Role of Images on World Wide Web Readability

by

Ehsan Elahi

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in

Computer Science and Technology

Universidad Carlos III de Madrid

Advisor(s):

Ana María Iglesias Maqueda

Jorge Luis Morato Lara

Tutor:

Ana María Iglesias Maqueda

March 2023

I would like to dedicate this thesis to my family, especially my parents.

ACKNOWLEDGEMENTS

To begin, I want to thank God, who has been there for me during the happiest and most difficult seasons of my life. I want to express my gratitude to everyone who assisted me in developing my thesis. I'm going to close my eyes now and concentrate on recalling the sequence of events that led up to this point. When I was 18 years old, one of my goals was to pursue my dream of earning a doctoral degree. It's incredible how our dreams may lead to real-life changes. But I believe this to be a partial narrative. My parents had a healthy respect for the scientific community, and they raised us to value it as well. I will never be able to repay my parents for all the love, assistance, and support they have provided me throughout my life. Ammi! I was whining about how quickly I'd lost my motivation for a while, and it went on for a time. Then you told me the story of the horse that happened a year ago, about how even the most challenging horse may fall but would immediately get back up and continue running. That narrative worked its way into my head like a spell. I am grateful to you for being there for me in both the good and the terrible moments. Abu! You taught me that I should only do what I think is correct and that I shouldn't feel bad about changing my mind if it turns out that the reverse is true. Thank you for that valuable lesson. You must be brave and honest to live a life free from conflict. In addition, I want to express my gratitude to my brother, my sister, and my friends for the assistance and encouragement they provided.

My academic advisers, Dr. Ana Iglesias and Dr. Jorge Luis Morato Lara, are someone I will always be grateful to. Ana and Jorge! Over the course of the past three years, you have shown incredible tolerance, support, and wisdom to me, and I will never forget it. Since our first meeting, when I asked how we were going to work on my Ph.D. together, you have been really responsive and helpful and you said, "No! Only talk about your fears about getting a Ph.D. at this meeting." I knew I was in good hands. Your passion for research and how much you enjoy it are amazing. Your touch and ideas will always be a part of my future research. For me, you are a great example of a smart, strong person. Getting back to the beginning, I want to say thank you for letting me in.

I'd like to thank the academic committee that oversaw my thesis's pre-defense and final defense. They worked very hard to make sure this work was presented in a way that was scientifically sound. Everyone, thanks!

PUBLISHED AND SUBMITTED CONTENT

- This section includes researcher's publications that are included wholly and expanded too in the thesis:
  - Elahi, E., Iglesias, A., & Morato, J. (2022). Readability of Non-Text Images on the World Wide Web (WWW). IEEE Access, 10, 116627-116634.
    https://doi.org/10.1109/ACCESS.2022.3218632
    Chapter: 1—4.
    The material from this source included in this thesis is not singled out with typographic means and references.
  - Ehsan Elahi, Ana Iglesias & Jorge Morato. Impact of relevant graphical contents on the web readability. Universal Access in the Information Society, August 2022.
    Chapter: 1—4.
    The material from this source included in this thesis is not singled out with typographic means and references.

- This section includes researcher's publications that are included partially in the thesis:
  - Elahi, E., Lara, J. L. M., & Maqueda, A. M. I. (2022). Image Relevance on Websites and Readability. In Information Systems and Technologies: WorldCIST 2022, Volume 1 (pp. 286-295). Cham: Springer International Publishing.
    https://doi.org/10.1007/978-3-031-04826-5_28
    Chapter: 2—4.
    The material from this source included in this thesis is not singled out with typographic means and references.
  - Elahi, E., Iglesias, A., & Morato, J. (2022, June). Readability of Graphical Contents on World Wide Web (WWW). In 2022 17th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-4). IEEE.
    https://doi.org/10.23919/CISTI54924.2022.9820011
    Chapter: 1—4.
    The material from this source included in this thesis is not singled out with typographic means and references.
  - Elahi, E., Iglesias, A., & Morato, J. (2022, October). Web Images Relevance and Quality: User Evaluation. In Proceedings of the 5th International Conference on Computer Science and Software Engineering (pp. 66-69).
    https://doi.org/10.1145/3569966.3569984
    Chapter: 2—3.
    The material from this source included in this thesis is not singled out with typographic means and references.
  - Elahi, E., Maqueda, A. M. I., & Lara, J. L. M. (2023). Web Readability Challenges. In Data Science and Algorithms in Systems: Proceedings of 6th Computational Methods in Systems and Software 2022, Vol. 2 (pp. 446-454). Cham: Springer International Publishing.
    https://doi.org/10.1007/978-3-031-21438-7_35
    Chapter: 2.
    The material from this source included in this thesis is not singled out with typographic means and references.

ABSTRACT

As the Internet and World Wide Web have grown, many good things have come. If you have access to a computer, you can find a lot of information quickly and easily. Electronic devices can store and retrieve vast amounts of data in seconds. You no longer have to leave your house to get products and services you could only get in person. Documents can be changed from English to Urdu or from text to speech almost instantly, making it easy for people from different cultures and with different abilities to talk to each other. As technology improves, web developers and website visitors want more animation, colour, and technology. As computers get faster at processing images and other graphics, web developers use them more and more. Users who can see colour, pictures, animation, and images can help understand and read the Web and improve the Web experience. People who have trouble reading or whose first language is not used on the website can also benefit from using pictures.

But not all images help people understand and read the text they go with. For example, images just for decoration or picked by the people who made the website should not be used. Also, different factors could affect how easy it is to read graphical content, such as a low image resolution, a bad aspect ratio, a bad colour combination in the image itself, a small font size, etc., and the WCAG gave different rules for each of these problems. The rules suggest using alternative text, the right combination of colours, low contrast, and a higher resolution. But one of the biggest problems is that images that don't go with the text on a web page can make it hard to read the text. On the other hand, relevant pictures could make the page easier to read.

A method has been suggested to figure out how relevant the images on websites are from the point of view of web readability. This method combines different ways to get information from images by using Cloud Vision API and Optical Character Recognition (OCR), and reading text from websites to find relevancy between them. Techniques for preprocessing data have been used on the information that has been extracted. Natural Language Processing (NLP) technique has been used to determine what images and text on a web page have to do with each other. This tool looks at fifty educational websites' pictures and assesses their relevance. Results show that images that have nothing to do with the page's content and images that aren't very good cause lower relevancy scores. A user study was done to evaluate the hypothesis that the relevant images could enhance web readability based on two evaluations: the evaluation of the 1024 end users of the page and the heuristic evaluation, which was done by 32 experts in accessibility. The user study was done with questions about what the user knows, how they feel, and what they can do. The results back up the idea that images that are relevant to the page make it easier to read. This method will help web designers make pages easier to read by looking at only the essential parts of a page and not relying on their judgment.

Keywords

Image readability, Website readability, Relevancy, Extraction, User evaluation

Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

Tim Berners-Lee created the World Wide Web (WWW) to make it easier for researchers and scientists to share information. Its only purpose was to provide information to them (Berners-Lee et al., 2001). As the Web grew, the World Wide Web Foundation and Tim realized that it would be better if everyone could use the internet without asking for permission or paying a fee. They also wanted to ensure everyone had access to technology and change government and business policies for the better. The Web played a significant role in making education better by giving students, teachers, researchers, and other people access to a lot of information on many different subjects as shown in Figure 1.1. Over time, the Web has changed into something new and valuable, becoming one of the best learning tools. With the invention of the World Wide Web, it is much easier to learn, and research can be done more accurately because the information is already there. A device that can connect to the internet is all that's needed. Since there is a lot of information on the internet, it is easier to research there than to look for or buy books or encyclopedias with the same information about a certain topic. This information is published in books, references, and articles that are very helpful for students and teachers. The World Wide Web could change many things about education because it gives people instant access to these vast data collections. Information that is up-to-date on environmental issues, political issues, and other topics that change quickly on both a local and a global scale can help students and teachers learn from them and explore their own interests.

The World Wide Web has many benefits, but it also has many drawbacks, such as the loss of web pages, pages that take too long to load, text that is too small or too big, incorrect spelling or grammar, difficulty locating relevant information, pointing users in the right direction to find relevant information, and then providing users with more information than they can comprehend (Gradisar et al., 2006). One of the most significant challenges involves understanding the information presented in the written language and items on the website. One of the most critical problems is understanding what is written on the website. Today, most websites are written in English. People whose first language is not English have difficulty reading websites because they need to learn a great deal of the vocabulary, grammar, composition, and structure of sentences, graphs that explain themselves, and use of abbreviations or scary content display (Yu et al., 2010). This causes many problems with the readability of websites. Web readability can be defined as “a combination of reading comprehension, reading speed and user satisfaction in terms of reading comprehension, dictionary, thesaurus and existing online tools and browser add-ons”. Readability may also be defined as “how easily a person can read and understand written materials”. Website readability is an indicator of the overall difficulty level of a website (Lau et al., 2006).

Pakistan is a country with many different languages. English and Urdu are the two official languages. Urdu is also the primary language of the country. Pakistan also has two major regional languages: Saraiki and Kashmiri. There are also four major provincial languages: Punjabi, Pashto, Sindhi, and Balochi. Since the 1990s, people in Pakistan have been able to use the Internet.

Pakistan has been following an aggressive IT policy to help it modernize its economy and build a software industry that can be exported. Even though English is an official language and is taught as a foreign language, it is not the most commonly spoken language, and only some people in Pakistan can speak it well. Only 10.9% of people in the country speak English. So, when people try to read English information on the web, they need help with web readability.

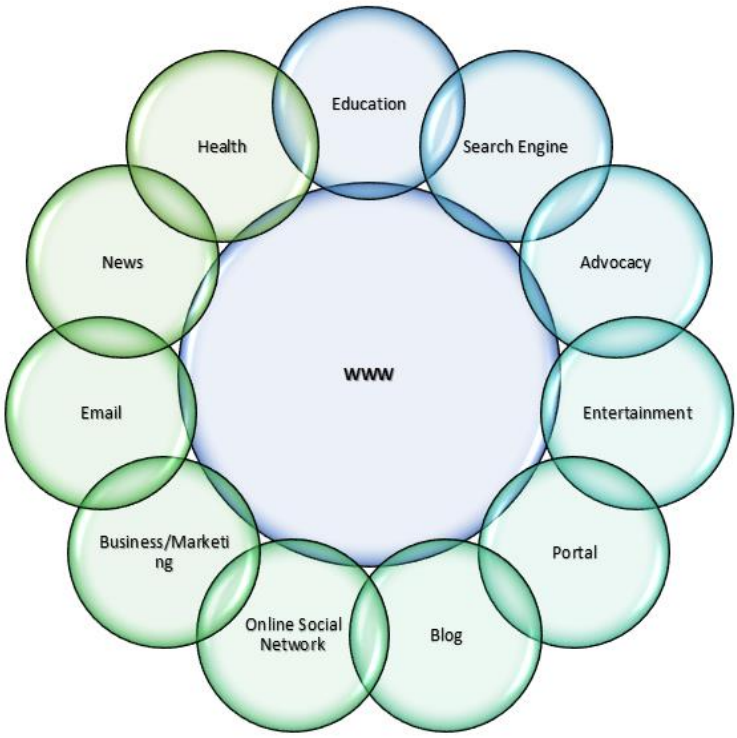


Figure 1.1 World Wide Web

Images are an important part of the website's content as a whole. Since we are all consumers in some way, it's easy to see how pictures can help us decide what we like and don't like. But you can't just go with your gut. There is a lot of data that shows why images are important on websites (Cyr et al., 2009):

- Almost 65% of people learn best by seeing
- About 28% of the words on a page are never read
- People remember only 20% of what they read but 80% of what they see
- Those who use visual communication to persuade are 43% more successful than those who only use words
- The brain can recognize images seen for as little as 13 milliseconds, which shows that vision helps us quickly identify ideas and decide where to put our attention when there are many choices

Images on websites are essential for more than just how our brains process them. Using images in web design also has several other pros:

- People on social media share content with pictures more often than they share content without images. 63% of the content on social media is visual, and almost half of all people who use the internet have shared a photo or video they saw online. Adding content images to your website makes it more likely that you will connect with and attract new users
- Using images on websites improves the user experience because they are easier to understand and remember
- Using search-friendly metadata and giving your web graphics good names and captions will make it easier for search engines to index your site, which will help your on-page SEO. Also, up to 27% of U.S. web searches are for images on Google. Therefore, using good images increases the chance that a searcher will end up on your website
- Images can bring up memories and ideas, affecting how we feel and connect with others. Seeing the right picture at the right time can make us do things we wouldn't have done otherwise
- Sites with pictures get almost twice as many views as sites without images

Not all images help people understand and read the text they go with. For example, images that are just for decoration or that were picked by the people who made the web page should not be used. Also, poor image resolution, bad aspect ratio, the wrong colour combination of the image itself, font size, etc., can make it hard to read images on the web. Different ways to fix these problems are suggested by the Web Content Accessibility Guidelines (WCAG) (Li et al., 2012). The rules say to use low contrast, additional text, the right combination of colours, and a higher resolution. But one of the most severe problems is that images that don't have anything to do with the text on a website can make the pages hard to read. When judging the readability of websites, the researchers only looked at the text and suggested different ways to measure it (Hall et al., 2004) (Patel et al., 2021). But relationships between the images and readability of the web pages have been never measured yet.

