Effects of Using Arabic Web Pages in Building Rank Estimation Algorithm for Google Search Engine Results Page

Mohamed Almadhoun
School of Computer Sciences, Universiti Sains Malaysia, Malaysia
mdmadhoun@ucas.edu.ps

Nurul Malim
School of Computer Sciences, Universiti Sains Malaysia, Malaysia
nurulhashimah@usm.my

Abstract: *Search Engine Optimization (SEO) aims to improve a website's reputation and user experience. Without effective SEO strategies, it requires significant investment in paid advertisements. Search Engines (SEs) use algorithms to rank results, assessing on-page and off-page factors for relevance. Machine learning techniques have been used to build classifiers for estimating page rank. However, no research has compared rank estimation with other languages or analyzed the effects of different languages on performance or differences between SEO factors. The study aims to improve rank estimation algorithms for Arabic web pages on desktop devices using a new multi-category dataset from Google Search Engine Results Page (SERP). The experimental findings suggest that Arabic web pages are more suitable than English ones for training a model to estimate the ranking of Arabic web pages. Machine learning models were applied to two datasets. SE scraping was used to collect URLs, descriptions, and other data from the Google SE. Data preprocessing steps were taken before using the datasets for rank estimation algorithms. Experiments were conducted to assess the implications of using Arabic and English web page datasets.*

Keywords: *Web data mining, search engine optimization, search engine results pages, rank estimation, scrap search engine, machine learning, digital marketing.*

Received June 21, 2023; accepted October 8, 2023
https://doi.org/10.34028/iajit/20/6/15

1. Introduction

These days, Search Engines (SEs) serve as the primary information source for users in need of information retrieval [4, 29]. SEs improve opportunities to identify and access the appropriate websites that employ a technique known as Search Engine Optimization (SEO). Its goal is to raise the standard and reputation of a certain website and encourage a positive user experience [35]. A user can receive thousands of visitors and more attention with the aid of effective SEO [33]. The main solution for presenting the appropriate web pages based on the ranking is SEO [2]. SEO has been discovered to be more useful and advantageous than any other traditional marketing strategy [38].

71.33% of searches on a SE go to a website appearing on the first page of organic search results, according to Sistrix and Moz statistics, with the second and third pages generating 5,59% of all clicks. Not using effective SEO strategies to rank high in organic search results will require paying a lot of money for paid advertisements for particular keywords, but using SEO will save money [38].

SEO is used to raise a website's position for certain search terms by controlling inbound links and web page aspects [18]. SEs use the page ranking algorithm to sort websites according to their content, structure, and popularity [2]. The importance of web page factors and their effect on Search Engine Result Page (SERP) ranking has been the subject of several research papers. The bulk of research approaches are centered on looking at highly ranked websites, evaluating their attributes, and figuring out which criteria are most commonly connected with those attributes and have a big effect on page ranking [18].

Web pages should be built in accordance with search algorithms to be found. To assist in maintaining websites' top rankings, it's needed to understand what new SE standards are being employed now. Scraping SERP will help to look at the pages that can show up in the top SERP results to understand how SEs rank web pages. Web scraping, or gathering URLs from SERP, was a common research methodology.

Google SE was chosen for this study because, with a global market share of 92.48% across all platforms as of May 2022, Google leads the SE industry according to StatCounter Global Stats [30]. This study's main objective is to suggest a ranking algorithm that can estimate the Google SERP ranking of web pages for Arabic websites via desktop devices, utilizing a newly created multi-category dataset from SERPs, concerning various categories and types of search phrases. Moreover, it will examine the effects of utilizing datasets in Arabic and English on enhancing SERP rank estimation. The experimental findings of this study will demonstrate that when estimating the rank of Arabic web

pages, it is preferable to utilize a dataset of Arabic web pages rather than English ones for training the SERP rank estimator.

The rest of this paper is organized as follows. Section 2 provides a scientific background about SEO. Section 3 reviews the related work on web page ranking estimation and SEO. Section 4 presents the experimental setup, methodology, and the results of the comparative analysis. Section 5 discusses the implications and findings of the experiments. Section 6 concludes the paper, summarizes the main contributions, and suggests some directions for future work.

2. Background

In terms of frequency, web browsing is second only to e-mail. The information that is accessible on the enormous internet is identified by a SE [21]. Crawling, indexing, sorting, determining relevance, and retrieval are some of SEs' most crucial tasks [2].

Crawling is the process of visiting websites to scan the pages and copy them so that a SE index may be created. The crawler also visits the pages that have hyperlinks on the page that's crawled. In the course of its work, a crawler gathers keywords, photos, and other information that could be included in user queries, and then copies the resulting web pages. The crawler typically begins its operation from the URL list, visits each page, filters the links on each page using content analysis, and then deletes redundant URLs after downloading each page. Finally, it gathers any new links it finds and adds them to the URL list. Pages are kept in the index during the crawling process. The SE examines the content found on the website, records it all in its index, and gives it a sort so it will be displayed appropriately on the search results page [5].

The user's query is matched by a SE with related web pages in the web database using result matching. Their result ranking determines the order in which the user will see the search results. In a perfect world, the user would discover results that are interesting to him on the first pages. A sorting algorithm is used by SEs to rank the results. SEs analyze the content after the web has been crawled to create an index that points to the relevant result [21]. The Google pagerank algorithm defines a numerical score that assesses how relevant a web page is to a given query. Due to the high PageRank score value that defines the list of SEPR for matching searches, it is significant [1].

On-page optimizations primarily refer to the technical effort carried out on the website to incorporate target keywords in various places [22]. It is a technique for making it easier for the SE to understand the website's content [3]. The text on a web page, text in meta tags, links, images, obvious navigation, page title, use of H tags, URL, and HyperText Markup Language (HTML) code are all examples of on-page SEO factors, and they are all completely within the webmaster's control [18]. According to Ziakis and Vlachopoulou [38], on-page criteria include the use of an SSL certificate, specific keywords, responsive design for mobile devices, and website loading time [34]. According to Matosevic *et al*. [18], keywords are search queries made up of one or more words used when looking for information on SEs. They are crucial to the SEO process. The focus keywords for a web page should be included in both on-page and off-page SEO strategies, according to some experts.

On the other hand, off-page SEO aims to manage an outside factor that affects the site's ranking independently of the website [3]. Off-page optimizations are primarily based on the work done on other websites [22]. The use of hyperlinks from other websites that have been optimized for SEs to the connected pages, the quality of incoming links and their relevance to the website's specialization, and recommendations from social networking sites are all examples of off-page factors. Web pages with more links were therefore seen to be more significant and ought to show up higher in search results [18, 34]. In addition, website design, meta tags, and keywords are three elements that affect internal website optimization; public domain, social media, and linking are three aspects that affect external website optimization [34].

Machine learning is an interdisciplinary field that uses algorithms to enhance the effectiveness of data. It involves databases, statistics, and data science. The goal is to teach computers how to recognize patterns, identify structures, and predict variable values. The two main forms of machine learning are supervised and unsupervised, depending on whether the data is tagged or not [18].

