others. The result showed improved performance compared to LexRank and TextRank, with an F score of 0.6799. Chouigui et al. [16] created an Arabic News Texts (ANT) dataset for text summarization consisting of news articles collected from five sources: Al-Arabiya, BBC, CNN, France24, and SkyNews. The study applied some of the techniques used for ATS, including LexRank, TextRank, Luhn keyword-based method [17], and Latent Semantic Analysis (LSA) to the ANT dataset. The results show that the LexRank technique provides the best results, with a *BLEU* score of 0.690 and a *ROUGE* score reaching 0.972. Elayeb et al. [33] continue the work of [16] using ANT and EASC [22] datasets. The study proposed two new techniques for text summarization that depend on analogical proportions and achieved a *ROUGE-1* score of 0.75 and a *BLEU* score of 0.47. A summary of selected related work is presented in Table 1 ordered by publishing date.

Table 1. Summary of selected extractive text summarization studies.

Reference	Dataset	Language	Method	Evaluation	Results	Limitations
[11]	25 Arabic articles gathered from different websites on different subjects	Arabic	Graph-based method, PageRank, and MMR to filter redundant sentences	Human-generated summaries	Precision: 0.79 Recall: 0.72 F-measure: 0.75	Limited dataset, not using the standard <i>BLEU</i> and <i>ROUGE</i> metrics
[2]	Wikipedia pages and articles from some popular Arabic newspapers.	Arabic	Deep learning, variational auto-encoder (VAE) model with graph-based and query-based approaches	Human-generated summaries	<i>ROUGE-1</i> : 0.561–0.660 on 45% summary size	Focused on the news domain with a relatively small dataset
[34]	Hotel reviews from the Tripadvisor website	Arabic	Feature-base and sentiment-based summaries were generated using TF-IDF and cosine similarity	Subjective evaluation	Accuracy: 72.84%	Using subjective evaluation only
[4]	Reviews of two hotels from the Tripadvisor website	English	Clustering, Gaussian Mixture Model (GMM) algorithm	Comparing different clustering methods using statistical analysis	GMM method showed a more significant “usefulness” grade than the other methods	English only, focused on comparing two approaches using statistical analysis only

Table 1. Cont.

Reference	Dataset	Language	Method	Evaluation	Results	Limitations
[7]	EASC dataset (153 Arabic articles) [35]	Arabic	PageRank with the number of nouns in the sentence as starting value for each node	Five human summaries for each document, generated using Mechanical Turk	Precision: 0.6875 Recall: 0.7294 F-measure: 0.6799	Single document, focusing on clean news articles, not replicable with noisy reviews or tweets
[33]	EASC [35] and ANT [16]	Arabic	Analogical reasoning, an AI approach for learning by analogy	Human-generated summaries	<i>ROUGE-1</i> : 0.75 <i>BLEU</i> : 0.47	Single document, focusing on news, depends on keywords
[19]	Hotel reviews from the Tripadvisor website	English	Two methods: 1. Selecting relevant sentences for the summary base on (TF-IDF) score 2. Pairing adjective to the nearest noun and considering the polarity	Human-generated summaries	Method 1: <i>ROUGE-1</i> : 0.2101 <i>BLEU</i> : 0.7820 Method 2: <i>ROUGE-1</i> : 0.0670 <i>BLEU</i> : 0.03672	The polarity-based method showed poor performance
[3]	Restaurant reviews from the Tripadvisor website	English	Topic-based and sentiment-based summary using TextRank	Subjective evaluation using informativeness, clarity, helpfulness, and likes indicators	According to users' evaluation, the summaries provided sufficient and precise information about the target restaurant	Using subjective evaluation only
[31]	8385 news articles from Aljazeera.net news website	Arabic	Graph embedding abstractive summarization	<i>ROUGE</i> and subjective evaluation	F1 score of 0.047	Low performance. Focused on well-written articles, unreproducible with Web 2.0 noisy nature
[32]	The Arabic Headline Summary (AHS) [31] and The Arabic Mogalad_Ndeef (AMN) [28] datasets, both comprise news articles	Arabic	Neural networks, including GRU, LSTM, and BiLSTM	<i>ROUGE</i> and <i>BLEU</i>	<i>ROUGE-1</i> score of 51.49	Focused on news only data, requires high computing power

Many attempts have been made in text summarization in Arabic and English. Many methods were applied, and different document types, including web pages, Wikipedia pages, and reviews. Many Arabic datasets for document summarization research were found in the reviewed literature, such as EASC [35] and ANT [16]; also, Elsaid et al. [5] stated in their comprehensive review the availability of 13 Arabic news datasets created for document summarization. Yet, no Arabic dataset was found that targets Arabic review summarization research. Moreover, studies focusing on Arabic review summarization are

limited; only one study was found in this area [34]. Thus, this research investigates a graph-based approach, TextRank, in the review summarization task. Graph-based summarization methods are language-independent; therefore, no linguistic resources are needed, which are scarce in the Arabic language [11]. Additionally, we propose to combine TextRank with sentiment analysis to improve overall performance and readability. The dataset consists of hotel reviews collected for sentiment analysis purposes [36]. In the next section, we elaborate on the methods applied and the dataset characteristics.

4. Methodology

We applied polarity classification and the TextRank algorithm in this study to summarize the hotel reviews. Five main conceptual stages are followed: data collection, preprocessing, polarity classification, summarization, and evaluation. Figure 1 provides an overview of these stages, which are explicitly discussed in the following subsections.

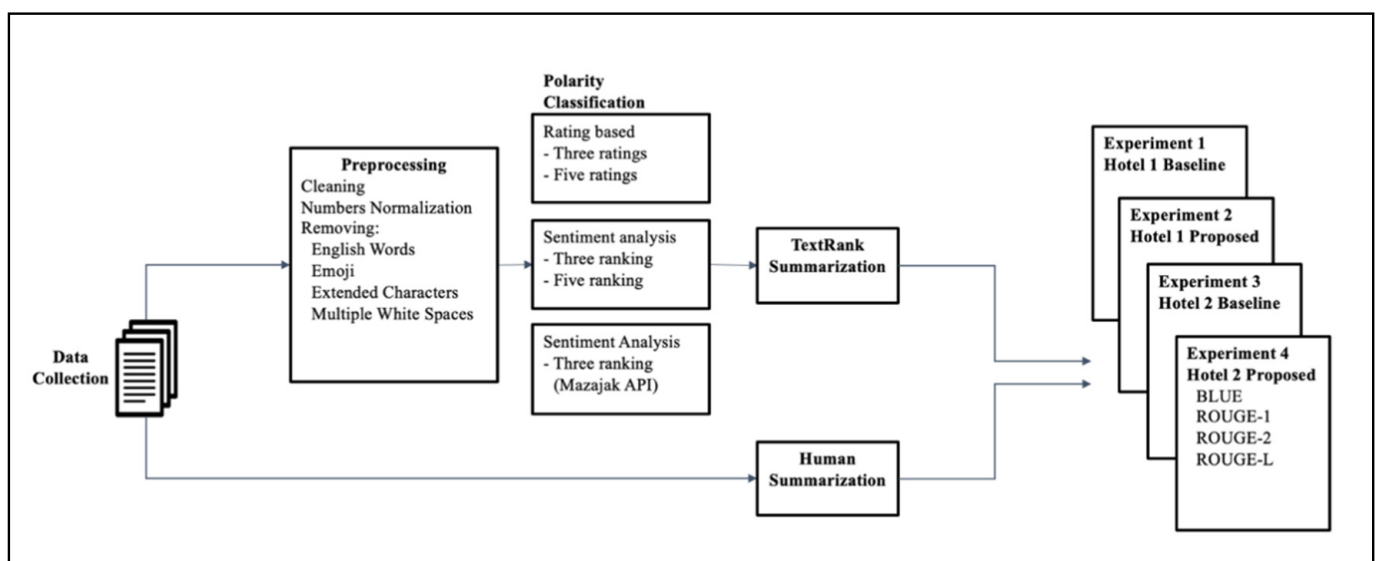


Figure 1. Proposed methodology stages.

