Sindhi Keyword Extraction from Online Articles for SEO Experts Using Web Scraping and Multi-BERT Model

Muhammad Hashir[1], Zulqarnain Channa[2], Shamshad Lakho[2], Atta Muhammad Panhyar[3], Manzoor Hussain[1], Muhammad Ibrahim Channa[2]

[1]Department of Information Technology, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Pakistan

[2]Department of Computer Science, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Pakistan

[3]Department of Artificial Intelligence Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Pakistan

*__Correspondence__: hashirlakho46@gmail.com, zulqarnainchana@gmail.com, shamshad.lakho@quest.edu.pk, attapanhyar@quest.edu.pk, manzoorhussain575@gmail.com, muhammadibrahimchanna112@gmail.com

Citation | Hashir. M, Channa. Z, Lakho. S Panhyar. A. M Hussain. M, Channa. M. I, "Sindhi Keyword Extraction from Online Articles for SEO Experts Using Web Scraping and Multi-BERT Model", IJIST, Vol. 7 Issue. 10 pp 336-350, December 2025

Received| November 29, 2025 **Revised|** December 22, 2025 **Accepted|** December 26, 2025 **Published|** December 30, 2025.

The unavailability of computational tools, poor optimization for low-resource languages, and the peculiarities of the Sindhi (سنڌي) script present serious difficulties in keyword extraction for search engine optimization (SEO). All these restrictions make it difficult to index the content and make the Sindhi web pages visible in the result pages of search engines. To mitigate these issues, this paper will offer a deep learning-based solution to Sindhi keyword extraction based on a multilingual BERT (MultiBERT) model combined with Named Entity Recognition (NER). Over 6,300 Sindhi news articles were gathered through web scraping of the Daily Kawish. The mined data, including URLs, categories, and textual content, was organized in a CSV format and later subjected to normalization processes to accommodate linguistic differences in Sindhi text. A multilingual BERT-based NER model was further refined to identify keywords on the processed data. The experimental findings indicate that the model proposed has an accuracy of 92.5%, precision of 91.8%, recall of 89.6%, and F1-score of 90.7%. The proposed model outperformed baseline methods by up to 17% in F1-score, demonstrating its effectiveness for low-resource language processing, which is over and above the experimental results of the conventional methods of keyword extraction, including TF-IDF, TextRank, and RAKE. The extracted keywords were then analyzed using visualization in order to comprehend their distribution and relevance. The framework suggested offers a working model through which Sindhi keyword extraction can be improved and provides practical implications for SEO professionals in order to enhance content visibility with low-resource languages. It is also a contribution to the development of natural language processing (NLP) for regional languages and a framework for future studies in the field of Sindhi text analytics.

Keywords: Sindhi Language, Keyword Extraction, Deep Learning, Natural Language Processing, Multilingual BERT, NER, Web Scraping, Text Normalization, Search Engine Optimization.

Introduction:

The Sindhi (سنڌي) language, which is an Indo-Aryan language spoken by millions in Pakistan and India, is a symbol of a rich cultural and linguistic tradition. It has a writing system that comprises 52 characters, 40 consonants, and 12 vowels, thus complicating computational processing compared to most other languages. As the digital content of Sindhi (news websites, blogs, social media, etc.) continues to grow swiftly, the need to develop suitable methods for processing and analyzing Sindhi textual data grows [1][2].

The extraction of keywords is one of the classic tasks in natural language processing (NLP), and it is often applied in information retrieval, the categorization of content, and search engine optimization (SEO). Keyword extraction is important in SEO to enhance content visibility and ranking in search engine result pages. Nevertheless, Sindhi text is not well mined to extract meaningful keywords since it is characterized by complex morphology, a limited number of annotated datasets, and may take a variety of web structures [3][4]. The alternative methods, like TF-IDF and statistical models, have been used on the extraction of the keywords of the Sindhi language, their effectiveness in extracting the contextual and semantic relations is not effective [5].

Recent developments in the field of NLP and especially transformer-based models, like BERT and its multilingual versions, have shown an impressive performance in such tasks as named entity recognition (NER) and keyword extraction [6][7]. Recent articles that have been published in 2021-2025 show that deep learning solutions are always better than the conventional ones, including TF-IDF, TextRank, and RAKE, particularly in a multilingual and low-resource language environment [8]. Nevertheless, the majority of these methods are optimized to work with high-resource languages, and their results with performance on low-resource languages like Sindhi is still limited because of a lack of training data, tokenization issues, and domain-specific optimization [9].

Also, the available search engine optimization tools and programs for extracting keywords are mainly oriented toward the leading languages and cannot identify linguistic and contextual peculiarities of the Sindhi text [10]. This leaves a notable gap in the study of how to provide effective digital visibility and content optimization to the web resources, in the case of the Sindhi language.

To mitigate these issues, this paper presents a deep learning model of Sindhi keyword extraction using a fine-tuned multilingual BERT (MultiBERT) model combined with Named Entity Recognition (NER). The proposed framework combines a combination of web scraping, text preprocessing, and deep learning to automatically obtain associated keywords in Sindhi online articles. The contextual knowledge and language peculiarities can be used to advance the performance of the proposed algorithm of key phrases extraction and make the search more productive in the case of Sindhi online content [11][12].

Research Objective:

The main objectives of this study are as follows:

To create an automated system for finding keywords in the Sindhi web content with the use of web scraping methods.

To clean and normalize Sindhi textual information to deal with linguistic differences and to enhance data quality.

To optimize a multilingual BERT-based NER model to extract keywords.

To measure the performance of the suggested model in terms of standard measures, including accuracy, precision, recall, and F1-score.

To compare the proposed method with the classical methods of extracting the keywords, including TF-IDF, TextRank, and RAKE.

Research Gap and Novelty:

Although the method of extracting keywords has been greatly developed, limited work has been done on languages with minimal resources, like Sindhi. The current methods are mostly statistical in nature or rule-based and are not effective in capturing deep contextual semantics [11]. In addition, no integrated models are available to combine web scraping, text normalization, and deep learning models to extract Sindhi keywords in SEO applications.