3. Related Work

In several studies, classifiers for estimating a web page's rank were built using machine learning techniques, and the best combinations of features were discovered by applying statistical analysis to the chosen features such as Portier *et al*. [23], Salminen *et al*. [26], Su *et al*. [32], Arora and Bhalla [4], Manohar and Punithavathani [17]; Jayaraman *et al*. [14], and Matosevic *et al*. [18]. To demonstrate the feasibility of their results and offer recommendations, others offered strategies and conducted tests utilizing particular techniques such as An and Jung [3] and Roslina and Nur Shahirah [25]. Other studies looked at the web rankings of academic institutions to determine the connection between the popularity of academic institutions and the SEO rating of their websites such as Halibas *et al*. [13], Ziakis *et al*. [39], Dalvi and Saraf [7], Schilhan *et al*. [27], Vállez and Ventura [36], Shahzad *et al*. [28], Özkan *et al*. [19], and Giannakoulopoulos *et al*. [10]. Others studied tourism websites such as Pan [20] and Vyas [37], or news and media websites such as Giomelakis *et al*. [11], Karyotakis *et al*. [16], Giomelakis and Veglis [12], Prawira and Rizkiansyah [24], and Dick [8].

The top SEO variables for ranking a web page on the first page of SERP were specified by Portier *et al*. [23] using a variety of machine learning techniques. They gathered their data by using a series of search queries to scrap the Google SE. After scanning the websites of the SERP results' URLs, they retrieved the relevant elements.

Salminen *et al*. [26] employed a learn-to-rank machine learning system to predict where websites will appear in Google search results. Their information was created by scraping SE and collecting various SEO elements from SERP sites. They employed hyperparameter random optimization to obtain the ideal parameters for the fundamental model. Also, they calculated the significance values for the extracted SEO variables. They used keywords in Finnish, and the top 10 results for each search phrase were compiled. To gather 23 factors from each web page, they developed a Python script.

Drivas *et al*. [9] proposed a predictive model for finding the most efficient combinations of characteristics that boost the visibility of organic SE results by analyzing the SEO elements that were derived from a set of 171 cultural heritage websites using a set of big data analytics techniques.

With a cap of 100 results per SE for each of the submitted search queries, Joglekar *et al*. [15] gathered their data by crawling the four SEs: Google, Yahoo, Bing, and DuckDuckGo. All of the URLs on the pages accessed by their crawler were taken from search results. They compiled the textual content of web pages into a data dictionary. They built an unsupervised machine learning algorithm for ranking web pages based on content quality.

3.1. Extraction of SEO Factors

Salminen *et al*. [26] created a Python script to download web page HTML, compute a set of factors, and utilize desktop software called "netpeak checker" to extract additional information from web pages. To create a dataset from scratch, Portier *et al*. [23] combined features from two SEO software applications and utilized a customized scraper designed for their purposes to extract features. "Porter stemmer" was used by Matosevic *et al*. [18] to extract keyword frequencies. Drivas *et al*. [9] used the "google search console API" to retrieve information about the size of the study websites, used the "checkbot API" to extract a set of 55 technical SEO factors related to crawling, speed, and security, and used the "similar web API" to retrieve a set of behavioral SEO factors for the pages of websites, such as visit duration, number of clicks, and bounce rate. Strzelecki [31] retrieved data from "google search console" for a full 15 months, including information on searches, clicks, impressions, locations, devices, and Click Through Rate (CTR), using the desktop application "clusteric search auditor".

3.2. Machine Learning

Portier *et al*. [23] classified web pages as either being on the first page of SERP (top 10) or not using a binary classification technique. They applied four classification models. Using metrics based on the confusion matrix, measured performance. The top 10 characteristics and the least significant five features for each classification model were determined using feature weighting. MLP neural network was used by Banaei and Honarvar [6] to forecast the position of a web page in Google search results. With category 1 having the greatest rank and category 5 having the lowest rank, they separated the rankings of the URLs in their dataset into 5 groups. Matosevic *et al*. [18] entered their data through decision trees, naive bayes, k-nearest neighbors (k-NN), Support Vector Machine (SVM), and logistic regression, five classification methods. They used the hold-out approach and 10-fold cross-validation to assess each classifier's accuracy. They also used hyperparameter tuning to get the best results possible from each classifier. None of the earlier studies compared classifiers created using various languages. Arabic keywords were not used in any research. The content of websites in different languages varies and there can be variations in SEO parameters like keyword occurrences.

3.3. Findings and Results

The average Cross-validation Normalized Discounted Cumulative Gain (NDCG) scores for Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM) by Salminen *et al*. [26] were 0.852 and 0.848, respectively. From the weighting of features by the SHapley Additive exPlanations (SHAP) algorithm, they discovered that more internal links increase the rank, but not by a large amount, that mid to low response times improve ranking, that long H1 causes lower ranking, that a high number of H tags improve ranking, and that a low image count increases rank.

The best factors, according to Portier et al. [23] are Alexa rank, the number of backlinks, keyword repetition, the total number of words in the content, the number of internal links, keyword density, and the number of pages indexed by Bing SE. They disregarded the use of SSL or a keyword in the domain name as well as the fact that a web page serves as the homepage. Nevertheless, utilizing Alexa rank was not a good choice for them because Amazon has discontinued using it and it is seen as a dependent variable.

4. Methodology and Results

In this study, machine learning models were applied to two datasets, each of which provided new information and a basis for an algorithm to estimate SERP rank. Using a set of predetermined keywords, a collection of web pages was obtained from SE results. Using the on-page SEO criteria for these web pages as input, datasets

will be built. Algorithms for estimating SERP rank can be created using machine learning. A significant number of web page parameters must be extracted for this procedure to be successful and precise. The act of collecting URLs, descriptions, and other information from SEs is known as SE scraping. The methodology employed in this research involves scraping a SE using keywords from English and Arabic languages. Once the web pages have been obtained, a range of SEO tools will be used to extract further information about them. Several data preprocessing steps will then be taken before utilizing the datasets to create rank estimation algorithms and conduct data mining and analysis to uncover insights.

The main goal of this study approach is to assess the implications of using datasets from Arabic or English web pages to develop rank estimation algorithms for labeling Arabic and English web pages. Based on how the SE responds to queries used from different languages, the research will provide SERP rank prediction algorithms for web pages. For the SE-scraped web pages, the methodology will extract information from on-page SEO factors. Figure 1 summarizes the process of research.

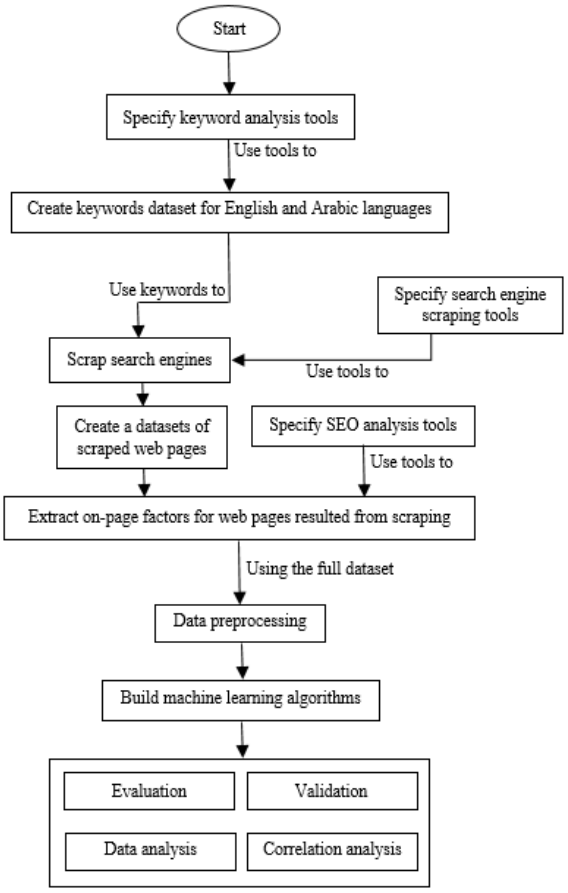

Figure 1. Process of data extraction and preparation.