1.1 Motivation

Images are often easier to understand than just words (Miniukovich et al., 2019). It is well known that the brain can process images faster than it can process written or spoken information. Figuratively and literally, a picture says more than a bunch of words. We tend to pay more attention to pictures than we do to just a bunch of words. When put together well, an image can give you excitement and information in an instant that will stay with you much longer than words on a page. But images on a website can only have their full effect if they are used the right way (Xu et al., 2020).

But irrelevant images on the web also make it hard to read because they take the reader's attention away from what they're reading (Elahi et al., 2022A) (Elahi et al., 2022B). Most of the time, researchers only thought about the text on websites when judging how easy they were

to read. They suggested different tools for judging readability, such as FOG, SMOG, Automatic Readability Index, etc. (Ojha et al., 2021). But some past work has focused on how relevant images are on the web, and it has never been measured how the image affects how easy it is to read the web page. The thesis's main goal is to determine how relevant or low-quality images affect how easy it is to read a webpage. This study will assist web designers in improving readability by focusing only on the important parts of a page and not relying on expert opinion. To check if images are relevant, they came up with a new way to automatically figure out if they are relevant. This method combines different ways to get information from images and read text from web pages in order to find connections between them. This method is used to look at fifty different educational websites and figure out whether or not their pictures are relevant. With different kinds of questions, a user study has been done in Pakistan to test the proposed methodology. The results back up the fact that images that are relevant to the page make it easier to read.

1.2 Objectives

The main aim of this thesis is to propose and evaluate the relevancy of images on the World Wide Web (WWW) for educational websites. Therefore, the secondary objectives of the thesis related to the main aim are described as follow:

- To analyze the current solutions for measuring the relevancy of images for the websites.
- To analyze the relationship between the relevancy of the images and the website readability.
- To propose a measure of images relevancy and guidelines to increase the readability of websites.

1.3 Hypothesis

The main hypothesis of the thesis is that images could increase the readability of the web pages when images are relevant to the text of the web page. Images being used on the web page should contribute to conveying the context of the page in a more effective way. Only quality relevant images can play this role. Poor quality images and irrelevant images could negatively affect the readability of web pages. This hypothesis could be split into two sub-hypothesis:

- The use of images relevant to the text in the webpage could increase the readability of the web pages.
- The excessive use of images irrelevant to the text on the webpage contributes to poor readability.

1.4 Research Questions

To validate our hypothesis, we will answer the following research questions:

QR1**:** Is any program able to measure the image relevancy of the websites?

QR2**:** Do relevant images enhance web readability?

QR3: Do irrelevant images influence web readability?

1.5 Main Ph.D. Contributions

The expected contributions of my research work are:

- Study of image relevancy measures in websites.
- Analysis of the factors which could influence the readability of the images.
- Propose an image relevance measure and design a methodology to evaluate the relevancy of images on the web.

Chapter 2: Literature Review

In this chapter, we have conducted research into the various aspects that contribute to the readability of web pages. When evaluating the readability of websites, the vast majority of currently available readability metrics concentrate solely on textual aspects of the World Wide Web. But to our best knowledge no work on the image's relevancy on websites from a readability perspective.

2.1 Readability

Readability is a way to measure how well information is shared with a large group of people when they are trying to access it (Oydanich et al., 2022). Readability analysis is a growing area of research, with contributions from numerous dissimilar angles. Many people should be able to use the information on web pages. The pages need to be easy to read and understand for everyone to be able to use them. These things need to be taken into account, along with technical accessibility. The main goal of "Easy-to-Read" on the World Wide Web is to gather short, up-to-date suggestions and to bring attention to issues that people, particularly those with cognitive disabilities, are having (Ferrari et al., 2022).

Readability has been talked about a lot when figuring out how hard a text is to understand. Even though there isn't yet a final and fully representative way to measure readability that can provide computational criteria for figuring out how hard a text is to understand. Researchers came up with a way to rank readability using machine learning methods. This method was based on the idea that readability is based on how easy it is for a reader to understand text structures. In order to make an educated guess as to how challenging reading will be for individuals with minor intellectual disabilities, a Latent Trait Model is used to a subset of the factors gleaned from an experiment on reading. A device for automatically evaluating readability is used to accomplish this goal. A text may need to be simplified and made easy to read in order to be comprehended by people with cognitive disabilities because it may already be complicated and difficult to read (Gulbrandsen et al., 2022).

Researchers thought that crowds could be used to measure how simple a text is. To prove that the crowd may be used to judge how understandable a document is, we analyzed 2500 crowd annotations. For this reason, one of the studies aims is to develop a web-based automated system that may be used to rewrite written material so that those with intellectual disabilities can comprehend it. This system will take into account how hard it is for these people to understand written and spoken information. Researchers looked at Deaf people in a study. Because sign language is based on seeing things, it was hard for deaf people to understand text-based web documents. A system was made that turns complicated sentences into simple ones and shows Deaf people how they relate to each other with a graphic. This makes web documents easier to read (Man et al., 2022).

Researchers did research on the readability, accessibility, and web page rankings based on the results of various readability and accessibility tests. They also look for links between sites to find out how easy it is to get to, how easy it is to read, and how high a site ranks. Another study was done on the readability and accessibility of the homepages of Indian universities. For this, different evaluation tools like wave tool, achecker and Gunning fog were used (Ismail et al., 2018).

The Bangla language was used as a test subject to see if the English readability metrics could be used for other languages (a popular language spoken in India). It wasn't true, as it turned out (Sinha et al., 2014). Machine learning techniques such as regression, support vector machines, and support vector regression were used for this investigation. Readability has seen significant technical advancements in recent years, prompting Rebekah George Benjamin to assess the topic and provide recommendations for ongoing and future studies. This is because professionals in virtually every field of education seek methods for predicting a text's difficulty level (Benjamin et al., 2012).

A tool called GUI Evaluator is shown that uses metrics to evaluate the complexity of user interfaces based on their structure. Language's role in getting information has been looked at, and it seems that language may be a double barrier. Crawlers were used to get information about web hosts and links, and log file analysis was used to get information about website users. The findings were compared to the Revised Hierarchy Model and the Information Foraging Theory. When reviewing interfaces for people who are fully blind, a set of protocols is recommended to help professionals uncover issues and features. The goal is to find usability problems. By having users take part in the design process, interface designers should be able to figure out what users need (Alemerien et al., 2014).

2.2 Analysis of plain text readability

2.2.1 Classical measures

Table 2.1 shows a list of the readability tools and models we've looked at and studied. The models that different researchers made were made to figure out how easy it is to read different kinds of writing. We found that most of the formulas guessed the level of readability in terms of the US grading scale. The primary and secondary grade levels are different depending on where you live and what the environment is like. It is still not clear if these formulas work everywhere in the world. Researchers came up with the Fernandez Huerta Index, Djoko formula, Kandal and Moles Index, and Al-Heeti grade level as ways to predict how easy it would be to read Indonesian, Spanish, French, or Arabic text. The formula developed for non-English texts showed a correlation with people's reading abilities, but the scores from tools for the English language did not match what was expected (Fernández Huerta, 1959) (Biddinika et al., 2016) (François et al., 2012).

Most of the earlier readability formulas took into account things like the length and number of words, sentences, syllables, and complex words (Crossley et al., 2011). This means that even

nonsense could get a good readability score with these formulas. After the 1980s, tools like the Read-X, ATOS, Lexile Framework, Coh-metrix, and the new Dale-Chall readability formula were made. These tools measure readability by taking into account things like cognitive-structural elements, semantic units, and the complexity of syntactic structures.

The Lexile Framework is a popular way to figure out how hard a text is, but texts written in the 1980s are usually hard to understand. The Lexile Framework is a unique way to test how well you can answer questions about what you've read. In this framework, both the score for the reader and the lexile score for the text are made. The reader can answer comprehension questions correctly if he or she has a correct matching score for the text. It uses actual reading tests instead of age or grade levels to figure out how well someone can read. How well you understand what you read depends on how well you know the meanings of the words and how the sentences are put together. It looks at how often a word is used to determine its purpose and how long a sentence is to figure out how hard it is to put together (Stenner et al., 2023).

Table 2.1 Readability Tools and Models

Sr#	Tool/Model	Inferences
1	The Flesch Reading Ease Readability Formula	One of the right ways to judge a school text. The text is easier to read the higher the score, and it gets harder as the score goes down. Score is between 0 and 100.
2	Dale-Chall Read-ability Formula	The FRE-inspired Dale-chall score takes into account how hard the words are and how long the sentences are. A text with a score below 4.9 is easy for a US fourth-grader to understand, and a score above 10 is easy for a college graduate to understand.
3	Flesch-Kincaid Grade Level Readability Formula	Score is a modified version of FRE. It shows the level of education needed to understand text in the US.
4	FOG	This formula comes from research on daily newspapers and magazines. Text with a score of 7–8 on the Fog Index is thought to be ideal, while a score of 12 or higher is too tough for maximum users.
5	Forcast	Number of words with one syllable. Thought of as the perfect formula for text material with multiple-choice questions in the US. Strictly not to be used to judge books for young readers.
6	Fry	Used to make sure that regulatory purposes can be read easily. Most likely, the differences are within a single grading scale.
7	PSK	Number of the syllables, gl = grade level ra = Reading Age. This is the best way to find out what a sample text for a US grade level looks like. It works best for kids in elementary school, and kids older than 10 shouldn't use it.
8	Automatic Readability Index	Gives an estimate of the grade level needed to understand the text. For instance, a US grade level 1 is comprehensible by kid's ages 6 to 8, and a US grade level 12 is understandable by a 17-year-old. It is based on the number of words and characters.
9	CLI	Based on words rather than syllables, it takes about a US grade level to understand the text. A grade level of 10.6 is easy for students in 10th or 11th grade to understand, while a level of 14 is for students in college.
10	BRI	Using the Dale-Chall word list, BRI checks text samples for word knowledge (Dale and Chall, 1948). It's like the Dale-Chall formula. The main difference is that it utilizes characters instead of syllables and averages easy words instead of complex words.
11	LIX	It is the formula used to figure out how easy it will be to read a French text. A lix score between 20 and 25 is considered very easy, while a score of 60 is very hard.

12	Raygor Estimate Graph	The grade level is shown by the point where the X and Y axes meet. If the point where the lines meet is inside the parallel lines, the grade level is right. Grade level ranges between three and fourteen.
13	Djoko Formula	The Djoko formula, which is based on 13 parts of a text, is used to measure how easy it is to read Indonesian text (paragraphs, words, and sentences). The range of the criteria is based on the variance within easy and hard text.
14	Pisarek's Index	It's like the FOG index, which looks at how long a sentence is on average and how many complex words it has.
15	The Mistrik Formula	There are three things you need to know about a text to figure out how easy it is to read. As = Average number of syllables per word Av = Sentence length and how hard it is to understand I = the number of words
16	Fernandez Huerta Index	Still a popular way to figure out how easy it is to read Spanish text. It is a change from FRE. In its original form, the Huerta formula cannot be scaled.
17	Kandal & Moles Index	It is a change from the FRE text to the French text. Lp = Number of words on average in a sentence Lm stands for the average number of syllables in a word
18	Al-Heeti Grade Level	The score given by the Al-Heeti readability formula shows the grade level needed to understand an Arabic text.
19	SMOG	Dale-Chall predicts two grades higher than SMOG. It's thought to be good for kids in middle school. A text with a polysyllabic count of 1–6 is at grade level 5, while one with a count of 211–240 is at grade level 18.
20	Spache	Spache is similar to the Dale-Chall formula, but it doesn't work well for advanced texts (above grade 4).
21	Read-X	Read-X analyzes the readability of text on the web in real time. It does this by doing a web search, filtering the results by category level, and grouping the results by theme.

Using the reading assessment database and the massive book, Renaissance Learning Inc. and Touchstone Applied Science Associates Inc. came up with two formulas. The formulas are called ATOS for readability of books and ATOS for readability of text. Traditional variables for both formulas are the length of a word or sentence and the grade level of the words. The length of a book is one factor that affects how hard it is, so the formula for books needs to take that into account. When making the formula for matching books, the weaknesses of earlier formulas are taken into account. The following areas were found to need improvement and were fixed (Benjamin et al., 2012).

- An improvement has been made to a readability formula's semantic component when more words are added to the corpus.
- This method can use different kinds of texts that need to be made longer.
- Possible changes need to be thought about because some words are used more than once in the text.

Research has shown that the ATOS readability formula is an effective tool for assisting students in selecting books that are at the appropriate reading level for them. ATOS has been validated as a valid and accurate measure of text complexity because it considers the most important factors

that can be used to predict the difficulty of a text. Read-X is a web search program that allows you to locate and analyze reading content online. This program searches the web for text or a keyword given by the user. It pulls text from web pages without HTML code and checks how easy it is to read using well-known formulas. It sorts the results into groups based on their themes and gives the results and the extracted text in a format that can be changed (Benjamin et al., 2012).

