Web Content Analysis and Design Trends Identification through Automated Web Scraping

K Kushal , Mohith Krishna V , M Pranavkrishnan, *Ms. Sangita Khare
Dept. of Computer Science and Engineering
Amrita School of Computing, Bangalore, Amrita Vishwa Vidyapeetham, India
bl.en.u4cse21078@bl.students.amrita.edu, bl.en.u4cse21125@bl.students.amrita.edu,
bl.en.u4cse21113@bl.students.amrita.edu, k_sangita@blr.amrita.edu

***Abstract*—This project uses automated online scraping methods to find popular design trends by methodically analyzing web information. It concentrates on important components including UI elements, color palettes, and fonts. This study aims to provide a thorough insight of modern web design techniques by gathering a large amount of data from a variety of websites. After that, the gathered data is visualized using a variety of methods, such as graphs and charts, to show these design tendencies in an understandable way. Web designers, marketers, and UX experts may quickly grasp and utilize the analysis's conclusions thanks to this method. In essence, this study uses data analysis and automated site scraping to provide a thorough understanding of online design. It not only reveals current design trends but also gives industry professionals insightful information that helps them make visually appealing and user-centered web experiences.**

Keywords—Web content analysis, Design trends, Automated web scraping, Font preferences, UI elements, Dataset collection, UX design, Current user expectations

I. Introduction

In the rapidly evolving digital world, the role that website design plays in attracting and retaining user interaction is crucial. This research uses python packages like Beautiful Soup and Selenium for automated web scraping. The main objective is to do a methodical examination of web content, scraping essential design elements from a wide range of websites. These elements include links, typeface families, colour schemes, and a plethora of other user interface components that work together to provide the visual and interactive experience that characterises online platforms.

A vast array of information is retrieved and examined, it includes basic components like the indexing status and page loading time, as well as more subtle ones like background colour, font faces, and the proportions of text, photos, and videos on web pages. This thorough analysis allows for a comprehension of the dominant design trends in the digital space.

For a variety of reasons, SEO (Search Engine Optimization) optimization is crucial in today's digital environment. It helps in increasing a website's exposure in search engine results pages (SERPs) is mostly dependent on SEO optimization. A website that is optimized for search engines is more likely to appear higher in the results when potential users look for pertinent products, services, or information. More organic, targeted traffic as a result of this enhanced visibility presents chances to draw in and interact with a larger audience.

SEO optimization improves usability and user experience. Improving a website's content quality, navigation, and structure are frequently necessary for effective optimization. Not only do these enhancements satisfy search engines, but they also improve user experience, decreasing bounce rates and raising conversion or engagement rates.

This study's main goal is to compile thorough design and user interface (UI) insights that may be used to help construct websites with design components that are closely aligned with those seen on websites that rank highly. Additionally, by carefully examining keywords, this study aims to optimize information posted on many websites, making it more visible and relevant in internet searches. This study attempts to offer useful advice for site developers and content producers looking to improve user experience and optimize the impact of their online presence by looking at design principles and keyword optimization tactics.

II. Related Work

A thorough investigation of web scraping and data analysis is presented by David Mathew Thomas et al. [1]. Using Python programs and the web crawler Scrapy, the writers pull data from a variety of sources, including Reddit and other websites. Extensive testing confirms that the extracted data is relevant, and the findings are shown in the form of a pie chart that highlights the most popular material on the tested website. Their combined effort tackles issues related to the retrieval of data from hidden webs and suggests a comprehensive fix. This includes an efficient data extraction method based on site structure, indexing, query processing, and an intuitive search interface. A prototype for completely automated and domain-dependent data extraction and integration behind search forms is the result of their combined efforts.

The study conducted by Dr. Anand Singh Rajawat et al. [2] focuses on problems associated with handling and analyzing big amounts of multidimensional data, or "raw data sets." The methodology they suggest entails creating a model for intelligent information personalization that is grounded in the Semi-Supervised Support Vector Machines (S3VM) idea. Improving user preference recommendation models is the aim, especially when it comes to the categorization and suggestion of online newspaper content. The suggested technique is simulated and evaluated by the authors using the machine learning program WEKA.

