Artificial Intelligence Review (2024) 57:144
https://doi.org/10.1007/s10462-024-10780-9

Exploring the impact of SEO-based ranking factors for voice queries through machine learning

Zafar Saeed[1] · Fozia Aslam[2] · Adnan Ghafoor[2] · Muhammad Umair[2] · Imran Razzak[3]

Accepted: 24 April 2024 / Published online: 16 May 2024

Abstract
The use of voice search is proliferating and expected to grow into the foreseeable future; this is why websites increasingly optimize their content associated with voice-based search to improve their ranking. In this era of rapid growth in voice search technology, it is a topical matter that needs research. Moreover, many predictions about its future excite the subject and require systematic investigation. This research aims to analyze important features that contribute to the SEO of webpages. Therefore, there is a need to examine various ranking factors that improve the ranking of the webpages for voice search queries on the Search Engine Results Page (SERP). This study consists of two phases. The first phase comprises systematic data acquisition and identifying important SEO-based ranking factors. The second phase includes a longitudinal case study to evaluate the impact and significance of identified factors. To achieve this goal, we conduct experiments on methodical combinations of features through machine learning algorithms such as Support Vector Machine, Logistic Regression, Naive Bayes Classifier, K-Nearest Neighbors, Decision Trees and Random Forest. Comparing results for multiple feature designs evaluates the contributing nature of specific features in SEO-based optimization for ranking. Results suggest the importance of the newly identified feature set (FF) outperforms baselines (EF and EFN) by a significant margin. A longitudinal case study on a blog over four months confirms that optimizing these features improves page ranking; therefore, webmasters must optimize these features while preparing the webpage.

Keywords Search Engine · Optimization · Voice Search · Voice Queries · Page Ranking · Information Retrieval

Extended author information available on the last page of the article

1 Introduction

Search engines help their users find the appropriate and most relevant information they are looking for by retrieving a ranked list of results that matches the query. Once results are displayed in SERP, users decide whether to click on the results presented on the first page or need to explore more results on later pages. There are many search engines, but among all search engines, Google[1] is the top search engine with two main competitors Yahoo[2] and Microsoft's Bing.[3] Search engines work in three phases: crawling, indexing, and ranking. Crawlers are robots, also called bots or spiders. They continuously search for new or updated content on the internet. The new content can be a document, webpage, image, or video. Once the Google crawler discovers new webpages, it analyzes what each page is about (e.g., content, images, or videos). The latest content is processed and indexed using Google Caffeine,[4] to retrieve it later when a user generates a query. This process is called indexing. The last phase of the information retrieval process is ranking, which happens after the information is found using the Google index against a user's query. The search algorithm considers numerous ranking factors to determine the order of search results. Ranking organizes the results based on relevancy, webpage quality, and authority as well. However, higher rankings depend on Google's algorithm functionality.[5] Search Engine Optimization (SEO) is a process to improve website performance to achieve high visibility in search engines. High visibility or high ranking leads to more visitors' attention, improving the quantity of website traffic. Unlike Google's paid ads, you cannot pay Google for a higher rank to get organic search traffic. SEO is further divided into two categories: on-page SEO and off-page SEO. On-page SEO includes all practices that ensure webpage optimization for users and search engines. Common on-page SEO factors include meta-tag optimization, content uniqueness, internal linking, keywords and image optimization, URL optimization, mobile-friendliness, and code minification (Patil and Patil 2018b). Off-page SEO plays a vital role in a successful ranking in the Google SERP and higher traffic. It involves a holistic approach to building your website's reputation, authority, and online presence through inbound links. For instance, if a reputable industry blog links to your site as a reference, it boosts your site's credibility and increases the chances of web users landing on your webpage. Standard practices for off-page SEO include acquiring backlinks from authoritative websites to boost page credibility via blog posts, social media engagement, online reviews, content marketing, brand mentioning, making classified ads, forum posts, sharing images, infographics, and videos related to the page.

1.1 Motivation and research Gap

The Google ranking algorithm serves as a pivotal mechanism for the assessment and ranking of websites. In pursuit of maintaining the integrity of its search results, Google

[1] www.google.com
[2] www.yahoo.com
[3] www.bing.com
[4] https://web.archive.org/web/20230904051047/https://googleblog.blogspot.com/2010/06/our-new-search-index-caffeine.html Access Date: 07-02-2024.
[5] https://web.archive.org/web/20231003111956/https://developers.google.com/search/docs/fundamentals/how-search-works Access Date: 07-02-2024.

consistently improves the ranking algorithm,[6] notable among them being 'Penguin,' 'Panda,' and 'Hummingbird'. These updates are designed to identify and penalize websites that fail to align with Google's prescribed search optimization guidelines. Google's algorithm uses over 200 factors to rank webpages. These factors have been intentionally kept confidential; however, content quality, keyword density, and engagement on social media platforms are an open secret (Khan and Mahmood 2018). Due to the undisclosed ranking algorithm, the existing research studies propose SEO-related important factors through reverse engineering; hence, they quickly become outdated due to the continuous change in Google's ranking algorithm. Formulating an adaptable framework is necessary to identify significant factors and comprehend their influence on webpage ranking. After reviewing the available literature on search engine optimization trends, SEO factors, and analysis, it has been observed that the previous studies have several limitations. First, they did not introduce the crucial factors related to speed-related factors (mobile and desktop speed), h1 length, content word count, domain authority (DA), page authority (PA), image links, and website categories. Thus, the need arises to collect data about these factors and investigate the impact of the existing and newly collected factors through different classification methods of machine learning, identifying the contributing factors that can help webmasters achieve higher rankings. Second, the existing literature concerns the significant ranking factors of the search engine, which collects the dataset of webpages through different text queries. Advancements in technology have enabled users to search and browse online content through their hand-held and voice-activated devices such as mobile phones. It is easier for such users to generate voice queries; hence, the SEO factors contributing to the voice-generated queries are an important area that requires investigation. So, there is a need to collect the webpages through voice search queries and analyze the factors that influence Google to select the top-ranked pages. Therefore, there is a need to explore the impact of voice search on SEO by collecting the voice search-related factors such as snippet type, question words in page title/voice search-related terms, structured data: markup size, schema markup usage, and question words in URL/voice search-related term.

1.1.1 Research questions

The existing research gap under the motivational scenarios leads to the following Research Questions addressed in this research study:

RQ1: Can a framework be designed to express an adaptive process to acquire voice query-based data and evaluate the impact of SEO-related factors?
RQ2: What are the top factor(s) involved in the SEO process?
RQ3: Can optimizing significant factors identified through machine learning improve webpage ranking?

[6] https://web.archive.org/web/20230906123723/https://developers.google.com/search/updates/core-updates Access Date: 07-02-2024.

1.1.2 Research contributions

This research aims to analyze the factors that influence the ranking of webpages over voice search queries. The study results would help webmasters understand the factors they must consider to get a higher search rank. The Key contributions in the underlying studies are as follows:

- We devised a systematic design to formulate voice-based search queries.
- We acquire data and develop a benchmark based on voice-based search queries for evaluation.
- We identify on-page SEO-related factors that have not been investigated for voice search queries.
- We perform feature engineering and identify significant factors contributing to the page ranking.
- We conduct a longitudinal case study on a blog to evaluate the ranking impact when optimizing SEO-related factors identified in this research.

The rest of this work is organized as follows: Sect. 2 describes the background knowledge, then categorically discusses the recent trends and use of machine learning in SEO. Section 3 discusses the data acquisition and feature characterization in detail. Section 4 discusses the framework and methodology. Section 5 discusses the experimental setup, evaluation measures, and feature designs. Section 6 discusses the results and significance of the proposed features. Section 7 discusses a longitudinal study of a Blog and verifies the impact of identified SEO factors. The study concludes with Sect. 8 and describes possible future directions in Sect. 9.

2 Related work

The following section briefly describes the background of SEO, its trends, usage of machine learning, ranking algorithms, and discusses the previous research in the preceding literature to analyze the major ranking factors of the Google search engine.