4.1. Selection of Search Keywords and Scraping SERP

In order for machine learning to construct efficient classifiers, it is crucial to have a variety of keywords, whether the Arabic or English web pages. An example subset of the keywords used in this study is shown in Table 1. The search query that will be submitted to the SE is listed in the "keyword" column, and the "search volume" column lists the number of times per month that a keyword is submitted to the google seglobally. the "keyword ideas" tool in the "google ads" web system can be used to determine search volume.

Table 1. Sample subset of keywords dataset.

Keyword	Search volume (global)	Language
قصص اطفال	49,500	Arabic
تعريف المحاسبة	880	Arabic
يوم الزراعة العربي	70	Arabic
أنواع التجارة الإلكترونية	140	Arabic
وحدة قياس الطاقة	880	Arabic
البيئة والانسان	1000	Arabic
الصحة للاطفال	100	Arabic
ريادة الأعمال	10K – 100K	Arabic
tax return	246,000	English
what is agriculture	90,500	English
car manufacturers	74,000	English
travel news	74,000	English
ecommerce website templates	12,100	English
fashion store	60,500	English
gifts for men	823,000	English
best antivirus	60,500	English

Two keyword datasets, one with 100 English keywords and the other with 100 Arabic keywords, were chosen. Both datasets had unique keywords, with the Arabic dataset having 12 categories of Arabic subjects and the English dataset having keywords chosen from 16 different categories. To choose keywords, the website keywordtool.io was used. The chosen keywords had a variety in popularity and importance as demonstrated by the average monthly search volumes, which ranged from hundreds to thousands and millions.

To submit each keyword to the SE and obtain results, apify.com's automatic Google Search Results Scraper tool was used. The input settings were set to 3 maximum pages per search phrase, 10 results per Google page, and desktop results, with country US and language as default. The chart of SE scraping on apify is shown in Figure 2. Results were exported as a Comma-Separated Values (CSV) file for additional data preparation techniques, the most essential data columns were: "position" which is the order of some URL in the SERP (its values are from 1 to 30), "search result page" which is the page number of SERP that URL appeared in (its values are from 1 to 3), "type of result" which specify if result record is organic or paid, "search keyword" which is the submitted search term to SE, and "URL" which is the address of the web page that appeared in SERP.

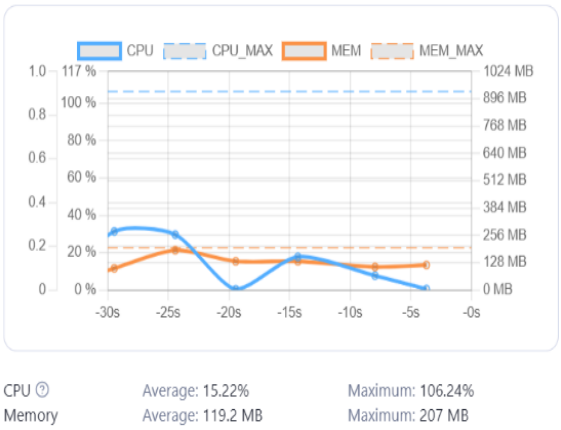

Figure 2. Chart and stats of google search scraper.

Duplicate results are wrong and shouldn't be shown because Google never returns the same Link twice on the same result page for a single query. A collection of duplicate URLs for the same search query was found and eliminated from the apify scraper's output data. Both paid and organic search results were included in the output data, however, data shouldn't include the sponsored results because their positions in SERP don't adhere to the SE ranking methodology. Moreover, entries with a missing search phrase or URL data were removed. To get rid of duplicates, paid results, and missing numbers, a procedure using the RapidMiner program was created (Figure 3).

After making use of the process in Figure 3 for each of the two datasets, the number of instances that were produced was as follows: The Arabic dataset included 2621 examples, and the English dataset had 2596 examples. The variance in the number of examples was caused by the different percentages of duplicates, missing values, and sponsored URLs in the two datasets.

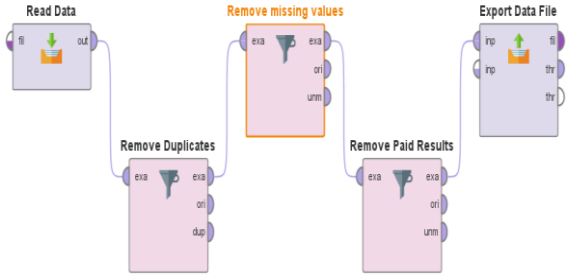

Figure 3. Process by RapidMiner to eliminate duplicates, missings, and paid results.

4.2. SEO Feature Extraction

To implement SERP rank estimation based on classification, classifiers must be trained using a dataset of SEO factors for a set of web pages. On-page factors require distinct methods to be extracted. To collect on-page factors, websites can be crawled, and their HTML content can be parsed. Many internet SEO tools can do this, for example, semrush.com or ahrefs.com.

The tool chosen in this research was ScreamingFrog, a desktop program, from among the SEO factor extraction tools that were researched online. ScreamingFrog was chosen because it can process 500 URLs for free in each run, runs can be repeated indefinitely, it is a quick and accurate tool, and it can extract numerous on-page features with the option to export the output as a CSV file. A screenshot of an output dataset is shown in Figure 4 which is the result of execution of the crawl operation.

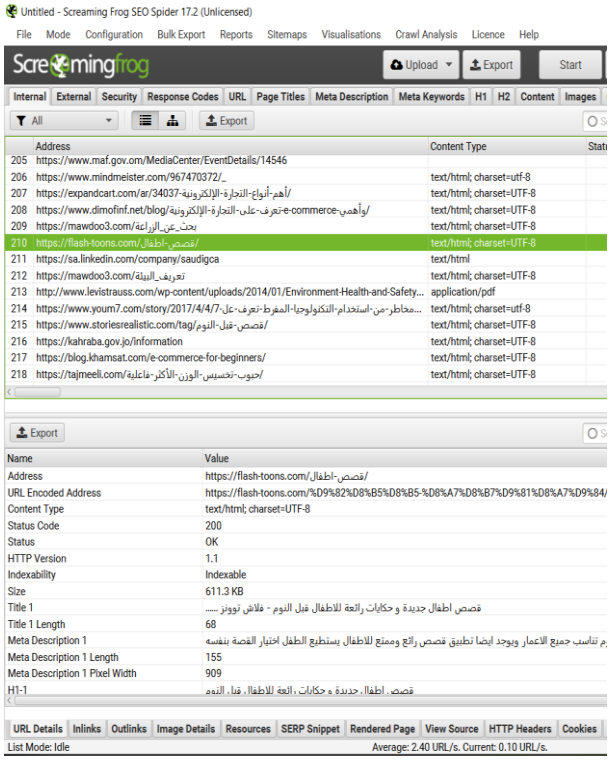

Figure 4. A sample screamingfrog crawl operation run.

There were more than 70 attributes in the output dataset from ScreamingFrog, however not all of them will be used since some of them have text values such as title or description, are unique such as crawling time, have only one value such as HTTP version 1.1, and are identities such as Hash.