4.1. Data Collection

In this study, we used a part of the dataset introduced by [36], where a large dataset was compiled from the hotel reservation site Booking.com during June/July 2016. The dataset contains 409,312 reviews for 1173 hotels written in Modern Standard Arabic and dialectal Arabic. The dataset comprises the attributes: hotel name, rate (reviewer's rating out of 5), user type (family, single, couple), room type, nights (number of nights stayed), and review. The dataset reviews rating frequency is presented in Figure 2. At the same time, the distribution of the number of reviews per hotel is shown in Figure 3. Most hotel reviews range between 7 and 147 reviews. The median number of reviews equals 224 reviews. To conduct the experiments, a summary should be generated for one hotel at a time, so we selected two hotels with a review count equal to the median value of 224.

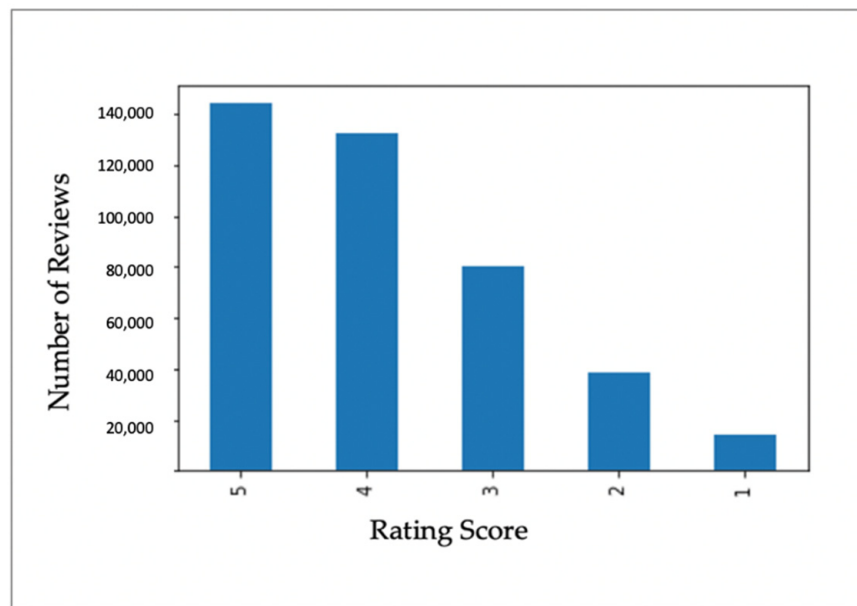


Figure 2. Reviews class frequency chart according to the users’ rating.

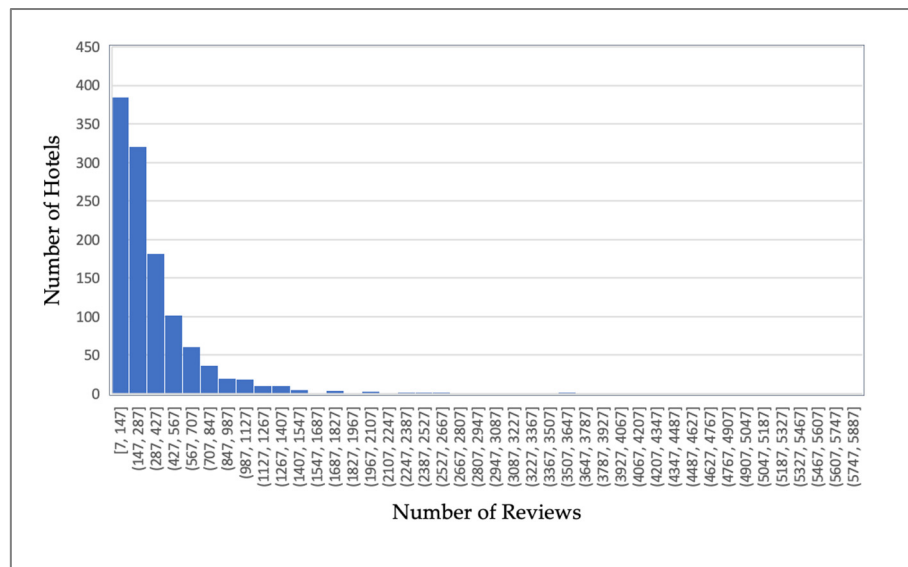


Figure 3. Dataset distribution according to the number of reviews per hotel.

4.2. Data Preprocessing

Before the summarization process, we prepared the text data using specific preprocessing steps to reduce noise and normalize the dataset. Table 2 presents a sample review before and after preprocessing. Figure 4 illustrates the major preprocessing procedures conducted in this work. First, we eliminated all null values, duplicates, hashtags, punctuations, and diacritics. Then, we normalized all numbers as most Arab countries use Indian numerals, while north African Arab countries use Arabic numerals. Normalization involves converting all numbers to Arabic numerals. See Table 2 for an example. Additionally, we removed English words to ensure that the algorithm will only use Arabic words. Finally, we removed Emojis, multiple white spaces and extended characters to reduce the noise on the dataset.

Table 2. Preprocessing applied to a sample review.

Before Preprocessing	After Preprocessing	Translation
<p>amazing'وننا اااااااا هالمنتج جمييييييييييييييييل لايفوتكم . "ماقصر العموم بكل الخدمات والميزه والادب والادب والاحترام @@@ ماراح نخليكم دوماً اليكم من الرياض #الرياض .الى .المنتج تسلمووووووون ١٠ من ١٠ .</p>	<p>وناسه هالمنتج جميل لايفوتكم ماقصر العموم بكل الخدمات والميزه والادب والاحترام ماراح نخليكم دوما اليكم من الرياض الى المنتج تسلمون 10 من 10</p>	<p>Fun and beautiful resort. All the staff did a great job in the service and dealt with politeness and respect. We will always get back to you from Riyadh to the resort. You get 10 out of 10</p>

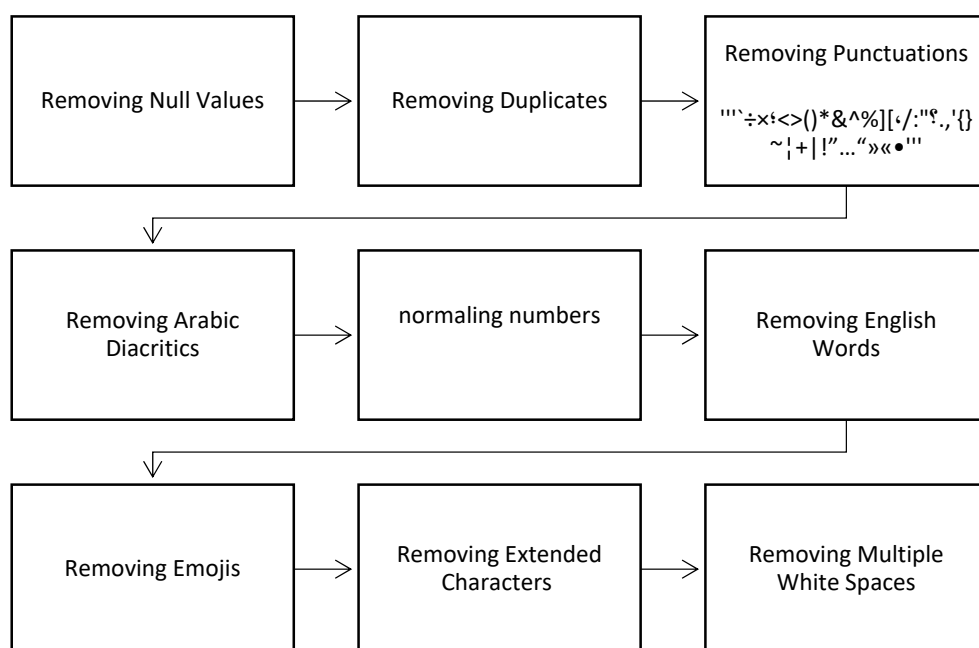


Figure 4. Text preprocessing steps applied in this work.