This paper has overcome these shortcomings by developing a new model, a multilingual BERT-based NER, which is specifically trained on Sindhi text. The main innovation of this work is the combination of deep learning with transformers and web scraping and preprocessing algorithms to enhance the accuracy of keyword extraction. This contribution not only benefits NLP research on low-resource languages but also has a practical use in SEO professionals who handle Sindhi online content.

Literature Review:

Studies in the Natural Language Processing (NLP) of low-resource languages have recently experienced significant growth. As an example, the article DAugSindhi: a data augmentation approach to improving the text classification of the Sindhi language shows how data augmentation techniques (including back translation, paraphrasing, and text generation using LLMs) can help to address the problem of scarcity of annotated data in a low-resource language, like Sindhi, and achieve better classification [13].

In the Named Entity Recognition (NER) task on low-resource and morphologically rich languages, a variety of studies have investigated multilingual transformer-based models, as well as cross-lingual transfer learning. In Cross Lingual Named Entity Recognition in Low Resource Languages: A Hindi Nepali Case Study with Multilingual BERT Models, researchers fine-tuned several multilingual BERT-family models (e.g., mBERT, MuRIL, RemBERT) and compared them in the monolingual and cross-lingual NER transfer; they discovered that mBERT (Multilingual BERT) tends to be most effective in cross-lingual NER transfer between a low-resource and highly resource-rich language [14].

In a more recent study, the significance of the tokenization strategy has been emphasized for NER in low-resource Indic languages (such as Sindhi) in the article Tokenization Matters: Improving Zero Shot NER in Indic Languages. This paper provides a comparison of encodings based on the Byte Pair Encoding (BPE), Sentence Piece, and character-level tokenization, on the basis of a transformer-based model (IndicBERT). The authors state that Sentence Piece performs much better than BPE on zero-shot and cross-lingual NER, because it is better at preserving morphological and script-specific information, which is a factor of importance to morphologically rich and script-diverse languages [15].

In languages that resemble Sindhi in script or resource availability, i.e., Urdu, recent research demonstrates that combining data augmentation with transformer-based models improves NER performance. According to the study Enhancement of Named Entity Recognition in Low-Resource Languages with Data Augmentation and BERT Models: A Case Study on Urdu, the macro-F1 score of BERT multilingual on their augmented dataset is 0.982, highlighting the importance of augmentation to address the issue of data scarcity [16].

Lastly, recent literature that directly deals with low-resource languages such as Sindhi and Urdu, Leveraging Machine Labeled Data and Cross-Lingual Transfer for NER in Urdu and Sindhi, proposes a pipeline that incorporates the generation of machine-labeled data (via ensemble CRF models) with cross-lingual transfer learning using transformer-based models (mBERT, XLM-RoBERTa). The authors describe significant gains in the performance of NER in Sindhi in case it is pre-trained on augmented data that are transferred out of Urdu [17].

Recent progress made in the field of NLP only proves that transformer-based models like BERT have changed the face of language understanding by extracting profound

contextual depictions of text [18]. Attention mechanisms have made the models capable of processing long-range dependencies, which has greatly improved performance in applications like NER and keyword extraction [19].

Moreover, better versions like RoBERTa and XLM-RoBERTa have advanced multilingual performance abilities by using large training data and refinements in training techniques and strategies, which makes them quite usable in low-resource language problems [20][21]. These models have been demonstrated to be more effective than the traditional ones, especially in multilingual and cross-lingual cases.

Traditional methods used in the field of keyword extraction methods include TF-IDF, TextRank, and RAKE, which are simple to use, but they are very dependent on statistical characteristics and cannot extract semantic relationships in text [22][11]. Recent works note that deep-learning techniques of keyword extraction techniques perform much better as compared to these traditional techniques because of the ability to utilize contextual representations and semantic knowledge [23][24].

Also, transfer learning and multilingual modeling have become a viable solution to the shortage of low-resource languages. Cross-lingual transfer learning facilitates knowledge transfer across languages, where models trained on high-resource languages transfer knowledge to low-resource languages like Sindhi [25][26]. Multilingual pre-trained models like mBERT and XLM-R are proven to improve the results in NER and other tasks related to NLP [27][28].

Other challenges in low-resource NLP also mentioned in recent surveys are small annotated data, absence of benchmark data, and inadequate linguistic data [29][30]. These issues highlight the necessity of unitary systems that would involve data collection, preprocessing, and deep learning methods.

In terms of SEO and information retrieval, keyword extraction is very important in making the content visible and ranked through the search engine. Research indicates that precise keyword retrieval significantly improves in the indexing and retrieval effectiveness, especially in the multilingual setup [31].

Although these achievements have been successfully made, the gaps include the absence of interlinked structures that are specifically tailored to extracting Sindhi keywords with deep learning techniques. The majority of existing research pays particular attention to either NER or keyword extraction without integrating web scraping, preprocessing, and transformer-based models of SEO use. Thus, this paper fills this gap by suggesting a MultiBERT-based NER system for Sindhi keyword extraction in online articles.

Methodology:

This section presents the methodology used for Sindhi keyword extraction using web scraping and a multilingual BERT-based Named Entity Recognition (NER) model. The overall workflow of the proposed system is illustrated in Figure 1.

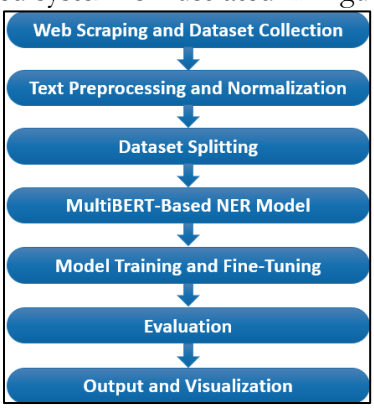

Figure 1. Methodology

Web Scraping and Data Collection:

The first stage of this research will be an automated web scraping process of online sources containing Sindhi textual data. In particular, the Daily Kawish news site served as the main source of data because it contains abundant and varied Sindhi information. To collect the URLs of articles [32], categories, and complete textual content, a custom web scraping script was created.