All stages of the data mining cycle—from data collection and pre-processing to pattern identification, assessment, and real-time application—are covered in this thorough analysis on Web personalization as a data mining application. The method emphasizes adaptability by using a variety of data sources and newly found models in an automated personalization system. Data preparation, integration from numerous sources, and different pattern finding approaches including clustering, association rule discovery, sequential pattern mining, Markov models, probabilistic mixture, and hidden variable models are among the tasks and strategies covered. Hybrid data mining frameworks and recommendation algorithms for tailored content distribution are also investigated. The paper, written by Bamshad Mobasher, [3] ends by outlining problems and the need for fixes to create the next wave of efficient and scalable Web personalization and recommender systems.

The work by Estrella Díaz et al. [4] is about the assessment of hotel chain websites' persuasive elements. The study uses latent class segmentation to further categorize hotel chains based on the persuasiveness of their websites after classifying them as luxury, midscale, or economy. A sample of 229 hotel chain websites is used to assess the six elements of online persuasiveness: informativeness, usability, credibility, inspiration, participation, and reciprocity. The findings show notable variations in persuasiveness between hotel types, with upscale accommodations often receiving higher marks. Using latent class cluster analysis, the study divides hotel chains into four groups according to persuasiveness factors. These clusters include engaging hotel chains, informative hotel chains with promotion efforts, reputable and easy-to-navigate hotel chains, and hotel brands that are rarely compelling.

Shilpa Chaudhari et al.'s study [5] suggests an algorithm for online scraping and ingredient-based recipe searches. The difficulty of keeping a balanced diet in the face of a hectic schedule and erratic eating patterns is what spurred the research. They discuss how hard it is for chefs to prepare dishes when they don't know the amounts and negative effects of diet ingredients. To extract and save recipe details, the suggested approach makes use of Python, MongoDB, and web scraping. A smartphone and online application for a "smart chef" is created, enabling users to keep up a balanced diet with an assortment of recipes. Recipe websites are scraped, extracted recipes are stored in a MongoDB database, data is retrieved for a smart chef application, and an interactive interface allowing users to seek recipes based on particular ingredients is implemented. The paper's conclusion discusses the performance of the web app and the results of web scraping recipe websites. It also makes suggestions for future improvements to the web app's performance using machine learning algorithms for automated web scraping and recipe searches.

The study, written by Sumit Sakarkar et al. [6], focuses on web use mining and web scraping methods for web personalization. The main goal is to assess suggestions for learners who are actively using the resources by looking at their usage habits, history, and preferred methods of learning. In order to give consumers with relevant and convenient information based on their preferences, the research employs ideas such as web scraping and web use mining for personalized web search results. The authors go over the tasks that the suggested system needs to do, such as data scraping, data analysis and preprocessing, result classification based on user requirements, and using appropriate algorithms to improve system performance. The report also presents the addition of speech recognition to improve the system's usability. The authors acknowledge the difficulties and constraints associated with online personalization, but they also stress how important it is for producing consistent and fulfilling user experiences. In order to improve users interactions with online resources, the suggested system's overall goal is to effectively extract, preprocess, and personalize web material for users.

A three-phase process for extracting and evaluating product data from e-commerce websites—Flipkart and Snapdeal in particular—is presented in the study by V. Srividhya et al. [7] using Python modules (requests and BeautifulSoup4), the method incorporates web scraping to extract product details and save the data in CSV format. Numpy helps with statistical data processing, enabling operations like mean, standard deviation, and more. Bar charts are used to accomplish visualization; they show the frequency of Apple iPhone pricing, the price ranges of various items, and a comparison of the prices on Flipkart and Snapdeal. The findings show that Flipkart typically has lower pricing. In handling large amounts of unstructured data, the research emphasizes the usefulness of online scraping and visualization tools.

In this study, Thulasi Bikku et al. [8] address the difficulties in evaluating extensive and insightful datasets for use in decision-making systems. The authors suggest using Hadoop, a Map-Reduce framework created for big data processing efficiency, to get over computational errors in large-scale dataset analysis. In addition to identifying problems with class imbalance and dimension reduction in typical large data classification techniques, the study presents a strategy that combines rough-set feature selection with Hadoop-based classification models. The usefulness of the suggested technique is demonstrated by experimental findings, which were done on a Hadoop framework utilising several classification models, including Genetic Algorithm, Neural Networks, ELM-Tree, and Random Forest. In comparison to conventional models, the study emphasizes the improvement in true positive rates and decrease in time complexity for intrusion detection. All things considered, the study offers insightful information about decision-making models and feature selection in the context of Hadoop-based large data processing.