4.3. Polarity Classification

In this work, we propose to perform polarity classification before the automatic summarization to mimic the human way of writing summaries. Traditionally, it is assumed that user sentiment, expressed in text reviews, is reflected by the score rating. However, by skimming the reviews, we noticed that sometimes the sentiment of the review disagrees with the score value given by the user. A few examples of these reviews are presented in Table 3. To find whether the number of contradicting reviews is significant, we calculated the percentage of these reviews for each hotel, the result of this step is presented in Table 4. The result shows that the disagreement between the rating-based and content-aware sentiments is 18.7% in the first hotel and 5.35% in the second. This inaccuracy could affect the performance of the sentiment classifier negatively. Thus, to provide more accurate labeling, we considered text mining approaches to find the sentiment of each review. Formally, we have applied the following five methods:

1. Classifying is based on the original users’ rating, which includes ratings from 1 (lowest rating) to 5 (highest rating).
2. Classifying based on users’ positive, negative, or neutral ratings. To perform this approach, the users’ ratings of 5 and 4 were considered positive, three as neutral, and 2 and 1 as negative reviews.
3. Classifying based on sentiment analysis of all the reviews in the dataset, i.e., 409,312 reviews for the 1173 hotels. This classification was also based on the users’ original ratings; however, it considers the linguistic features and patterns modeled via

the training of deep learning architecture. This procedure is illustrated in Algorithm 1 below. The deep learning architecture is discussed in the following subsection.

4. This approach is similar to the previous approach and is based on sentiment analysis of all reviews in the dataset using deep learning modeling. However, the classification model classifies the reviews into positive, negative, and neutral using the same method described above for the second approach.
5. Instead of using the rating score for sentiment in this experiment, we relabeled each review automatically using a pretrained model Mazajak [37], a deep learning model trained on Arabic Twitter data using CNN-LSTM models, more about Mazajak is presented below. This API labeled each review as positive, negative, or neutral.

Table 3. Examples of the disagreements between the original reviews rating and the rating of the developed deep learning-based sentiment analysis model.

Review	Translation	Rating Based on the Proposed Deep Learning Model	Original Rating
يستحق التجربة مكان رائع كل شيء رائع	Worth the experience, great place, everything was great	5	2
تقلمي بطئ تنفيذ الخدمات	Slow services	3	4
رائع جدا انصحكم وبشدة الراحة والهدوء والفخامه	Wonderful, I highly recommend it, comfort, calm, and luxury	5	4
الفندق جميل للغاية تأخير في الفطور	Very nice hotel, late breakfast	4	5
غير مرضي الغرفة سيئة أقل من 3 نجوم	Unacceptable, the room was terrible, less than three stars	2	3
جيد الإفطار ممتاز	Good, breakfast was excellent	4	3

Table 4. Disagreements between the original reviews rating and the rating of the developed deep learning-based sentiment analysis model (left) and the rating distributions of the two hotels (right).

Hotel	Total Reviews	Disagreements	Percentage	Rating Distribution
1	224	42	18.7%	
2	224	12	5.35%	

Algorithm 1: The proposed polarity classification based on sentiment analysis.

Procedure: Classifying reviews based on sentiment analysis

Input:

R : a collection of labeled review $R = \{R_1, R_2, \dots, \dots, R_m\}$

Output:

R^U : a collection of updated labeled reviews $R = \{R_1^u, R_m^u, \dots, \dots, R_m^u\}$

Construct the review polarity classification model:

- Add embedding layer
- Add Bidirectional layer
- Add dense layer
- Add dropout layer # to avoid overfitting
- Add softmax layer to include all possible class labels
- Use categorical_crossentropy loss function

Train the the review polarity classification model

- validation_split = 0.1,
- number of epochs = 5,
- batch_size= 128

Plot Accuracy (training and validation) learning curves

- Finetune the model

Evaluate and test the model

Classify reviews and update R^U

4.3.1. Deep Learning for Polarity Classification

Recently, advances in NLP research have shown that deep learning (DL) techniques are cutting-edge technology for various text classification tasks [38]. For example, the study in [39] showed that the deep learning-based approach outperformed other algorithms, such as Naive Bayes, for text classification. Thus, in this work, we trained a DL model to predict the rating of the reviews. Figure 5 shows the proposed model general architecture. Model complexity is depicted by Table 5, which shows the number of parameters in each layer. The input features will be based on word embedding. An embedding is a dense representation in which similar words have an identical encoding [40]. Encoding values are trainable parameters learned from a dataset when performing a specific task. In this study, the input is represented as a 32-dim encoding vector. A bidirectional layer is then used; it trains two Long Short-Term Memory (LSTM) layers on the input sequence. The first is on the input sequence, and the second is on a reversed copy of the input sequence. This learning can provide additional context to the model and result in better and faster learning. We also used a dropout layer to control overfitting the training set. The output layer consists of 5 units representing the five categories of rating using the softmax function:

$$S(X_i) = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}} \tag{1}$$

where $S(X_i)$ is the softmax, X_i is the input vector, e^{x_i} is the standard exponential function for the input vector, e^{x_j} is the standard exponential function for the output vector, and K is the number of classes.

We trained the model using Adam optimizer with ten epochs and ReLu activation function:

$$f(x) = \max(x, 0) \tag{2}$$

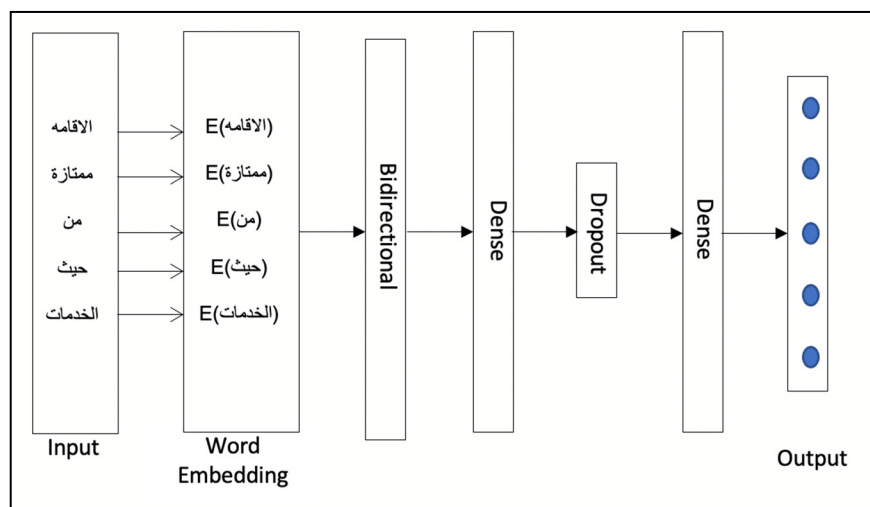


Figure 5. The proposed review classification model.

Table 5. The proposed deep learning model’s layers and the corresponding number of parameters.

Layer (Type)	Number of Parameters
embedding (Embedding)	9,227,072
bidirectional (Bidirectional)	12,672
dense (Dense)	2080
dropout (Dropout)	0
dense_1 (Dense)	165
Total params: 9,241,989	
Trainable params: 9,241,989	
Non-trainable params: 0	

4.3.2. Mazajak API for Polarity Classification

In the fifth experiment, we used Mazajak API [37], an open-source online Arabic sentiment analyzer. This model is based on a deep learning model, namely CNN-LSTM, and was trained on several Arabic dialect datasets. These datasets include SemEval 2017, ArSAS, and ASTD. SemEval is a series of NLP workshops, and in 2017 Arabic was first introduced. The dataset is composed of more than 3000 Arabic-labelled tweets. ArSAS is a dataset of 21 K human-labeled tweets from a range of topics in several different Arabic dialects. Finally, ASTD, Arabic Sentiment Twitter Dataset, is a collection of approximately 10K tweets in Egyptian dialect [41] on many other trending topics at the time of the collection.

4.4. TextRank Summarization

TextRank algorithm includes three steps: calculating the similarity value between sentences, calculating sentence weight, and ranking sentences. The similarity value between two sentences is calculated in TextRank based on Equation (1) below, where K represents sentences and w is the word:

$$sim_{ij} = \frac{|\{w_k | w_k \in k_i \ \& \ w_k \in k_j\}|}{\log(|k_i|) + \log(|k_j|)} \tag{3}$$

where sim_{ij} represents the similarity value between sentences k_i and k_j .