A system called Coh-Metrix is used to figure out how cohesive and coherent written and spoken texts are. Coh-Metrix is used to figure out how hard a piece of writing is for the audience it is meant for. Here, "cohesion" means the parts of the text that help the reader make mental connections between the ideas in the piece. Coh-Metrix is able to comprehend human speech with the aid of computational linguistics tools such as part-of-speech classifiers, dictionaries, syntactic parsers, and latent semantic analysis (Graesser et al., 2004).

2.2.1.1 Similarity methods

Text matching is the method of recognizing and locating specific text matches in raw data. This is a vigorous section in an essential procedure and practical applications in some areas. In addition, certain dynamic approaches have been introduced in this area with the intention of simplifying the process of pattern formation based on the words. There are four primary categories of text similarity evaluation measures (Alqahtani et al., 2021).

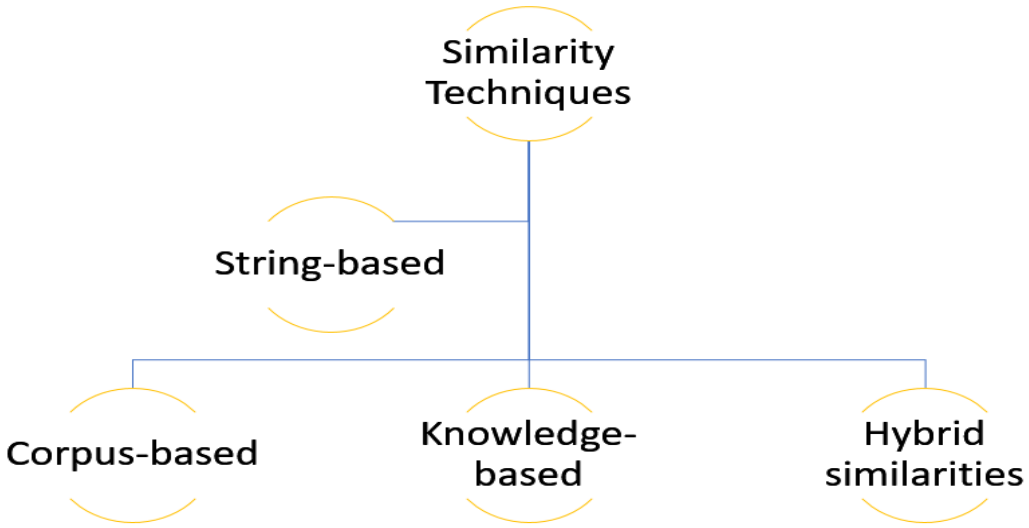


Figure 2.1 Similarity Techniques

One of the oldest ways is the string-based similarity measure (Vijaymeena et al., 2016). Character-based similarity and token-based similarity are the two main types of string-based similarity. The corpus-based similarity is based on how words are used (Mihalcea et al., 2006). Corpora are collections of text that can be written, spoken, or stored digitally, and this technique helps determine the degree of similarity between two concepts using this data. These strategies keep track of sentences together with their foreign-language equivalents. The objective is for the

translated texts to be an exact match for the source texts in the corpus. Information is compared using a set of semantic measures derived from semantic networks, which make up the knowledge-based similarity measurements. The goal of this kind of information is to figure out how similar words are to each other. Semantic relatedness and semantic similarity are two parts of knowledge-based similarity (Makvana et al., 2016). Similarity measures for hybrid classifications do not make a separate group. They combine parts of the previous methods to try to get the best of each (El Desouki et al., 2019). These methods use recursive steps to get around the problems with the other methods.

2.2.1.1.1 String-based Technique

Similarity measurements based on strings are the most common, widely used, and established method available. This metric takes string sequences and character order into account. The two most common kinds of string similarity functions are those that compare characters and those that compare tokens (Lara-Clares et al., 2022).

2.2.1.1.2 Corpus-based Technique

A semantic approach is used for corpus-based similarity. This similarity method uses information taken from a large corpus to figure out how similar two ideas are to each other. A corpus, or a group of corpora, is a large collection of written or spoken text that is stored electronically. Corpus is a collection of sentences that have already been written and their translations into another language. The goal is to match the text that is typed in with the text that is in the corpus (Kadupitiya et al., 2016).

2.2.1.1.3 Knowledge-based Technique

A knowledge-based similarity measure is a way of figuring out how similar two words are by using information from semantic networks. The two components that make up knowledge-based similarity are semantic similarity and semantic relatedness. Researchers all over the world have talked a lot about these ideas. Similarity describes two ideas that can be used interchangeably, while relatedness ties ideas together semantically (Akerkar et al., 2020).

2.2.1.1.4 Hybrid Technique

In addition to the three families already mentioned, there are still a few measures of similarity that can't be put into any of the three families. This strategy is based on the concept of combining corpus-based similarity, string-based similarity, and knowledge-based similarity to create a more effective metric by capitalizing on the advantages of each (Metzler et al., 2007).

In this section, we looked at the different ways that similarity is used to measure relevance. We will use a string-based similarity method to figure out how relevant images are on websites. This method is mostly language-independent, so it works well for languages from all over the world.

2.2.2 Analysis of Multimedia Readability introducing the problem of images

2.2.2.1 Images on Websites

Images are the best way to get a lot of information across quickly. In just a few seconds, the right picture can tell a user what you do, how big you are, how good your work is, how you do it, etc. "A picture is worth a thousand words ", as the saying goes. Research shows that people who use websites don't read them, they just look at them. Images are the best way to show complicated information without making the person look it up. They set the scene and send a message that you choose. When images can be seen and read on the web, we can get a real idea of what they are. The word "image" refers to any two-dimensional picture, like a painting, a picture, a map, a diagram, a chart, or a graph. When we refer to something as a "visual representation," we mean the same thing. For example, it is possible to take photos of artificial optical tools (cameras, microscopes, etc.) and not man-made natural phenomena (like the eye or the surface of water). Created manually, as in the case of drawings and paintings, or mechanically, as in the case of printing and computer graphics technology. (Conway et al., 2010). There are different types of images on the websites, for instance text images, non-text images, logos, diagrams etc. Text Image is the term for when readable text is shown inside an image. This includes text that has been shown in a fixed image form to get a certain look as shown in Figure 2.2. Non-text Images that contain no text at all are shown in Figure 2.3.

Some research has been done on how easy it is to see images on websites. The World Wide Web Consortium accessibility guidelines say that each image should have a textual equivalent. The alt and longdesc are properties of the HTML img tag. Standard makes it easy to add this alternative text. Another non-standard method for providing alternate text is through the use of the title element. This content is readable by a wide variety of accessibility tools, including screen readers and refreshable Braille displays, amongst others. One reason the web isn't accessible is that people who make websites don't give enough alternative text. Choosing the correct alternative text is more art than science, making it harder to build and check the alternative text. Images critical to understanding a page or finding your way around it should have alternative text. Images only there to make a page look better should have an alt attribute with a length of 0 to clarify this. If you don't follow these accessibility rules, most web pages will be harder to use.

Figure 2.2 Input Text Image

Figure 2.3 Input Non-Text Image

Images that do something (like links or buttons) or have more than one colour or are bigger than a certain size are especially dangerous. Images can be inaccessible to some users if they don't have accompanying alternate text. The demonstrated WebInSight solution aims at such pivotal pictures and provides a mechanism for automatically adding appropriate substitute text. It handles incoming web requests and dynamically modifies the returned pages to achieve this. In addition, it coordinates three custom-built image-labeling modules as part of the overall transformation process. These components rely on OCR, human labeling, and better online context analysis to perform their tasks. But there is no any work on image relevancy on websites from a readability perspective, because relevant images enhance web readability (Elahi et al., 2022A) (Elahi et al., 2022B).

2.2.2.2 Readability of Multimedia Content

Web content comes from many different places, and there are many ways to make sure it is easy to understand. Researchers presented a common method that can be used to support author tools

in multiple languages (Nietzio et al., 2014). If you asked readers for feedback, you could get realistic results for testing how easy something is to understand, but this wouldn't work for web content. Readability indices assess the difficulty of reading something based on its length and the number of syllables, words and sentences. The core problem is that many readability methods and indices have been established for various situations, but none are suitable for web content use. There are numerous issues with readability indices. They were designed for standard text, but when writing E2R (Easy-to-Read) text, sentences become longer and contain more words, making them more difficult to read. Readability indices can only be used to test web content in the development stage. To deal with limitations, you need to think about how sentences are put together. You can do this by using style and grammar checkers.

Readability prediction, which has many ways to determine if a text is easy to read, is hard to do with web content because it is written differently. The blog, comments, search engine results, and online ads contribute to the non-traditional nature. Images, audio, video, and other elements with a rich layout can also be included. One way to improve existing content discovery is to add metadata to web pages that indicate how easy it is to read them. Labeling metadata on web pages with readability estimation, which is also helpful for basic web search, has led to several surprising and new uses. One of the most popular ways to find information is through a web search engine, but the people who make them need to pay more attention to how easy they are to read (De et al., 2014).

In 2010, Gyllstrom and Moens came up with the idea of proposing an algorithm that would provide binary labels for web publications. That algorithm was called Age Rank. The walk algorithm is applied in order to determine which pages are suitable for youngsters and which are suitable for adults. The Walk algorithm is founded on Google's PageRank algorithm, which is used to determine the significance of a webpage and is used by Google. AgeRank uses things like the colour of the page, the size of the font, and other sources like hypertext to label pages. When machine learning algorithms are combined with web graph, non-vocabulary and vocabulary features, it gives a good estimate of how easy it is to read (Collins-Thompson, 2014).

Statistical language modeling was used to come up with readability methods: Recent research on how to make web documents easier to read has shown that captions, punctuation mistakes, and sidebar menus create noise. Traditional formulas could have worked better when they were used to look at web documents. When a new language for statistics was added, it made new things possible. Improvements in computer science and statistical models led to Statistical Language Models and Support Vector Machines, which led to a new study. The Statistical Language Models method is based on how likely a word or words are in a language model for a particular grade level or stage. The Support Vector Machines technique helps us find grammar and pattern features that are common in third-grade texts. Together, these two techniques build a grade level text model and figure out how likely it is that the generated text fits into that model (Si et al., 2011).

In 2004–2005, using a sizable corpus and categorizing web texts over 12 levels of difficulty, the authors refined the process by which this information can be obtained. The Fry Short Passage readability formula was suggested for pupils in the fourth grade and up. Adding Statistical Language Models to grammatical feature sets yields just a marginal performance improvement, according to the research on web writing. However, when feature sets based on grammar were augmented with the help of context-free grammar parsers, the features alone did a decent job of predicting web content. In a standard search, the Support Vector Machines technique performed admirably. It employed machine learning techniques to determine the user's reading level and evaluate the complexity of web text. With more and more students reading online, the Online-Boost algorithm can help them better understand what they're reading. In addition, the online-boost method may check for readability updates and assess the reader's level of understanding. Experiments have shown that the method proposed using this algorithm helps improve learners' comprehension (Collins-Thompson et al., 2005).

Correlation between advances in cognitive theory and readability methods: During the 1970s and 1980s, when new theories about how humans store and retrieve information were being developed, Text processing researchers discovered that the factors considered in traditional/classical models do not contribute to readability as much as the coherence and relationship between the text's parts do (Benjamin, 2012). Because of this, researchers started looking at how hard it was to understand a text. They also took into account how theories in cognitive science had changed, and as a result, they came up with a number of methods and variables. Examples:

The Proposition and Inference Model (Kintsch et al., 1978), Prototype Theory, and Latent Semantic Analysis, and Semantic Networks use high-level parameters like the reader's cognitive abilities, cohesion, and organization. A system called Coh-Metrix is used to figure out how cohesive and coherent written and spoken texts are. Coh-Metrix is used to figure out how hard it is for a certain audience to understand a piece of writing. In this case, "cohesion" refers to the parts of a piece of writing that help the reader understand its meaning by making mental connections between its ideas. Coh-Metrix uses parts of computational linguistics like part-of-speech classifiers, lexicons, syntactic parsers, latent semantic analysis, and more.

The Delite software can tell how hard a text is based on things like its morphology, vocabulary, syntax, semantics, and discourse. A syntactic-semantic parser is used to look at German text. It uses machine learning algorithms to improve its performance and to normalize the values of the parameters. Traditional formulas don't match user predictions as well as Delite software does, and it acts as a bridge between methods based on cognition and methods based on statistical language modeling. Textual web accessibility can be measured by a site's lexical quality, which is how well the text on the site represents the subject matter. Lexical quality is a broad term for the quality of words in a text, such as spelling mistakes, typos, etc. It is related to how easy it is to read a website (Magnini et al., 2011).

Several researchers are working hard to figure out how easy it is to read a web page. This research is meant to improve the formulas that were used to estimate how easy it is to read a web page before. Article presented a statistical model to predict how easy it would be to read a webpage. The model presented combines the statistical model with the readability of text. The model looks at both the content and the language of the text. It found that the language model is a more important factor than sentence length in figuring out how easy it is to read a webpage.

Researchers have devised a method to evaluate the readability of web documents by taking a look at the textual characteristics and structural elements of HTML. Textual characteristics are extracted from the individual text strings that are contained within a web document. Statistics and information regarding the construction of characters are both included in the text features category. HTML features are components of a web browser that can be customized, and can include headings, fonts, paragraphs, character sizes, and line spacing, among other things. Documents are organized into groups and represented as vectors in order to facilitate machine learning. As learning data, the web documents that have been categorized and grouped together are employed (Yamasaki et al., 2014).