2.1 Search engine optimization (SEO)

In order to get a better rank in search engines, a set of techniques called search engine optimization (SEO) is applied (Sharma and Verma 2020). According to Lemos and Joshi (2017), the overall SEO process comprises six different phases that webmasters should carry out for effective optimization of the webpages (see Fig. 1). These phases include keyword research, goal setting, content building, webpage optimization, link building, and creating progress reports. In (Patil and Patil 2018b, a), the authors researched the techniques to provide users with reliable and ranked results. Generally, if a webpage's rank is higher, it will be visited by a large number of users. The primary purpose of SEO is to improve the quality and quantity of website traffic and ensure that the webpages appeal to search engines. Patel and Atkotiya (2020) divided SEO ranking factors into two categories: on-page and off-page optimization. On-page factors include Meta tags,

Fig. 1 SEO Process

Heading tags, Sitemaps, and Robot files. Off-page optimization comprises link building, social sharing, comments, business listings, and blogging. Roslina and Shahirah (2019) performed testing by creating a website using cPanel and WordPress after implementing SEO. Their results show an improvement in the keyword ranking on Google's first page. Vyas (2019) evaluated tourism websites using search engine optimization tools. They use seven SEO tools to analyze the six tourism websites: traffic estimate, Twitter search, Google trends, Alexa, similar websites, SEMrush, SEO analyzer, and Moz. They use all these tools for different purposes, such as searcher preference, judging keywords, global rank, organic search traffic, domain authority, page authority, and how a website ranks in SERP. Mittal and Sridaran (2019) shows multiple factors that need to be considered to improve the site performance, e.g. page size, page requests, page speed, browser caching, page redirects, compression, render blocking, responsive design, website viewport, page title, meta description, headings, sitemap and SSL certification.

2.2 Voice queries Vs text queries

Generally, voice queries differ from how people make regular queries, i.e., searching through text. Regarding text queries, users choose keywords to translate their information needs and browse the results to find relevant information. However, when a user performs a voice search, they expect a direct answer from the search engine that best describes their information needs (Strzelecki and Rutecka 2020a). The language of voice queries is closer to natural language than text queries. The research by (Guy 2016) found that the average query length for voice search is significantly longer. Voice queries mainly include question words and natural language such as what is, in the, show me, I am looking for, etc. For instance "Looking for a restaurant that serves oysters in San Francisco". Meanwhile, the text queries, such as "oyster's restaurant sf", are very specific. Moreover, they concluded that only 13.1% of the voice queries were identical to the text samples.

2.3 Trends in SEO

Moreover, SEO is an ongoing process, and Google engineers continuously update their algorithms for the best user experience. The major and core updates can be seen on Google's official forum.[7] On the other hand, the growth of voice-enabled devices and voice search is also undeniable. The number of devices with virtual assistants is also on the rise. Voice search is simply a matter of speaking the search query. Voice search is a concept that started with mobile phones but quickly spread to smart speaker devices, in the car, and televisions (Strzelecki and Rutecka 2020b). Regarding voice search on mobile, user experience is considered the most essential ranking factor. And Google prefers mobile-friendly websites that must meet the requirements of mobile users. In 2016, for the first time, it was analyzed that mobile traffic surpassed desktop traffic, accounting for 51.3%. In 2019, mobile devices claimed the top position among devices browsing networks, comprising 51.6% of the total. It shows a prominent increase in mobile phone users, leading to a huge number of voice search queries. Therefore, there is a need for information retrieval researchers to understand this new medium of search and its differences from traditional text searches. With this, there is also a need to make a web structure mobile-friendly and optimized for voice search.

The proliferation of the use of voice search has quickly gained attention over the last decades and is poised to continue through 2021. We can't ignore its rapid growth and adoption during the COVID-19 pandemic. In July 2019, Adobe[8] released a survey that shows that 48% of users use voice search for general web queries. Along with voice search, there are also other trends to consider that will affect SEO. These trends include artificial intelligence in future SEO, the effect of voice search on search queries, quality content, featured snippets, image optimization, videos, and local listings.[9] The usage of featured snippets cannot be ignored. Featured snippets represent the most recent and popular type of snippet. Search engines extract pieces of information from webpages to display them in the form of a box alongside organic results. A featured snippet is also known as the direct answer or Google answer box (Strzelecki and Rutecka 2020a). The studies (Strzelecki and Rutecka 2019, 2020b) show that snippets appear in several different formats, e.g. paragraph, list, table, ordered and unordered list. Snippets taxonomy shows that most of the snippets in Google SERP are the results of long-tail keywords. These are often 3-5 words long and are likely to be used during searches. Long-tail keywords are mostly informational keywords that are used to find specific information, e.g., How far is California from San Jose? Long Tail keywords are combinations of words that represent a very long and specific search query (Jerkovic and Warrior 2009; Anuradha et al. 2021).

[7] https://web.archive.org/web/20230904124839/https://moz.com/google-algorithm-change Accessed Date: 04-09-2023.

[8] https://web.archive.org/web/20231006164214/https://blog.adobe.com/en/publish/2019/07/22/voice-assistant-statistics-trends-2019, official: https://blog.adobe.com/en/publish/2019/07/22/voice-assistant-statistics-trends-2019 Access Date: 07-02-2024.

[9] https://web.archive.org/web/20230523163625/https://www.semrush.com/blog/seo-trends/ Access Date: 07-02-2024.

2.4 Machine learning in SEO

The machine learning approach helps search engines understand the page ranking criteria. Nowadays, most search engines like Google, Yahoo, Bing, Ask, and many more use machine learning techniques for webpage classification and webpage ranking. To sort the search results, Google uses the "RankBrain" algorithm. RankBrain is a machine learning algorithm that helps Google understand and process search queries. It is considered the third most essential algorithm that understands what users are asking and allows them to provide results by bringing up similar results in response to the search query. The first primary task of RankBrain is to understand the search queries, then measure the user experience and anticipate how users interact with the web results (Sunny 2020). Google's current figure confirms approximately 4.2 billion active webpages (Hingoro and Nawaz 2021). PageRank (PR) plays a crucial role in ranking the most relevant results in response to search queries. PR is a ranking algorithm that evolves a set of rules. The PR algorithm's working depends on the link structure of the webpage. At the same time, the rank score of the webpage given by PR is based on the backlinks of the webpage. A webpage linked by many high authority pages receives a high PR score (Selvan et al. 2012), and the range of PR is from (0 to 10) (Jadav and Shrivastava 2021). However, Google claims to use more than 200 parameters in its ranking algorithm (Khan and Mahmood 2018; Su et al. 2014; Hingoro and Nawaz 2021). The ranking algorithm employs various factors. While most of these factors are published in Google guidelines,[10] the exact role in ranking and the method to attain page ranking have not been precisely specified (Matošević et al. 2021). These ranking factors are kept confidential by Google (Khan and Mahmood 2018), and only some of them are revealed, e.g. website popularity, the density of keywords, quality of content, page speed and website security. Numerous efforts have been made to reveal the important ranking factors of the search engine. The study conducted by Su et al. (2014), aims to determine the important ranking factors influencing the search engine to rank the webpages. To conduct their study, they collected the top 20 pages against 60 keywords and analyzed 17 ranking features for each webpage. They use the SVM-rank implementation with linear and polynomial kernels to train the ranking functions. The dominant ranking factors they revealed are page rank, keyword in the title tag, keyword in the meta description, keyword in the hostname and keyword in the URL's path. A similar study has been conducted to investigate the contributing factors that help to increase organic traffic through various SEO factors. The authors analyzed 171 cultural heritage websites along with their keyword ranking performance and user experience. The five focused factors that they explored in their research are website size, loading speed (LD), SEO crawling, website security (HTTPs), organic traffic (TF) and user behavior. They developed a diagnostic exploratory model derived from linear regression to analyze the cause-and-effect relationship between the five factors. This statistical approach was developed to understand the impact of each factor on the organic traffic change. Their analysis concluded that SEO crawling, website security, website size, and user behavior significantly impact the increase in organic search traffic. Among all these influencing factors, user behavior seems to have the highest impact on increasing organic search traffic. A detailed literature analysis was conducted by (Ziakis et al. 2019) on SEO factors that influence rankings on SERP. They recorded the features of 24

[10] https://web.archive.org/web/20230904051047/https://googleblog.blogspot.com/2010/06/our-new-search-index-caffeine.html Access Date: 07-02-2024.

websites, with the most significant factors involving on-page and off-page SEO factors. They gathered on-page SEO factors from previous research. They conducted a statistical analysis on each factor through the Spearman correlation coefficient and concluded that the most significant factors were the quantity and quality of backlinks, the bounce rate, and the SSL certificate. Moreover, their study also confirmed that the website's loading time (LD), URL length (ULC), and use of target keywords in the title (KWT) do not participate in the ranking of the website. Moreover, the study (Matošević et al. 2021) shows that machine learning can be used to classify the sample webpages and detect those that need improvements to comply with SEO guidelines. They extracted the target on-page factors through expert knowledge and machine learning. The importance of features was estimated through different statistical methods, including correlation, information gain, chi-square, relief, and random forest. The classification methods used in their research to evaluate the results were decision tree, naive Bayes classifier, k-nearest neighbor, support vector machine, and logistic regression. The results of this study concluded that the important factors that webmasters must consider while preparing a webpage are keywords in the meta title, keywords in the meta description, keywords in the heading tags and keywords in the webpage's body.