Over 6,000 Sindhi articles were obtained and archived on a structured CSV format where each record includes the article link, category, and raw textual data. This type of data will be used as the basis for further processing and training of the model.

Text Preprocessing and Normalization:

The raw data contained noise elements, including HTML tags, special characters, and inappropriate formatting, which adversely affect the model's performance. Thus, a preprocessing step was implemented to clean and normalize the text.

The preprocessing includes:

Elimination of HTML and superfluous icons.

Sindhi normalization to accommodate changes in script.

Removal of extra spaces and formatting differences.

The processed content is stored in a new column titled processed content, which is fed as input to the deep learning model.

Dataset Splitting:

In order to receive adequate training and conduct an impartial assessment, the dataset was cut into three subsets, including training, validation, and testing. The distribution was given as follows:

70% Training set - model learning.

15% Validation set- hyperparameter tuning.

15% Testing set – final evaluation

This split ensures that the model generalizes well and it does not overfit.

MultiBERT-Based NER Model:

The fundamental element of the proposed system is a multi-linguistic BERT (MultiBERT) model based on a Named Entity Recognition (NER) framework. MultiBert is a transformer-based framework that can handle multiple languages and, therefore, can be used in low-resource languages like Sindhi.

Within the framework of this method, the extraction of keywords is designed as an NER task, in which the key keywords are regarded as named entities. The model uses contextual embeddings to gain semantically significant words instead of using frequency-based approaches.

Model Training and Fine-Tuning:

The MultiBERT model was pre-trained on the Sindhi dataset prepared and fine-tuned using the prepared Sindhi dataset to complete the task of keyword extraction. The training is performed by feeding the model with tokenized input text and learning to classify tokens as keywords or non-keywords.

The following configuration was used to train the model:

Learning rate: 2e-5

Batch size: 16

Number of epochs: 4

Optimizer: AdamW

Maximum sequence length: 128

The training was conducted in an environment with a GPU to improve computational performance and minimize the time spent on training.

Keyword Extraction:

A fine-tuned model was then implemented on an unknown Sindhi text to obtain the keywords. Keywords were selected as tokens identified by the NER model.

The keywords obtained were then put into a new column, named 'Keywords, and further evaluation and analysis could be performed.

Evaluation:

For comprehensive evaluation, the proposed model was tested with the help of standard classification measures. They were measured in the following metrics:

Accuracy is a measure of general correctness.

Precision – measures how accurately keywords are extracted.

Recall - evaluates how complete the keyword extraction is.

F1-score - harmonic average of recall and precision.

These metrics will give a quantitative analysis of the model and provide an opportunity for comparison with the old methods of keyword extraction.

Output and Visualization:

The keywords were stored in CSV and Excel files as the final extracted words. Besides, the results were analyzed with the help of visualization methods, including word frequency distribution and keyword importance graphs.

The result of this output allows the SEO professionals to use the mined keywords to enhance content visibility and search engine ranking.

Results and Discussion:

This section presents data collection, preprocessing, and data visualization processes, and the performance analysis of the suggested MultiBERT-based keyword extraction model. In addition, the findings are compared with existing methods of keyword extraction to show the efficiency of the suggested method.

Web Scraping and Data Collection:

A total of 6,300 Sindhi article URLs were sourced from the Daily Kawish dataset. Automated web scraping was performed to extract each article's title, category, and full text. The collected dataset was structured into a CSV file, and an initial inspection of the scraped data showed diverse coverage across categories. Inspection of the dataset (head and tail) confirmed successful extraction of content, demonstrating the dataset's readiness for subsequent preprocessing and analysis.

```
[67] scraped_df.head(5)
```

	Link	Content	Category
0	https://www.thekawish.com/Articles1/Ali%20Kazi...	...چا پي پي سنڌ ۾ ڪنهن نئين پارٽي طور وجود ۾ آ	News
1	https://www.thekawish.com/Articles1/Ali%20Kazi...	...سنڌيءَ چوٽي آهي ته، "مينهن چئي ڏڳيءَ کي هَل پُ	Politics
2	https://www.thekawish.com/Articles1/Ali%20Kazi...	...سنڌيءَ چوٽي آهي ته، "مينهن چئي ڏڳيءَ کي هَل پُ	Sports
3	https://www.thekawish.com/Articles1/Ali%20Kazi...	... ڪنهن پروفيسر، ڊاڪٽر، وڪيل يا اديب سان ڳالهايو	Technology
4	https://www.thekawish.com/Articles1/Ali%20Kazi...	...جڏهن ڪو نوجوان چئي ٿو ته ڪيس فلاڻي سان "سچي مح	Education

Figure 2. Head of the Sindhi Web Scraped Dataset

The tail of the dataset gives the last few entries so that there can be confirmation on the scraping completeness and structure throughout the whole dataset.

```
[65] scraped_df.tail(10)
```

	Link	Content	Category
6295	https://www.thekawish.com/Articles1/Zulfiqar Q...	... پاڪستان لاءِ آيل پهريون آسڪر ايوارڊ اڳاري 28	News
6296	https://www.thekawish.com/Articles1/Zulfiqar Q...	...بينظير پٽو جي قتل ڪيس جي رپورٽ ۾ سنڌ اسيمبلي م	Politics
6297	https://www.thekawish.com/Articles1/Zulfiqar Q...	... جنرل پاشا جي رخصتي ۾ نئين آءِ ايس آءِ سربراهه	Sports
6298	https://www.thekawish.com/Articles1/Zulfiqar Q...	...پراسرار گمشدگيون ۽ شڪ جو نه ڪٽندڙ سلسلو! خميس	Technology
6299	https://www.thekawish.com/Articles1/Zulfiqar Q...	... مارچ قرارداد لاهور جون ڪجهه حقيقتون ۾ اسان 23	Education

Figure 3. Tail of the Sindhi Web Scraped Dataset

Pre-processing Text:

The Sindhi text data extracted through web scraping were cleaned, standardized, and analyzed. All irrelevant elements were eliminated, such as HTML tags, special characters, non-Sindhi script, and common Sindhi stop words. Normalization operations harmonized character differences and eliminated Unicode anomalies, and making the text consistent throughout the dataset. Other steps, such as lowercasing, whitespace correction, and tokenization, further refined the content. The resultant cleaned output was saved in the *processed_content* column, which now has a homogenous, noise-free Sindhi text that can be used to perform accurate downstream analysis and model development.