According to a recent study (Palotti et al., 2016), researchers have developed a method to score health-related websites based on their usefulness and readability. They made search engine results more relevant by emphasizing readability. Syntactic and lexical aspects were measured using surface measurements, such as the number of characters, words, syllables, and sentences. To determine how challenging a text is, we look at how frequently it uses specific words. There is a correlation between these indicators and a person's vocabulary in general. Regarding lexical aspects, we counted the occurrences of numbers, stop words, and frequently used words. We relied on indicators specific to the scientific realm and medical terminology for our lexical and morphological features. Researchers found that search engine results are better when retrieval features and readability features are used together.

GUI Evaluator is a tool for assessing the complexity of graphical user interfaces depend on information complexity structural measures such as alignment, size, grouping, density and balance. This method can be used to assess a website's visual aspects, and graphics' impact can also be considered when predicting how easy something is to read. GUI Evaluator looks at the Screen Layout Complexity, Alignment, Size, Balance, Density, Grouping, and Grouping Density (Alemerien et al., 2014).

In this part, a lot of research on readability indices and tools that are used and published in the field of making web content easy to understand is looked at. These readability formulas and tools were made to figure out how hard a traditional text is to understand. But there are a lot of things that need to be looked at when measuring the readability of the web, like how meta-data on the same page relates to other meta-data, how reading levels vary across different domains and web pages, and how relevant images are on websites. But there isn't a way to figure out if an image on a website is relevant or not.

2.2.3 Web Readability Evaluation Methodologies and Methods

Different evaluation techniques in the literature review have been conducted mostly based on the textual content of the web pages as shown in Table 2.2. Research on the website's automated and manual usage of thirty-nine readability criteria was presented by (Miniukovich et al., 2019). A group of dyslexic and typical readers participated in this study, and eye tracking was used to determine how difficult it is to read a set of fifty web pages. According to the findings, there is a connection between twenty-two different rules and readability. Furthermore, the contrast between the results generated by the computer and those caused by humans revealed another intricate pattern: computers are better or just as good as humans at judging website pages based on specific rules, especially those about low-level details like readability and how text is organized. However, there are a few guidelines that necessitate the use of human discretion in order to decode and comprehend the content of a website page. These findings contribute to the elaboration of a description of a guideline that establishes the foundation for upcoming methods of evaluating design.

We compared how effective and efficient heuristic evaluation and user testing were when we looked at four commercial websites (Nova et al., 2022). The findings indicated that heuristic evaluation and user testing tackled distinct usability difficulties in their respective approaches. An examination of the gravity of the difficulties, as well as a model illustrating the connection between the number of newly discovered problems and the number of users and evaluators who participated in the study, were both presented as examples. The big changes that were found between these two approaches suggested that they should work together and not against each other. Another study was done to see how easy it was to read and how good the websites were that gave potential patients information about clear aligners (Meade et al., 2020). We investigated thirty websites that teach people how to execute strokes, using criteria such as readability, responsibility, and consistent quality throughout the sites. Eleven health professionals and fifteen customers evaluated six websites in terms of their information, designs, and convenience of use. The website's pages have always adhered to the responsibility models, but their quality scores have consistently been poor, and their content is straightforward. Consumers' opinions were always more positive than those of health experts, but their scores showed that they were more likely to like explicit pages, especially when it came to design. When designing and suggesting site pages, it's important to think about what customers want (Griffin et al., 2004).

Also, a study was done to see how serif and san serif text styles affected the readability of Malay text on websites. This study looked at how screen text styles and print text styles were grouped. Because of this, four different text styles were chosen: For the first two respondents, the fonts used were Georgia (serif) and Verdana (sans serif), while for the third and fourth respondents, the fonts used were Times New Roman (serif) and Arial (sans serif). Both Georgia and Verdana were explicitly designed for use on computer displays. Times New Roman and Arial, on the other hand, were made for print at first. 48 students took a test of their ability to understand what was on a PC

screen. The results indicated no significant difference between the readability of text written in serif and sans serif styles for either the screen show class or the print show class. This was the case for both classes. Both the exploration findings and the writing outline point to Verdana and Georgia as preferable options when displaying lengthy material on websites. Additionally, and as was to be anticipated, Times New Roman and Arial are excellent text styles for print media because they are simple to read. This makes them the printing text style class (Ali et al., 2013).

Researchers looked at how useful the sites for advanced education in Asia were. At first, a web-based Google application review structure was planned to use Google Forms and be used to measure how easy it was for students to use the web and how they responded. After a lot of research, a small model called the "Web Usability Evaluation Model" was made to judge how easy it is to use educational websites (WUEM). In this test, the ten best design schools in Asia were measured against the criteria in the WUEM. The evaluation research shows that the educational sites are about half usable in terms of their instructional design, navigation, and weak unavailability. The evaluation gives a point-by-point overview of what needs to be fixed on these sites to make them easier to use. The proposed WUEM helps web designer's rate sites in a way that is both convincing and easy. The test will help academic web designers make their sites easier to use by taking into account simple things like those listed in WUEM (Manzoor et. al, 2012).

In another study, researchers looked at how to make web pages easier to read for people of different ages. This study focused on eight factors that have always been important to readability, such as shading contrast, blank space, line spacing, text style, text size, text width, headings, designs, and liveliness. By changing these eight factors, it is possible to see how people of different ages use web applications (Rayner, 1986).

Table 2.2 Web readability Evaluations

Sr#	***Paper Name with Year***	***Types of Users***	***Number of Users***	***Main aim***	***Websites***	***Types of Website***	***Input is given to the Users***	***Type of Questions***
1	Guideline-Based Evaluation of Web Readability (Miniukovich et al., 2019)	Dyslexia/Experts	79/35	To evaluate the readability based on the guidelines	50	news, non-profit, and governmental organizations	User could randomly selected 5 web pages with guidelines	User was asked to select one of the following ● Rate out of seven to describe how well the guidelines are followed by the Website ● User doesn't understand the guideline ● Checkbox this rule doesn't apply
2	Web evaluation: Heuristic evaluation vs. user testing (Tan et al., 2009)	Users/Experts	12/9	To see how easy it is to read based on the rules	4	commercial web sites	Websites	Users were provided a set of severity criteria to assess the severity of problems. The three distinct severity ratings consist of ● Severe ● Medium ● Mild Problems

3	Web-based information on orthodontic clear aligners: a qualitative and readability assessment (Meade et al., 2020)	Normal Users	Not Mentioned	To evaluate the readability of webpages as well as their overall quality	50	orthodontic clear aligners websites	Websites	Rate out of five on the following ● Information regarding reliability ● information related to treatment choices
4	Stroke Education Materials on the World Wide Web: An Evaluation of Their Quality and Suitability (Griffin et al., 2015)	Consumers/ Health professional	15/11	To assess the readability stroke education websites	30	stroke education websites	Websites	Rate out of ten on the following ● Design and aesthetics o Layout o Presentation o Graphics o Appeal o Diversity of broadcasting ● Ease of use o Usability o Navigability o Functionality ● Content of website o Helpfulness of information o Easiness of understanding o Range of information o Satisfactory information for requirements of user o The precision of information o The impartiality of information
5	Reading on the Computer Screen: Does Font Type has Effects on Web Text Readability? (Ali et al., 2013)	Undergraduates Students	48	To check the effect of font type on web text readability	1	Real	Two Paragraphs	Speed and accuracy were considered as questions
6	A Web Usability Evaluation Model for Higher Education Providing Universities of Asia (MANZOOR et al., 2012)	Graduated Students	30	To assess the usability according to the guidelines	10	Real	Websites	Rate out of hundred on the following ● Do you know what the homepage looks like? ● Do you need help getting around on the website? ● Do you find it easy to remember the website addresses for your school? ● Are the words on the pages easy to read? ● Does your school keep you up to date on the latest news and events? ● Does your website support more than one language? ● Does your website have the correct

								headings and titles for each page? ● Does your website keep a consistent design style?
7	Eye Movements and the Perceptual Span in Beginning and Skilled Readers (Rayner, 1986)	Children's/Skilled Readers	Not Mentioned	To check the readability of text	Several	Real	Paragraph	Speed and accuracy were considered as questions
8	Web Readability Factors Affecting Users of All Ages (Hussain et al., 2011)	Children, Teenage and old age users	Not Mentioned	To evaluate the readability based on content, style, design, and structure.	Several	Real	Website	Rate the content on four categories ● Content ● Style ● Design ● Structure (text width, font size, headings, white space, colour contrast, line spacing, font style, graphics and animation)
9	Improvement and evaluation of readability of Japanese health information texts: An experiment on the ease of reading and understanding written texts on disease (Sakai, 2011)	College students	91	Enhancement and evaluation of Japanese health information text readability	Several	Real	Paragraph	Speed and accuracy were considered as questions

In summary, in this section we have analyzed many techniques that evaluate web readability. Most researchers have worked on evaluating text on the web from different points of view, such as figuring out how easy it is to read based on guidelines, figuring out how good a website is and how easy it is to read, figuring out how easy it is to read stroke education websites, and figuring out how easy it is to read based on content, style, design, and structure. None of the research done before has focused on figuring out how relevant images on a website are from a reading point of view.

2.3 Relationship between readability and images in multimedia documents

In order to check image relevancy on websites, we have presented the practical tools on extracting contents from the websites in this section. Lastly, we have discussed the image processing techniques to check quality of images.

2.3.1 Extracting Contents from Web and Image Processing

2.3.1.1 Extracting Content from Web

Researchers face a huge challenge when they try to find useful information on the World Wide Web. This process of using data mining techniques to find useful information is called "Web mining." Figure 2.4 shows how web mining can be broken down into five smaller tasks. As shown

in Figure 2.5, there are three types of web mining: web content mining (WCM), web structure mining (WSM), and web usage mining (WUM).

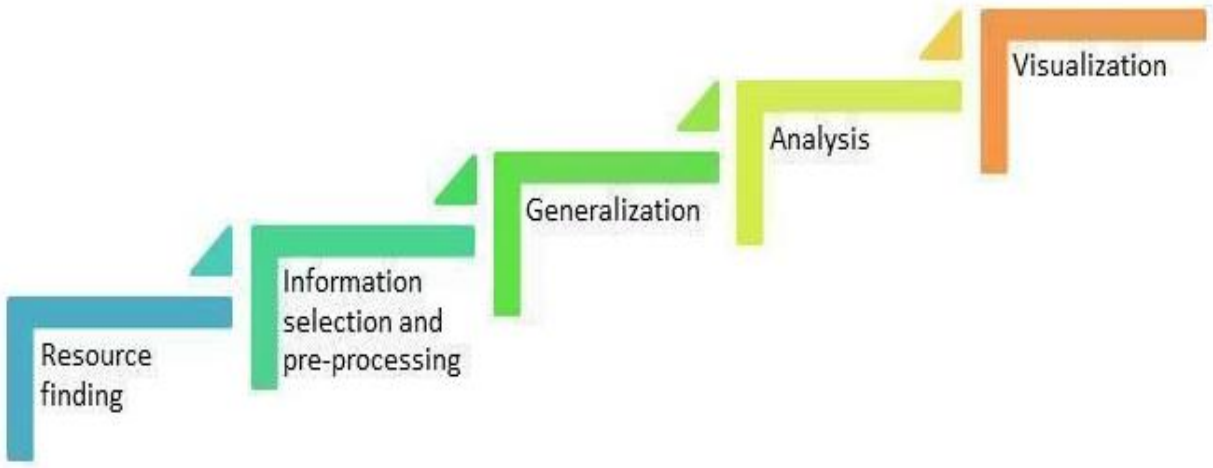


Figure 2.4 Web Ming Tasks

WCM is the process of getting user-specific information from images, text, audio, or video files that are already on a website (Bharanipriya et al., 2011). This process is also called "web text mining," because text content is the most researched thing on the World Wide Web. WSM is another way that graph theory is used to study the structure of a website's nodes and links. Web structure mining has been split into two groups based on the type of web structure data. The first one is figuring out patterns from web links (da Costa et al., 2005). The second one is mining the structure of the document. This includes using the tree-like structure to look at the HTML or XML tags on the page and define them. The goal of WUM is to use logs of how people use the Web to find patterns. Here, we talk more about content mining on the Web.

2.3.1.1.1 Web Content Mining (WCM)

WCM finds the valuable information from the contents or data on the websites. Nevertheless, such data in its wider form has to be additionally narrowed down to the valuable information. In this section, we begin with two key methods of Web Content mining and describe how it varies from Data Mining.