Fig. 2 Effected Site (mediaite.com)

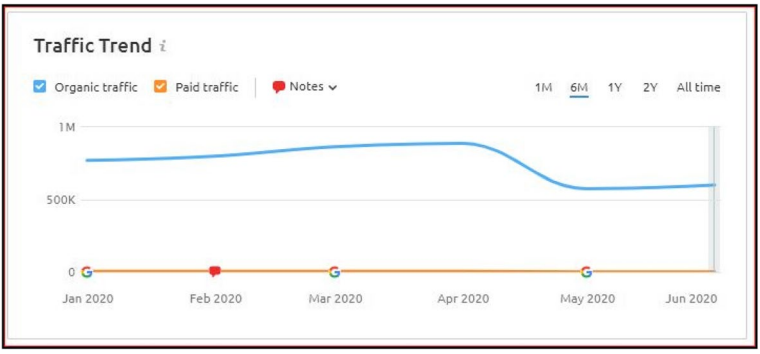

Fig. 3 Effected Site (rfi.fr)

Table 1 Technology Domains

Sr.	Domain	Sr.	Domain	Sr.	Domain	Sr.	Domain
1	Technology	5	Virtual Reality	9	Digital Marketing	13	Artificial Intelligence
2	Social media	6	Cloud Computing	10	Cyber Security	14	Robotics
3	IT Security	7	Block chain	11	laptop	15	Machine learning
4	Video Games	8	Web and Internet	12	Mobile	16	IT

3 Dataset collection & characterization

We collected the latest dataset for voice search analysis since we only found a dataset from one preceding literature related to the information about featured snippets (Strzelecki and Rutecka 2020a), which couldn't help us identify the SEO trend. Moreover, Google started rolling out the May 2020 core update (04/05/2020), and it was the second hottest update after the August 2018 medic update. The recent May core update was significant, and many websites were affected by Google SERPs. Therefore, we are motivated to collect the latest dataset and analyze its ranking factors. Each time Google updates, its algorithm improves the user experience and shows the most relevant results to the searcher. Below are images taken from the SEMRUSH tool that show the down traffic of affected sites after the Google May 2020 core update. The x-axis in Figs. 2 and 3 shows the months, while the numbers in the y-axis show the amount of traffic, and the line trend shows the prominent drop in traffic after the May 2020 core update.

3.1 Voice query generation

Existing literature (Strzelecki and Rutecka 2020a) confirms that voice queries have a long tail structure. Voice search queries are more extended than text search queries and involve question words, e.g. (What, When, How, Why, Where, Did, Who, etc.). The first step of this research is the formation of voice search queries. The primary list of keywords associated with each of the 16 domains is collected from search engine suggestions and auto-completion tools. Later, we used AnswerThePublic[11] and keyword.io[12] to formulate extended search queries with the seed keyword. These tools generate long-tail queries relevant to our target seed keywords and retain a structure similar to the voice search query. The list of search queries comprises common questions and keywords selected from the well-known areas under the technology domain (see Table 1). The long-tail queries were transformed/synthesized from text-to-speech to simulate the procedure of asking questions for voice search queries. We added "Hey Google" command with each query, fed the queries to Google Text-to-Speech, and then converted the output into an audio file. In our setup, Google Assistant took 5–10s to process each query. Therefore, we modified the audio file using Audacity's audio editor and added a 10-seconds silent buffer after each query, ensuring that each query audio was correctly aligned with the Google Assistant. By

[11] https://web.archive.org/web/20240207095709/https://answerthepublic.com/ Access Date: 07-02-2024.
[12] https://web.archive.org/web/20240206093920/https://www.keyword.io/ Access Date: 07-02-2024.

the end of this process, a list of voice search queries was finalized and ready to be used for the data acquisition.

3.2 Voice search and data acquisition

The process of asking queries and acquiring data involves the following steps.

3.2.1 Device setup for voice search

In this research, the device that we use for asking questions to Google Assistant is "Google Home Mini," a smart speaker that is powered by Google Assistant. After the Google Home Mini device's setup, we automate the process by feeding audio files (voice queries) into the Android phone.

The audio file is then played on the Android phone, which programmatically asks all the queries one by one to Google Home Mini. We extract the search results answered by Google Assistant and URLs of webpages displayed against the voice search queries in Google SERPs. The whole process is executed by logging through Google My Activity.[13]

3.2.2 Dataset

The dataset of URLs is collected through the extraction bookmarklet code. After collecting URLs, we formulated the dataset into a CSV structure containing feature estimation for each item (page) in the dataset along with the query. For each query, we collected 20 URLs segregated into the top 1-10 and bottom 91-100 webpages. A total of 80 webpages were dropped due to Google Assistant's inability to interpret some of the voice queries correctly. Finally, the dataset consists of 2960 webpages, extracting 31 features for each webpage. To collect this dataset, approximately the top 50% of queries were selected. We asked 148 voice search queries to Google Assistant to build a benchmark dataset. The quality of data is essential when training a machine learning model. Therefore, we calculated the percentage of outliers in the data to understand the data quality. We use Interquartile Range (IQR) as a statistical measure to assess the outliers in the dataset. It is defined as the difference between the third quartile (Q3) and the first quartile, i.e. $IQR = Q3 - Q1$. Where Q1 and Q3 are the first (25^{th} percentile) and third (75^{th} percentile) values in the dataset, respectively, sorted in ascending order. Lower Bound (LB) and Upper Bound (UB) Outlier values are defined as $Outlier_{LB} = Q1 - 1.5 \times IQR$ and $Outlier_{UB} = Q3 + 1.5 \times IQR$, respectively. Estimating outliers was not straightforward. We split the data into two parts; top-10 is separated from the bottom-10 ranked pages. Intuitively, the feature values would have a visible gap between these two groups, resulting in greater outliers. However, these data points are equally important for the learning process of classifiers. Therefore, removing outliers from the data may lead to incorrect conclusions in this particular scenario. Outliers from each group are estimated separately. The complete data with 31 features has a total of 4.688% outlier values. However, in the final/filtered feature set, the outliers have reduced to 2.67%. Moreover, Meta Description Length in Characters (MDLC), Traffic, and backlinks have the highest outlier values 12.11%, 9.73%, and 7.87%, respectively. The outliers values in these features are comprehensible because several webpages optimize MDLC with extensive descriptions for including relevant keywords. On the other hand, top-ranked webpages

[13] https://web.archive.org/web/20240207001614/https://myactivity.google.com/ Access Date: 07-02-2024.

Table 2 Statistical detail of data acquisition

S.No	Item	Value
1	No. of voice queries	300
2	Total pages collected	3040
3	Webpages dropped	80
4	Total features per page	31
5	Outlier percentage in dataset	4.68%
6	Outlier percentage in the filtered feature set	2.67%
7	Top features with highest outlier ratio	MDLC (12.11%), Traffic (9.73%), and Backlinks (7.87%)
8	Missing values in the dataset	None

have massive traffic; therefore, the outliers are relatively high in these features. Similarly, backlinks were once considered an essential factor for SEO; therefore, some webpages use backlinks in huge numbers. However, our feature selection methodology proved that backlinks are not among the top influential factors for improving page ranking. Further details about the dataset are shown in the Table 2 and publicly available on GitHub repository.[14]

3.3 Features collection

The features were collected using a Python script where we implemented the BeautifulSoup library and used SEO tools, including Alexa, Google page speed insight tool, Moz, ahref, Semrush and Netspeak Spider. While collecting the features, we selected the top-10 organic results from Google SERP and the results from the bottom page. For feature analysis, we collected the ranking position of each page in the Google SERP. Table 3 shows the list of features for each webpage we contained against the queries to conduct our research. The existing features (EF) are taken from the literature (Khan and Mahmood 2018; Drivas et al. 2020; Matošević et al. 2021; Su et al. 2014). Additionally, a list of newly analyzed features (NAF) was also collected to investigate their influence on Google ranking.

4 Methodology

To address the research question (RQ1), the following section describes the methodology derived through the framework shown in Fig. 4. We used top keywords related to 16 different domains and used them as seeds to formulate extended/long-tail questions like queries (see Sect. 3.1) through a systematic process. On average, each domain's top 20 questions are voice synthesized and recorded as queries. Later, these voice queries are simulated to ask questions from Google Assistant. The retrieved results through the search engine are segregated according to their ranks. A detailed feature analysis process is conducted to collect features set for the resultant webpages (see Table 3). The feature selection methods

[14] https://github.com/Zafar-Saeed/SEO_Dataset Accessed on Date: 10-10-2023.