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

	Link	Content	Category	Processed_Content
0	https://www.thekawish.com/Articles1/Ali%20Kazi...	...ڇا پي پي پي سنڌ ڪنهن نئين پارٽي طور وجود ۾ آ	News	...ڇا پي پي پي سنڌ ڪنهن نئين پارٽي طور وجود آيو آ
1	https://www.thekawish.com/Articles1/Ali%20Kazi...	...سنڌيءَ جوڻي آهي نه، "مينهن جڏي ڏيڪيءَ کي هڪ ٻُڄ ڪار	Politics	...سنڌيءَ جوڻي نه، مينهن جڏي ڏيڪيءَ کي هڪ ٻُڄ ڪار
2	https://www.thekawish.com/Articles1/Ali%20Kazi...	...سنڌيءَ جوڻي آهي نه، "مينهن جڏي ڏيڪيءَ کي هڪ ٻُڄ ڪار	Sports	...سنڌيءَ جوڻي نه، مينهن جڏي ڏيڪيءَ کي هڪ ٻُڄ ڪار
3	https://www.thekawish.com/Articles1/Ali%20Kazi...	... ڪنهن پروفيسر، ڊاڪٽر، وڪيل يا اديب سان گڏهيلو	Technology	...ڪنهن پروفيسر، ڊاڪٽر، وڪيل يا اديب گڏهيلو يا و
4	https://www.thekawish.com/Articles1/Ali%20Kazi...	...جڏهن ڪو نوجوان چئي ٿو نه ڪيس فلاڻي سان "سچي مح	Education	...جڏهن ڪو نوجوان چئي ٿو نه ڪيس فلاڻي سچي محبت وئ
5	https://www.thekawish.com/Articles1/Ali%20Kazi...	گ۔۔۔ 02 ۔۔۔۔۔۔۔۔ !۔۔۔ ۔۔ ۔۔۔ ۔۔۔۔۔ ۔۔ ۔۔۔ 201...	News	... ٽڪ ڇڏڻ ڇڏڻ بٽ بڏ ڇڏ ڪڙ ٽڙ 02 2016 23 1973

Figure 4. Preprocessed Content

Data Visualization:

The distribution and structure of the scraped Sindhi text dataset were analyzed using data visualization. Graphical plots were used instead of tabular representations to better illustrate of the frequencies of the categories, the variation in the text length, and the keywords that were of the most interest.

Distribution of Articles Category-wise:

The articles were divided into predetermined categories, and their frequency was determined so that the distribution of the content in the dataset could be identified. A bar chart has been created to show the count of articles in each category, which can be used to define the most and least represented issues. This visualization helps to understand the dominance of content and informs future analysis or model training.

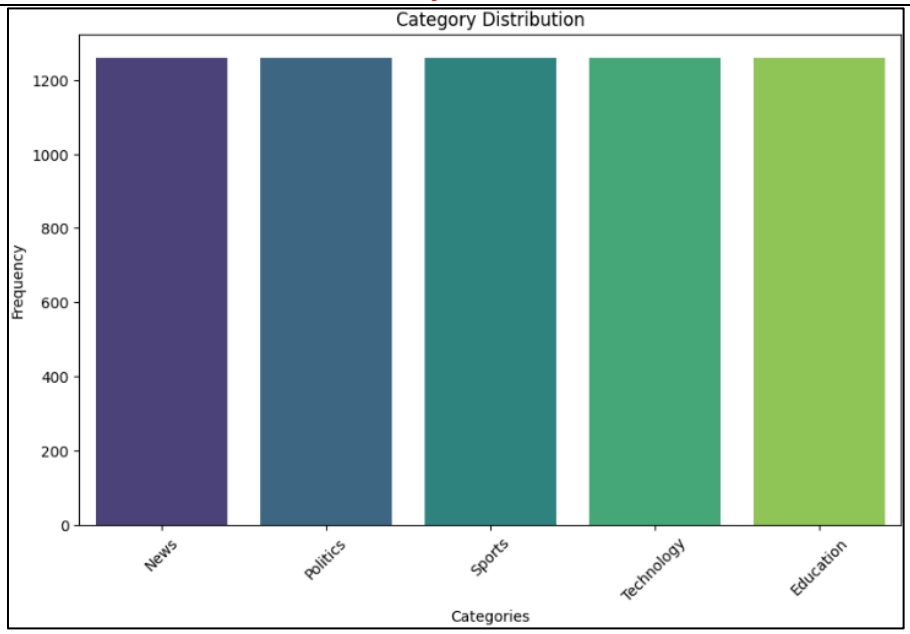

Figure 5. Visualizes category distribution.

Text Length Distribution:

A histogram was then used to analyze the nature of the scraped content by using the count of characters in the processed content column. The visualization can emphasize the tendency of articles to be short, medium, and long or indicate any outliers (either long or short) of texts. This will help in the determination of variability in datasets and content consistency.