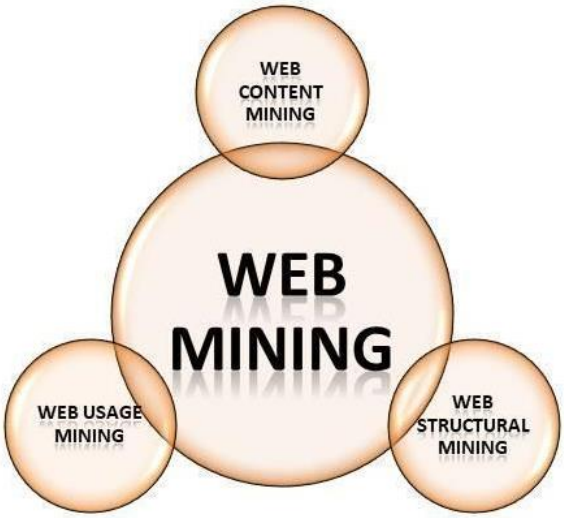


Figure 2.5 Categories of Web Mining

Structured data, like the information in tables, is one type of web content data. Unstructured data, like free texts, is another. There are two main methods in the WCM which are given below:

- **Unstructured Text Mining** Web content data is quite a bit of unstructured text data. Knowledge discovery in texts (KDT) or text data mining is the study of how to use data mining techniques on unstructured text. So, text mining could be seen as a type of Web content mining (Rajpathak et al., 2013). For any structured data, the preprocessing steps are done by extracting information, putting the text in the right order, or using NLP techniques.
- **Structured and Semi-Structured Data Mining** Structured data on websites are often very important because they show what their host pages are about. This is why it is important and common. When compared to unstructured texts, structured data is also easier to pull out. For the Web and database networks, semi-organized data is when the way documents were set up before is mixed in with the new data. Relational tables with numbers and strings give way to tables with numbers and strings that can be used to describe real complex things like books, papers, movies, etc., without making the application writer twist and turn (Asai et al., 2004). The Object Exchange Model differs from new ways of describing semi-structured data, such as the XML (OEM). Data is stored in the OEM as single objects or groups of objects. Atomic objects can be numbers or strings, and labeled edges on compound objects can point to other objects.

2.3.1.1.2 Difference among Web Content Mining, Text Mining and Data Mining

Web content mining uses data mining techniques. In contrast to the structured data that is managed by data mining, the data on the web is typically unstructured and only loosely organized. Data mining, on the other hand, organizes and maintains such data. In addition to this, it is related to text mining because the majority of the information that can be found on the internet is text. The

difference between text mining and web content mining is that web content mining works with semi-structured data, whereas text mining works with unstructured data. Web content mining in this manner necessitates inventive uses of text mining and information mining strategies, as well as its unique methodologies.

2.3.1.2 Web Content Mining Tools

Web content mining is the process of downloading information from websites. This cycle is highly demanding and time-consuming. A computer could utilize the software or tools connected with web content mining to download the necessary data in order to grow such a cycle. This software is associated with web content mining and can be applied in order to expand such a cycle. It collects relevant information from websites that a user requests. Various types of Web content mining tools are detailed below:

- Mozenda: To extract web data effectively and to oversee it reasonably is valuable by using the Mozenda. Clients can set up operators that consistently extract, store and circulate data to a few objections by utilizing it. When data is in the Mozenda frameworks then clients can repurpose, arrange and blend the data to be utilized in other offline /online applications or as intelligence (Haddaway, 2015).
- Screen-Scraper: It enables content mining from websites, such as probing a database, SQL server, or SQL database that communicates with the software, to meet content mining requirements. Screen scrapers might also be accessed using computer languages, including PHP, Java, Visual Basic, .NET, and Active Server Pages (ASP) (Sirisuriya, 2015). (Herrouz et al., 2013).
- Web Info Extractor: It is useful in mining of web data, extricating web substance, and checking content updates. Prickly layout rules are not needed to be characterized (Sleiman et al., 2012).
- Web Content Extractor: Most remarkable and simple to utilize data extraction tool for web scratching, data mining or data recovery from the web will be Web Content Extractor (Weninger et al., 2016).
- Automation Anywhere: It is a web data extraction tool that is utilized for retrieving web data easily, screen scratch from webpages or use it for web mining (Sharma et al., 2012).

2.3.1.3 Extracting Images from the Web

Images are regularly the favored mode for showing the information over the website and you might need to save all the Images from the website. Be that as it may, you would think that it's somewhat hard to extract the pictures alone from the website as there are numerous other media on the website. We are presenting some tools following are:

2.3.1.3.1 Save all Images

Save All images is an image extractor helping you to download all the graphical contents in a given URL. It is exceptionally quick and simple to utilize. You could review the images before sparing them. It would likewise show you the size of the images, which would enable you, to better, conclude if to download the images.

2.3.1.3.2 Online Webpage Image Downloader and ImageInfo Grabber (OWDIG)

With OWIDIG service you can snatch pictures from any website or webpage, see them, filter them, get information about them or download them. You may incorporate or avoid pictures which are rehashed. Additionally, CSS pictures (situated inside style labels, outer templates (by means of connection or import) and furthermore inside inline styles) are upheld also. Pictures can be gotten from websites with standard HTTP protocol just as secure HTTPS protocol. There are two fundamental modes - the LIST mode and the TABLE mode. In the rundown mode all pictures are recorded individually, one next to the other. We may tap on them to open them in another window in their unique measurements. You can even share them through interpersonal organizations or email. In the table mode we can see pictures in two tables - one table for standard HTML pictures, another for pictures inside CSS. By tapping on section headers you can sort recorded pictures concurring picked property. The sea size of pictures in this mode may or probably won't be restricted (as a matter of course it is restricted to 330 pixels in both width and tallness). Another element is a profound snatching. You might need to get pictures from the entire site including every one of its subsections.

OWIDIG empowers clients to snatch pictures up to connect profundity of 10 for site links that have a similar space as the first URL. Obviously, the profound getting may now and then take a lot of time, if there are numerous subsections. For getting from a single URL there is one more choice to get pictures created with JavaScript. This can be utilized when pictures aren't in the first source code, however are in the code produced with JavaScript. OWIDIG gives heaps of sifting choices to snatch simply those pictures you truly need to get. You can pick whether to incorporate either norm and CSS pictures, or just ones of them. You can filter pictures by their dimensions, URL, type, ordinal number and filename. You can set user-agent, intermediary, referrer and treats to impact returned got picture content. Lastly we come to downloading images. You can simply list all pictures in another window and afterwards utilize the sparing abilities of the browser to spare them. You may likewise utilize an applet. This applet requires client endorsement as it composes data - the pictures - to your hard drive. Utilizing this applet you can set filenames and ways of downloaded images. In certain programs (Opera and Firefox uphold, halfway Chrome uphold - for more modest documents) you can likewise make ZIP files from images.

2.3.1.4 Processing images

After extracting graphical contents from the web it's necessary to extract information from the images to check the web readability. For this purpose, we will use image processing (Chitradevi et al., 2014) in order to compute the relevance of images on the websites. The goal of image processing is to recognize the information and design segments in images, and to extract the expected data as a human would. There are some techniques to extract information from images:

2.3.1.4.1 Theoretical Models

In this section, From the Library and Information Science (LIS) perspective, we have shown the primary standards and theoretical models for investigating and showing pictures. First, the discipline of LIS and the role of visual elements in this field of study are explained. Also, clear theoretical standards exist for how information is shown in pictures, and explicit hypothetical models are used to classify and describe images.

2.3.1.4.1.1 Library and Information Science

It is the field "devoted to applying theory and technology to the creation, selection, organization, management, preservation, dissemination, and use of collections of information in all formats" (Reitz, 2004). Another definition is "the professional knowledge and skill with which recorder information is selected, acquired, organized, stored, maintained, retrieved, and shared" (Enser, 2008).

Dissimilar actions of a picture can be distinguished in Library and Information Science:

- Normally, in Library Science a picture is considered as a report itself (a visual record), as on account of a verifiable photograph in a computerized document. The result of this approach was a limit of concentration on the connections between the picture and its unique situation, as a result of separating the picture's content from its context. However, this old-style method has changed into an additional supple methodology that considers the client, space information, and the context of the image. For instance, the report is handled differently depending on whether it belongs to a document or a library (Bates, 1999).
- In some cases, a picture isn't viewed as a record, yet a straightforward representation. Thus, the picture is considered exclusively for its reasonable use; for instance, a picture utilized as a thumbnail that outwardly supports the decision of a reference in a list (Guo et al., 2021)
- In other contexts, the image is viewed as supplementary information, especially when dealing with images in logical spaces. For example, a head scan in a medical record is typically considered additional to the primary history (Feather et al., 2003).

Traditional picture theory in LIS was primarily influenced by the breadth of the humanities, focusing on studies of social and artistic images. These kinds of pictures have prevailed in the writing of portrayal, ordering and recovery of pictures and have been utilized to approach hypothetical foundations for picture portrayal studies. Still, pictures from logical areas, like

clinical, structural and designing, have not customarily been subject of examination all alone, because of the way that these sorts of pictures have been normally treated as assistants to parent record. As of late, logical pictures have been perceived as significant data objects by their own doing. This change has prompted the production of specific assortments for investigation and preparation purposes (Enser, 2008).

2.3.1.4.1.2 The Process of Information Representation in Images

The documentary procedure for the data portrayal in pictures arranges and sorts out the perusing of picture components in progressive stages, to reveal the various layers of their importance and reason. In the LIS discipline, the examination and translation of pictures is separated into two tasks:

i. **Formal investigation of information:** The interaction for making passages for an index is at present called bibliographic depiction. The bibliographic depiction manages the ID and the portrayal of the physical and bibliographic attributes of the picture: title and proclamation of obligation (creator, supervisor, writer, and so forth), subtleties of distribution and appropriation (spot of creation, date of creation, and so on), the actual depiction (including the arrangement of the picture, the type of picture, and so on), the series, the notes, the standard number, and the terms of accessibility (such as the DOI and the cost) (Kosslyn, 1975). This accurate portrayal of the component concludes the title(s) and name(s) to be utilized as passageways in the list; however, this decision does not affect the task of subject headers. This appropriate inspection is based on the standards of recording, which vary according to the field of use (ISBD and the Old English American Classifying Rules, Second Edition (AACR2) for libraries; ISAD G for documents; CDWA and VRA for exposition halls, and so on).
ii. **Content inspection:** It determines the picture's significance and its consequence on its listeners. Numerous frameworks have been developed for examining report content (visual and literary). So, take the Valle Gastaminza model as an example (del Valle Gastaminza, 1999), the examination of photographs for the most part starts with the recognizable proof of the visual components of the picture (for example variety, surface, spatial conveyance and areas) and go on with the investigation of the importance of the picture with regards to social and other relevant information, like philosophical, strict or tasteful convictions, references to political and social codes.

After this investigation, a bunch of thoughts and ideas illustrative of the content of the picture has been obtained. The procedure of appearance of these ideas and ideas is called portrayal: this is the content depiction, the literary amalgamation of the data sent through the picture. The portrayal of the records' subject includes a bunch of tasks that produce optional data mirroring the most considerable components of the report's content. These tasks are annotation, arrangement, and abstracting.

- Annotation is a key strategy used to prepare information for computer vision. With the goal for machines to see objects in their environmental factors, commented on pictures are expected to prepare Machine Learning calculations to figure out how to consider them to be as we do. Comment in Machine Learning is basically the way toward marking information in the different modes of pictures, text or video. The names are generally foreordained by an Artificial Intelligence architect or computer vision researcher and are picked to give the computer vision model data on objects portrayed in a picture. There are various kinds of picture comments like Bounding box, Polygon explanation, Line annotation and Point comment. Semantic Segmentation is the errand of isolating a picture into different areas and ordering each pixel in each fragment to a comparing class name of what it addresses (walker, vehicle, and light post). This gives machines an extensive comprehension of each pixel of a scene in a picture (Belkin et al., 1976).
- The content component portrayals are summarized in a theoretical manner. The theoretical is the portrayal of the picture in text design and in normal language at various degrees of fulfillment, as per the client's profile and needs. Conceptual is extremely significant in the narrative course of portraying pictures, since the visual substance of the picture should be converted into words for further development. In contrast to the changed compositions that address the content of messages, which are governed by an ISO standard, there is no standard model for producing updated works for images (Hlava, 2015). Instead, a model for the unique picture has been proposed by (Pinto Molina et al., 2004), in which a collection of highlights represents the concept of a photograph according to a predetermined order (variety, shot point, sort of lighting, component in first shot, objects addressed, social structure of the photograph, and so on.) The meaning of an ISO standard that controls the production of edited compositions for various kinds of pictures could be additionally valuable to further develop webpages availability.
- Grouping is the method involved with partitioning articles or ideas into consistently various leveled classes and subclasses in light of qualities they share and those that recognize them. For the most part consists of doling out a code to the picture from a current characterization; a framework with a calculated and methodical design of classes connected with one another as per a bunch of normal qualities. Instances of arrangement frameworks for libraries are the Dewey Decimal Characterization (DDC) or the Library of Congress Grouping (LCC).

2.3.1.4.1.3 Theoretical Models for the Image Analysis

There are a few Theoretical models that propose a methodical way to deal with the investigation of the picture content. They address the conventional foundation for picture depiction and ordering. In this review, four of the most agent models are introduced, because of their association with the target of postulation and number of references in the Library and Information Science educational area.

2.3.1.4.1.3.1 Iconographic Model

The model of the art historian Panofsky (Morgado, 1993), which has figured unmistakably in the writing, has the target of officially breaking down Renaissance workmanship pictures. It groups the content of craftsmanship pictures as indicated by three unique degrees of portrayal: an essential topic ("pre-iconography"), a subsequent topic ("iconography") and a tertiary topic ("iconology").