Table 3 Complete list of features uses in this research

Feature	Description	Value Type	Status
ST	Snippet Type	Categorical	NAF
HTTP(s)	URL with HTTP or HTTPs	Categorical	EF
TLC	Title Length in Chr.	Continuous	EF
DA	Domain Authority	Continuous	NAF
MDLC	Meta Description Length in Chr.	Continuous	EF
PA	Page Authority	Continuous	NAF
H1LC	H1 Length in Chr	Continuous	EF
OL	Number of Outgoing links	Continuous	EF
ULC	URL length in Chr.	Continuous	EF
IL	Number of internal Links	Continuous	EF
KWT	Keyword count in Title	Continuous	EF
IMGL	Number of images links	Continuous	NAF
KWMD	keyword count in Meta Description	Continuous	EF
Backlinks	Links from other websites	Continuous	EF
KWURL	Keyword count in URL	Continuous	NAF
Traffic	Traffic	Continuous	EF
WCIMD	Word Count in Meta Description	Continuous	EF
AR	Alexa Rank	Continuous	EF
QAT	Question words in Page title	Continuous	NAF
SMR	SEMRUSH Rank	Continuous	NAF
QWURL	Question word count in URL	Continuous	NAF
RTS	Robot.txt Status	Categorical	EF
CWC	Content Word Count	Continuous	NAF
DSX	Domain Suffix	Categorical	NAF
CT	Use of Canonical Tags	Categorical	EF
WC	Website Category	Categorical	NAF
MS	Loading speed on mobile	Continuous	NAF
SDMS	Structured Data: Markup Size	Continuous	NAF
DS	Loading speed Desktop	Continuous	NAF
SMU	Schema Markup Usage	Categorical	NAF
LD	Website fully load time	Continuous	EF

based on Information Gain and SelectKBest (Chi-square as a scoring function) are used to identify the expressive features. An ensemble feature selection approach ranks and filters top features (see Sect. 5.2), later used for experimental design. Two existing feature designs were compared as baselines with four newly analyzed feature designs (see Sect. 5.6). We then used well-known machine learning approaches (SVM, Logistic Regression, Naive Bayes, K-Nearest Neighbours, Decision Tree, and Random Forest) to identify whether the selected feature set contributes to a higher page rank. Finally, to cross-verify the impact of outperforming feature design, we optimized the identified factors of an online blog over four months and retested the rank results (see Sect. 7). The convincing improvement in page ranking verifies that the identified factors should be prioritized when optimizing for SEO.

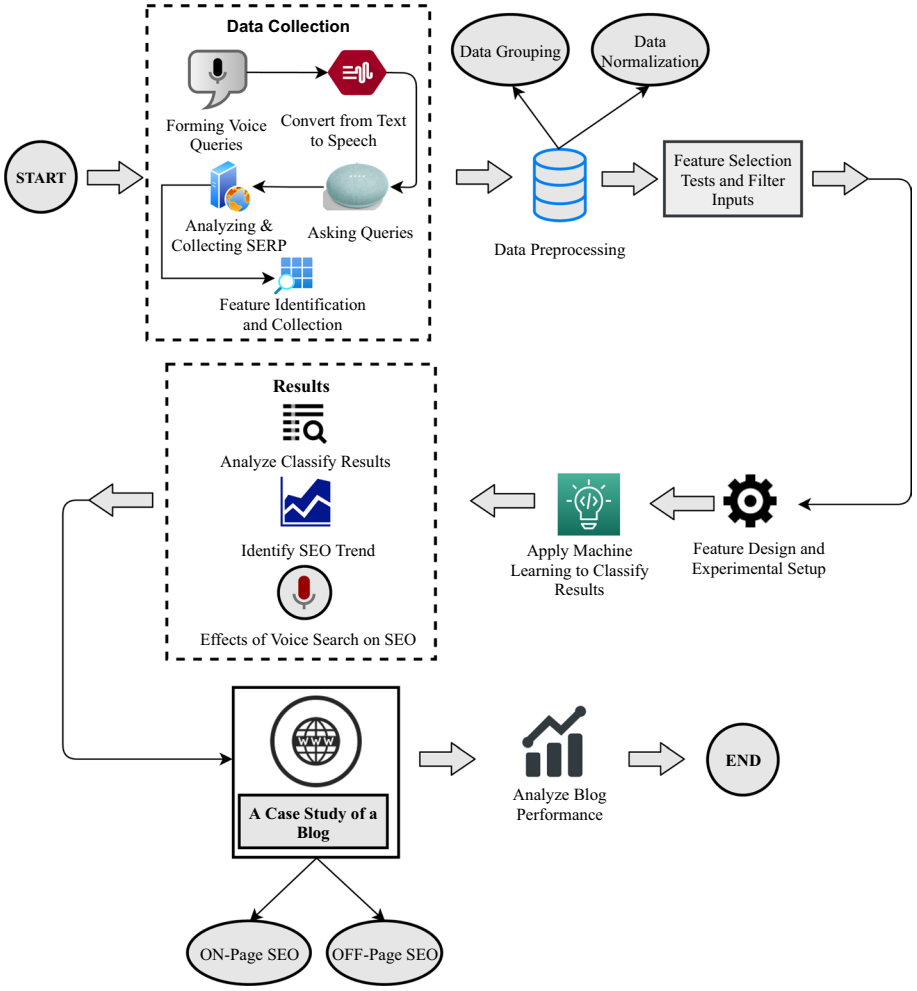

Fig. 4 The proposed Framework for the research methodology including the data acquisition process

5 Experimental setup

This section presents a detailed experimental setup, including various combinations of feature design with data normalization techniques. The section also discusses the results of the proposed research methodology.

5.1 Data preprocessing

Concerning the statistical tests and classification, we employ a 2-steps process in data preprocessing:

1. Data normalization
2. Label Encoding

Before we apply any classification algorithm, preprocessing through normalization was performed on our dataset because the features we collected had different numerical ranges. There is high variance in the dataset, i.e. some of the features such as *Traffic*, *Backlinks*, and *CWC* have high numeric ranges compared to other features such as *WCIMD*, *LD*, and *TLC*. To correct this bias and get better results for the classification, we employed min-max normalization in our dataset. Normalization changes the value of the dataset to a standard scale. Our dataset uses min-max normalization with a scale ranging from 1 to 100.

$$z_i = \frac{(x_i - X_{min})}{(X_{max} - X_{min})} \times (Max - Min) + Min \tag{1}$$

In Eq. 1, z_i is the normalized value in the dataset, x_i is the i^{th} value in the dataset, minimum *Min* and maximum *Max* value range is defined as 1 and 100. respectively, for continuous value type features. Furthermore, label encoding (Cerda and Varoquaux 2020) is used to transform the categorical features into numeric values.

5.2 Feature selection

The feature engineering reveals the importance of various features in the dataset. The two approaches we use in our research are based on information gain (IG) and univariate selection. By measuring the results of these tests, we identify the important features and drop those with low IG. To calculate the IG, we first calculated the entropy of the entire dataset and every single feature. Entropy measures disorder in the data and ranges between 0 and 1. If data is entirely homogeneous concerning the target class, entropy is 0. On the other hand, if data is uniformly divided concerning the target class, the entropy is 1. Information Gain estimates how much information a feature tells about the target class.

$$Entropy = -\sum_{i=1}^{c} p_i \log_2(p_i) \tag{2}$$

Where p_i is the probability of the element i in the data we can select the features in descending order of their score once we calculate the IG. It is measured by subtracting the entropy of particular features from the entropy of the entire dataset by comparing the entropy before and after the data split using a particular feature, as shown in Eq. 2.

$$Information\ Gain = Entropy(Entire\ Dataset) - Entropy\ (Features) \tag{3}$$

Univariate feature selection is also widely used in machine learning. It selects the top features with the most vital relationship with the output variables to better understand the data. We apply a univariate selection to the normalized dataset. SelectKBest (see Algorithm 1) is used to choose the top-K features according to the highest scores. We have a multi-variable classification problem; therefore, the chi-square scoring function is used to rank the continuous (independent) variables according to categorical (dependent) variables.