The importance of recommender systems in enhancing consumer buying experiences is the main topic of the study written by Saravanan S. et al. [9]. The authors compare the effectiveness of Spark and Hadoop-based recommendation systems while introducing improved MapReduce-based content-based recommendation algorithms and data preparation. The research shows that Spark-based content-based recommendation algorithms digest information more quickly than Hadoop-based algorithms through tests on an Amazon co-purchasing network metadata dataset. A graphical user interface (GUI) for using the recommender system is also provided by the authors. Ubuntu 16.04, 8 GB RAM, and an Intel Core i7 CPU make up the experimental configuration. The outcomes demonstrate how much better Spark performs in terms of suggestion generation time.

Dr. Tripty Singh et al. [11], explore the current design frameworks in several sectors, emphasizing the difficulties brought forth by fresh and machine-generated data. The integration of Apache Hadoop into contemporary data architecture is the main topic of discussion, along with its function in distributed computing and large-scale dataset querying for dependable and quick results. The authors discuss the exponential rise of data, projecting that it would reach 40 Zeta Bytes by 2020, with new data kinds responsible for a large share of this increase from 2.8 Zeta Bytes in 2012. The article highlights Hadoop's potential for managing unstructured and semi-structured data by offering distributed processing, storage, and scalable analytics.

Software requirements specifications (SRS) documents are essential to the software development life cycle. In the paper by Kici D at al. [12] discuss the difficulties in correctly interpreting and categorizing SRS documents. Because SRS papers are frequently written in colloquial language, subjectivity and ambiguity may be introduced, which increases the risk of project failure, cost overruns, and schedule delays. The research suggests a transfer learning strategy for better multi-class text categorization on SRS data that makes use of the DistilBERT language model in order to address these problems. This method is evaluated on two datasets: PROMISE-NFR and DOORS Next Generation. It is compared against various deep learning techniques, such as LSTM and Bi-LSTM. The experimental results indicate the potential of transfer learning to improve text classification in software engineering domains with minimal labelled data, as DistilBERT performs exceptionally well, especially in tasks linked to Priority and Severity classification. In addition, the work provides intriguing avenues for future research, such as investigating domain-specific pre-trained models for SRS classification tasks and doing more in-depth error analysis. These approaches could greatly enhance the productivity of software development projects by automating software development processes.

Effective multi-label text classification models are critical to research and educational institutions' identification of scientific papers linked with the 17 Sustainable Development Goals (SDGs) of the United Nations 2030 Agenda. In this study, by RC Morales-Hernandez et al. [13] the performance of two models—a transfer learning model employing DistilBERT and a standard combination of Label Powerset (LP) and Support Vector Machine (SVM)—is compared for this task. The assessment makes use of a dataset that includes 31,434 scientific publications from the Organic Agriculture 3.0 field that have been labelled with the 17 SDG classifications. The models are tested in five distinct dataset scenarios with varied label balances. The outcomes show that, even in situations where the data is unbalanced, LP-SVM is still a competitive option for multi-label text categorization. DistilBERT works similarly, although some institutions have trouble with its resource-intensive design. The paper highlights the intricacy of pre-trained models and proposes further research on hyperparameter adjustment to enhance performance, as well as the possibility of employing this dataset for knowledge management and data mining.

This work tackles the major problem of automatic text classification in the context of the Internet of Things (IoT) and the large amount of textual data produced by Web 2.0 platforms and IoT devices. Using a variety of datasets, R Silva Barbon et al. [14] evaluate the performance of two cutting-edge language models, BERT and DistilBERT, in the context of text classification tasks for the English and Brazilian Portuguese languages. Their results show that DistilBERT, a more condensed and resource-efficient model, has significant advantages over BERT and performs on par with it in text categorization. DistilBERT is a strong option for situations with constrained computational resources because it is around 40% smaller and finishes training about 45% faster. Notably, it preserves approximately 96% of language comprehension abilities, which is especially noticeable in datasets that are balanced. This work highlights the great potential that lightweight models such as DistilBERT have in facilitating successful and economical text classification, thus demonstrating the adaptability and usefulness of transfer learning approaches in the field of data analysis and natural language processing.