4.3. Preparing Data

To create a single dataset with all attributes collected from both ScreamingFrog and apify, the output datasets from both tools should be combined. For each of the two datasets (English and Arabic), a RapidMiner process was developed to connect the results from apify and ScreamingFrog (SERP output and SEO extracted attributes). Figure 5 demonstrates the sequence of the RapidMiner process, which includes an operator to remove any duplicate URLs in the output of ScreamingFrog because they will have the same features. However, it's not correct to remove duplicate URLs in apify data because the same URL may appear in search results for numerous search keywords. The URL parameter was used as the join's key attribute since it is the common key value between the apify and ScreamingFrog datasets in Figure 5's join operator,

which performs a left join with the apify dataset on the left. Through this process, datasets are ready for the subsequent data preprocessing, analysis, and classification.

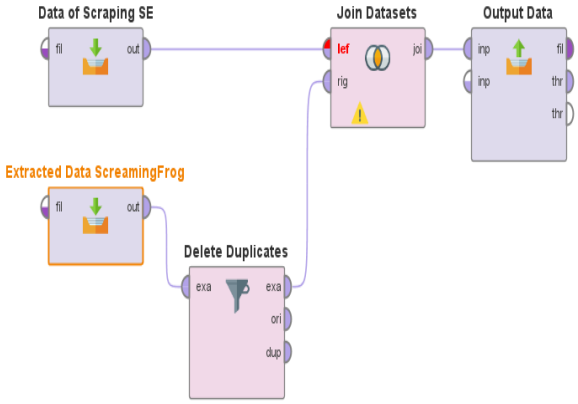

Figure 5. Apify and screamingfrog output data combined in the rapidminer process.

A RapidMiner sub-process was developed and used on each of the two input datasets (Arabic and English) to get the data ready for classification. The sequential operators used in pre-processing are depicted in Figure 6. Table 2 lists details of the process steps.

Table 2. Steps of data preparation sub-process.

Step #	Step name	Step description
1	Set role	For performing learning and testing the classification model, the operator "set role" will select the "search result page" property as the class label.
2	Remove redundant variables	Eliminating unnecessary variables that won't be used during any pre-processing or classification phases (Shown in the left list of Figure 7).
3	Generate attributes	Create new attributes for data analysis relating to the use of search keywords in titles, meta descriptions, and meta keywords (Figure 8 shows the derived values).
4	Remove paid	Remove the URL records from the sponsored search results as they are not organic results.
5	Select target attributes	Decide which features should be candidates for classification. (Shown in the right list of Figure 9).
6	Type change of status code	Change the "status code" property from numerical to poly-nominal because its numeric values are codes and not numerals for calculations (e.g., values are: 200, 302, 403).
7	Type change of search page	Because it is a requirement of the classification model, change the attribute "search result page" from numeric to poly-nominal (Its values are: 1, 2, and 3).

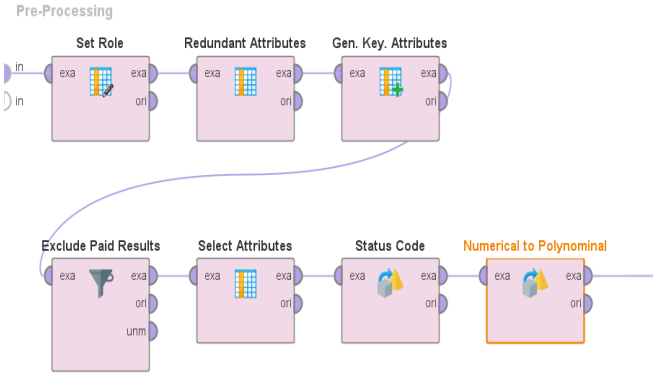

Figure 6. RapidMiner sub-process for data pre-processing.

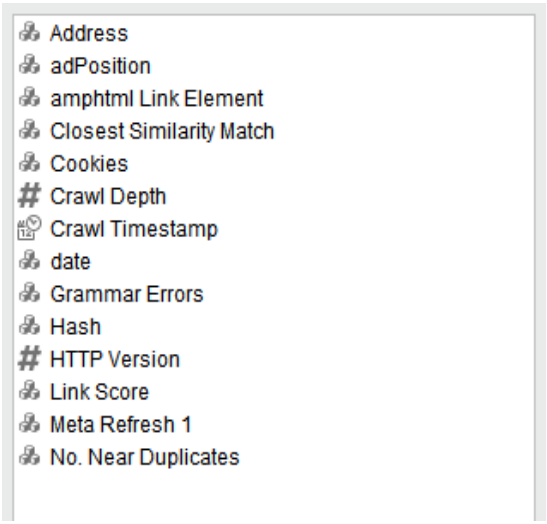

Figure 7. The eliminated unnecessary features.

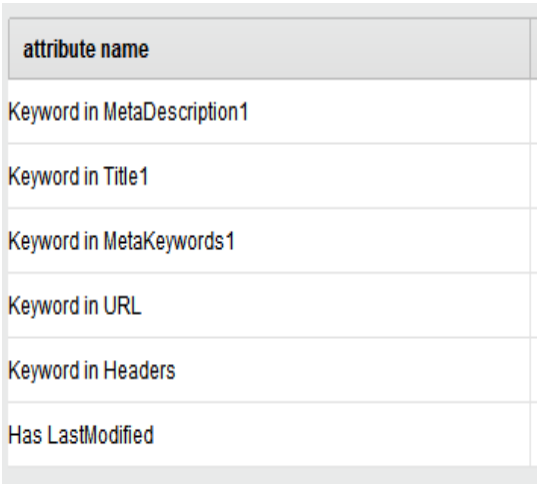

Figure 8. Derived features for keyword reccurances.

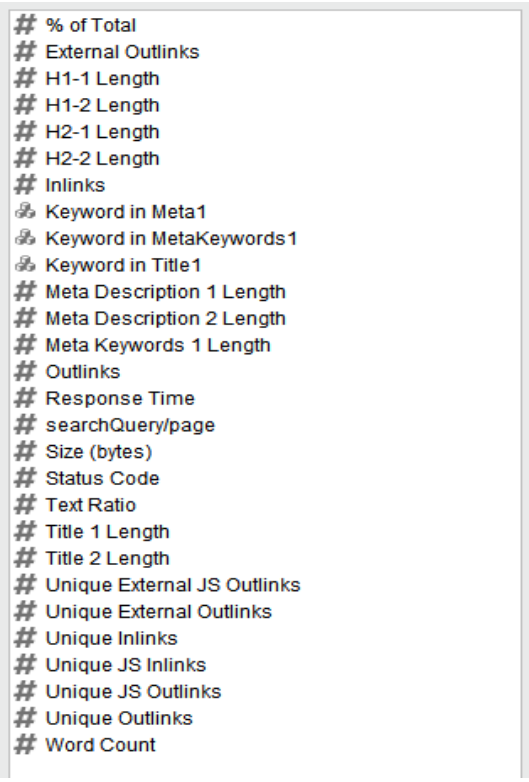

Figure 9. The features chosen for classification.

Table 3. Accuracy of all deployed classifiers.