Algorithm 1 SelectKBest Algorithm for Feature selection and ranking

Input:
- Dataset D consists of feature matrix with n samples, and m features
- Number of top k features to be selected

Output:
- Dataset $\overline{D}$ with top k selected features.
- A list of sorted feature set Γ
- A list of scores Φ corresponding to Γ i.e., $\Phi_i \in \Gamma_i$

function SELECTKBEST(D)
  $\Phi \leftarrow [\,]$ ▷ Initialize an empty list for *scores*
  $\overline{D} \leftarrow [\,][\,]$
  for i from 1 to m **do**
    score $\leftarrow$ Calculate a score for $D[:, i]$ ▷ scoring function (chi-squared)
    $\Phi_i \leftarrow score$ ▷ Append the score to Φ
  end for
  $\Gamma \leftarrow Sort(D, \Phi)$ ▷ Sort the features in descending order of their scores
  for i from 1 to k **do**
    $\overline{D}[:, i] \leftarrow \Gamma_i$
  end for
  return $\overline{D}, \Gamma, \Phi$
end function

The features are sorted from highest to lowest values concerning Information Gain and SelectKBest, as shown in Figs. 5 and 6. We identified these expressive features for their estimated values in the dataset. Moreover, some of them are also recommended in previous studies (Drivas et al. 2020; Matošević et al. 2021; Su et al. 2014). Although we used two different approaches, the feature scores in both analyses are correlated. The features with the highest scores are ST (Snippet type), MS (Mobile speed), MDLC (Meta description length in characters), KWMD (Keyword in Meta description), KWURL (Keyword in URL), DS (Desktop Speed), and KWT (keyword in title). We drop the less important features based on their scores and literature review (Khan and Mahmood 2018). We exclude AR and SEMRUSH Rank from the feature set because the rank evaluated by these two is based on the top-level domain instead of the distinctive webpage (Khan and Mahmood 2018), which means the information about the distinctive pages is not inherent to these features. Moreover, DSX and WC features are also dropped because they are based on nominal values. DSX tells about domain suffixes like.com,.org,.edu, etc. The feature WC gives information about website categories, e.g., blogs, services, eCommerce, etc. However, we plan to use these features in our future work.

5.3 Feature filteration

Feature filtration is based on aggregation. The aggregated method fuses the ranking score from IG and SelectKBest to get a ranked feature set, as shown in Fig. 7 and Table 5. Furthermore, during the experiments, we successively dropped features from the tail-end of the ranked list until the classifiers' performance (accuracy) started degrading. The features

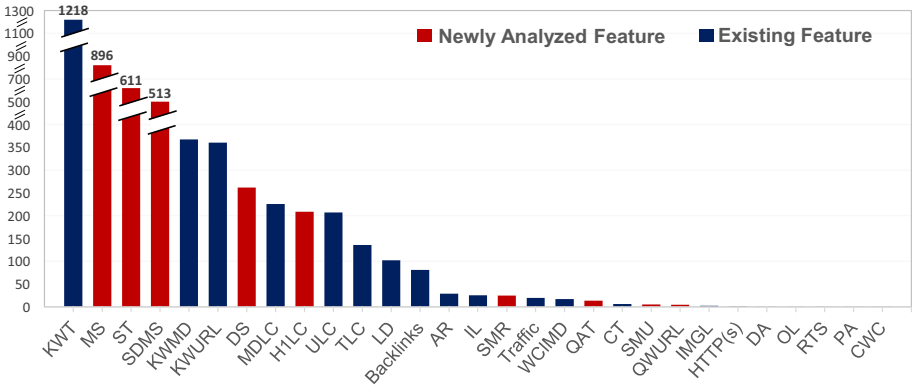

Fig. 5 Ranked list of all features based on SelectKBest algorithm

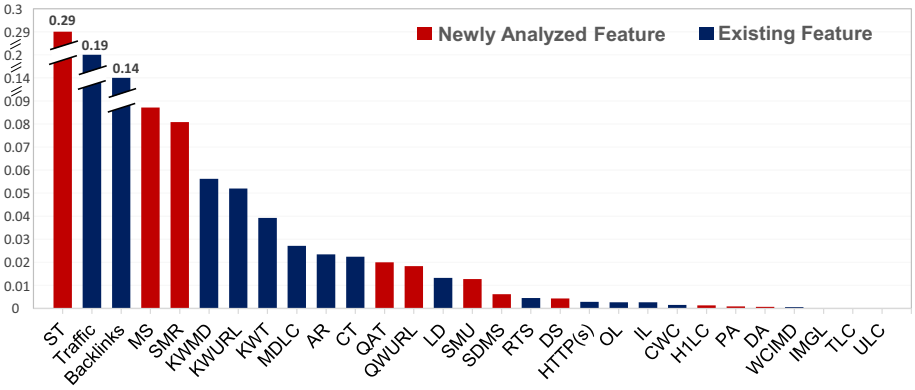

Fig. 6 Ranked list of all features based on Information Gain

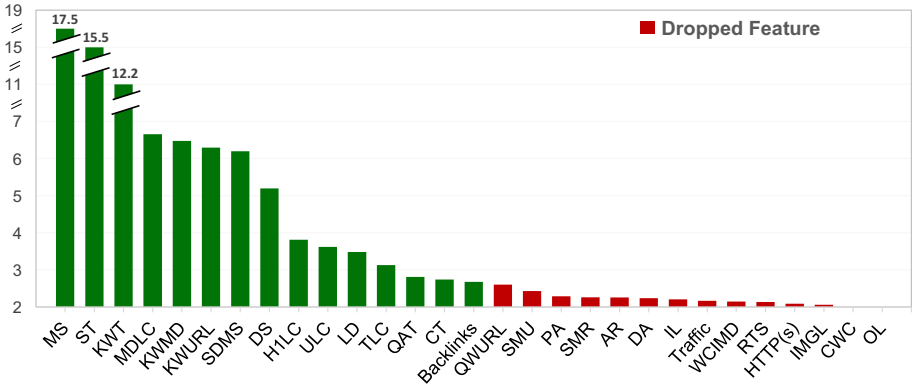

Fig. 7 Feature Filtration using Aggregated Method

Table 4 Machine learning classifiers used for the experimentation

Abbreviation	Algorithm
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
LR	Logistic Regression
DT	Decision Tress
NB	Naive Bayes
RF	Random Forest

we reduced during this process are CWC, OL, IMGL HTTP(s), RTS, DA, PA, WCIMD, SMU, Traffic, IL, and QWURL.

5.4 Machine learning algorithms

In our research, we use the supervised machine learning approaches that are widely used for classification problems. Classification is the process of predicting the target class (dependent variable) based on the given data points (independent variables). In this study, we have used six different classifiers to evaluate the effectiveness of various feature designs by comparing the results based on new and existing feature sets. The list of classifiers used in this study is given in Table 4. Python's Scikit-learn library[15] was used for all the classification algorithms. The k-fold method is used for cross-validation with the value of K = 10 for reliable assessment and avoiding overfitting.

5.5 Evaluation measures

Evaluating the feature design is an essential part of this research. The evaluation is performed by comparing the accuracy of classification methods. A classification model's accuracy (as shown in Eq. 4) is a ratio between the number of correct predictions and all the predictions. It is defined as the ability to correctly identify all true cases and reject all false cases. Finally, the average accuracy for each classifier is estimated to compare and evaluate the performance against each featured design.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}, \tag{4}$$

where *TP* and *TN* are positive and negative instances respectively that are correctly classified. However, *FP* and *FN* are positive and negative instances that are incorrectly classified.

5.6 Feature design

We designed six combinations to evaluate and compare the impact and contribution of the selected feature set. The feature design for the experimental setup is defined as follows:

[15] https://scikit-learn.org/stable/supervised_learning.html#supervised-learning Access Date: 07-02-2024.

- Complete Feature set (CF)
- Complete Feature set Normalized (CFN)
- Filtered Feature set (FF)
- Filtered Feature set Normalized (FFN)

The complete feature set (CF) includes all the newly analyzed features and existing features (Drivas et al. 2020; Matošević et al. 2021; Su et al. 2014). In comparison, the Complete Feature Set Normalized (CFN) is the normalized version of CF. Through the feature selection process, We drop the less important features according to their scores and literature review (Khan and Mahmood 2018) and then perform classification on various feature sets. The list of selected and dropped features is listed in Table 5.

The selected features are marked as "✓", while the dropped features are marked as "✗". As for the FF, the features we exclude from the evaluation are AR (Alexa Rank), CWC (webpage content word count), DSX (Domain suffix), SMR (SEMrush Rank) and WC (Website category) from the newly analyzed features and OL (outgoing link) from the previous features (see Sect. 5.2 for detail). In this research, the following two settings of the existing feature sets (i.e., EF and EFN) collected through the literature were considered as baselines for the comparison with four settings of the new feature sets (i.e. CF, CFN, FF, and FFN).