This study by Qasim et. al. [15] explores the field of text classification, an important task with applications in content filtering, marketing, healthcare, and entertainment. Utilising the capabilities of machine learning and natural language processing (NLP), the research focuses on categorizing textual data produced on social media sites. For testing, three different datasets are used: COVID-19 related false news, COVID-19 English tweets, and extremist-non-extremist material. Using a variety of cutting-edge models, including BERT, RoBERT, DistilBERT, and others, the study investigates the efficacy of transfer learning strategies by evaluating how well they perform in text categorization tasks. Metrics including accuracy, precision, recall, and F1-score are used to assess the models. The results show that transfer learning classifiers are superior, with some models, such as RoBERT-base and BERT-base, obtaining exceptionally high accuracy. This paper highlights future research topics for further investigation in the discipline and illuminates the potential of transfer learning in text classification.

The gap in all the papers in the literature review was the lack of specific datasets related to the domain of web design analysis, the lack of proper data visualizations when it comes to design and UI features is also apparent. The papers also talk about the possibility of developing domain-specific pre-trained models or fine-tuning existing models for specific industries or fields could lead to even more accurate and context-aware text classification systems.

Through this research, we aim to counter those shortcomings by making the specific datasets necessary, for both visualization of design and UI features. For fine-tuning existing models to identify and classify SEO keywords we have for our field, a new dataset generated via web-scraping and collecting keywords. Creation of these datasets have ensured that the insights generated via this research can be used as a blueprint for web designers.

III. Proposed Methodology

A. Methodology for Data Analysis

The proposed methodology uses Python library such as BeautifulSoup, Selenium package for automated web scraping, which will be used for systematic web content analysis and design trends identification. Important design components like as colour-palettes, font-families , links and other user interface elements are extracted from a wide variety of websites. This approach seeks to identify popular design trends, providing insightful information on font choices, theme selections, and the frequency of user interface elements in web content. To evaluate different websites, a web scraping program is used.

1) Dataset Description

The dataset is a web scraped data which includes data, covering a wide range of topics including travel, health, e-commerce, and more. The dataset has 789 instances in total and 19 distinct attributes providing unique features of the popular websites. Table 1 displays the extracted features from these websites which contains the feature and the description.

TABLE .1 EXTRACTED FEATURES

Feature	Description
Indexing Status	Verifies whether canonical tags are present and whether the website is permitted in robots.txt.
Page Loading Time	This indicator shows how long it takes a page to load.
Background-Colour	This function returns the website's background colour.
Font Faces	This utility extracts font families from websites.
Percentages of pictures, Videos, and Text	Determines the proportion of pictures, videos, and text on the page.
Links both Internal and External	Indicates how many links both internal and external there are.
Domain	The domain name of the website is listed
SSL Certificate	Verifying if HTTPS is being used by the website and gives information about the SSL certificate's state.
Secure Protocol	This data is extracted to evaluate the website's security setup.
SSL Issuer	This data facilitates the identification of the Certificate Authority (CA) in charge of certificate validation and issuance.
Keywords	This data includes the various kinds of keywords extracted from a website.

2) Data Preprocessing

A thorough data preprocessing was done on the Dataset. Which includes removal of missing data. Filling in the missing values using the median of the feature . Furthermore, suitable default values were used to fill in the missing values in other category fields, such as domain, SSL Certificate, and others. To comprehend the distribution of numerical properties, descriptive statistics were calculated. In addition, data types were transformed correctly to guarantee correctness and consistency.

B. Methodology for SEO optimization Analysis

DistilBERT is a variant of the popular Transformers Bidirectional Encoder Representations (BERT) model that is smaller and faster while retaining much of the performance of BERT. It was introduced by Hugging Face researchers in his 2019 to make large transformer models more accessible to use on devices with limited computing resources.

DistilBERT is a compact version of BERT with approximately 40 times fewer parameters, making it suitable for resource-constrained devices like mobile and edge devices. It also reduces training time significantly. While it doesn't match BERT's performance in some tasks, it strikes a good balance between model size and task performance. Pre-trained DistilBERT models can be fine-tuned for specific NLP tasks and are easily accessible through the Hugging Face Transformers library. In essence, DistilBERT offers efficiency in size and training time without sacrificing practical NLP performance, making it valuable for various applications in constrained environments. The Figure 1 describes the model architecture for the LLM model.