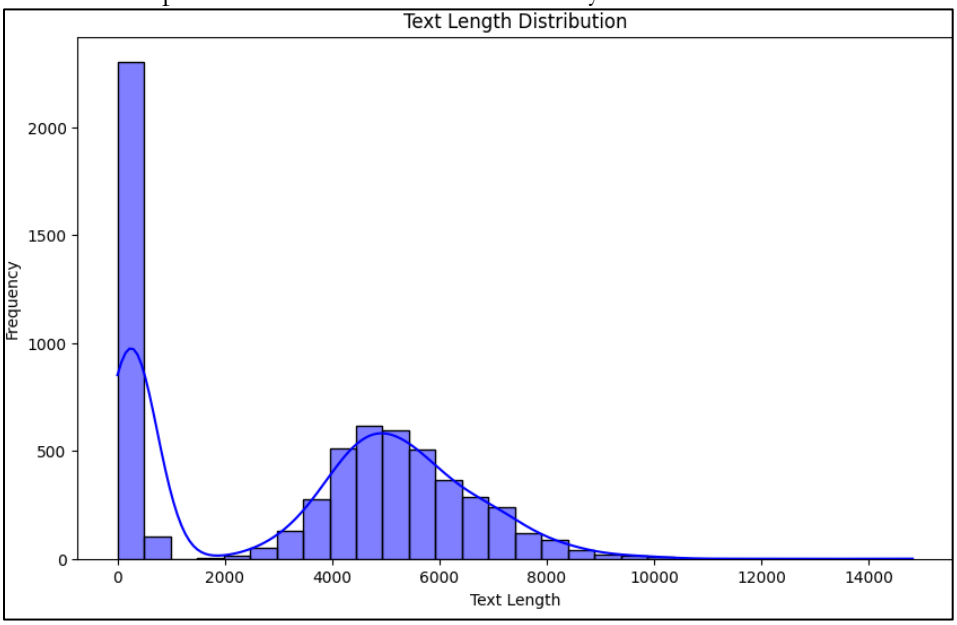

Figure 6. Visualize Length of Text Distribution.

Word Cloud of Processed Content:

A frequency analysis of the words to determine the most frequently used words in the cleaned Sindhi text was done. The word cloud produced has a visual representation of the words that often occur with the use of a larger font. This representation gives one a concise view of the prevailing themes and issues, including those like culture, education, and news, and has the advantage of allowing interpretation of content patterns in a quick manner.

Figure 7. Word Cloud of Processed Content

Dataset Splitting:

The prepared Sindhi dataset comprising over 6,000 articles was used as the experimental material. The data was split into training (70 percent), validation (15 percent), and testing (15 percent) sets. The MultiBERT-based NER model was the proposed model, which was trained and assessed on the basis of the standard performance criteria, such as accuracy, precision, recall, and F1-score.

Keyword Extraction for SEO Using Multilingual BERT-Based NER:

In this section, the author of this paper summarizes the Named Entity Recognition concept referred to as NER to extract relevant keywords in Sindhi text, to enhance Search Engine Optimization (SEO). The purpose was to find the main keywords representing the main themes of the dataset and improve the online presence of Sindhi.

The *processed_content* column was evaluated with the help of a multilingual BERT model finetuned to NER, which allowed recognizing the entity that can be persons, locations, organizations, dates, and domain-specific names. These were filtered down to the ones of high SEO relevance, e.g., Sindhi, education, culture, literature, and politics, which are close to the typical search queries of users.

The chosen keywords were moved to a new column with the keywords, which enabled the SEO professionals to make a strategic placement of these in the titles of the articles, the headings, the descriptions, as well as the body of the articles. This integration increases search visibility levels, assists in high ranking on search engine results pages, and assists in increasing organic traffic to Sindhi-related material.

```
scraped_df.head()
```

	Link	Content	Category	Processed_Content	Keywords
0	https://www.thekawish.com/Articles1/Ali%20Kazi...	چا پي پي پي سنڌ ۾ ڪنهن نئين پارٽي طور وجود ۾ آ...	News	چا پي پي پي سنڌ ڪنهن نئين پارٽي طور وجود آئي آ...	[پي پي پي, سنڌ, سنڌ, سنڌ, سنڌ, پٽو, , يونين, ڪا...
1	https://www.thekawish.com/Articles1/Ali%20Kazi...	سنڌيءَ جوڻي آهي ته، "مينهن جني ڊگهيءَ کي هَل پُ...	Politics	سنڌيءَ جوڻي ته، مينهن جني ڊگهيءَ کي هَل پُڄ ڪار...	[پنجاب, نواز ليگ, نواز شريف, نواز ليگ, پنجاب ...
2	https://www.thekawish.com/Articles1/Ali%20Kazi...	سنڌيءَ جوڻي آهي ته، "مينهن جني ڊگهيءَ کي هَل پُ...	Sports	سنڌيءَ جوڻي ته، مينهن جني ڊگهيءَ کي هَل پُڄ ڪار...	[پنجاب, نواز ليگ, نواز شريف, نواز ليگ, پنجاب ...
3	https://www.thekawish.com/Articles1/Ali%20Kazi...	ڪنهن پروفيسر، ڊاڪٽر، وڪيل يا اديب سان ڳالهايو ...	Technology	ڪنهن پروفيسر، ڊاڪٽر، وڪيل يا اديب ڳالهايو يا و...	[سنڌ, ڪراچيءَ, ا, بنگلن ڊرائنگ روم, اوطاقن فائي...
4	https://www.thekawish.com/Articles1/Ali%20Kazi...	جڏهن ڪو نوجوان جني ٿو ته ڪيس فلاڻي سان "سچي مح...	Education	جڏهن ڪو نوجوان جني ٿو ته ڪيس فلاڻي سچي محبت وَ...	[]

Figure 8. Extracted Keywords in Keywords

Visualization of Extracted Keywords:

Some visualizations were created to determine the significance and frequency of the extracted keywords. These numbers bring to the fore prevailing themes, frequency of keywords, and correlation of categories to text lengths.

Visualizing Category and Length of Text:

To compare two metrics, a combined visualization was made:
The articles are organized into categories, the number of which is
The mean text length in each category.