- Existing Feature set (EF)
- Existing Feature Set Normalized (EFN)

Table 5 Complete Feature Set, whereas set of selected features with (✓) represent filtered feature set

Complete Feature Set			
Newly Analyzed Features		Existing Features	
Feature	Selected/Drop	Feature	Selected/Drop
ST	✓	WCIMD	✗
MS	✓	TLC	✓
DS	✓	MDLC	✓
QAT	✓	HTTP or HTTPs	✗
H1LC	✓	Backlinks	✓
CWC	✗	RTS	✗
DA	✗	KWURL	✓
PA	✗	KWMD	✓
IMGL	✗	KWT	✓
DSX	✗	ULC	✓
SMR	✗	LD	✓
SDMS	✓	IL	✗
SMU	✗	Traffic	✗
QWURL	✗	CT	✓
WC	✗	AR	✗
		OL	✗

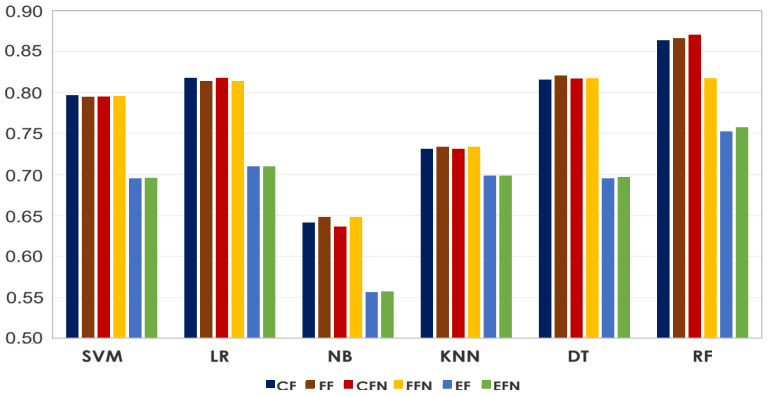

Fig. 8 Comparison of newly analyzed feature sets (CF, CFN, FF, and FFN) with baseline feature sets (EF and EFN)

We formed a list of existing features based on previous studies (Drivas et al. 2020; Matošević et al. 2021; Su et al. 2014) (see Table 5). The six configurations of the feature sets discussed above are used as the classifier's input. Each feature set incorporates a different number of features. We use the classification results to understand the role of the newly analyzed features we extracted in this research. Hence, we compare the performance of each classifier for different combinations of feature sets. The performance of each feature set is measured through a confusion matrix and average accuracy. Through the above feature design, we evaluate the SEO trend by comparing the results of CF, FF, and EF.

6 Results and discussion

In this section, we compare the newly analyzed and existing features using machine learning. Figure 8 compares the feature set (CF, CFN, FF, and FFN) with the feature set (EF and EFN) using SVM, LR, NB, KNN, DT, and RF. The feature sets CF, FF, CFN and FFN comprise our newly analyzed features as defined earlier, while the EF and EFN contain the features recommended by existing studies. The x-axis and y-axis represent the classification methods and their performance in terms of accuracy, respectively.

6.1 Performance comparison

The results show that the RF outperforms all other classifiers using CF and CFN, with an accuracy of 0.86 and 0.87. The performance of SVM, LR and DT is comparable using CF (0.80, 0.82, 0.82) and CFN (0.80, 0.82, 0.82), respectively. However, NB and KNN have the lowest accuracy using CF (0.64, 0.73) and CFN (0.64, 0.73), respectively. On the other hand, if we compare the performance of classifiers using FF and FFN, the results show that RF outperforms all other classifiers using FF and FFN, with an accuracy of 0.87 and 0.82, respectively. The performance of SVM, LR and DT is comparable using FF (0.79, 0.81, 0.82) and FFN (0.80, 0.81, 0.82), respectively. NB and KNN remain the lowest accuracy

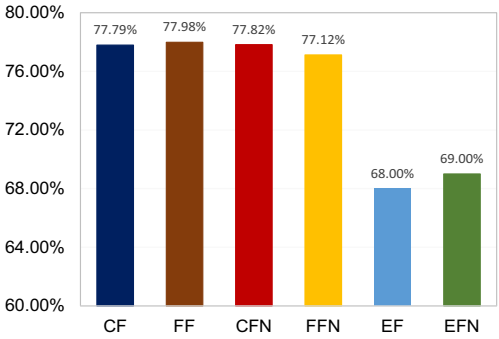

Fig. 9 Average accuracy comparison using newly analyzed (CF, FF, CFN, and FFN) and baselines (EF and EFN) feature sets

using FF (0.65, 0.73) and FFN (0.65, 0.73), respectively. Figure 8 compares the results of all classifiers against different feature sets. The results show a significant improvement with the newly analyzed features compared to baseline feature sets. The average accuracy results using newly analyzed features (CF, CFN, FF, and FFN) significantly outperform baseline feature sets (EF and EFN). However, no significant difference was observed among the CF, CFN, FF, and FFN results. Figure 9 compares the average accuracy using each feature set. The x-axis and y-axis represent the feature set and average accuracy, respectively. The comparison of average accuracy using (CF, FF, EF) and (CFN, FFN, EFN) shows (77.79%, 77.98%, 68%) and (77.82%, 77.12%, 69%), respectively, leading to a conclusion that normalization of data does not have a significant effect on the performance of classifiers. The average accuracy using FF (77.98%) remains the best among all other feature sets.

The average classification accuracy we achieved for CF, CFN, FF and FFN is 77.79%, 77.98%, 77.82% and 72.12%, respectively, while the average classification accuracy for EF and EFN was 68% and 69%, respectively. The results emphasize that FF consists of important factors that webmasters must consider while designing webpages. It confirms the trend that the newly analyzed feature set (FF) is the most essential and will affect the webpage's ranking, hence satisfying the research question (RQ2).

6.2 Heatmap

To achieve a high-level comparison of the results for the proposed feature design, we utilized heatmaps for the side-by-side comparison (see Fig. 10). A heatmap provides a visual summary of all feature designs and their level of performance using different classifiers. There are 20 unique classes (ranks); each heatmap visualizes 20×20 confusion matrix. The scale range is between 0 (blue) and 1 (red), reflecting the percentage of correctly classifying the page to their actual ranks. The heat signature of existing features (EF, EFN) is scattered in all classifiers. The signature is scattered because most webpages are classified incorrectly (true negative). While in the decision tree, the heat signature is relatively better than others, but after the 10^{th} position, the signature is scattered. Overall, the signature against the filtered feature (FF) set shows better results as the heat signature is concentrated on the diagonal, which shows more accurate predictions. There is a clear distinction and segregation in the heat signals; the top-10 page rank is predicted more accurately by most of the classifiers compared to the bottom-10 pages. In FF, the narrow signature of the top-10 pages guarantees that the FF feature design is correctly optimized and is significant. The heatmap of CF and FF are distinct from the EF. CF and FF separate the top-10 and

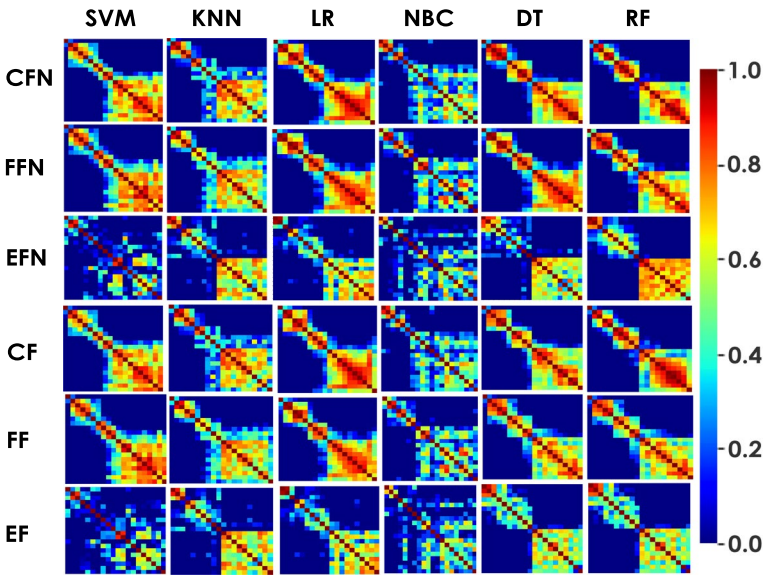

Fig. 10 Heatmap: A visual summary of the classifiers' performances for different feature sets

bottom-10 more accurately. That guarantees the correct rank position of the top-10. However, the signature of CF is a bit scattered in some classifiers. Hence, it shows that FF produces relatively better results. Intuitively, we can conclude that these features are optimized in the top-10 results, which is why Google ranks them at the top. However, the bottom-10 pages are not optimized properly, which is why the signature of the heatmap is scattered at the bottom. The heatmap signature concluded two things: First, the features contained by FF are significant. Second, if these factors are correctly optimized, the overall page rank could improve, and the page can get more visibility in the user search.