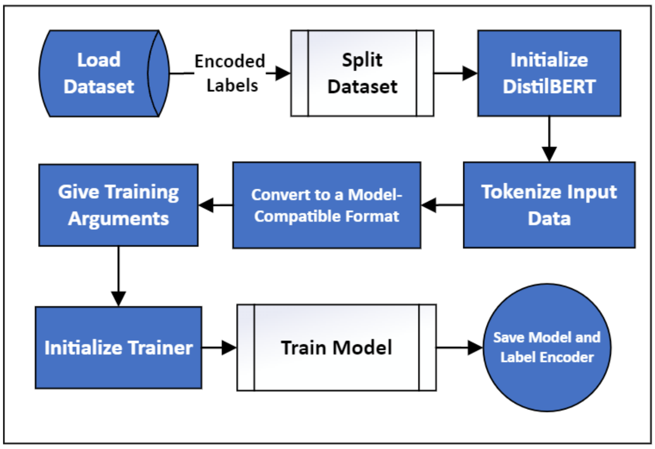

Fig. 1. LLM Model Architecture

1) Dataset Description

The dataset is a novel dataset, comprising of keywords scraped from hundreds of websites and existing SEO keywords. It has two columns, namely 'Keyword' and 'Label.' The keyword column contains all the optimized words on that website and the label contains information on the type of website it is. The dataset contains over 12000 instances over 8 kinds of labels (E-commerce, Ed-tech, Finance, News, Health, Streaming, Travel, social media) This ensures that good SEO insights can be generated across multiple types of websites.

2) *Dataset Preprocessing*

The processing steps include removing Null Values from dataset to improve the accuracy of our model. Removing Stop Words that are repetitive and lower the effectiveness of our model. Adding Labels manually as websites don't have labels inbuilt

C. *Data Storage*

The retrieved Datasets are then stored using Hadoop and PySpark to store the scraped data in a distributed and scalable manner throughout the data storage stage of our web scraping and analysis workflow. PySpark is a Python library for Apache Spark that is used because it can easily interface with the Hadoop Distributed File System (HDFS) and handle big datasets in a distributed and parallel fashion. The gathered information which includes link counts, design elements, indexing status, and page load times is converted into an organized format that can be saved and used for additional study.

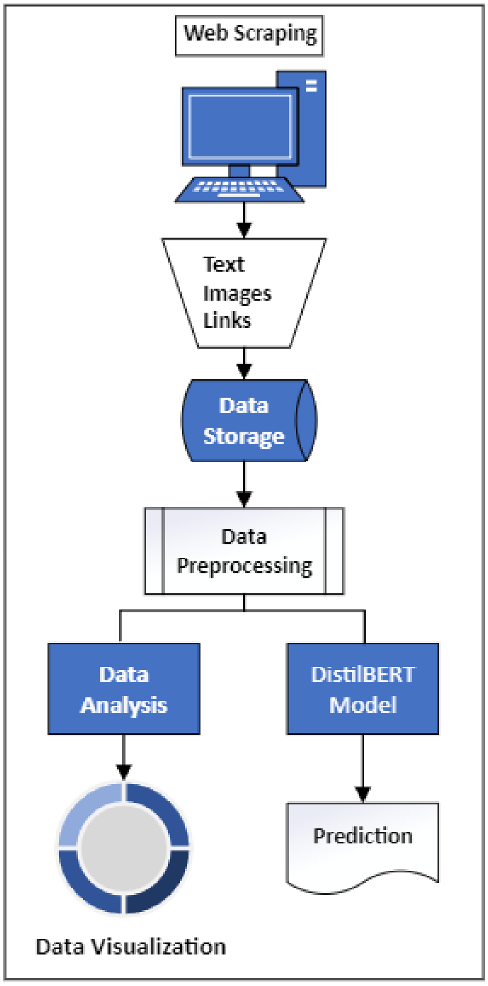

Fig. 2. Proposed Methodology

IV. Results And Discussion

The results present a detailed analysis derived from the data generated in this study, the valuable insights of various aspects of website design and performance are included. The following are the key findings trends, patterns, and considerations observed in the dataset.

A. *Data Analysis*

1) Popularity of Colour: The bar graph in Figure 3 reveals significant insights into the top 10 background colors used on websites, offering valuable considerations for designing a new website.