Dataset	Decision tree	K-NN	Naïve Bayes	Gradient boosted trees	XGBoost	Random forest	Deep learning	W-logistic regression
English	71.2%	34%	38.79%	**77.58%**	72.96%	69.69%	41.37%	70.8%
Arabic	57%	37.1%	41.2%	58.5%	**60%**	58.68%	37.7%	54.18%

4.4. Classification and Results

In this research, four sets of experimental comparisons were applied between Arabic and English datasets. The first investigated the performance of different classifiers, the second exchanged training and testing datasets between Arabic and English, and the third and fourth were applied using binary classification. Eight machine learning models (decision tree, k-NN, naïve bayes, gradient boosted trees, XGBoost, random forest, deep learning, and W-logistic regression) were built using the training dataset, the Rapidminer default parameter configuration for each classifier was used. Each classifier was tested particularly to record its accuracy and compare it with other classifiers. The classification accuracy was the evaluation measure. The same training and testing datasets were used with all classifiers.

The first experiment utilized a comparison between the two datasets (English and Arabic). For each dataset, a RapidMiner process was developed, going through pre-processing, classification, and 10-fold cross-validation, and repeating this procedure with eight classifiers for each dataset (16 tests). The outcomes of this procedure are listed in Table 3, with the best accuracy value for each dataset using cross-validation being highlighted. By this comparison, similarity was regarded where each dataset was created from 100 search terms and passed through the same pre-processing using the same variables and the class label "search result page." The same classification was also utilized, with the same parameter values being specified.

In the second set of experiments, the classifier was trained and tested using English and Arabic datasets reciprocally. Testing data would be chosen from a different dataset than the dataset that was used for training. Given that the XGBoost classifier achieved the maximum classification accuracy with the Arabic dataset, it will be used. The comparison was conducted by utilizing split validation, local random seed, and stratified sampling. The four tests that were run are listed in Table 4. Both Arabic and English datasets were divided in the first and second tests into training and testing portions, respectively, of 70% and 30%. Due to the different training and testing datasets, the full datasets were used in the third and fourth experiments. The sub-process used to train and test the classifier using data from the same dataset is illustrated in Figure 10 including split validation (same language). But when the classifier is trained and tested using data from the two datasets (Arabic and English) the split validation sub-process is different Figure 11.

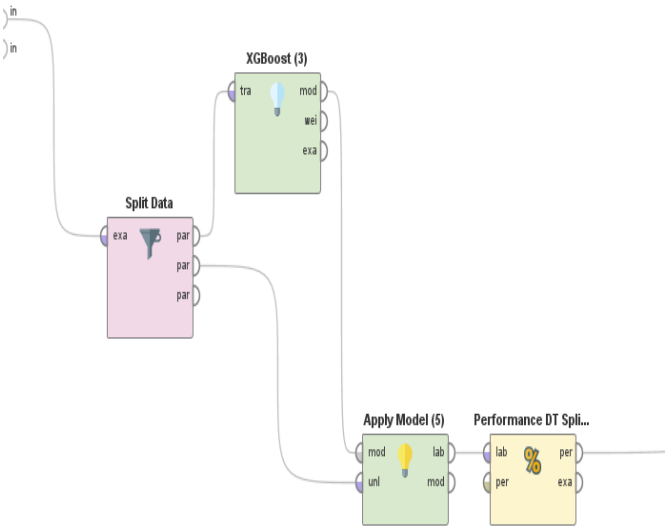

Figure 10. RapidMiner's split-validation subprocess (training and testing data from the same language).

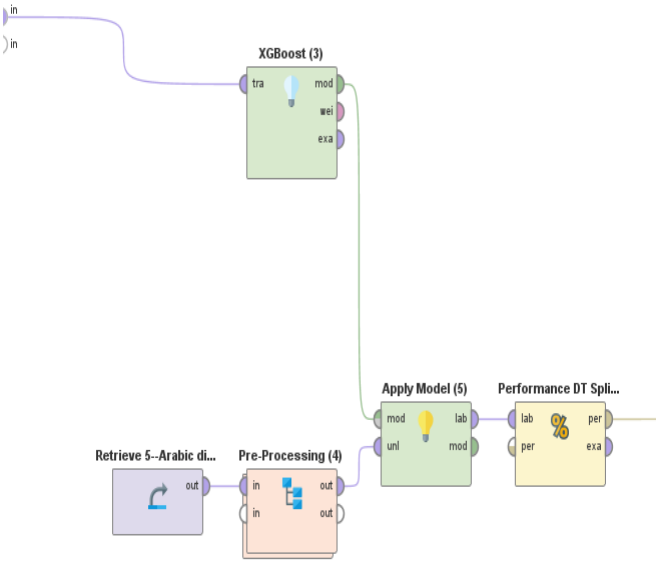

Figure 11. RapidMiner's split-validation subprocess (training and testing data from different languages)

Table 4. Accuracy of English and Arabic classification (class label: page number).

Experiment	Training data	Testing data	Accuracy
1	Arabic	Arabic	**58.52%**
2	English	English	**73.68%**
3	Arabic	English	41.76%
4	English	Arabic	49.56%

The third comparison employed binary classification. The two values "top10" and "not top10" were present on the newly utilized class label. Each of the classifiers described in Table 5 underwent 10-fold cross-validation as part of this series of tests.

Table 5. Performance of binary classification in terms of accuracy.

Dataset	Decision tree	K-NN	Naïve Bayes	Gradient boosted trees	XGBoost	Random forest	Deep learning	W-logistic regression
English	70.88%	59.75%	65%	**86.56%**	86.1%	68.57%	67.99%	70.88%
Arabic	85.31%	62.11%	63.42%	85.54%	**86.53%**	86.15%	69.52%	85.31%

The final comparison between Arabic and English datasets used binary classification with the label "top10," split validation with stratified sampling, and local random seed. Classification using XGBoost was used because it produced the highest accuracy rate with the Arabic dataset. Table 6 provides a list of performance outcomes.

Table 6. Performance of binary classification on Arabic and English datasets

Experiment	Training data	Testing data	Accuracy
1	Arabic	Arabic	**85.01%**
2	English	English	**85.11%**
3	Arabic	English	58.24%
4	English	Arabic	79.55%

5. Discussion and Analysis

The need for a unique SERP rank estimation classifier for each language is investigated in this study. This section will describe the series of comparisons that were mentioned earlier by presenting findings and discussing outcomes. These comparisons' main objectives were:

- Find the most effective classifier for estimating the rank of Arabic and English web pages on Google SERP.
- Obtain proof for the need to create a specific classifier for English and Arabic SERP rank estimation to get better performance.

By creating datasets for training the classifier of SERP rank estimation, the first set of comparison experiments mentioned in the previous section attempted to identify the top classifiers for ranking English and Arabic web pages. It also sought to determine whether it was necessary to create a separate classifier specifically for Arabic web pages in addition to the one for English. According to the accuracy rates obtained from the initial performance comparison and presented in Table 3, it can be concluded that in contrast to the English dataset, which was classified with the highest accuracy by gradient boosted trees (Figure 12), XGBoost had the highest accuracy rate for the Arabic dataset, making it the best classifier (Figure 13). This indicates that Arabic websites have different characteristics than English websites. This supports the theory that Arabic websites differ from English websites in terms of their characteristics and how SEs rank them, leading to the development of a custom classifier specifically for Arabic websites to estimate ranks.