To calculate the sentence weight, Google’s PageRank utilizes a web link structure to calculate a quality ranking for each webpage. PageRank normalizes the number of links on a page rather than equally counting the links from all pages. PageRank is computed by Equation (2) below [42]:

$$PR(X) = (1 - d) + d \left(\frac{PR(T1)}{C(T1)} + \frac{PR(T2)}{C(T2)} + \dots + \frac{PR(Tn)}{C(Tn)} \right) \tag{4}$$

$PR(X)$ represents the PageRank of page X , while $PR(T1)$ represents the PageRank of page $T1$, which refers to page X . $C(T1)$ represents the number of outbound links on page $T1$. A damping factor is defined as d , which can be assigned a value between 0 and 1, which is the click-through probability used to prevent sinking pages with no outgoing links [42]. PageRank computation is applied to the TextRank text summarization algorithm to calculate the node weight. PageRank and TextRank are based on graphs; the difference lies in the definition of node and edge. In PageRank, the node represents the web page, and the edge represents the page’s link. While in TextRank, the node represents the sentence, and the edge represents the similarity value between sentences. In TextRank, sentence weight is calculated using Equation (3) below:

$$WS(V_i) = (1 - d) + d \sum_{V_j \in \text{In}(V_i)} \frac{sim_{ij}}{\sum_{V_k \in \text{Out}(V_j)} sim_{jk}} WS(V_j) \quad (5)$$

$WS(V_i)$ represents the weight for sentence i . sim_{ij} represents the similarity value of the sentence i and j . $\text{In}(V_i)$ are all edges pointing to sentence $i(V_i)$, while $\text{Out}(V_j)$ is all the edges pointing outward from a sentence $i(V_i)$, d represents the damping factor. Finally, after determining the weights of the sentences, the last step in obtaining the summary is to rank the sentences according to their weight, from highest to lowest, and the summary represents a certain percentage of the top-ranked sentences, depending on how long the summary is required to be. The Genism Python library [43] was used in this study to apply TextRank. The summaries ratio was set to 0.05 for the combined summary and 10% for each polarity group summary. The reason for choosing these parameters is to create a summary size close to the human-authored summary. Different ratios were tested, and those ratios generated the best results due to the size similarity.

4.5. Evaluation

To evaluate the quality of the generated summary, we used two metrics, Bilingual Evaluation Understudy (*BLEU*) and Recall-Oriented Understudy for Gisting Evaluation (*ROUGE*). *BLEU* is a precision-based metric for evaluating a generated sentence with respect to a reference sentence [42]. The range of the *BLEU* score is between 0 and 1. The higher the value, the higher the similarity of the candidate text to the original text. In this study, we used human summarization provided by the two native Arabic speakers in place of the original text. *BLEU* uses a modified precision formula that assesses the similarity between the human summary and the generated summary. Equation (4) defines the modified precision P_n as follows:

$$P_n = \frac{\sum_{C \in \{\text{Candidates}\}} \sum_{n-gram \in C} \text{countClip}(n - gram)}{\sum_{C' \in \{\text{Candidates}\}} \sum_{n-gram \in C'} \text{count}(n - gram)} \quad (6)$$

$\text{Count}(n - gram)$ represents the $n - gram$ number for the candidate in the test set, and $\text{CountClip}(n - gram)$ is the clipped $n - gram$ number for the candidate sentences. To maximize precision, *BLEU* avoids too short summaries by using a brevity penalty *BP* factor, which is computed by Equation (5) below:

$$BP = \begin{cases} 1 & \text{if } |c| > |r| \\ e^{(1 - \frac{|r|}{|c|})} & \text{if } |c| < |r| \end{cases} \quad (7)$$

where $|c|$ is the length of the summary, and $|r|$ is the length of the human summary. Then, the *BLEU* score is computed by Equation (6) as follows:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^N w_n \log P_n\right) \quad (8)$$

where N is the number of $n - gram$, w_n is the weight for each modified precision, the baseline is set to $N = 4$, w_n is $1/4 = 0.25$, and P_n is the modified precision [42].

ROUGE is a recall-based evaluation metric developed by Lin [44] to compare the automatically generated summary with the gold standard or the human-written summary. The algorithm used by *ROUGE* count the number of units such as $n - gram$, word sequences, and word pairs that appeared in both summaries under comparison [44]. *ROUGE* has many variations; the most used ones are *ROUGE - N* and *ROUGE-L*. *ROUGE - N* is an $n - gram$ recall between a generated summary and the gold standard. *ROUGE - N* is computed as follows:

$$ROUGE - N = \frac{\sum_{S \in (Ref)} \sum_{n-gram \in S} countMatch(n - gram)}{\sum_{S \in (Ref)} \sum_{n-gram \in S} count(n - gram)} \tag{9}$$

where *Ref* represents the reference summaries and n is the $n - gram$. $countMatch(n - gram)$ is the maximum number of $n - gram$ in both the reference summaries and the corresponding candidate summary. For example, *ROUGE-1* refers to the overlap of unigrams between the generated summary and the human reference summary; similarly, *ROUGE-2* denotes the overlap of bigrams between the generated summary and the human summary. *ROUGE-L*, longest common subsequence (*LCS*), is the common maximum length sequence between two given sequences, X and Y [44]. The larger the value of *ROUGE-L*, the more similar the two summaries. To evaluate summaries using *ROUGE-L*, the F-measure is computed to estimate the similarity between two summaries, X of length m and Y of length n , where X is a human summary sentence, and Y is the generated summary sentence. Thus, *ROUGE-L* is computed by finding F-measure as follows:

$$R = \frac{LCS(X, Y)}{m} \tag{10}$$

$$P = \frac{LCS(X, Y)}{n} \tag{11}$$

$$F = \frac{(1 + \beta^2) R P}{R + \beta^2 P} \tag{12}$$

Such that $LCS(X, Y)$ is the length of the longest common subsequence of X and Y , while $\beta = P/R$ when $\partial F/\partial R = \partial F/\partial P$.

To summarize, *BLEU* focuses on precision, or how much the words or $n - gram$ in the system-generated summary appear in the human summary. In contrast, *ROUGE* focuses on recall, i.e., how much the words or $n - gram$ in the human summary appear in the system summary. In other words, *BLEU* calculates the percentage of $n - gram$ in the candidate translation overlapping with the references [44]. To illustrate the working of both metrics, we ran the algorithms over simple sentences, as presented in Table 6. The stemming, returning the word to its root, improves the score significantly in both metrics, as shown in Tables.

As presented above, to evaluate a system-generated summary using *BLEU* or *ROUGE*, the summary should be compared against a gold standard or human summary. Two fluent Arabic speakers summarized the 224 reviews for the two hotels under the experiment. The two reviewers summarized the review using two different approaches as one reviewer rewrote and organized the information in a new way with his writing style. At the same time, the other selected unique reviews and kept the writing style unaltered. Doing so helps highlight differences between human reviewers and, thus, provides a baseline for evaluating the automatic review summarization.

Table 6. Examples of *BLEU* and *ROUGE* scores of simple Arabic reviews illustrating the stemming effect.