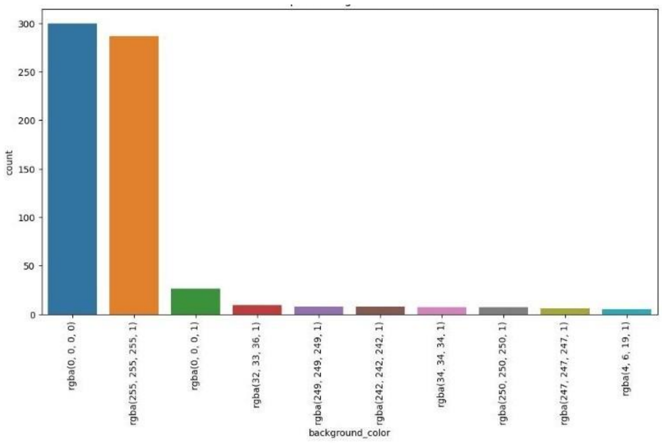

Fig. 3. Top 10 Background colours

a) Colour Preferences: Two colours stand out as significantly more popular, likely neutral tones such as whites and light greys for readability and navigation.

b) Neutral Tones: The prevalence of neutral colours suggests a preference for clean and minimalist aesthetics, imparting a modern and professional look.

c) Website Speed: From the websites analysed, the website loading time is also dependent on the website's background. It is observed that a plain background is preferred in comparison to complex background patterns.

2) Page Load Time Variability

The box plot in the figure 4 shows significant variability in page load times across different types of websites. E-commerce sites, in particular, show a wide range of page load times, indicating that there are both very efficient and very slow-loading e-commerce sites.

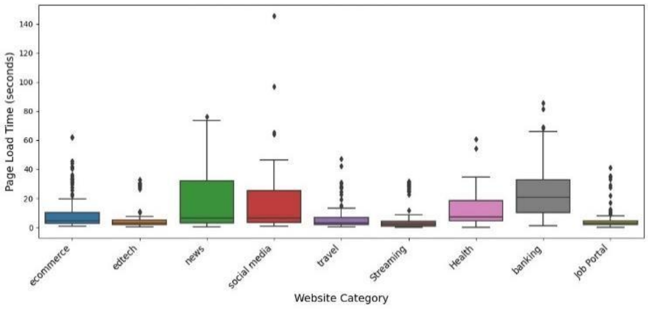

Fig 4 Box Plot of Page Load Time by Website Category

a) Performance Comparison :Some website categories, like news and health, appear to have more consistent and generally lower page load times. This suggests that these sectors may have optimized content delivery or less bulky site designs.

b) Design and Content Considerations: For design, the choice of colors and fonts should be consistent with the brand and easy on the eyes. Additionally, balancing text, images, and videos is essential to maintain an engaging yet fast-loading site.

c) SEO Considerations: The box plot shows that even within the same category, sites can have different load times, which may influence their search engine ranking.

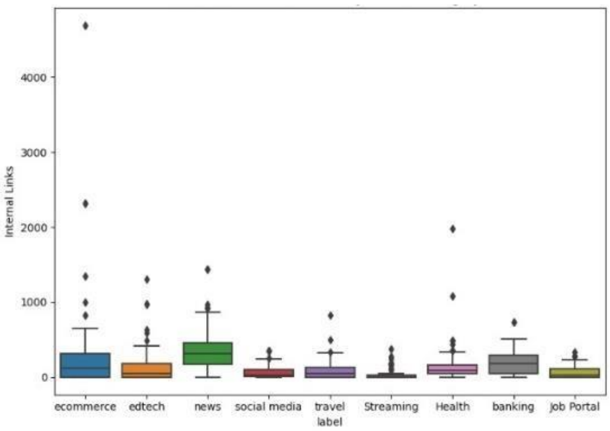

Fig. 5. Internal Links Distribution by Website Category

The dataset presents a varied assortment of websites in several categories, hence providing valuable perspectives on crucial facets of online development such as design, performance, and security. A noteworthy finding is the considerable variance in the use of internal links between the various categories, with news and e-commerce websites using them at a higher rate as shown in Fig. 5. Although internal links aid in effective navigation and search engine optimization, a careful balance must be struck to prevent overwhelming consumers. It is advised to put in place a planned internal linking strategy to maximize user experience without sacrificing SEO advantages.

Websites with quicker page loads are more likely to keep people on them, since page load time has shown to be a significant determinant in user engagement. In order to keep up with the performance of the most popular websites, new websites should aim for a page load time of less than three seconds. Compressing pictures, minimizing CSS and JavaScript, and using content delivery networks (CDNs) are some suggestions for doing this. For best website speed, load times should be regularly monitored, particularly after adding new features.