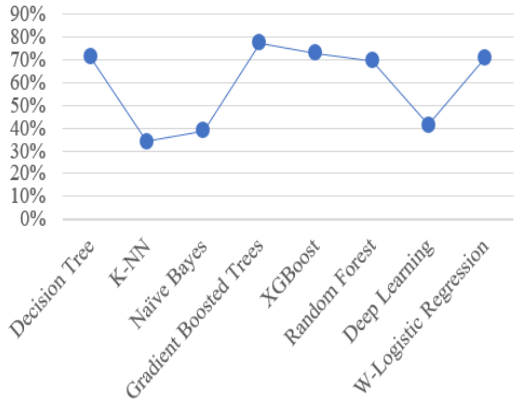

Figure 12. Performance of classifiers with English dataset.

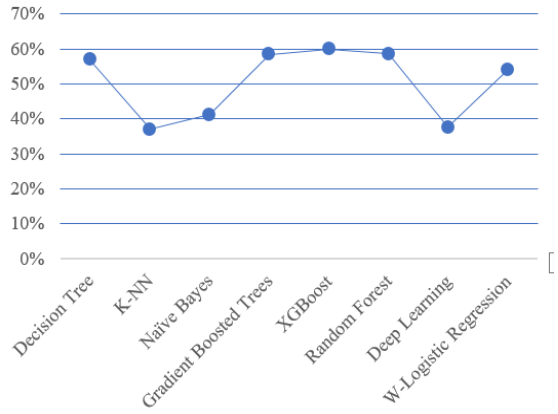

Figure 13. Performance of classifiers with Arabic dataset.

The objective of the second set of comparison tests was to determine whether a classifier developed using a dataset of English web pages could successfully classify Arabic web pages. The performance scores shown in Table 4 reveal two conclusions:

1. When using Arabic web pages in testing two classifiers, one trained on a dataset of Arabic web pages and the other on a dataset of English web pages, it was found that the classifier that was trained on the dataset of Arabic web pages performed better. This was demonstrated in Figure 14, which shows the results of tests 1 and 4, where the classifier trained on a dataset of Arabic web pages showed an improvement in accuracy of more than 10% over the classifier trained on a dataset of English web pages. The difference in accuracy rate is thought to be sufficient to support the claim that it is preferable to employ a classifier specifically designed for classifying Arabic websites rather than one that was trained on a dataset of English websites.
2. When the classifier is trained with a dataset of English web pages, the rank estimation of English websites will also be improved. This was examined by

obtaining a performance boost of more than 30% when classifying English web pages using a classifier trained on English web pages rather than Arabic web pages (tests 2 and 3 in Figure 14).

So, all of Table 4's findings support the idea that to improve performance and increase estimation accuracy, the SERP rank estimation classifier of web pages should be dedicated to a single language. For each language, a unique classifier should be created.

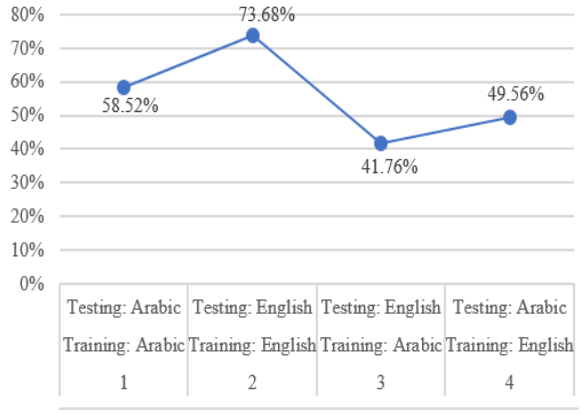

Figure 14. Performance of classification with datasets in Arabic and English.

In the third set of comparison tests, the binary classification method used by Portier *et al*. [14] for rank estimation was replicated. Nevertheless, as they are regarded as dependent variables, the bing index and Alexa rank were not utilized in these tests, and Alexa is no longer running. The following conclusions can be drawn from the outcomes of these tests, which are revealed in Table 5:

1. Binary classification has some higher accuracy rates. It may be difficult for the multi-class classifier to distinguish between websites on the second and third pages of SERP due to the smaller number of classes used with binary classification or the high degree of similarity between web pages on the second and third pages of SE results.
2. A similar result had been reached for the Arabic dataset which had XGBoost as the best classifier, unlike the English dataset which had Gradient Boosted Trees as the best classifier. This conclusion demonstrates the need for an additional classification model for Arabic websites that is separate from the one being developed for English websites.
3. By using binary classification with XGBoost, a performance boost of more than 25% was obtained with the Arabic dataset (Figure 15), indicating that the Arabic web pages listed on the second and third pages of the Google SE results are similar to one another and differ from the web pages listed on the first page (websites on the second and third pages were binary classified as belonging to the same class, "non-top10"). On the other hand, the English dataset achieved better performance with binary classification, but with less performance boost (Figure 16).

Figure 15. Performance of multi-class classification against binary classification for the Arabic dataset.

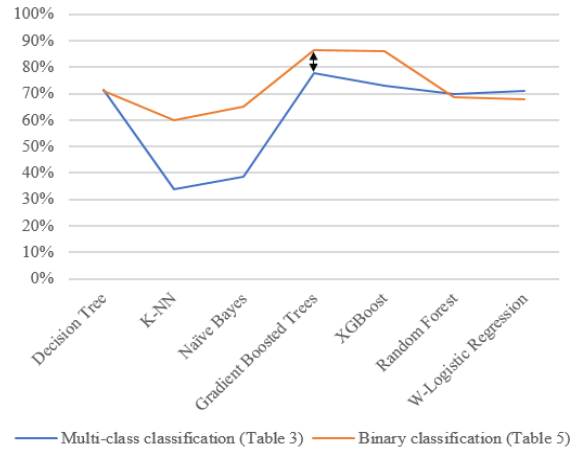

Figure 16. Performance of multi-class classification against binary classification for the English dataset.

The fourth set of comparison tests used binary classification and was comparable to the second set of comparison tests. These tests aimed to validate the results of the second comparison. The results shown in Table 6 allow us to draw the following conclusions:

1. Using Arabic and English training datasets, there was a 6% performance difference in classifying Arabic web pages (tests 1 and 4 in Figure 17). The same thing, but with a difference of over 25%, for the difference between tests 2 and 3. This supports Table's findings, which state that it is essential to estimate the SERP ranking of a web page using a classifier that's trained with web pages in the same language.
2. Experiments 1 and 3 in Figure 17 show that the performance of a classifier trained on the Arabic dataset decreased by more than 25% when the testing dataset was changed from Arabic to English, whereas experiments 2 and 4 show that the performance of a classifier trained on the English dataset decreased by about 6% when the testing dataset was changed from English to Arabic. This indicates that, in terms of SEO considerations, English-language websites are more informational than Arabic-language websites. This

might prompt experimentation with a hybrid classification model in the future, either in the classification model stages or in the training dataset.

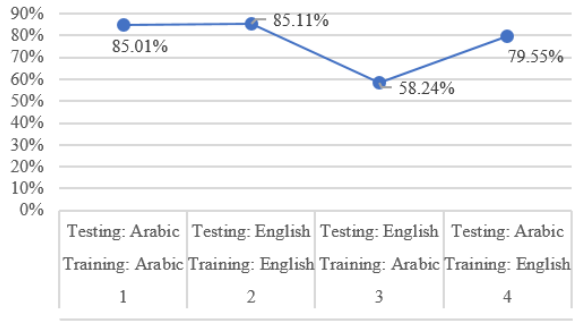

Figure 17. Binary classification performance of Arabic and English datasets.