- The essential topic is the portrayal of the topical or inherent content of the picture. It incorporates the distinguishing proof of the visual natives, similar to variety, surface, and shapes. It manages the conventional components of the picture and it doesn't need interpretative abilities.
- The subsequent topic is the traditional content wherein explicit subjects and ideas (coherent or determined highlights, articles, exercises, and occasions) are put. It requires the watcher to decipher the picture.
- The tertiary theme encompasses the inherent relevance of the image (inductive translation, distinctive aspects) and necessitates an undeniable level of viewer semantic deduction. At this level, the observer can discern the norms of a nation, a verified period, a class, and their strict or philosophical beliefs.

2.3.1.4.1.3.2 Shatford Model

Panofsky's approach is applied to the ordering of images by Shatford (Christensen, 2017), who renames Panofsky's words as conventional, explicit, and dynamic. These three tiers can be broken down into their constituent parts, each of which includes the Who, What, Where, and When. Some studies have used the 3x4 framework created in this way to analyze both static and moving images; this framework is known as the Shatford/Panofsky model. The model also includes a distinction between what an image is "of" (objective things, either nonexclusive or explicit) and what the image is "about" (more emotional or unique implications). In addition to the conventional elements (such as protagonist, location, activity, and context) that make up a picture's "topic," Shatford also takes into account the non-visual information an image can convey:

- Anecdotal: the qualities coming about because of the investigation of the set of experiences or life of the picture: date and spot of creation, title, limitation, cost, and so on.
- Exemplified: the characteristics coming about because of the investigation of the pictures considered as a specific sort of item. For instance, a picture might be a photo or a banner.
- Relationship: the characteristics coming about because of the investigation of how the picture connects with different pictures or text.

2.3.1.4.1.3.3 Syntactic and Semantic Model

The theoretical groundwork for the development of the pyramid model for visual portrayal was laid by the work of Panofsky, Shatford, and others. This model was proposed by Jaimes (Jaimes et al., 1999). The target of this calculated system is the ordering of various parts of visual data.

2.3.1.4.1.3.4 Eakins/Graham Model

The model of Graham and Eakins (Eakins et al., 1999) proposes a differentiation like the past model, yet centers on inquiries instead of on files, and its particular objective is to work on the recovery of visual reports. It recognizes three degrees of picture questions, which reflect various levels of data necessities and relate to various elements of the picture:

- Queries rely upon the crude ascribes, for example, tone or shape. The crude (low level) questions are variety, surface, shape, or the spatial area of picture components. They are both goal and logical from the actual pictures, and they needn't bother with references to any outside information base for deciphering. Instances of such questions could incorporate "find pictures with long slim dull items in the upper left-hand corner," "find pictures containing yellow stars organized in a ring" - or most normally, "find me more pictures that seem to be this"(Eakins, 2001).
- Queries in view of coherent (at times known as determined) highlights, for example, the items portrayed. Sensible questions incorporate some level of coherent surmising about the personality of the items portrayed in the picture and typically require a reference to some outer wellspring of information.

In this section, we have presented the fundamental standards and theoretical models on the inspection and depiction of pictures according to the perspective of the LIS discipline. The LIS, first and foremost, discipline and the role of visual content in the study are introduced. Furthermore, theoretical standards for the portrayal of data in pictures and theoretical models utilized in the cataloging and annotation of pictures are definite.

2.3.1.4.2 Probabilistic Methods

The method uses probabilistic modeling to figure out how likely it is that the image content matches the annotations. Using techniques based on probabilistic modeling, the image without labels is broken up into several image segments. We next determine which labels are most likely to be the picture annotations by calculating the probabilities of the labels transferred to the image segments. Mori finds the co-occurrence by looking at the sub-images and the labels that go with them (Mori et al., 1999). First, the picture is broken up into several smaller pictures. Second, it pulls out low-level features that can be used for clustering. The next step is to figure out how often each cluster and its labels show up together. Even though the process takes less time than classification-based methods, it is less accurate than those. Kuric looked at both the local and

international parts of the pictures (Kuric et al., 2015). Both local and global features are taken from the parts of the images. The regions for clustering are then represented by locality-sensitive hashing (LSH). For an unlabeled image, a similar area in the dataset is chosen, and the weight of each label is calculated to update the probability of labeling the unlabeled images. Zhang came up with ObjectPatchNet, which is a combination of the BVWand probability (Zhang et al., 2014).

ObjectPatchNet figures out how often each cluster appears together and how likely it is that an image patch and a label go together to show the relationship. The downsides of methods based on probabilistic modeling are that low-level features don't have meaning and that the low-level features of the same individual objects are the same. But people think that different orientations are different. Methods based on probabilistic modeling are better than methods based on classification, and they can be used on social platforms like Flickr. In general, the concrete expression for images allows you to define category labels using instance classes. However, two photographs of the same item from different angles or orientations are interpreted as two separate objects due to low-level features that lack semantic information. Hong connects semantic concepts using data from commercial image engines and a large amount of data (Hong et al., 2014). The relationship between each pair of ideas is put into a certain category. Image recognition is the basis for all of the above techniques, but image features can't define a lot of abstract ideas, like where something is. There's no question that it's hard to get accurate results because the pictures don't show abstract ideas. That is, if the datasets aren't precisely described with domain knowledge, the resulting retrieval results will be incomplete. Because of this, it employs ontology theory to learn about various definitions, attributes, and the connections between people in order to ensure that the labels used for picture annotation have semantic meaning (Im et al., 2015).

2.3.1.4.3 AI Methods

These methods use scale-invariant feature transformations to make a bag of visual words that can be used to recognize objects. In the first step of bag-of-visual-words, the scale-invariant feature transform keypoints of the training images are extracted. Then, using k-means, the keypoints are grouped into several groups. Then, for each image, it figures out how many scale-invariant features are needed to change the keypoints in each cluster and turns that number into a vector that describes the image again. Using supervised learning techniques like a support vector machine, a classifier is trained for each category. But there are thousands of scale-invariant features that change keypoints in an image, so training classifiers takes a long time. Noises can also change how well classification works. Kesorn showed a way to improve the quality of words that can be seen (Kesorn et al., 2011). The plan was to combine the close keypoints and eliminate the cluster with many documents that don't fit into any other categories statistically. Lu developed Laplacian regularization-based semantic regularized matrix factorization to improve how well bag-of-visual words are trained (Lu et al., 2015). In addition to the bag-of-visual-words model, Su et al. made the Annotation by Image-to-Concept Distribution Model (AICMD) to make different models to represent the images. First, clustering is used by AICMD to find patterns from six low-level

features (Su et al., 2011). Then, entropy, tf-idf, and association rules are used as parts of the images in the patterns. Then, a support vector machine uses all the features to train the classifiers. Some researchers use Hidden Markov Models instead of a support vector machine for this (Li et al., 2003).

Classification-based methods are very effective, but they take a long time to train. Still, it's hard to tell if the object is an instance class or not. Also, users might tag different things with the same word. This is called an ambiguity problem. Because of this, Feng ranks tags in descending order of how relevant they are to the given image. This reduces the learning space and makes a hard problem much easier to solve (Zhang et al., 2014). Zhang and Xia came up with the idea of refining and expanding the vague tag words separately as a way to solve the problem of ambiguity (Xia et al., 2014). Zhang used the Random Walk with Restart (RWR) algorithm to improve the CTSTag tag, which was a rough description of the query image. The exact tags that are made help connect different images with tags that are similar. Xia also used the idea of ontology to improve the accuracy of the tags in the image social networking service in order to improve the image tag. As for the hierarchical idea, Yuan made a hierarchical image annotation system to make tags for images that are based on the hierarchy (Yuan et al., 2015). Fang suggested a hierarchical ontology of ideas and their connections to make it easier to understand the meaning of information (Fang et al., 2016). Yi-Hao also suggested a system called Diffscriber that would help blind or visually impaired people work together with slide-authoring assistants by identifying and describing changes to the design of visual presentations (Peng et al., 2022).

2.3.1.4.4 Optical character recognition (OCR) Tools

In this section, we will talk about the different OCR tools that are used to find text in scanned articles, photos, ads, and other digital images. It is widely used as a tool for entering information, and it is capable of extracting valuable information from scanned articles like printed forms (that users fill out), automatic receipts, invoices, business cards, bank statements, passport documents, emails, and any other document that is suitable (Neudecker et al., 2021). There are many other applications comprising searching within institutional repositories, processing cheques in the banks, automatic number plate recognition, identifying barcodes, scanned legal articles, testing text-based captcha codes, etc. We have divided OCR tools into three categories such as Open Source (Tesseract, Ocrad, GOCR and OCRopus), Proprietary (ABBYY FineReader, Transym OCR, Readiris and Adobe Acrobat) and Online (ABBYY Cloud OCR, Google Docs, Free-Online-OCR and Online OCR) shown in the Figure 2.6.

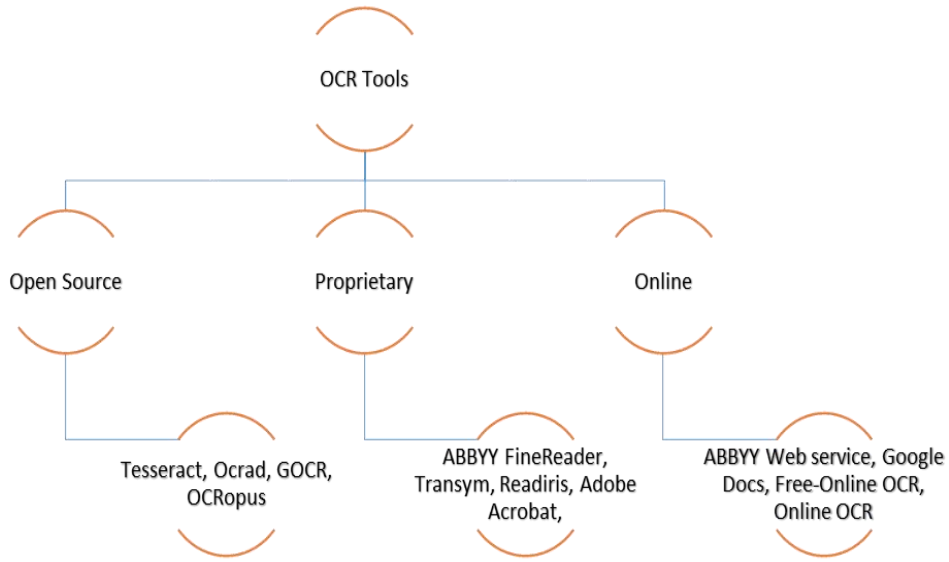


Figure 2.6 Optical Character recognition Tools

2.3.1.4.4.1 Open-source OCR Tools

The best way to control open-source OCR tools is through their command-line interfaces, and most of them don't have GUIs (Grădinaru et al., 2022). This section talks about some of the most used open-source OCR tools.

- Tesseract: It comes with a command line tool called "tesseract" that is easy to use. With Tesseract's API, it can be added to C++ or Python code. OCRFeeder, FreeOCR, PDF OCR X, YAGF, gImageReader, QTesseract, Lector, SunnyPage, VietOCR, and Lime OCR are all desktop programmes that use Tesseract as their text recognition engine. Tesseract is also used by the web apps CustomOCR, WeOCR, NewOCR, and i2OCR (Smith, 2007).

- Ocrad: This tool lets you choose which character sets to look for when doing a character search. It quickly recognizes characters, but it is also very sensitive to mistakes, and it is hard to change this tool so that it can recognize new characters. When characters are a minimum of 20 pixels high or when the image is scanned at a resolution of at least 300 dots per inch (dpi), the best results are achieved (Krejcar, 2012).

- GOCR: It can function either as a stand-alone console application or as the optical character recognition engine for other programs. It's written in C, and you have to look at it twice to figure out what it is. The whole document was read the first time, and only the unknown characters were read the second time (Dhiman et al., 2013). People say that GOCR can work with single-column sans-serif fonts that are between 20 and 60 pixels tall. Problems have been reported with italic fonts, serif fonts, noisy images, slanted fonts, multiple columns, small fonts, coloured images, different fonts, handwritten text, large angles of

skew, overlapping characters, tables, complex layouts, and text in a language other than Latin (Jain et al., 2021).

- OCRopus: with a user interface that is based on a command line. Its design is very modular, which means that each step of OCR (such as page layout analysis, binarization, text line recognition, etc.) can be done on its own with independent procedures; many modules are available for use by the user (Breuel, 2008).

2.3.1.4.4.2 Proprietary OCR Tools

Developers typically pay for and support proprietary OCR tools. In addition, they usually have an excellent graphical user interface (Romanov et al., 2017). There are some popular Proprietary OCR tools discussed in this section:

- ABBYY FineReader: with high-quality images in English, this tool achieves up to 100% word-level accuracy and has been the clear choice for layout analysis. Researchers (Heliński et al., 2012) say there are two ways to get to it: ABBYY Online and ABBYY FineReader SDK.

- Transym OCR (TOCR): It was carefully made so that it would be easy to work with other software. This tool can read characters that are blurry, hard to see, or even broken. It has a very light and easy-to-use graphical user interface (Feng et al., 2004).