Security is emphasized as a best practice in web development and is demonstrated via SSL certificates. SSL certificates are essential for protecting sensitive data in addition to fostering user confidence. Due to their handling of sensitive user data, industries including e-commerce, banking, and healthcare prioritize security when it comes to SSL certificate uptake, as seen by the distribution of certificates across various categories. Most of the top websites have SSL certificates as shown in Fig. 6. The general advice is to start using SSL right away to ensure online security and gain from SEO.

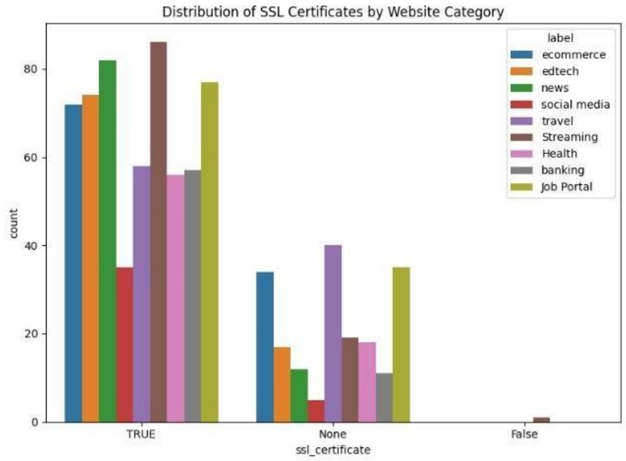

Fig. 6. Bar Graph Representing Distribution of SSL Certificates by Category

User attention and retention rates are impacted by font selections, which are an essential component of online design. To draw in users, it's important to consistently employ understandable typefaces that complement your brand. However, it's critical to strike a balance between readability and aesthetics and to take into account how font files affect how quickly pages load. Most websites use sans-serif font as shown in Fig.7. To guarantee the best readability and load speeds possible, websites should be tested across a variety of devices and browsers. This will help to increase user retention rates. Finally, this thorough examination offers insightful suggestions for the creation of new websites, including internal linking, load speed optimization, security precautions, and typeface concerns. Finally, this thorough examination offers insightful suggestions for the creation of new websites, including internal linking, load speed optimization, security precautions, and typeface concerns.

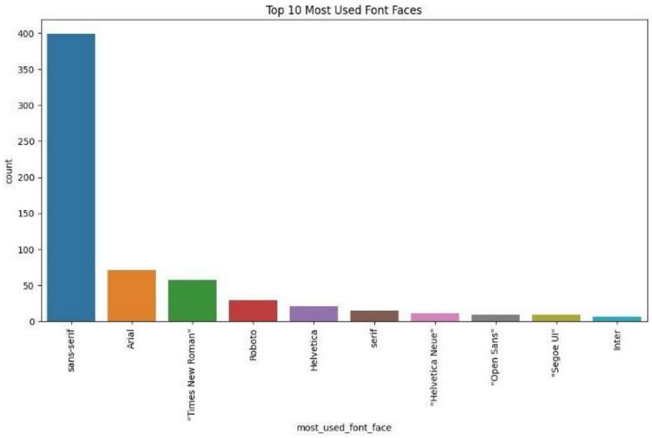

Fig. 7. Top 10 most used Font Faces

B. *Large Language Model Analysis*

TABLE .2 Extracted Features

Website	Label	SEO Optimization (%)
Flipkart	E-comm	72.97
Coursera	Ed-tech	32.11
Groww	Finance	44.21
Apollo Hospital	Health	83.33
Hotstar	Streaming	75.00

The table 2 shows the predicted SEO optimization for some top websites, the model's predictions are mostly reflective of the website type, as shown in the table. However, the model's effectiveness reduced when one website has multiple labels possible. The accuracy in the table 2 is a measure of how many keywords in a given website match with the given label to that website and gives a percentage score based on the predicted labels and given label.

V. CONCLUSION

In conclusion, the data analysis and SEO optimization study has given important new perspectives on a range of website performance, design, and SEO strategy-related topics. Important trends emerged from the data analysis, including the prevalence of neutral background colours for websites, the effect of design complexity on page load times, and the significance of internal linking for efficient navigation and search engine optimization. The disparity in page load speeds among various website categories highlighted the necessity of carefully weighing content and design components in order to strike a balance between loading speed and user engagement. DistilBERT's SEO optimization analysis demonstrated how well it could forecast SEO scores for various website categories, with a considerable degree of accuracy. But there were problems when websites featured more than one label.