6. Conclusions

Web pages with high SE rankings are typically more visible to users. For webmasters, increasing the ranking has taken precedence over all other considerations. In an effort to increase customer visits, SEO has grown into a sizable industry. A website can be made SEO-friendly by measuring its rank and optimizing it, this can help SEs become more capable of finding the needed information and websites will become more consistent. Combining machine learning and SEO can improve e-business performance and give management access to a different tool for making decisions.

Building classification models using datasets of characteristics extracted from the web pages that appear on SERP after scraping SERP with a set of search keywords is a good way to understand how SE ranking work. Therefore, the nature of training datasets will have an impact on how well the SERP rank estimation algorithm performs. Based on that, this research proposed an improved SERP rank estimation methodology through a dedicated dataset to train classifiers for Arabic and English web pages.

As a result, the methodology used in this study involved submitting Arabic and English keywords to the SE to scrap SERPs, crawl the resulting web pages, and then extract SEO factors to create training datasets. A collection of machine learning models was created to find the best model after applying data pre-processing tasks. A comparative approach was used to compare the performance of classifiers when training them using web pages of the same or different language about the testing dataset. To replicate the experiment of prior studies, a binary and multiple-class classification was used.

Results show that a good enhancement occurred in classification performance when using Arabic web pages rather than English web pages for estimating the rank of Arabic web pages, and vice versa for English web pages. The classifier trained on a dataset of Arabic web pages showed an improvement in accuracy of more than 10% over the classifier trained on a dataset of English web pages. This can support the hypothesis that the SERP rank estimation classifier of web pages should be dedicated to a single language to improve performance and increase estimation accuracy. So, for each language, a unique classifier should be created. By comparing 8 machine learning algorithms, the English dataset generated the highest accuracy of classification by gradient boosted trees which was 77.58% with multi-class classification and 86.56% with the binary classification, while XGBoost had the highest accuracy rate for the Arabic dataset with 60% by multi-class classification and 86.53% with the binary classification.

This research contributes to the field of web engineering and SEO by proposing an improved SERP rank estimation methodology that considers the language of the web pages. In future works, it will be beneficial to extract off-page SEO factors to increase the robustness of classifiers and boost performance. This will help to improve classification accuracy. Additionally, it will be preferable to use more search terms in order to have a larger training dataset and find classifiers that are more thorough and capable of ranking web pages from all anticipated categories or contents. Also, using the hyper-parameter configuration can improve the classifier's ability to give better performance. In our next experiments, we are going to use data mining techniques such as association rules or clustering to discover the best SEO practices for Arabic and English web pages.

Authors' Contributions

M.A. conceived and designed the study and methodology, performed the experiments, wrote the original draft of the paper, and reviewed and edited the final version. N.M. supervised the project and reviewed the paper.

References

[1] Al-Kabi M., Alsmadi I., and Wahsheh H., "Evaluation of Spam Impact on Arabic Websites Popularity," *Journal of King Saud University-Computer and Information Sciences*, vol. 27, no. 2, pp. 222-229, 2015. https://doi.org/10.1016/j.jksuci.2014.04.005, 2014. DOI:10.14445/22312803/IJCTT-V12P140

[2] Al-Mukhtar F., Mahmoodd N., and Kareem S., "Search Engine Optimization: A Review," *Applied Computer Science*, vol. 17, no. 1, pp. 69-79, 2021. DOI:10.23743/acs-2021-07

[3] An S. and Jung J., "A Heuristic Approach on Metadata Recommendation for Search Engine Optimization," *Concurrency and Computation Practice and Experience*, vol. 33, no. 3, pp. 1-10, 2019. https://doi.org/10.1002/cpe.5407

[4] Arora P. and Bhalla T., "A Synonym Based Approach of Data Mining in Search Engine Optimization," *International Journal of Computer Trends and Technology*, vol. 12, no. 4, pp. 201-

205, 2014. DOI:10.14445/22312803/IJCTT-V12P140

[5] Attia M., Abdel-Fattah M., and Khedr A., "A Proposed Multi Criteria Indexing and Ranking Model for Documents and Web Pages on Large Scale Data," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 10, pp. 8702-8715, 2022. https://doi.org/10.1016/j.jksuci.2021.10.009

[6] Banaei H. and Honarvar A., "Web Page Rank Estimation in Search Engine Based on SEO Parameters Using Machine Learning Techniques," *International Journal of Computer Science and Network Security*, vol. 17, no. 5, pp. 95-100, 2017. https://www.researchgate.net/publication/317543658

[7] Dalvi A. and Saraf R., "Inspecting Engineering College Websites for Effective Search Engine Optimization," *in Proceedings of the International Conference on Nascent Technologies in Engineering*, Navi Mumbai, pp. 1-5, 2019, DOI:10.1109/ICNTE44896.2019.8945823

[8] Dick M., "Search Engine Optimisation in UK News Production," *Journalism Practice*, vol. 5, no. 4, pp. 462-477, 2011, https://doi.org/10.1080/17512786.2010.551020

[9] Drivas I., Sakas D., Giannakopoulos G., and Kyriaki-Manessi D., "Big Data Analytics for Search Engine Optimization," *Big Data and Cognitive Computing*, vol. 4, no. 2, pp. 1-22, 2020. https://doi.org/10.3390/bdcc4020005

[10] Giannakoulopoulos A., Konstantinou N., Koutsompolis D., Pergantis M., and Varlamis I., "Academic Excellence, Website Quality, SEO Performance: Is there a Correlation?," *Future Internet*, vol. 11, no. 11, pp. 1-25, 2019. DOI:10.3390/fi11110242

[11] Giomelakis D., Karypidou C., and Veglis A., "SEO Inside Newsrooms: Reports from the Field," *Future Internet*, vol. 11, no. 12, pp. 1-15, 2019. https://doi.org/10.3390/fi11120261

[12] Giomelakis D. and Veglis A., "Investigating Search Engine Optimization Factors in Media Websites: The Case of Greece," *Digital Journalism*, vol. 4, no. 3, pp. 379-400, 2016. https://doi.org/10.1080/21670811.2015.1046992

[13] Halibas A., Cherian A., Pillai I., Reazol L., Delvo E., and Sumondong G., "Web Ranking of Higher Education Institutions: An SEO Analysis," *in Proceedings of the International Conference on Computation, Automation and Knowledge Management*, Dubai, pp. 411-415, 2020. DOI:10.1109/ICCAKM46823.2020.9051481

[14] Jayaraman S., Ramachandran M., Patan R., Daneshmand M., and Gandomi A., "Fuzzy Deep Neural Learning Based on Goodman and Kruskal's Gamma for Search Engine Optimization," *IEEE Transactions on Big Data*, vol. 8, no. 1, pp 268-277, 2022. DOI:10.1109/TBDATA.2020.2963982

[15] Joglekar B., Bhatia R., Jayaprakash S., Raina K., and Mulchandani S., "Search Engine Optimization Using Unsupervised Learning," *in Proceedings of the 5th International Conference on Computing, Communication, Control and Automation*, Pune, pp. 1-5, 2019, DOI:10.1109/ICCUBEA47591.2019.9129011

[16] Karyotakis M., Lamprou E., Kiourexidou M., and Antonopoulos N., "SEO Practices: A Study about the Way News Websites Allow the Users to Comment on their News Articles," *Future Internet*, vol. 11, no. 9, pp. 1-13, 2019. https://doi.org/10.3390/fi11090188