Sentence 1	Sentence 2	Before Stemming			After Stemming				
		<i>BLEU</i>	<i>ROUGE-1</i>	<i>ROUGE-2</i>	<i>ROUGE-L</i>	<i>BLEU</i>	<i>ROUGE-1</i>	<i>ROUGE-2</i>	<i>ROUGE-L</i>
فندق جميل	فندق جميل	1	F: 0.9 P: 1.0 R: 1.0	F: 0.9 P: 1.0 R: 1.0	F: 0.9 P: 1.0 R: 1.0	1	F: 0.9 P: 1.0 R: 1.0	F: 0.9 P: 1.0 R: 1.0	F: 0.9 P: 1.0 R: 1.0
فندق راقى وجميل	الفندق جميل وفخم	0	F: 0 P: 0 R: 0	F: 0 P: 0 R: 0	F: 0 P: 0 R: 0	0.6	F: 0.66 P: 0.66 R: 0.66	F: 0 P: 0 R: 0	F: 0.66 P: 0.66 R: 0.66
فندق راقى وجميل	الفندق جميل وراقى	0	F: 0 P: 0 R: 0	F: 0 P: 0 R: 0	F: 0 P: 0 R: 0	0.6	F: 0.66 P: 0.66 R: 0.66	F: 0 P: 0 R: 0	F: 0.66 P: 0.66 R: 0.66
استقبال الفندق جيد	الاستقبال جيد	0.3	F: 0.39 P: 0.5 R: 0.33	F: 0 P: 0 R: 0	F: 0.39 P: 0.5 R: 0.33	0.6	F: 0.79 P: 0.66 R: 1	F: 0 P: 0 R: 0	F: 0.79 P: 0.66 R: 1
استقبال سيء	الاستقبال سيء	0.5	F: 0.49 P: 0.5 R: 0.3	F: 0 P: 0 R: 0	F: 0.49 P: 0.5 R: 0.5	1	F: 0.9 P: 1.0 R: 1.0	F: 0.9 P: 1.0 R: 1.0	F: 0.9 P: 1.0 R: 1.0
كانت التجربة رائعة و مميزة	تجربتنا كانت مميزة و رائعة	0.8	F: 0.59 P: 0.6 R: 0.6	F: 0 P: 0 R: 0	F: 0.39 P: 0.4 R: 0.4	1	F: 0.9 P: 1.0 R: 1.0	F: 0 P: 0 R: 0	F: 0.39 P: 0.4 R: 0.4

5. Experimental Results

We carried out several experimental evaluations to demonstrate the proposed approaches' potential. First, summarization was performed without sentiment grouping, i.e., using the combined mixed reviews. Second, summarization was performed based on polarity grouping, which means grouping the reviews according to their polarity before feeding the TextRank algorithm. These experiments were divided further into five sub-experiments, including five and three rating grouping, five and three sentiments grouping, and grouping based on the sentiment analysis of the reviews using Majazak API [37]. Thus, the total number of experiments for the two hotels' summaries for both reviewers before and after applying the proposed evaluation method was 48. We have calculated the *BLEU*, *ROUGE-1*, *ROUGE-2*, *ROUGE-L*, precision, recall, and F1 scores for all the experiments.

To demonstrate the difference between summarizing all reviews and summarizing by sentiment based on rating score, a hotel with 55 reviews was selected. The reviews were summarized by both techniques, with a summary size of 10%. The result of this experiment is shown in Table 7. Positive reviews are underlined, and negative reviews are in bold. Positive and negative reviews are shared with the combined reviews; however, that is not always the case, as the neutral review is not part of the combined summary. Thus, the two techniques produce different results, so combining the three polarity classes does not yield the combined summary, as noted by the neutral summary. This explains the difference in the evaluation scores, as shown in Tables 8–11. The example highlights other essential aspects of the summarization techniques. First, TextRank summarizes by selecting a whole sentence ending with a period, so if a sentence contains many sentences without a period, it is considered one sentence. Second, the sentiment depends on the rating score, so it may not be very accurate, as explained in Section 4.3.

Table 7. Example of baseline and the proposed approach.

Summary Type	Summary
Combined reviews	<p>جيد جدا المكان نظيف وجود ملهى ليلي وعدم ذكر ذلك في موقع البكونغ. ضعيف جدا ضيق المواقف رائحة الفندق سيئه جدا صوت الديسكو مزعج لدرجة اني في الدور السادس ولا استطيع النوم خدمة تنزيل الاغراض لسيارة بطيئه البوفيه سيئ جدا. ممتاز طاقم العمل والنظافة ومواقف مجانيه وسهولة الوصوله اليه و يبعد عن سوق نايف 5 كيلو فقط والفطور ممتاز ورائع عبارة عن بوفيه مفتوح انا سكنت بالطابق الخامس هادئ ولا يوجد ازعاج من الملهى الليلي سوى هزات خفيفة جدا وسعر الفندق مناسب لا يوجد نت مجاني. اسوأ قائمة فندق نزلت فيه منذ 25 عام الموقع جيد الموظفين غير ودودين موظفي الاستقبال غير الكفاء الافطار سيء وموعده غير مناسب الاصوات عالية من الازعاج شديد من الديسكوتيك الفندق غير مناسب للأطفال الافطار هندية فقط لا يوجد اختيارات جيدة للطعام جميع القنوات بالتليفزيون هندية. مخيب للأمل فقط المرافق متهالكة لا يستاهل نجمة واحدة يوجد حشرات بالحمام الغرف المجاورة المدير الواي فاي بفلوس وبسعر مبالغ فيه الساعه ب 10 درهم الاستقبال تعاملهم سيء المواقف صغيره جدا ومعقده.</p>
Translation	<p>Very well. The place is clean. The presence of a nightclub and not mentioned it on the booking website. Very weak. Tight parking. The smell of the hotel is nasty. The disco sound is so annoying that I'm on the sixth floor and can't sleep. Car service is slow. The buffet is very bad. Excellent staff, cleanliness, free parking, easy access to it, and it is only 5 km from Naif market, and the breakfast is excellent and wonderful, which is an open buffet. I lived on the fifth floor, it was quiet, and there was no disturbance from the nightclub except for very light trembles, and the hotel price was suitable. There is no free internet. Worst hotel I've stayed in in 25 years The location is good, The staff is not friendly, the reception staff is unskilled, breakfast is bad and the time is not suitable, the breakfast menu is Indian only There are no good choices of food, the hotel is not suitable for children The noise is very loud from the disco from the neighboring rooms, there are insects in the bathroom, the facilities are dilapidated not worth a star, only Indian TV channels. Disappointing only the manager, WIFI is overpriced, and the hour is 10 dirhams. The reception is bad. Parking is very small and complex.</p>
Positive reviews	<p>ممتاز طاقم العمل والنظافة ومواقف مجانيه وسهولة الوصوله اليه و يبعد عن سوق نايف 5 كيلو فقط والفطور ممتاز ورائع عبارة عن بوفيه مفتوح انا سكنت بالطابق الخامس هادئ ولا يوجد ازعاج من الملهى الليلي سوى هزات خفيفة جدا وسعر الفندق مناسب لا يوجد نت مجاني</p>
Translation	<p>Excellent staff, cleanliness, free parking, easy access to it, and it is only 5 km from Naif market, and the breakfast is excellent and wonderful, which is an open buffet. I lived on the fifth floor, it was quiet, and there was no disturbance from the nightclub except for very light trembles, and the hotel price was suitable. There is no free internet.</p>
Negative reviews	<p>مخيب للأمل فقط المدير الواي فاي بفلوس وبسعر مبالغ فيه الساعه ب 10 درهم الاستقبال تعاملهم سيء المواقف صغيره جدا ومعقده</p>
Translation	<p>Disappointing only the manager, WIFI is overpriced, and the hour is 10 dirhams. The reception is bad. Parking is very small and complex.</p>
Neutral reviews	<p>مناسب اذ وجدت سعره رخيص الموقع وجود سكورتي في المواقف للمساعدة في الوقوف تعامل الموظفين. تم سحب المبلغ مرتين من الفيزا وطلبت زيادة يوم اخر فقالوا لا بد من دفع 120 درهم اضافي فطلبت المبلغ المسحوب كاش فوافقوا وطلبت تعويضي عن فارق العملة ورسوم البنك لانه خطأهم ويتحملون جميع تبعاته فرفضوا وغضبت فوافقوا على اعطائي يوم اضافي بنفس المبلغ المسحوب وهو نفس مبلغ الحجز من بوكينج كانوا يحاولون التفاوض بطريقة فاشلة الماء في ارضية الحمام يتراكم المواقف ضيقة اصناف الفطور قليلة ومتواضعة القنوات قليلة وهندية على الاغلب</p>
Translation	<p>Convenient, as I found its price cheap. The location and the presence of security in the parking lots to help stand up and treat the staff. The amount was withdrawn twice from the visa, and I asked for an increase for another day. They said that I must pay an additional 120 dirhams, so I asked for the withdrawn amount in cash, so they agreed and asked to compensate me for the currency difference and the bank fees because it is their fault, and they bear all the consequences. They were trying to negotiate in an unsuccessful way, the water on the bathroom floor was accumulating, the parking was cramped, the breakfasts choices were few and modest, the channels were few and mostly Indian.</p>

The experimental results for both hotels with three-class polarity grouping are presented in Table 8 and with five-class polarity grouping in Table 9. The best scores are highlighted in bold. Rows H1 and H2 show the agreement between the human summaries 1 and 2.