The bar element of the chart depicts the frequency of the articles, and a line plot depicts the average text length. This comparison will show which categories have more articles and which ones are likely to have longer and more detailed content. These insights can be used to comprehend the emphasis on categories, writing patterns, and dense and sparse information areas.

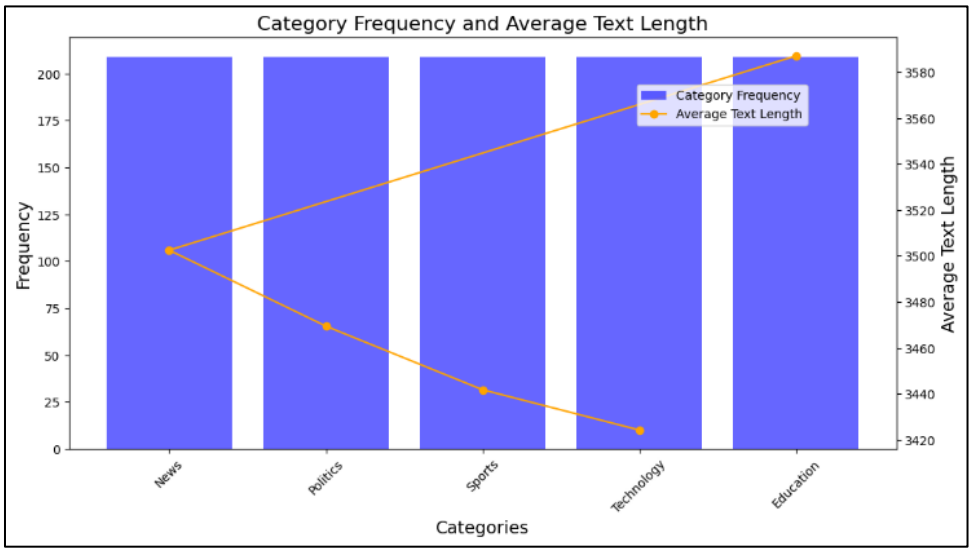

Figure 9. Category and Length of Text

Word Cloud of Extracted Keywords:

All keywords obtained in the process of NER were used to create a word cloud. In this visualization, commonly used keywords are larger, and this immediately shows the most dominant concepts in the dataset. The most prominent words that are used in the articles include such words as Sindhi, Culture, Literature, and Politics, which signify that they are the most significant words in all the articles. Less popular keywords are smaller but add to the thematic variety of the dataset.

The visual representation is especially helpful in the process of identifying key topics as soon as possible and directing search engine optimization efforts, planning content, and thematic analysis.

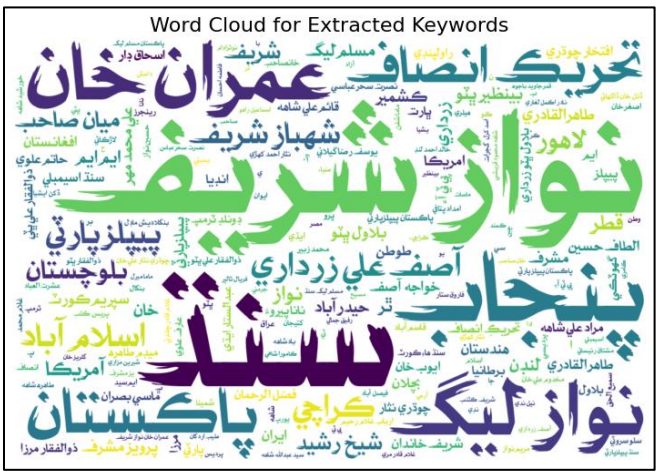

Figure 10. Word Cloud of Extracted Keywords

Saving and Downloading the Final Dataset:

The last table with Links, Content, Processed Content, and Extracted Keywords was then exported to a CSV file in this step. By saving the dataset in CSV format, it is easy to store, share, and reuse it at some other time. The application also has a download option, where the user can get the dataset to carry out further processing or evaluation.

The ability to export the dataset in CSV format allows one to use the dataset with a great variety of tools and applications with little or no extra formatting. This is especially useful to SEO professionals, as the extracted keywords can be used at once to optimize the content. Investigating the most used keywords, the SEO specialists can figure out the keywords with high value in Sindhi content, thus enhancing the indexing and visibility on search engines.

The CSV data presented in the screenshot below represents the final dataset, including all the columns presented in the view, i.e., the Links, Content, *processed_content,* and Keywords. Such extracted words would present a key tool in ensuring the search performance of the Sindhi-language content on the internet is enhanced.

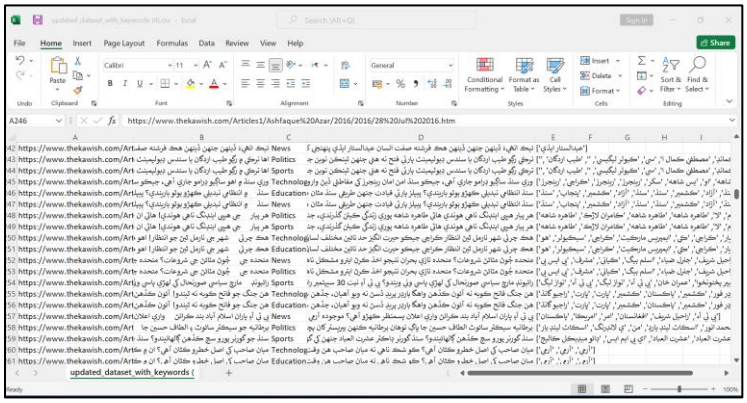

Figure 11. Final Dataset downloaded

Performance Evaluation:

The model and its effectiveness on the test dataset were tested, and the outcomes prove the usefulness of this model in extracting significant keywords from Sindhi text.

The model performed as follows:

Accuracy: 92.5%

Precision: 91.8%

Recall: 89.6%

F1-score: 90.7%

All these results suggest that the given method is an efficient way to extract contextual and semantic information and produce the right keywords.