[17] Manohar E. and Punithavathani D., "Effective Preprocessing and Knowledge Discovery in Web Usage Mining," *Middle-East Journal of Scientific Research*, vol. 23, no. 10, pp. 2433-2439, 2015. DOI: 10.5829/idosi.mejsr.2015.23.10.22480

[18] Matošević G., Dobša J., and Mladenić D., "Using Machine Learning for Web Page Classification in Search Engine Optimization," *Future Internet*, vol. 13, no. 1, pp. 1-20, 2021. https://doi.org/10.3390/fi13010009

[19] Özkan B., Özceylan E., Kabak M., and Dağdeviren M., "Evaluating the Websites of Academic Departments through SEO Criteria: A Hesitant Fuzzy Linguistic MCDM Approach," *Artificial Intelligence Review*, vol. 53, no. 2. pp. 875-905, 2020. https://doi.org/10.1007/s10462-019-09681-z

[20] Pan B., "The Power of Search Engine Ranking for Tourist Destinations," *Tourism Management*, vol. 47, pp. 79-87, 2015. https://doi.org/10.1016/j.tourman.2014.08.015

[21] Pant P., Joshi P., and Joshi S., "A Comparative Study of Search Engines Results Using Data Mining and Statistical Analysis," *International Journal of Statistics and Applied Mathematics*, vol. 5, no. 5, pp. 30-33, 2020. https://www.mathsjournal.com/pdf/2020/vol5issue5/PartA/5-4-20-929.pdf

[22] Portier W., Li Y., and Kouassi B., "Improving Search Engine Ranking Prediction Based on a New Feature Engineering Tool," *in Proceedings of the 4th International Conference on Vision, Image and Signal Processing*, Bangkok, pp. 1-6, 2020. https://doi.org/10.1145/3448823.3448878

[23] Portier W., Li Y., and Kouassi B., "Feature Selection Using Machine Learning Techniques Based on Search Engine Parameters," *in Proceedings of the 3rd International Conference on Signal Processing and Machine Learning*, Beijing, pp. 28-34, 2020. DOI:10.1145/3432291.3432308

[24] Prawira I. and Rizkiansyah M., "Search Engine Optimization in News Production Online

Marketing Practice in Indonesia Online News Media," *Pertanika Journal of Social Sciences and Humanities*, vol. 26, no. T, pp. 263-270, 2018. http://www.pertanika.upm.edu.my/pjtas/browse/regular-issue?article=JSSH-T0727-2018

[25] Roslina A. and Nur Shahirah M., "Implementing White Hat Search Engine Technique in E-Business Website," *in Proceedings of the 10th International Conference on E-Education, E-Business, E-Management and E-Learning*, Tokyo, pp. 311-314, 2019, https://doi.org/10.1145/3306500.3306533

[26] Salminen J., Corporan J., Marttila R., Salenius T., and Jansen B., "Using Machine Learning to Predict Ranking of Webpages in the Gift Industry: Factors for Search-Engine Optimization," *in Proceedings of the 9th International Conference on Information Systems and* Technologies, Cairo, pp. 1-8, 2019. DOI:10.1145/3361570.3361578

[27] Schilhan L., Kaier C., and Lackner K., "Increasing Visibility and Discoverability of Scholarly Publications with Academic Search Engine Optimization," *Insights*, vol. 34, pp. 1-16, 2021. DOI: 10.1629/uksg.534

[28] Shahzad A., Nawi N., Sutoyo E., Naeem M., "Search Engine Optimization Techniques for Malaysian University Websites: A Comparative Analysis on Google and Bing Search Engine," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 8, no. 4, pp. 1262-269, 2018. DOI:10.18517/ijaseit.8.4.5032

[29] Sharma P. and Yadav D., "A Novel Architecture for Search Engine using Domain Based Web Log Data," *The Internatonal Arab Juornal of Information Technology*, vol. 20, no. 1, pp. 92-101, 2023. https://doi.org/10.34028/iajit/20/1/10

[30] StatCounter Global Stats, "Search Engine Market Share Worldwide," https://gs.statcounter.com/search-engine-market-share, Last Visited, 2023.

[31] Strzelecki A., "Google Web and Image Search Visibility Data for Online Store," *Data Descriptor*, vol. 4, no. 3, pp. 1-10, 2019. https://doi.org/10.3390/data4030125

[32] Su A., Hu Y., Kuzmanovic A., and Koh C., "How to Improve your Search Engine Ranking," *ACM Transactions on the Web*, vol. 8, no. 2, pp. 1-25, 2014. https://doi.org/10.1145/2579990

[33] Sujatha P. and Kavitha K., "Proficient Data Mining Approach for Search Engine Optimization," *Journal on Science Engineering and Technology*, vol. 2, no. 3, pp. 190-194, 2015. http://jset.sasapublications.com/wp-content/uploads/2017/10/6702647.pdf

[34] Tsuei H., Tsai W., Pan F., and Tzeng G., "Improving Search Engine Optimization (SEO) by Using Hybrid Modified MCDM Models," *Artificial Intelligence Review*, vol. 53, no. 1, pp. 1-16, 2020. https://doi.org/10.1007/s10462-018-9644-0

[35] Ullah A., Nawi N., Sutoyo E., Shazad A., Khan S., and Aamir M., "Search Engine Optimization Algorithms for Page Ranking: Comparative Study," *International Journal of Integrated Engineering*, vol. 10, no. 6, pp. 19-25, 2018. DOI:10.30880/ijie.2018.10.06.003

[36] Vállez M. and Ventura A., "Analysis of the SEO Visibility of University Libraries and How they Impact the Web Visibility of their Universities," *Journal of Academic Librarianship*, vol. 46, no. 4, pp. 102171, 2020. https://doi.org/10.1016/j.acalib.2020.102171

[37] Vyas C., "Evaluating State Tourism Websites Using Search Engine Optimization Tools," *Tourism Management*, vol. 73, pp. 64-70, 2019. https://doi.org/10.1016/j.tourman.2019.01.019.

[38] Ziakis C. and Vlachopoulou M., "Web Content Management Systems Used by Search Engine Optimization Experts for Top Rankings in Search Engine Result Pages," *Wseas Transactions on Computers*, vol. 20, pp. 207-216, 2021. DOI:10.37394/23205.2021.20.22

[39] Ziakis C., Vlachopoulou M., Kyrkoudis T., and Karagkiozidou M., "Important Factors for Improving Google Search Rank," *Future Internet*, vol. 11, no. 2, pp. 1-12, 2019. DOI:10.3390/fi11020032.

Mohamed Almadhoun is a Lecturer at the University College of Applied Sciences, Palestine, holding a bachelor in Computer Engineering and Master of Information Technology, and currently pursuing PhD in computer sciences at Universiti Sains Malaysia. His current research interests include machine learning, data mining, search engine optimization, and image processing. He has experience as a web developer, analyst, and project manager, he was the head of the computer center, and assistant vice rector for administrative affairs at UCAS.

Nurul Malim received her PhD in 2011 from The University of Sheffield, United Kingdom. She is currently an Associate Professor at the School of Computer Sciences, Universiti Sains Malaysia, Malaysia. Her current research interests include Big Data Analytics, Machine/Deep Learning, Data Mining, Sentiment Analysis and Chem/Bio/Neuro-Informatics.