Table 8. Performance evaluation of summarizing both hotels' reviews with TextRank, and the proposed method with three-class polarity grouping by the three suggested approaches: rating score, the DL-based sentiment, and using Mazajak API, against human-generated summaries 1 and 2 (H1 and H2).

Experiment	Hotel 1				Hotel 2				
	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	
Human 1 and 2 agreement score	0.4744	R: 0.4586 P: 0.3881 F: 0.4204	R: 0.1295 P: 0.1289 F: 0.1292	R: 0.4173 P: 0.3531 F: 0.3825	0.2677	R: 0.6590 P: 0.3917 F: 0.4914	R: 0.263 P: 0.1373 F: 0.1806	R: 0.6206 P: 0.3690 F: 0.4628	
H1	TextRank	0.2983	R: 0.3462 P: 0.5000 F: 0.4091	R: 0.1187 P: 0.2307 F: 0.1568	R: 0.3026 P: 0.4370 F: 0.3576	0.4030	R: 0.6791 P: 0.4965 F: 0.5736	R: 0.4395 P: 0.2950 F: 0.3530	R: 0.6604 P: 0.4829 F: 0.5578
	Rating grouping	0.4327	R: 0.4241 P: 0.4300 F: 0.4270	R: 0.1756 P: 0.2058 F: 0.1895	R: 0.3620 P: 0.3671 F: 0.3645	0.4807	R: 0.6666 P: 0.5148 F: 0.5809	R: 0.3435 P: 0.2646 F: 0.2989	R: 0.6430 P: 0.4965 F: 0.5604
	Sentiment grouping	0.3213	R: 0.3707 P: 0.4965 F: 0.4245	R: 0.1385 P: 0.2466 F: 0.1773	R: 0.3159 P: 0.4230 F: 0.3617	0.5469	R: 0.6648 P: 0.5603 F: 0.6081	R: 0.3644 P: 0.3119 F: 0.3361	R: 0.6432 P: 0.5421 F: 0.5883
	API grouping	0.3484	R: 0.3881 P: 0.5034 F: 0.4383	R: 0.1446 P: 0.2398 F: 0.1804	R: 0.3450 P: 0.4475 F: 0.3896	0.5766	R: 0.6414 P: 0.5785 F: 0.6083	R: 0.3778 P: 0.3412 F: 0.3585	R: 0.6212 P: 0.5603 F: 0.5892
F: 0.3866H2	TextRank	0.3355	R: 0.3535 P: 0.6033 F: 0.4458	R: 0.1152 P: 0.2250 F: 0.1524	R: 0.3171 P: 0.5413 F: 0.3999	0.4147	R: 0.3676 P: 0.4521 F: 0.4054	R: 0.1057 P: 0.1360 F: 0.1189	R: 0.3271 P: 0.4022 F: 0.3608
	Rating grouping	0.3909	R: 0.3448 P: 0.4132 F: 0.3759	R: 0.0810 P: 0.0954 F: 0.0876	R: 0.3034 P: 0.3636 F: 0.3308	0.3771	R: 0.3716 P: 0.4827 F: 0.4199	R: 0.1096 P: 0.1619 F: 0.1307	R: 0.3421 P: 0.4444 F: 0.3866
	Sentiment grouping	0.3283	R: 0.3368 P: 0.5330 F: 0.4127	R: 0.0902 P: 0.1613 F: 0.1157	R: 0.2976 P: 0.4710 F: 0.3647	0.3504	R: 0.3540 P: 0.5019 F: 0.4152	R: 0.0986 P: 0.1619 F: 0.1226	R: 0.3297 P: 0.4674
	API grouping	0.3647	R: 0.3557 P: 0.5454 F: 0.4306	R: 0.1105 P: 0.1840 F: 0.1381	R: 0.3288 P: 0.5041 F: 0.3980	0.3613	R: 0.3535 P: 0.5363 F: 0.4261	R: 0.1097 P: 0.1900 F: 0.1391	R: 0.3131 P: 0.4750 F: 0.3774

As shown in the tables, the agreement between the two human summaries in the *BLEU* score ranges from 0.26 to 0.4, which is relatively low, as each reviewer created the summary differently. Generally, summaries are challenging to rate by similarity. Other factors should be considered, such as the summary's quality and whether it covers all the aspects mentioned in the original text. The proposed approach achieved the best scores in most cases, except in a few instances. A compiled summary of the best *BLEU* scores is presented in Table 10, and the best ROUG-1 F scores are in Table 11. TextRank outperformed our approach on a few occasions; however, the average shows that our method performs better, using both metrics.

Table 9. Performance evaluation of summarizing both hotels' reviews with TextRank, and the proposed method with five-class polarity grouping by the two suggested approaches: rating score and the DL-based sentiment against human-generated summaries 1 and 2 (H1 and H2).

Experiment	Hotel 1				Hotel 2			
	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Human 1 and 2 agreement score	0.4744	R: 0.4586 P: 0.3881 F: 0.4204	R: 0.1295 P: 0.1289 F: 0.1292	R: 0.4173 P: 0.3531 F: 0.3825	0.2677	R: 0.6590 P: 0.3917 F: 0.4914	R: 0.263 P: 0.1373 F: 0.1806	R: 0.6206 P: 0.3690 F: 0.4628
H1	TextRank	R: 0.3462 P: 0.5000 F: 0.4091	R: 0.1187 P: 0.2307 F: 0.1568	R: 0.3026 P: 0.4370 F: 0.3576	0.4030	R: 0.6791 P: 0.4965 F: 0.5736	R: 0.4395 P: 0.2950 F: 0.3530	R: 0.6604 P: 0.4829 F: 0.5578
	Rating grouping	R: 0.3726 P: 0.4860 F: 0.4218	R: 0.1476 P: 0.2511 F: 0.1859	R: 0.3324 P: 0.4335 F: 0.3763	0.5743	R: 0.6744 P: 0.5899 F: 0.6294	R: 0.4218 P: 0.3738 F: 0.3964	R: 0.6536 P: 0.5717 F: 0.6099
	Sentiment grouping	R: 0.3692 P: 0.4790 F: 0.4170	R: 0.1310 P: 0.2149 F: 0.1628	R: 0.3180 P: 0.4125 F: 0.3592	0.5730	R: 0.6666 P: 0.5831 F: 0.6221	R: 0.4189 P: 0.3693 F: 0.3925	R: 0.6484 P: 0.5671 F: 0.6051
H2	TextRank	R: 0.3535 P: 0.6033 F: 0.4458	R: 0.1152 P: 0.2250 F: 0.1524	R: 0.3171 P: 0.5413 F: 0.3999	0.4147	R: 0.3676 P: 0.4521 F: 0.4054	R: 0.1057 P: 0.1360 F: 0.1189	R: 0.3271 P: 0.4022 F: 0.3608
	Rating grouping	R: 0.3190 P: 0.4917 F: 0.3869	R: 0.0784 P: 0.1340 F: 0.0989	R: 0.2895 P: 0.4462 F: 0.3512	0.3699	R: 0.3567 P: 0.5249 F: 0.4248	R: 0.1067 P: 0.1814 F: 0.1343	R: 0.3255 P: 0.4789 F: 0.3875
	Sentiment grouping	R: 0.3234 P: 0.4958 F: 0.3915	R: 0.0772 P: 0.1272 F: 0.0961	R: 0.2938 P: 0.4504 F: 0.3556	0.3643	R: 0.3541 P: 0.5210 F: 0.4217	R: 0.1008 P: 0.1706 F: 0.1268	R: 0.3203 P: 0.4712 F: 0.3813

Table 10. Summary of the best BLEU scores achieved for both hotels.