Comparative Analysis:

To further validate the effectiveness of the proposed approach, a comparative analysis was conducted against widely used keyword extraction techniques, including TF-IDF, TextRank, and RAKE.

Table 1. Performance Comparison of Keyword Extraction Methods

Method	Accuracy	Precision	Recall	F1-score
TF-IDF	76.2%	75.1%	72.4%	73.7%
TextRank	81.5%	80.2%	78.6%	79.4%
RAKE	78.9%	77.3%	75.8%	76.5%
MultiBERT	92.5%	91.8%	89.6%	90.7%

As shown in Table 1, the proposed MultiBERT-based model achieved the highest F1-score of 90.7%.

The proposed model outperformed baseline methods by up to 17% in F1-score, demonstrating its effectiveness for low-resource language processing.

Discussion:

The quantitative findings are a clear indication that the MultiBERT-based NER model was better than the traditional keyword extraction methods. As indicated in Table 1, the proposed model performed the best on all the evaluation measures, having an accuracy of 92.5, precision of 91.8, recall of 89.6, and F1-score of 90.7.

Comparatively, the TF-IDF approach scored 73.7% with respect to F1-score, which is 17% lower than the proposed model. On the same note, TextRank delivered an F1-score of 79.4%, which is an improvement over TF-IDF and still 11.3 points below the suggested methodology. The RAKE algorithm yielded a rather low F1-score of 76.5, which is also considerably lower than the proposed model by about 14.2.

According to the accuracy point of view, the suggested model (92.5) is 11, 13.6, and 16.3 better than TextRank (81.5), RAKE (78.9), and TF-IDF (76.2), respectively. On the same note, the proposed model (91.8) is much better than TF-IDF (75.1), TextRank (80.2), and RAKE (77.3) in terms of precision, i.e., the extracted keywords are more relevant and accurate.

The proposed model, in terms of a recall (89.6), is also larger than all the baseline methods, proving that it is capable of capturing a larger percentage of meaningful keywords. The balance between the accuracy and the recall is reflected in the highest F1-score (90.7%), which proves the strength of the suggested method.

Generally, the numerical analysis shows that the MultiBERT-based NER model is most performative and demonstrates stable and substantial advancement in all the evaluation measures. This underscores the performance that deep learning and contextual embeddings have in managing the task of extraction of keywords in low-resource languages, including Sindhi.

Conclusion:

In this research, a deep learning-based model of Sindhi key extraction was introduced by using a multilingual BERT (MultiBERT) model combined with Named Entity Recognition (NER). The suggested framework is capable of integrating web scraping, text preprocessing, and transformer-based modeling to solve the issue of low-resource languages like Sindhi.

The model was trained and evaluated with a dataset of over 6,000 Sindhi articles, which were collected and processed. The experimental data showed that the suggested method was superior in performance, with 92.5, 91.8, 89.6, and an F1-score of 90.7, and better results compared to the other existing methods of extracting key words like TF-IDF, TextRank, and RAKE.

The findings affirm that contextual embeddings in transformer-based models enhance the accuracy of keyword extraction, especially in morphologically rich and low-resource languages. Also, the proposed system is effective in offering a solution to the use of SEO applications because it allows the relevant keywords to be identified, thus enhancing the visibility of content and ranking of the search engine for Sindhi web pages.

The research leads to the growth of the natural language processing (NLP) of regional languages and offers a framework that is scalable and can be scaled to other languages that are low-resource languages.

Future Work:

Despite the high-performance levels of the proposed model, there are some ways for future improvement and expansion.

Increasing the size of the dataset with more varied Sindhi materials to enhance the generalization of the model.

Investigating the more powerful transformer architectures like XLM-RoBERTa and domain-specific language models to provide further performance improvements.

The inclusion of semantic ranking mechanisms to rank the extracted keywords according to their relevance and importance in the use of semantic tools in SEO.

Creation of the real-time keyword extraction systems that could be incorporated into the content management systems and SEO tools.

Generalizing the structure to other low-resource languages, making it possible to extract keywords in multiple languages, as well as optimizing cross-linguistic search engines.

Subsequent studies can also be done to enhance the quality of annotated data and create benchmark data in Sindhi NLP work, which will also consolidate research in the field.

References:

[1] Rubab Roshan, Irfan Ali Bhacho, "Comparative Analysis of TF–IDF and Hashing Vectorizer for Fake News Detection in Sindhi: A Machine Learning and Deep Learning Approach," *Eng Proc*, vol. 46, no. 1, p. 5, 2023, doi: https://doi.org/10.3390/engproc2023046005.

[2] Ali Nawaz, Muhammad Nawaz, "TPTS: Text Pre-processing Techniques for Sindhi Language," *Pakistan J. Emerg. Sci. Technol.*, vol. 4, no. 3, pp. 1–12, 2023, doi: 10.58619/pjest.v4i3.89.

[3] Irum Naz Sodhar, Muhammad Ibrahim Channa, Akhtar Hussain Jalbani, Dil Nawaz Hakro, "Identification of Issues and Challenges in Romanized Sindhi Text," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 9, 2019, [Online]. Available: https://thesai.org/Downloads/Volume10No9/Paper_29-Identification_of_Issues_and_Challenges_in_Romanized_Sindhi_Text.pdf

[4] Fatma Sezer Çırakoğlu, Özgün Koşaner, "Linguistic challenges in regional language SEO," *Telemat. Informatics Reports*, vol. 16, p. 100169, 2024, doi: https://doi.org/10.1016/j.teler.2024.100169.

[5] Feng Liu, Xiaodi Huang, "Performance Evaluation of Keyword Extraction Methods and Visualization for Student Online Comments," *Symmetry (Basel).*, vol. 12, no. 11, p. 1923, 2022, doi: 10.3390/sym12111923.