6.3 Statistical significance

Furthermore, we applied the independent two-sample t-test to compare the significance of feature design. A two-sample t-test uses the data points (results) and estimates the statistical difference in the sample mean (in either direction) of the data distribution(Easterling 2015, Chapter 3), as shown in Eq. 5. We use the average accuracy results of each feature design as a group sample across all classifiers.

$$t = \frac{\sqrt{s}(\mu_1 - \mu_2)}{\sqrt{2}\sigma_p} \tag{5}$$

Where μ_1, and μ_2 are the sample means of groups (feature designs) in comparison. s is the sample size. σ_p is the pooled standard deviation calculated using Eq. 6 with the degree of freedom $df = 2n - 2$.

Table 6 Comparison of the p-value for all six-feature designs results using the independent two-sample t-test

	EF				
CF	0.051	**CF**			
FF	**0.044**	0.968	**FF**		
EFN	0.962	0.057	**0.049**	**EFN**	
CFN	0.057	0.988	0.957	0.063	**CFN**
FFN	**0.048**	0.87	0.834	**0.049**	0.884

$$\sigma_p = \sqrt{\frac{(n-1)\sigma_1^2 + (n-1)\sigma_2^2}{df}} \tag{6}$$

Here, σ_1^2 and σ_2^2 are the variances of average accuracy for each feature design in comparison. The confidence interval is set to 95%; therefore, $p\text{-}value < 0.05$ shows a significant difference between the results of feature designs.

We test the significance of all the feature designs. However, we only discuss the test results of FF with the baselines (EF and EFN) to show the significance of the proposed features. The null hypothesis (H_0) and alternative hypothesis (H_a) for the t-test are described as:

$H_0 =$ "The performance of the FF feature design is equal to or lower than the baselines EF and EFN; hence, the newly analyzed factors are not significant."

$H_a =$ "The proposed FF feature design performs consistently better than the baseline (EF and EFN) and significantly improves webpage ranking."

Table 6 compares *p-values* for all the pairs of feature sets. The newly analyzed feature designs FF and FFN show significance compared to the existing features EF and EFN with p-values < 0.05.

Similarly, we performed these tests to compare the significance of classifiers used across each feature design. Our primary focus was to analyze the performance based on various feature designs; nevertheless, the results of the significance test show that random forest (RF) performs better in most cases. (see Appendix-A Table 9) with greater accuracy (see Fig. 8).

7 Case study of a blog

To address the research question (RQ3), we conducted a longitudinal study to optimize identified SEO factors and verify the impact by analyzing their visibility in Google SERP. Over four months, the improvements in the keywords rank position led to an increment in the number of users and sessions. Table 7 details the experimental blog we designed intending to increase the website's visitors and rank position in Google SERP.

Table 7 Experimental Blog Details

Specifications	Details
URL	https://www.joonse.com/
Goals	Improve the ranking of the selected keywords
	and increase number of visitors
Niche	Technology
	Health and care
	Entertainment
User's demographic	International Targeting

7.1 Checklist for technical SEO setup

The SEO activities started from 2 February 2021 to 15 May 2021. Before jumping into keyword-based research, we perform the technical SEO (see SEO Checklist in Fig. 11) to ensure that the website structure meets the guidelines of the Google SE Algorithm.

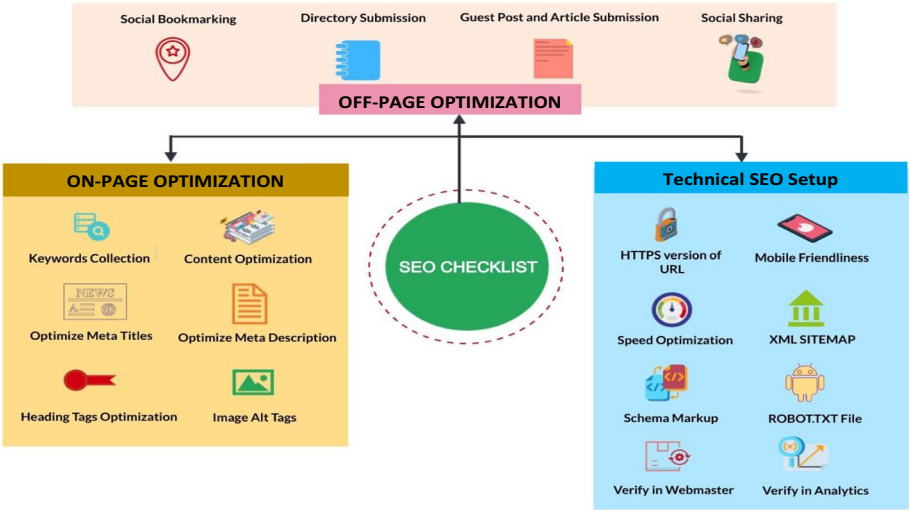

Fig. 11 Checklist for technical SEO

Table 8 Keywords Search Volume and Competition

SR	Keyword	Difficulty	Global Volume
1	Future of technology	55	3.2k
2	Best cpu for gaming	76	47000
3	iPhone SE 3 release date	70	22000
4	Earn money online without investment for students	23	20000
5	Egg white for hair	27	27
6	Target dresses for women	6	1200
7	How to make your hair healthy again	42	600
8	Egg yolk for hair growth	12	250
9	How to play keytar	0	150
10	What is a keytar	2	100
11	What is a keytar	31	100
12	Beginners guide to crossfit	13	100
13	Best fat burning drinks	26	90
14	Weight loss insanity	11	70
15	Do muslims celebrate valentines day	3	60
16	How to play a keytar	0	50
17	How do actresses lose weight fast	12	50
18	Is moxie based on a true story	2	40
19	Reasons to rebuy overwatch on switch	Not found	10
20	Year of robotics and artificial intelligence	Not found	10
21	How might travel in the future be different?	Not found	10
22	Things you dont know your mobile phone could do	Not found	not found

7.2 Keywords ranking in Google SERP

In January 2021, a blog[16] was analyzed for its influential ranking. The blog ranking was very low, which caused low traffic. To increase the blog's organic traffic, 23 keywords against 16 different webpages were selected with the help of Google Keyword Planner,[17] and the search volume was found through the Ahref tool.[18] The keyword research was conducted by considering the short-tail and long-tail queries, including question words for voice search. Table 8 shows the keywords' rank position in Google SERPs comparison for the last six months. The keyword rank position report is generated by the Google search console[19] and by searching for them in the Google search engine. The rank position of the previous six months is zero because the site was new at that time, and no SEO activity was performed. However, the rank position improves gradually.

[16] https://www.joonse.com/ Accessed Date: 04-09-2023.
[17] https://web.archive.org/web/20240130013322/https://ads.google.com/intl/en_au/home/tools/keyword-planner/ Accessed Date: 04-09-2023.
[18] https://web.archive.org/web/20240207051708/https://ahrefs.com/ Accessed Date: 04-09-2023.
[19] https://web.archive.org/web/20240204162237/https://search.google.com/search-console/welcome Accessed Date: 04-09-2023.

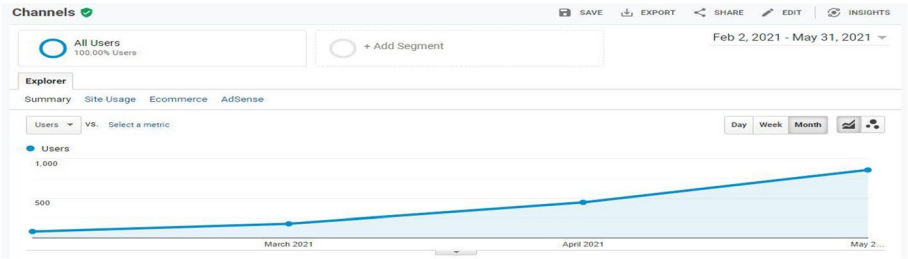

Fig. 12 Increasing traffic on the Blog designed for the case study

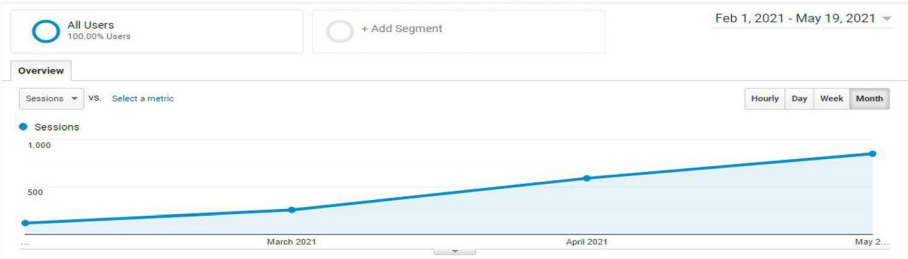

Fig. 13 Increasing user sessions on the Blog designed for the case study

7.2.1 Number of users and sessions

The results show the process of technical SEO setup to get the change in the number of users and sessions in Google Analytics. The period of SEO is from 2 February 2021 to 15 May 2021. The ranking and the number of users were observed through Google Analytics[20] and Google webmaster[21] on different dates. After implementing the identified SEO factors, there has been a prominent improvement in the number of users and sessions. Figure 12) shows the number of users on different dates. In February 2021, the number of users who visited the website was only 82, while in November 2021, the number of users increased to 857. Figure 13) shows the number of sessions that increased from 119 to 846. The statistics from the Google Analytics tool show the increment in the number of users and sessions.