- Readiris: is a tool that turns text from images, PDF files, or paper documents into fully editable files while keeping the page layout. It works with most scanners on the market, supports many dissimilar input formats, and has a graphical user interface that is easy to understand (Sharma et al., 2020).

- Adobe Acrobat: automatically turns scanned documents, image files, and PDF files into documents that can be edited and searched while keeping the format. Its accuracy is said to be high. Compared to ABBYY FineReader, this tool has fewer language options, but it is used more often because it is more business-oriented and less academic (Lund et al., 2011).

2.3.1.4.4.3 OCR Online services

When using OCR online services, you will not be required to install or download any OCR software at any point in the process. Instead, the user is just required to upload the input file, select a language, and select an output format (if they so wish). The output will then be generated (Markwood et al., 2017). Several of the most well-known OCR online services have been discussed in this article:

- ABBYY Cloud OCR: This tool is a Web OCR service that runs on Microsoft Azure infrastructure and does a great job of recognizing text. With a Web API, it is easy to add

this web service to your own program. Files that have been changed can be sent to DropBox, Google Docs, or Ever note (Metzger et al., 2019).

- Google Docs: is a tool that lets you store and edit documents in the cloud. It is part of Google's Google Drive service. Once an image or PDF file has been uploaded to Google Drive, you can convert it to OCR by right-clicking on it and choosing "Open with Google Docs." The text that was taken out can be downloaded in a form that can be changed (Tafti et al., 2016).
- Free-Online-OCR: even with low-quality documents, like fixes and screenshots, it can recognize them very well. With the help of a built-in dictionary, the accuracy is even better (Arief et al., 2018).
- Online OCR: this tool can turn pictures taken with a digital camera, faxes, and scanned documents into different formats that can be searched and changed. It is also possible to process documents written in more than one language. It lets you convert up to 15 images per hour for free as a guest without signing up, but signing up for free gives you access to more features (Isheawy et al., 2015).

Table 2.3 OCR Tools

Sr#	Tool	Available	Operating System	Features
1	Tesseract	Open Source	Windows, Linux, MAC OS X, Android	• Can use more than one language in one scan • Machine learning can be used to understand new languages, symbols, and fonts. • Doesn't work with GPUs.
2	Ocrad	Open Source	MAC OS X, Linux, BSD	• In the pre-processing step, you can cut, rotate, scale, and find the layout. • In the post-processing step, you can use both built-in and user-defined filters.
3	GOCR	Open Source	Windows, BSD, Linux, MAC OS X	• Graphical user interface (GUI). • No training data is needed (no neural network). • Barcodes can be read and translated.
4	OCRopus	Open Source	MAC OS X, Linux, BSD	• It can be taught to understand new languages and fonts, and Google Books uses it. • It also works with GPUs in OCRopus.
5	ABBYY Web service	Online	Platform independent	• Provides the highest level of data security by following all applicable data protection laws. • Keeps formatting. • Can also convert multi-page documents. • For a document written in more than one language, you can choose up to three recognition languages. • Maximum input file size: 30 MB.
6	Google Docs	Online	Windows, MAC OS X, Android, iOS,	• Automatically figure out what language the document is in, so you don't have to.

			ChromeOS, BlackBerry	● Maximum input file size: 50 MB. ● At the moment, OCR works best on documents that have been scanned well, have a high resolution, and use the most common fonts.
7	Free-Online OCR	Online	Browser-Based	● Maximum input file size: 200 MB. ● It automatically turns the pages, works with low-resolution images, keeps the original layout and formatting, and has a number of other features.
8	Online OCR	Online	Browser-Based	● For the best text recognition, images should be between 200 and 400 DPI. ● The maximum size of an input file is 200 MB. Images are automatically rotated (full-page de-skew) for better recognition. ● Colored areas that are not text are put back into the final document.
9	ABBYY FineReader	Proprietary	Windows, MAC OS X, Linux	● Techniques for pre-processing include getting rid of noise and fixing skew. ● Uses AI and ML to reconstruct documents more accurately and with more precision.
10	Transym OCR	Proprietary	Windows	● Automatically figure out which way the page or image is facing. ● Can find text with problems in the background ● Uses lexicon to make sure words are correct and reliable as much as possible.
11	Readiris	Proprietary	Windows, MAC OS X, iOS, Android	● Self-learning techniques based on neural networks are added to font-independent text recognition. ● Has its own dictionaries.
12	Adobe Acrobat	Proprietary	Windows, MAC OS X, iOS, Android	● Text can be changed in PDFs; a custom font is made. ● Create intelligent PDFs (only searching and copying capabilities without editing).

In this section, we looked at different content-based image retrieval methods. These methods look at the content of an image, pull out the features that describe it, and come up with annotations or labels for the image. Most of the time, an algorithm for machine learning is used to get these annotations. Existing machine learning algorithms are hard to use because you have to import a lot of training images and use a lot of CPU time. The application programming interface (API) for Google Cloud Vision has fixed these two problems. Cloud Vision API has been trained by Google, so it saves time when getting labels for images. We'll use Google Vision API to get information from images that don't have text in them so we can check if the images on websites are easy to read. On the other hand, we have analyzed the OCR tools and observed that Tesseract is much better than other tools. Its better precision and accuracy, open-source, efficiency and support for all three major operating systems (Windows/ Linux/ Mac OS) make this an ideal choice. Also, it has pre-processing techniques like detecting orientation and making minor corrections to skew. It can scan in more than one language at once. It can also recognize new languages, symbols, and

fonts with the help of machine learning. In this research, we have used Tesseract to extract information from text images on websites.

2.4 Discussion

In the literature review we have analyzed different readability techniques that make websites more readable for users, and measures the relationship between readability and images in multimedia documents. . In the first section, plain text readability has been studied in which classical and similarity methods were analyzed. In classical methods, different readability models and tools we've looked at and studied. The models that different researchers made were made to figure out how easy it is to read different kinds of writing. We found that most of the formulas guessed the level of readability in terms of the US grading scale. The primary and secondary grade levels are different depending on where you live and what the environment is like. Researchers came up with the Fernandez Huerta Index, Djoko formula, Kandal and Moles Index, and Al-Heeti grade level as ways to predict how easy it would be to read Indonesian, Spanish, French, or Arabic text. The formula developed for non-English texts showed a correlation with people's reading abilities. Most of the earlier readability formulas took into account things like the length and number of words, sentences, syllables, and complex words (Crossley et al., 2011). After the 1980s, tools like the Read-X, ATOS, Lexile Framework, Coh-metrix, and the new Dale-Chall readability formula were made. These tools measure readability by taking into account things like cognitive-structural elements, semantic units, and the complexity of syntactic structures.

While in similarity methods, different similarity similarity techniques have been analyzed. Character-based similarity and token-based similarity are the two main types of string-based similarity. The corpus-based similarity is based on how words are used (Mihalcea et al., 2006). The method helps figure out how similar two ideas are based on the information in their respective corpora, which is a group of written, spoken, or electronic text. These methods store a set of sentences and their translations into other dialects. The goal is to match the input text in the corpus with the final translations. The knowledge-based similarity measures are made up of a list of semantic measures taken from semantic networks and used to compare pieces of information. The goal of this kind of information is to figure out how similar words are to each other. Semantic relatedness and semantic similarity are two parts of knowledge-based similarity (Makvana et al., 2016). Similarity measures for hybrid classifications do not make a separate group. They combine parts of the previous methods to try to get the best of each (El Desouki et al., 2019). These methods use recursive steps to get around the problems with the other methods.

Some research has been done on how easy it is to see images on websites. The W3C accessibility guidelines say that each image should have a textual equivalent. The alt and longdesc attributes of the img tag in the HTML standard make it easy to add this alternative text. The title attribute is another non-standard way to give alternative text. Many accessibility tools, like screen readers and refreshable Braille displays, can read this text. One reason why the web isn't accessible is that people who make websites don't give enough alternative text. Many people think that choosing the

right alternative text is more of an art than a science, which makes it harder to build and check alternative text. Images that are important to understanding a page or finding your way around it should have alternative text. Images that are only there to make a page look better should have an alt attribute with a length of 0 to make this clear. If you don't follow these rules for accessibility, most web pages will be harder to use.

Images that do something (like links or buttons) or have more than one colour or are bigger than a certain size are especially dangerous. When there is no alternative text for these images, it can be challenging for some people to see them. The WebInSight system looked for these crucial images and allowed the correct alternative text to be added automatically. It handles web requests and changes the pages sent back on the fly. As part of the process of changing, it coordinates three new image-labeling modules built for this domain. These modules use methods based on improved web context analysis, optical character recognition (OCR), and human labeling. But there is no any work on image relevancy on websites from a readability perspective, because relevant images enhance web readability (Elahi et al., 2022A) (Elahi et al., 2022B).

Chapter 3: Methodology Proposal for Readability Evaluation of Multimedia documents

In this chapter we propose a methodology to measure the relevance of images on the website for readability purposes. In order to compute image relevancy, these are the fundamental steps followed:

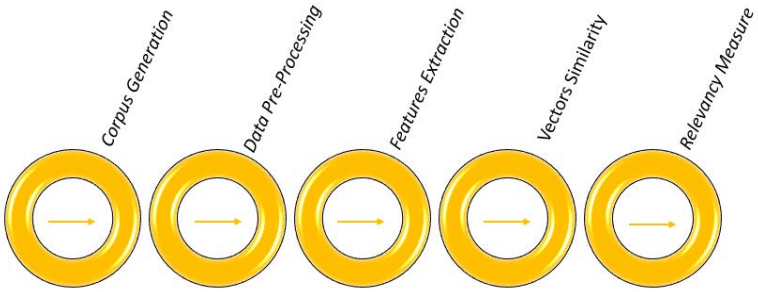


Figure 3.1 Methodology

3.1 Corpus Generation

Corpus generation is the first step in which text and images are extracted from the fifty different best educational websites in Pakistan. Educational institutes recognize the value of having a strong online presence today. The importance of an educational website for students through which they can easily study. An educational website that is easy to use is suitable for students and parents because it lets them see everything from their dashboard. It is essential to say what you want to say in a way that is clear and easy to understand. Your website shows what you look like online. Students can see how they are doing in school by logging into their accounts on the Organization website. Students learn much more when they have a website for their online classes. By taking online classes, they can learn whenever they want. So In recent years, it has become imperative to website design for educational institutions such as universities and colleges. The major obstacles faced by educational websites include achievement of readability and accessibility of websites. If a website is easy to read, visitors will need help to use all of its features. Educational institutions should ensure that websites are built according to rules for readability so that users are happy.

Table 5 contains a listing of the websites that were accessed in order to compile the corpus. A method known as picture web scraping was used, and the results were extracting images. We have collected around 500 images, of which 180 are without any accompanying text. After the images have been extracted, Google's vision AI services will classify and categorize any of them that do not contain any accompanying text. The artificial intelligence behind Google Vision assigns labels to visual input and efficiently organizes it into millions of predefined categories. This helpful tool locates objects and faces, deciphers printed and handwritten text, and generates vital metadata that

enables us to match the text on the image with the text on the webpage. For example, in Figure 3.2, information that was extracted from the image that did not contain words is “Smile, Trousers, Plant, Grass, Leisure, Recreation, Fun, Competition Event, Event, Lawn, Crowd, Team, Happy, Public Event, Academic Institution, Player, T-shirt, Sitting, Campus, University, Tourism ” with different confidence scores.