[6] Mohammed Abubaker, Hamza Sattuf, Bilal Babayigit, "BERT-based Models for Keyword Extraction from Arabic Scientific Articles," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 24, no. 10, 2025, [Online]. Available: https://dl.acm.org/doi/full/10.1145/3761805

[7] "(PDF) Deep learning based transformers for Keyword extraction." Accessed: Apr. 12, 2026. [Online]. Available: https://www.researchgate.net/publication/378923631_Deep_learning_based_transformers_for_Keyword_extraction

[8] M. K. Pasupuleti, "Multilingual NLP for Low-Resource Languages Using Transfer Learning," *Int. J. Acad. Ind. Res. Innov.*, vol. 05, no. 05, pp. 452–461, May 2025, doi: 10.62311/NESX/RPHCR7.

[9] "SEO Challenges and Strategies for Multilingual Websites | Cademix Institute of Technology." Accessed: Mar. 03, 2026. [Online]. Available: https://www.cademix.org/seo-challenges-and-strategies-for-multilingual-websites/

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

[11] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic Keyword Extraction from Individual Documents," *Text Min. Appl. Theory*, pp. 1–20, Mar. 2010, doi: 10.1002/9780470689646.CH1;PAGEGROUP:STRING:PUBLICATION.

[12] Partha Pakray, Alexander Gelbukh, "Natural language processing applications for low-resource languages," *Nat. Lang. Process.*, vol. 31, no. 2, 2025, [Online]. Available: https://www.cambridge.org/core/journals/natural-language-

processing/article/natural-language-processing-applications-for-lowresource-languages/7D3DA31DB6C01B13C6B1F698D4495951

[13] Raja Vavekanand, Bhagwan Das & Teerath Kumar, "DAugSindhi: a data augmentation approach for enhancing Sindhi language text classification," *Discov. Data*, vol. 3, no. 22, 2025, [Online]. Available: https://link.springer.com/article/10.1007/s44248-025-00040-8

[14] Dipendra Yadav, Sumaiya Suravee, Tobias Strauß, Kristina Yordanova, "Cross-Lingual Named Entity Recognition for Low-Resource Languages: A Hindi-Nepali Case Study Using Multilingual BERT Models," *MRL 2024 - 4th Work. Multiling. Represent. Learn. Proc. Work.*, 2024, [Online]. Available: https://aclanthology.org/2024.mrl-1.12/

[15] Priyaranjan Pattnayak, Hitesh Laxmichand Patel, Amit Agarwal, "Tokenization Matters: Improving Zero-Shot NER for Indic Languages," *arXiv:2504.16977*, 2025, [Online]. Available: https://arxiv.org/abs/2504.16977

[16] Fida Ullah, Alexander Gelbukh, "Enhancement of Named Entity Recognition in Low-Resource Languages with Data Augmentation and BERT Models: A Case Study on Urdu," *Computers*, vol. 13, no. 10, p. 258, 2024, doi: https://doi.org/10.3390/computers13100258.

[17] D. N. H. K.-U.-R. K. Z. B. Nazish Basir*, "Leveraging Machine-Labeled Data and Cross-Lingual Transfer for NER in Urdu and Sindhi," *J. Inf. Commun. btn btn-dark btn-xs btn-round)*, vol. 19, no. 1.

[18] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Apr. 20, 2025. [Online]. Available: https://arxiv.org/abs/1810.04805v2

[19] N. Vaswani, A., Shazeer, N., Parmar, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5998–6008, 2017.

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv:1907.11692*, 2019, [Online]. Available: https://arxiv.org/abs/1907.11692

[21] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, "Unsupervised Cross-lingual Representation Learning at Scale," *arXiv:1911.02116*, 2020, [Online]. Available: https://arxiv.org/abs/1911.02116

[22] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Text," 2004. Accessed: Jun. 05, 2025. [Online]. Available: https://aclanthology.org/W04-3252/

[23] Ricardo Campos, Vítor Mangaravite, "YAKE! Keyword extraction from single documents using multiple local features," *Inf. Sci. (Ny).*, vol. 509, pp. 257–289, 2020, doi: https://doi.org/10.1016/j.ins.2019.09.013.

[24] Corina Florescu, Cornelia Caragea, "PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents Corina Florescu, Cornelia Caragea," *ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, 2017, [Online]. Available: https://aclanthology.org/P17-1102/

[25] Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, Thomas Wolf, "Transfer learning in natural language processing," *Proc. 2019 Conf. North*, 2019, [Online]. Available: https://aclanthology.org/N19-5004/

[26] Shijie Wu, Mark Dredze, "Are All Languages Created Equal in Multilingual BERT?," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, 2020, [Online]. Available:

https://aclanthology.org/2020.repl4nlp-1.16/

[27] Guillaume Lample, Alexis Conneau, "XLM-R: Cross-lingual language model pretraining," *arXiv:1901.07291*, 2019, [Online]. Available: https://arxiv.org/abs/1901.07291

[28] Viktor Hangya, Hossain Shaikh Saadi, Alexander Fraser, "Improving Low-Resource Languages in Pre-Trained Multilingual Language Models," *Proc. 2022 Conf. Empir. Methods Nat. Lang. Process.*, 2022, [Online]. Available: https://aclanthology.org/2022.emnlp-main.822/

[29] Partha Pakray, Alexander Gelbukh, "Natural language processing applications for low-resource languages," *Nat. Lang. Process.*, vol. 31, no. 2, pp. 183–197, 2025, doi: 10.1017/nlp.2024.33.

[30] Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, Dietrich Klakow, "A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios," *Assoc. Comput. Linguist.*, 2021, [Online]. Available: https://aclanthology.org/2021.naacl-main.201/

[31] "Multilingual SEO Guide 2026: Ranking Across Languages |." Accessed: Apr. 12, 2026. [Online]. Available: https://phrase.com/blog/posts/multilingual-keyword-research/

[32] "Sindhi Kawish Articles Gallery URLs Dataset." Accessed: Mar. 03, 2026. [Online]. Available: https://www.kaggle.com/datasets/zulqarnainchanna/sindhi-kawish-articles-gallery-urls-dataset/data