[20] https://web.archive.org/web/20240203081720/https://analytics.google.com/analytics/web/provision/#/provision Accessed Date: 04-09-2023.

[21] https://web.archive.org/web/20240206013139/https://developers.google.com/search Accessed Date: 04-09-2023.

Table 9 Comparison of p-value for all classifiers' results using independent two-sample t-test

	SVM				
LR	0.564	**LR**			
NB	**0.000**	**0.000**	**NB**		
KNN	0.109	**0.042**	**0.000**	**KNN**	
DT	0.669	0.916	**0.000**	0.080	**DT**
RF	**0.047**	0.215	**0.000**	**0.004**	0.213

8 Conclusion

The emergence of more advanced mobile devices and voice search is incredibly increasing. Voice search is a natural addition and provides a new way to search the web. With this fast adoption of voice search technology, a need arises to understand the ranking factors that influence the Google search engine to rank webpages higher against voice search queries. SEO involves different techniques; the preceding literature attempts many types of research to identify some essential factors that affect ranking in search engines. There was a need to analyze additional ranking factors contributing to the search engine ranking against voice queries. Moreover, such studies must be updated due to continuous changes in Google's ranking algorithm. This research explored a machine learning approach to identifying the most dominant factors. We proposed a framework that can be adopted anytime Google's ranking algorithm updates by acquiring new webpage ranks and re-evaluating significant factors contributing to SEO. Four novel feature sets were designed and compared with existing ones. We used six widely used classifiers and trained on these features extracted from 2960 webpages. A unique feature design (i.e., FF) significantly improves webpage ranking prediction. FF feature set outperforms all other feature designs with an average accuracy of 77.98%, and CFN is the second best with 77.82%. The existing feature sets EF and EFN remain 68% and 69%, respectively. Furthermore, a longitudinal study on a blog over four months confirms that the proposed factors improve the webpage visibility and increase organic traffic. In conclusion, the SEO factors used in FF design will help webmasters achieve a higher ranking in search engines.

9 Future direction

Further experiments will be conducted to investigate the more influencing factors. For example, in this research, we only checked the quantity of backlinks. In the future, we would like to analyze the quality of backlinks with a scoring weight. In future, we aim to consider other domains for voice query generation and extend the benchmark dataset. Moreover, with the rapid growth of voice search technology, there is a great need to track voice search queries against a specific webpage and introduce a profile-based scoring function for weighing the voice queries against particular domains.

Appendix

Comparison of significance test

Acknowledgements This research study was partially supported by the project FAIR - Future AI Research (PE00000013), spoke 6 - Symbiotic AI(https://future-ai-research.it/), under the NRRP MUR program funded by the NextGenerationEU.

Author contributions ZS and FA conceived this research and designed experiments; ZS and AG contributed in developing the research idea; FA implemented the idea and conducted experiments; MU and IR worked on the problem formulation and presentation; ZS, FA, AG, MU, and IR wrote and edited the paper.

Funding Open access funding provided by Università degli Studi di Bari Aldo Moro within the CRUI-CARE Agreement. This research study was partially supported by the project FAIR - Future AI Research (PE00000013), spoke 6 - Symbiotic AI(https://future-ai-research.it/), under the NRRP MUR program funded by the NextGenerationEU.

Declarations

Conflict of interest There is no Conflict of interest or Conflict of interest to declare for this research.

Ethics approval Not applicable.

Conflict of interest The authors declare no Conflict of interest.

References

Anuradha T, Surekha TL, Praveena N et al (2021) Achieving more page views through search engine optimization. Advances in smart system technologies. Springer, Berlin, pp 315–324

Cerda P, Varoquaux G (2020) Encoding high-cardinality string categorical variables. IEEE Trans Knowl Data Eng 34(3):1164–1176

Drivas IC, Sakas DP, Giannakopoulos GA et al (2020) Big data analytics for search engine optimization. Big Data Cognit Comput 4(2):5

Easterling RG (2015) Fundamentals of statistical experimental design and analysis, 1st edn. John Wiley, Hoboken

Guy I (2016) Searching by talking: analysis of voice queries on mobile web search. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 35–44

Hingoro MA, Nawaz H (2021) A comparative analysis of search engine ranking algorithms. Int J. https://doi.org/10.30534/ijatcse/2021/1081022021

Jadav NK, Shrivastava S (2021) An analysis on incompetent search engine and its search engine optimization (seo). In: International Conference on Innovative Computing and Communications, Springer, pp 203–214

Jerkovic J, Warrior S (2009) Essential techniques for increasing web visibility. O'Reilly Media Inc., Sebastopol

Khan M, Mahmood A (2018) A distinctive approach to obtain higher page rank through search engine optimization. sādhanā 43(3):1–12

Lemos JY, Joshi AR (2017) Search engine optimization to enhance user interaction. 2017 International Conference on I-SMAC (IoT in Social. Mobile, Analytics and Cloud)(I-SMAC), IEEE, pp 398–402

Matošević G, Dobša J, Mladenić D (2021) Using machine learning for web page classification in search engine optimization. Future Internet 13(1):9
Mittal A, Sridaran R (2019) Evaluation of websites' performance and search engine optimization: A case study of 10 indian university websites. In: 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, pp 1227–1231
Patel M, Atkotiya K (2020) A review paper on SEO for ranking and effectiveness techniques in context of google search engine. Paripex Indian J Res. https://doi.org/10.36106/paripex/2705481
Patil AV, Patil VM (2018a) Search engine optimization technique importance. In: 2018 IEEE Global Conference on Wireless Computing and Networking (GCWCN), IEEE, pp 151–154
Patil VM, Patil AV (2018b) Seo: On-page+ off-page analysis. In: 2018 International Conference on Information, Communication, Engineering and Technology (ICICET), IEEE, p 1-3
Roslina A, Shahirah MN (2019) 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, pp 311–314
Selvan MP, Sekar AC, Dharshini AP (2012) Survey on web page ranking algorithms. Int J Comput Appl 41(19):1–7
Sharma S, Verma S (2020) Optimizing website effectiveness using various seo techniques. In: 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN), IEEE, pp 918–922
Strzelecki A, Rutecka P (2019) The snippets taxonomy in web search engines. In: International Conference on Business Informatics Research, Springer, pp 177–188
Strzelecki A, Rutecka P (2020) Direct answers in google search results. IEEE Access 8:103,642-103,654
Strzelecki A, Rutecka P (2020) Featured snippets results in google web search: an exploratory study. Marketing and smart technologies. Springer, Berlin, pp 9–18
Su AJ, Hu YC, Kuzmanovic A et al (2014) How to improve your search engine ranking: myths and reality. ACM Trans on Web (TWEB) 8(2):1–25
Sunny MNA (2020) Machine learning in search engines
Vyas C (2019) Evaluating state tourism websites using search engine optimization tools. Tourism Manag 73:64–70
Ziakis C, Vlachopoulou M, Kyrkoudis T et al (2019) Important factors for improving google search rank. Future Internet 11(2):32

Authors and Affiliations

Zafar Saeed[1] · Fozia Aslam[2] · Adnan Ghafoor[2] · Muhammad Umair[2] · Imran Razzak[3]

✉ Zafar Saeed
zafar.saeed@uniba.it

Fozia Aslam
l1s19mscs0016@ucp.edu.pk

Adnan Ghafoor
adnan.ghafoor@ucp.edu.pk

Muhammad Umair
muhammad.umair@ucp.edu.pk

Imran Razzak
imran.razzak@unsw.edu.pk

[1] Dipartimento di Informatica, Università degli Studi di Bari, 70125 Bari, Italy

[2] Faculty of Information Technology and Computer Science, University of Central Punjab, 54000 Lahore, Pakistan

[3] School of Computer Science and Engineering, University of New South Wales, Sydney, NSW 2052, Australia