	Method	Hotel 1	Method	Hotel 2	Average
H1	TextRank	0.2983	TextRank	0.4030	0.3507
	Three-class grouping: rating grouping	0.4327	Three-class grouping: API grouping	0.5766	0.5047
	Five-class grouping: rating grouping	0.3223	Five-class grouping: rating grouping	0.5743	0.4483
H2	TextRank	0.3355	TextRank	0.4147	0.3751
	Three-class grouping: rating grouping	0.3909	Three-class grouping: rating grouping	0.3771	0.3840
	Five-class grouping: Sentiment grouping	0.3342	Five-class grouping: rating grouping	0.3699	0.3521

Table 11. Summary of the best ROUG-1 F scores achieved for both hotels.

	Method	Hotel 1	Method	Hotel 2	Average
H1	TextRank	0.4091	TextRank	0.5736	0.4914
	Three-class grouping: rating grouping	0.4270	Three-class grouping: API Grouping	0.6083	0.5177
	Five-class grouping: rating grouping	0.4218	Five-class grouping: rating grouping	0.6294	0.5256
H2	TextRank	0.4458	TextRank	0.4054	0.4256
	Three-class grouping: API grouping	0.4306	Three-class grouping: API Grouping	0.4261	0.4284
	Five-class grouping: sentiment grouping	0.3915	Five-class grouping: rating grouping	0.4248	0.4082

6. Discussions

This research attempts to answer the question: can the Arabic automatic review summarization improve by considering the reviews' sentiment factor? The result confirms that creating summaries for each polarity category yields better scores than the traditional summarization method. TextRank extractive summarization technique creates a summary using the PageRank approach, selects the most important sentences, and includes them in the final output. Using TextRank with the five-class rating grouping achieved the best score and the best average score, with *ROUGE-1* of 0.6294 and 0.5256, respectively. However, the three-class rating grouping was the best on average. The rating grouping, in general, generated most of the high scores compared to the Mazajak API [37] and the sentiment-based approach (deep learning). However, in a few cases, the API grouping gave the best scores; in one case, the sentiment grouping yielded the best score, as seen in Tables 9 and 10. Human summary 1 generated better results, as the method used by this reviewer was selecting the sentences rather than rewriting new text; in contrast, reviewer 2 read all the reviews and rewrote the sentences in her writing style.

The dataset contains reviews of more than a thousand hotels, with some having more than 5000 reviews. We selected two hotels with the median number of reviews (224). The chosen summarization ratio was 5% and 10%, and the rationale was to keep the automatic summary length close to the human summary. We tested other higher ratios, up to 30%, and those chosen yielded the best results and provided better readability. The fact that 10% of the reviews were adequate to generate these high scores is encouraging and indicates that the proposed approach has potential for future applications.

To evaluate the summaries, we adopted the *BLEU* and *ROUGE* metrics. Although these metrics are widely used for summarization tasks, evaluating summaries needs some kind of human judgment, as done in [3,34]. Moreover, when we analyzed how these algorithms work, we found that they have some limitations, as discussed in Section 4.5. Stemming, however, mitigated the shortcomings of these evaluation metrics, as shown in Table 5.

Regarding Arabic review summarization, we only found one study in this area [34], where the study used hotel reviews from Tripadvisor and applied feature-base and sentiment-based summaries and evaluated the summaries subjectively, claiming 72% accuracy. Other Arabic studies applied graph-based methods for document summarization, such as [7] achieved an F-measure of 0.6799, [11] with an F-measure of 0.75, and [2] scored a *ROUGE-1* value of 0.561–0.660 on 45% summary size. Summarizing reviews, however, is more challenging than documents, as reviews are noisy, redundant, and contain many spelling and grammatical errors. Despite these hurdles, our approach achieved an average *ROUGE-1* score of 0.5256 on a 10% summary size.

Finally, reviews are vital for businesses and e-commerce; summarizing reviews could be helpful to stakeholders and customers alike. Websites selling products or services strive to provide summary statistics for their reviews; adding text summaries is a compelling feature that would add a competitive advantage to the business.

7. Conclusions and Future Work

This research provides a novel automatic text summarization framework for Arabic online reviews based on polarity. We used the Arabic hotel reviews dataset and TextRank graph-based summarization technique. Obtaining the sentiment for reviews, whether by considering the rating score or text mining, yielded better summarization performance scores than the baseline approach. The best average *BLEU* score achieved was 0.5047 using the three-class grouping, while the best average *ROUGE-1* score was 0.5256 using the five-class rating grouping. Both scores outperform TextRank, which reached a *BLEU* score of 0.3507 and a *ROUGE-1* score of 0.4914. Although many Arabic studies investigated document summarization techniques, only one study was found that targeted Arabic reviews; furthermore, no Arabic dataset was available for review summarization research.

The evaluation was conducted by selecting two hotels with the median number of reviews. Human summaries were authored by two native Arabic speakers for the two selected

hotels. Although more hotels with a different number of reviews should be considered to achieve better generalization, this work can be viewed as a seed for further research and development in Arabic reviews ATS. Additionally, applying other graph-based methods, such as LexRank, was challenging due to the lack of Arabic summarization tools.

The obstacles faced in this study opened many future research opportunities. There is a need for more Arabic open-source tools and Arabic review datasets for various products and services, including restaurants, hotels, books, and mobile apps. Review datasets should also include human summaries, preferably by more than one native Arabic speaker. Nevertheless, other qualitative methods should be considered for evaluating summaries, such as those suggested in [3,35]. Different approaches, including deep learning, could be investigated and compared with graph-based methods. Additionally, more focus is needed on multi-document summarization, as few studies have been conducted in this area, especially for the Arabic language. Additionally, review summarization could benefit from adding an aspect-based categorization [3]. Finally, combining text summarization with more text preprocessing, such as correcting spelling and eliminating duplicates, would improve the performance and summary readability. Although there is a scarcity of Arabic textual resources, tools, and datasets, we were able to build an Arabic sentiment-based review summarizer with remarkable results. We hope that the positive results achieved by this work encourage more research in the highlighted areas to enrich Arabic research and development.

Author Contributions: Conceptualization, A.A., G.A. and H.S.A.; methodology, G.A. and H.S.A.; software, H.S.A.; validation, A.A., H.S.A. and G.A., and formal analysis, H.S.A.; resources, G.A. and H.S.A.; data curation, H.S.A.; writing—original draft preparation, A.A., G.A. and H.S.A.; writing—review and editing, G.A. and H.S.A.; visualization, H.S.A. and G.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used is publicly available and was obtained from [36].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nayar, R. Role of Web 3.0 in Service Innovation. In *The Handbook of Service Innovation*; Springer: Berlin/Heidelberg, Germany, 2015.
2. Alami, N.; En-Nahnahi, N.; Ouatik, S.A.; Meknassi, M. Using Unsupervised Deep Learning for Automatic Summarization of Arabic Documents. *Arab. J. Sci. Eng.* **2018**, *43*, 7803–7815. [[CrossRef](#)]
3. Chen, Y.; Chang, C.; Gan, J. A template approach for summarizing restaurant reviews. *IEEE Access* **2021**, *9*, 115548–115562. [[CrossRef](#)]
4. Marzijarani, S.B.; Sajedi, H. Opinion mining with reviews summarization based on clustering. *Int. J. Inf. Technol.* **2020**, *12*, 1299–1310. [[CrossRef](#)]
5. Elsaid, A.; Mohammed, A.; Ibrahim, L.F.; Sakre, M.M. A Comprehensive Review of Arabic Text Summarization. *IEEE Access* **2022**, *10*, 38012–38030. [[CrossRef](#)]
6. Amoudi, G.; Albalawi, R.; Baothman, F.; Jamal, A.; Alghamdi, H.; Alhothali, A. Arabic rumor detection: A comparative study. *Alex. Eng. J.* **2022**, *61*, 12511–12523. [[CrossRef](#)]
7. Elbarougy, R.; Behery, G.; El Khatib, A. Extractive Arabic Text Summarization Using Modified PageRank Algorithm. *Egypt. Inform. J.* **2020**, *21*, 73–81. [[CrossRef](#)]
8. Suhara, Y.; Wang, X.; Angelidis, S.; Tan, W.-C. *OpinionDigest: A Simple Framework for Opinion Summarization*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 5789–5798.
9. Haque, M.M.; Pervin, S.; Begum, Z. Literature Review of Automatic Multiple Documents Text Summarization. *Int. J. Innov. Appl. Stud.* **2013**, *3*, 121–129.
10. Etaiwi, W.; Awajan, A. Graph-based Arabic NLP Techniques: A Survey. *Procedia Comput. Sci.* **2018**, *142*, 328–333. [[CrossRef](#)]
11. Alami, N.; Meknassi, M.; Ouatik, S.A.; Ennahnahi, N. Arabic text summarization based on graph theory. In Proceedings of the 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), Marrakech, Morocco, 17–20 November 2015; pp. 1–8.
12. Ibrahim, M.N.; Maria, K.A.; Jaber, K.M. Summarization Systems (AMD-SS). In Proceedings of the 2017 8th International Conference on Information Technology (ICIT), Amman, Jordan, 17 May 2017; pp. 1013–1022.