The results highlight the significance of consistent tracking of website performance metrics, observance of security protocols such as SSL certificates, and careful selection of fonts. All things considered, this research provides useful information for the creation of websites by suggesting the best possible designs, functions, and SEO tactics for the ever-changing online environment.

The focus for future work will be to create a more robust SEO model, which can handle keywords that relate to more than one label. This will increase the model's usability as most websites are not limited to a single label these days. Another area of focus will be to increase the number of labels present within our dataset, making increasing the types and the number of websites we can analyze with our dataset. A larger dataset would also improve the overall reliability of the SEO Optimization model.

For data analysis of website design features, the aim is to get insights into more features and therefore helping new website creators create better websites that will generate traffic.

Overall, this study provides valuable insights into the popular content on the investigated websites and confirms the effectiveness of Python-based web scraping for informative and transparent data analysis.

REFERENCES

[1] Thomas, D.M. and Mathur, S., 2019, June. Data analysis by web scraping using python. In 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 450-454). IEEE.

[2] Rajawat, A.S. and Upadhyay, A.R., 2020, February. Web personalization model using modified S3VM algorithm for developing recommendation process. In 2nd International Conference on Data, Engineering and Applications (IDEA) (pp. 1-6). IEEE.

[3] Mobasher, B., 2007. Data mining for web personalization. In The adaptive web: Methods and strategies of web personalization (pp. 90-135). Berlin, Heidelberg: Springer Berlin Heidelberg.

[4] Díaz, E. and Koutra, C., 2013. Evaluation of the persuasive features of hotel chains websites: A latent class segmentation analysis. International Journal of hospitality management, 34, pp.338-347.

[5] Chaudhari, S., Aparna, R., Tekkur, V.G., Pavan, G.L. and Karki, S.R., 2020, July. Ingredient/recipe algorithm using web mining and web scraping for smart chef. In 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) (pp. 1-4). IEEE.

[6] Sakarkar, S., Chaudhari, V., Gaurkar, T., Veer, A. and SCET, M.K., 2021, February. Web Personalisation based on User Interaction: Web Personalisation. In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) (pp. 234-238). IEEE.

[7] Srividhya, V. and Megala, P., 2019. Scraping and Visualization of Product Data from E-commerce Websites. IJCSE International Journal of Computer Sciences and Engineering.

[8] Bikku, T., Rao, N.S. and Akepogu, A.R., 2016. Hadoop based feature selection and decision making models on big data. Indian Journal of Science and Technology, 9(10), pp.1-6.

[9] KE, K., Balaji, A. and Sajith, A., 2018. Performance comparison of apache spark and Hadoop based large scale content based recommender system. In Intelligent Systems Technologies and Applications (pp. 66-73). Springer International Publishing.

[10] Subbulakshmi, S., Ramar, K., Omanakuttan, A. and Sasidharan, A., 2019. Automated Analytical Model for Content Based Selection of Web Services. In Advances in Signal Processing and Intelligent Recognition Systems: 4th International Symposium SIRS 2018, Bangalore, India, September 19–22, 2018, Revised Selected Papers 4 (pp. 309-321). Springer Singapore.

[11] Singh, T. and Darshan, V.S., 2015, October. A modern data architecture with apache Hadoop. In 2015 International Conference on Green Computing and Internet of Things (ICGCIoT) (pp. 574-579). IEEE.

[12] Kici, D., Malik, G., Cevik, M., Parikh, D. and Basar, A., 2021, May. A BERT-based transfer learning approach to text classification on software requirements specifications. In Canadian Conference on AI (Vol. 1, p. 042077).

[13] Morales-Hernández, R.C., Becerra-Alonso, D., Vivas, E.R. and Gutiérrez, J., 2022, October. Comparison Between SVM and DistilBERT for Multi-label Text Classification of Scientific Papers Aligned with Sustainable Development Goals. In Mexican International Conference on Artificial Intelligence (pp. 57-67). Cham: Springer Nature Switzerland.

[14] Silva Barbon, R. and Akabane, A.T., 2022. Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study. Sensors, 22(21), p.8184.

[15] Qasim, R., Bangyal, W.H., Alqarni, M.A. and Ali Almazroi, A., 2022. A fine-tuned BERT-based transfer learning approach for text classification. Journal of healthcare engineering, 2022.