13. Varade, S.; Sayyed, E.; Nagtode, V.; Shinde, S. Text Summarization using Extractive and Abstractive Methods. *ITM Web Conf.* **2021**, *40*, 03023. [[CrossRef](#)]
14. Erkan, G.; Radev, D.R. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *J. Artif. Intell. Res.* **2004**, *22*, 457–479. [[CrossRef](#)]
15. Brin, S.; Page, L. The anatomy of a large-scale hypertextual Web search engine BT—Computer Networks and ISDN Systems. *Comput. Netw. ISDN Syst.* **1998**, *30*, 107–117. [[CrossRef](#)]
16. Mihalcea, R.; Tarau, P. TextRank: Bringing Order into Text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25–26 July 2004; pp. 404–411.
17. Chouigui, A.; Khiroun, O.B.; Elayeb, B. An Arabic Multi-Source News Corpus: Experimenting on Single-Document Extractive Summarization. *Arab. J. Sci. Eng.* **2021**, *46*, 3925–3938. [[CrossRef](#)]
18. Luhn, H.P. The Automatic Creation of Literature Abstracts. *IBM J. Res. Dev.* **1958**, *2*, 159–165. [[CrossRef](#)]
19. Gunawan, D.; Harahap, S.H.; Rahmat, R.F. Multi-document Summarization by using TextRank and Maximal Marginal Relevance for Text in Bahasa Indonesia. In Proceedings of the 2019 International Conference on ICT for Smart Society (ICISS), Bandung, Indonesia, 19–20 November 2019; pp. 1–5.
20. Gabriella, N.H.; Siantama, R.; Amadea, C.I.A.; Suhartono, D. Extractive Hotel Review Summarization based on TF/IDF and Adjective-Noun Pairing by Considering Annual Sentiment Trends. *Procedia Comput. Sci.* **2021**, *179*, 558–565.
21. Al-Abdallah, R.Z.; Al-Taani, A.T. Arabic Single-Document Text Summarization Using Particle Swarm Optimization Algorithm. *Procedia Comput. Sci.* **2017**, *117*, 30–37. [[CrossRef](#)]
22. Qaroush, A.; Farha, I.A.; Ghanem, W.; Washaha, M.; Maali, E. An efficient single document Arabic text summarization using a combination of statistical and semantic features. *J. King Saud Univ. Comput. Inf. Sci.* **2021**, *33*, 677–692. [[CrossRef](#)]
23. El-Haj, M.; Kruschwitz, U.; Fox, C. Experimenting with Automatic Text Summarisation for Arabic. In Proceedings of the Language and Technology Conference, Poznan, Poland, 6–8 November 2011; pp. 490–499.
24. Fejer, H.N.; Omar, N. Automatic Arabic text summarization using clustering and keyphrase extraction. In Proceedings of the 6th International Conference on Information Technology and Multimedia, Barcelona, Spain, 1–3 April 2014; pp. 293–298.
25. Haboush, A.; Al-Zoubi, M.; Momani, A.; Tarazi, M. Arabic text summarization model using clustering techniques. *World Comput. Sci. Inf. Technol. J. ISSN* **2012**, *2*, 741–2221.
26. Al Qassem, L.; Wang, D.; Barada, H.; Al-Rubaie, A.; Almoosa, N. Automatic Arabic Text Summarization Based on Fuzzy Logic. In Proceedings of the 3rd International Conference on Natural Language and Speech Processing, Trento, Italy, 11–12 September 2019; pp. 42–48.
27. Elgamel, M.; Hamada, P.S.; Aboelezz, P.R.; Abou-kreisha, M. Better Results in Automatic Arabic Text Summarization System Using Deep Learning based RBM than by Using Clustering Algorithm based LSA. *Int. J. Sci. Eng. Res.* **2019**, *10*, 781–786.
28. Zaki, A.M.; Khalil, M.I.; Abbas, H.M. Deep Architectures for Abstractive Text Summarization in Multiple Languages. In Proceedings of the 2019 14th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, 17–18 December 2019; pp. 22–27.
29. Elmadani, K.N.; Elgezouli, M.; Showk, A. BERT Fine-tuning for Arabic Text Summarization. *arXiv* **2020**, arXiv:2004.14135.
30. Al-Maleh, M.; Desouki, S. Arabic text summarization using deep learning approach. *J. Big Data* **2020**, *7*, 109. [[CrossRef](#)]
31. Etaïwi, W.; Awajan, A. SemG-TS: Abstractive Arabic Text Summarization Using Semantic Graph Embedding. *Mathematics* **2022**, *10*, 3225. [[CrossRef](#)]
32. Wazery, Y.M.; Saleh, M.E.; Alharbi, A.; Ali, A.A. Abstractive Arabic Text Summarization Based on Deep Learning. *Comput. Intell. Neurosci.* **2022**, *2022*, 1566890. [[CrossRef](#)] [[PubMed](#)]
33. Elayeb, B.; Chouigui, A.; Bounhas, M.; Khiroun, O.B. Automatic Arabic Text Summarization Using Analogical Proportions. *Cognit. Comput.* **2020**, *12*, 1043–1069. [[CrossRef](#)]
34. El-Halees, A.M.; Salah, D. Feature-Based Opinion Summarization for Arabic Reviews. In Proceedings of the 2018 International Arab Conference on Information Technology (ACIT), Werdanye, Lebanon, 28–30 November 2018.
35. El-Haj, M.; Kruschwitz, U.; Fox, C. Using Mechanical Turk to Create a Corpus of Arabic Summaries. In Proceedings of the Language Resources and Evaluation Conference (LREC), Valletta, Malta, 17–23 May 2010; pp. 36–39.
36. Elnagar, A.; Khalifa, Y.S.; Einea, A. Hotel arabic-reviews dataset construction for sentiment analysis applications. *Stud. Comput. Intell.* **2018**, *740*, 35–52.
37. Farha, I.A.; Magdy, W. Mazajak: An online arabic sentiment analyser. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, Florence, Italy, 1 August 2019; pp. 192–198.
38. Alwehaibi, A.; Bikdash, M.; Albogmi, M.; Roy, K. A study of the performance of embedding methods for Arabic short-text sentiment analysis using deep learning approaches. *J. King Saud Univ. Comput. Inf. Sci.* **2021**, *34*, 6140–6149. [[CrossRef](#)]
39. Huan, J.L.; Sekh, A.A.; Quek, C.; Prasad, D.K. Emotionally charged text classification with deep learning and sentiment semantic. *Neural Comput. Appl.* **2022**, *34*, 2341–2351. [[CrossRef](#)]
40. Almuzaini, H.A.; Azmi, A.M. Impact of Stemming and Word Embedding on Deep Learning-Based Arabic Text Categorization. *IEEE Access* **2020**, *8*, 127913–127928. [[CrossRef](#)]
41. Heikal, M.; Torki, M.; El-Makky, N. Sentiment Analysis of Arabic Tweets using Deep Learning. *Procedia Comput. Sci.* **2018**, *142*, 114–122. [[CrossRef](#)]

42. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics—ACL '02, Philadelphia, PA, USA, 7–12 July 2002; p. 311.
43. Rehurek, R.; Sojka, P. Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valleta, Malta, 22 May 2010; pp. 45–50.
44. Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2004; pp. 74–81.