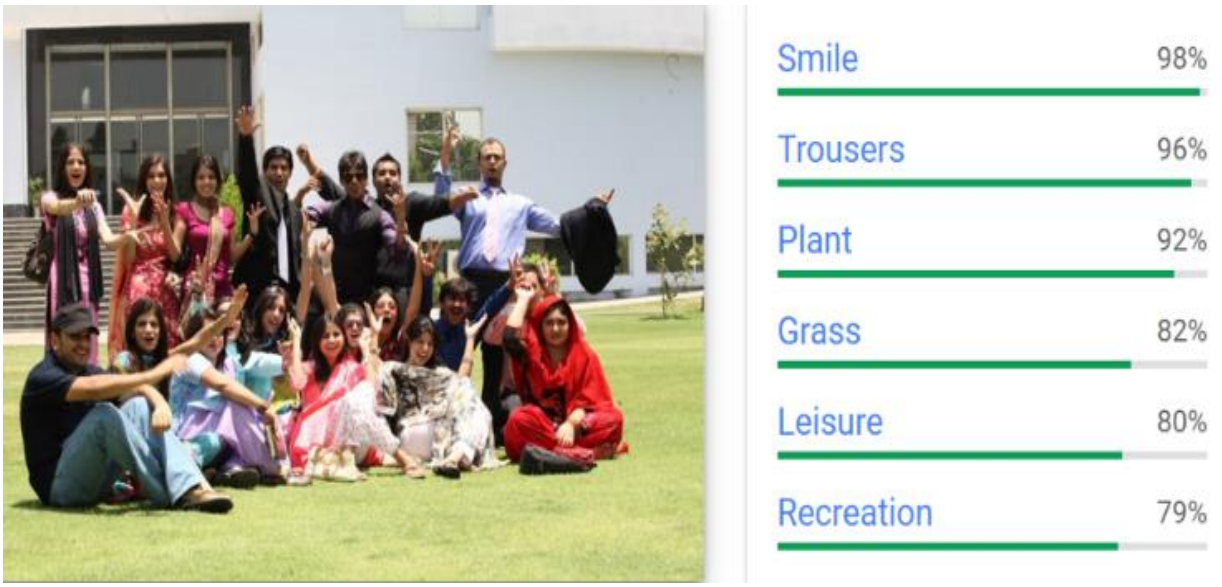


Figure 3.2 Information extraction from Non-text Image

On the other hand for text images, each individual text image is passed to an OCR (Optical Character Recognition) tool and the text is extracted from images. We have analyzed the performance of different OCR tools and observed that Tesseract is much better than other tools because it has better precision as well as accuracy than other tools, it is open-source, It takes less time in conversion than other tools, it works with Windows, Linux, and Mac OS and has pre-processing techniques like detecting the image's orientation and making minor adjustments to its skew. It can scan in more than one language at once. It helps recognize new languages, symbols, and fonts with machine learning. That’s why we have used the Tesseract tool in this thesis to extract information from the graphical contents on the websites. A console-based application using Tesseract was built in the .Net framework for this purpose. This application processes each image and extracts text along with a confidence score. The confidence score gives you an idea of how accurate you can expect the results to be. The accuracy of the extracted text goes up as the confidence score goes up. Figure 3.3 and Figure 3.5 show how an example image was processed, while Figure 3.4 and Figure 3.6 show the results. The confidence score gives you an idea of how accurate you can expect the results to be. The accuracy of the extracted text goes up as the confidence score goes up. For example, in figure 3.7(a), the confidence value is 0.90, which is very good, and the extracted text is "24 International Conference on Business, Management, and Social Sciences (ICBMASS-22)". Figure 3.7(b), on the other hand, has a bad OCR system because

the text is on top of the background image and there isn't enough difference between the colour of the text and the colour of the background.

Figure 3.3 Input text Images

```
Microsoft Windows [Version 10.0.19042.1165]
(c) Microsoft Corporation. All rights reserved.

C:\Users\HP>cd C:\Users\HP\Desktop\repos\repos\tesseract-samples\src\Tesseract.ConsoleDemo\bin\Debug\netcoreapp3.0

C:\Users\HP\Desktop\repos\repos\tesseract-samples\src\Tesseract.ConsoleDemo\bin\Debug\netcoreapp3.0>Tesseract.ConsoleDemo InputFiles
Detected 24 diacritics
Mean confidence: 0.25
Text (GetText):


Ny QS Ranking: NUST. ora es Aten ees ES CUIR:)
PLE eitsee)(olerlN7

Text (iterator):
<BLOCK>


<BLOCK>
Ny QS Ranking: NUST. ora es Aten ees ES CUIR:)
PLE eitsee)(olerlN7
```

Figure 3.4 Output text extraction

Figure 3.5 Input text Images

```
Microsoft Windows [Version 10.0.19042.1165]
(c) Microsoft Corporation. All rights reserved.

C:\Users\HP>cd C:\Users\HP\Desktop\repos\repos\tesseract-samples\src\Tesseract.ConsoleDemo\bin\Debug\netcoreapp3.0

C:\Users\HP\Desktop\repos\repos\tesseract-samples\src\Tesseract.ConsoleDemo\bin\Debug\netcoreapp3.0>Tesseract.ConsoleDemo InputFiles
Mean confidence: 0.62
Text (GetText):
By Accomplished Alumni to
Pursue
in USA

Learn how NUSTIAN USA` free coaching
program can help you pursue your
Ceutted tee tart

apecnc pres

c 9:00 pm PST

elat lise aes Cl 5)



Text (iterator):
<BLOCK>
By Accomplished Alumni to
Pursue
in USA
```

Figure 3.6 Output text extraction

The selection of the font size and aspect ratio may need some more attention. In this particular illustration, the level of confidence that the software possesses is 0%, and the text that was retrieved from the image is “ata Duss uroy te rN aaNet rol sels BN UN ake Cor err”. At the same time, the text from the webpage was extracted by our program.

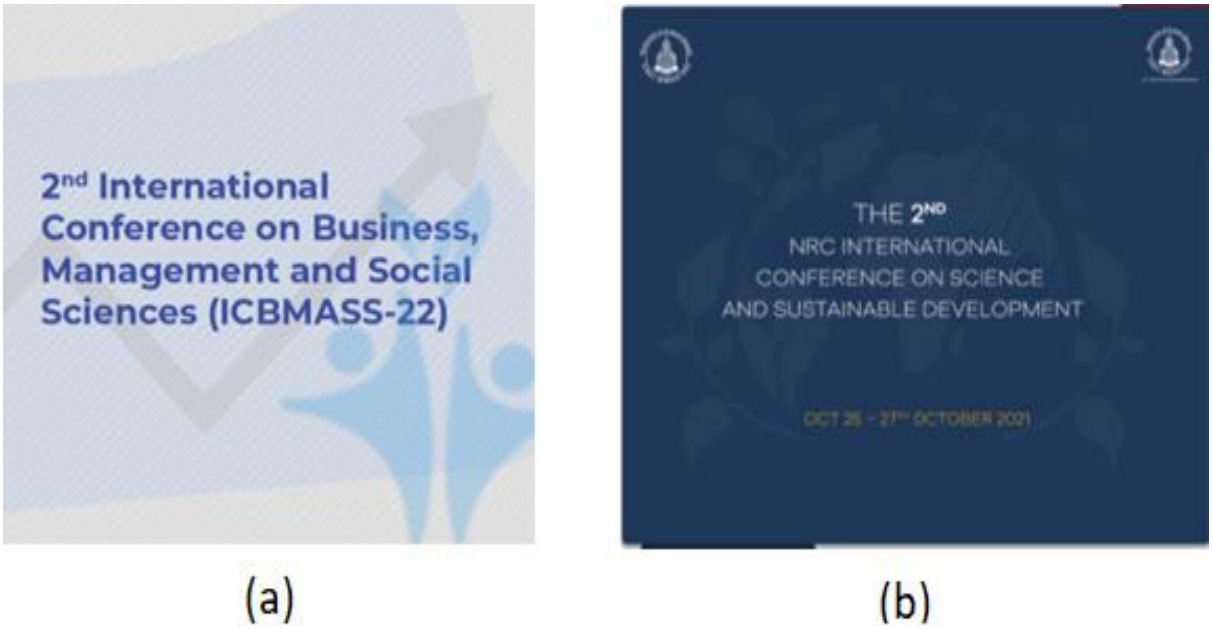


Figure 3.7 Input Images

3.2 Data Pre-processing

Once the information from the images and web page is taken out. Pre-processing has been done because the data needs to be cleaned up and put into a format that is easy to understand and can be analyzed for relevance evaluation. During the pre-processing, the following steps were taken:

3.2.1.1 Tokenization

Tokenization is a simple process that turns raw data into a string of data that can be used. Tokenization is well-known for its use in cybersecurity and the creation of non-fungible tokens, but it is also an essential part of the NLP process. For example, in natural language processing, tokenization breaks up paragraphs and sentences into smaller pieces that are easier to assign meaning to. So the first step in the NLP process is to get the information (a sentence) and break it down into parts that are easy to understand (words). Here's what a string of data looks like:

"Images on educational websites"

So that a machine can understand this sentence, the string is tokenized so that it can be broken up into its parts. This is what would happen with tokenization:

‘Images’ ‘on’ ‘educational’ ‘websites’

This may seem simple, but if you break a sentence into its parts, a machine can understand both the parts and the whole. This helps the program understand each word's meaning and how it fits into the whole text. This is especially important to find terms for image relevancy computation.

3.2.1.2 Remove Stop Words

Stop words are terms that are typically eliminated from a natural language before they are processed to be more easily understood. These are the most often used words in any language, such as articles, prepositions, pronouns, conjunctions, and so on; nonetheless, they contribute little to the meaning of the text. A few English stop words are "the," "a," "an," "so," and "what." There are a lot of stop words in every human language. By taking out these words, we get rid of the low-level information in our information so that the important information gets more attention. In other words, we can say that taking out these words doesn't hurt the model we are training for our task. Getting rid of stop words definitely cuts down on the size of the dataset, which in turn cuts down on the training time because there are less tokens to train on. We don't always take out the pauses. Getting rid of stop words depends a lot on the task we're doing and the goal we're trying to reach. For example, we might not take out the stop words when training a model to do sentiment analysis.

"The movie was not good at all," said the reviewer.

After taking out the stop words: "good movie"

We can tell from the review that the movie wasn't good. But when the stop words were taken out, the review became positive, which isn't true. So, getting rid of stop words can be hard in this case. Most of the time, stop words are not needed for tasks like text classification because the other words in the dataset are more important and give the general idea of the text. So, when we do tasks like this, we usually get rid of stop words.

3.2.1.3 Stemming and Lemmatization

This is another step before figuring out how relevant the text taken from images is to the text on a webpage. We did stemming in our application after removing stop words. For example, the roots of the words ``likes," "likely," and "liked" are all "like," which can be used as a synonym for all three words. So, an NLP model can figure out that all three words are similar and are used in the same way. Stemming lets we standardize words to their base stems, no matter how they are formed. This is useful for many things, like grouping or clustering text. Search engines use these methods a lot to give better results no matter how the words are spelt. We call it "over-stemming" when our program links different words that have nothing to do with each other to the same root. Even though the words ``universal," "university," and "universe" all have the same root word, they mean very different things. When we type these words into a good search engine, the results should be very different, not the same as if they were the same word. A mistake like this is called a "false positive." Under-Stemming is the opposite of overstepping. It happens when two or more words don't come from the same root, even though they should. The word "alumnus" refers to a former college student, and it is usually used to talk about men. "Alumnae" is the word for women, and "alumni " is a group of former college students. A basic search engine or other NLP program

should definitely treat these words as synonyms. But most Stemming algorithms don't cut it to their common root, which is a false negative error.

Lemmatization is an extension of stemming. It is the process of putting all the different forms of a word together so that they can be looked at as a single unit. Lemmatization is like Stemming, but it gives the words more meaning. So, it connects words that mean the same thing to one word. Most lemmatization algorithms also take positional arguments as inputs, like whether the word is a noun, verb, or adjective. When preprocessing text for natural language processing (NLP), we need both stemming and lemmatization. Sometimes, both of these words are used as if they mean the same thing, even though they don't. Most of the time, Lemmatization is better than Stemming because it looks at words in their context rather than using a hard-coded rule to cut off suffixes. But Lemmatization takes a lot more time if the text documents are very long, which is a major drawback. For Stemming and Lemmatization purposes, our application uses a third-party library called pluralize.net.

3.2.1.4 Uniform Case

Since the way a computer works with text depends on the case, all of the text needs to be changed to the same case. For example, Cat and cat are both the same word, but the capitalization is different. So, we should make all of the text the same case, preferably lowercase.

3.2.1.5 Remove punctuation Characters

Punctuation characters are $, ?, “, !, etc. C# function provides the list of punctuation. We have removed punctuation characters because they are not providing any information associated with semantic similarity.

3.3 Features Extraction

Once the text has been cleaned and images have been labeled and classified, the extraction of the main features of images and text is done. In this step, the representation of the text (sequence of sentences or words) into a numeric vector is calculated by using techniques of Natural Language Processing. Term Frequency and Word2Vec techniques have been utilized in this step, as well as the synonym search technique in the next order:

3.3.1 Expand the terms with their synonyms

A list of terms was expanded by including their synonyms. For this purpose, Word2Vec has been used in the application. This service takes the list of words and returns their synonyms. Therefore, the list of words obtained after stemming was passed as an input to Word2Vec and a list of their synonyms was obtained. Word2vec is a set of models that work together to make word embeddings. These models are shallow, two-layer neural networks that are taught to figure out the meanings of words from their contexts. Word2vec takes a large collection of text as input and turns

it into a vector space, which usually has several hundred dimensions. Each unique word in the collection is given a vector in the space. Word2vec can use either the continuous bag-of-words (CBOW) or the continuous skip-gram model architecture to make these distributed representations of words. In both architectures, as word2vec goes through the whole corpus, it looks at both individual words and a sliding window of context words around each word. In the continuous bag-of-words architecture, the model figures out what the next word will be based on the window of words around it. The order of words in a context doesn't change how you guess (bag-of-words assumption). In a continuous skip-gram architecture, the model uses the current word to make predictions about the words in the window around it. The architecture of skip-gram gives more weight to words close to each other in context than to words farther away. The authors' note said that CBOW is faster and skip-gram is better for words that don't come up very often. After the model has been trained, the learned word embeddings are placed in the vector space so that words with similar meanings and structures that appear in the same contexts in the corpus are close to each other. The words that are most different from each other are farther apart in space. This happens because words that affect the relative probabilities of other words in similar ways will have learned similar embeddings when the model is done. For example, you could think of the CBOW framework as a "fill in the blanks" task. The learned embedding for a word will show how it affects the relative chances that other words will appear in the "blank," or the spot in the middle of the context window, when that word appears in that window. So, semantically similar words should have similar effects on these probabilities, since semantically similar words should be used in similar ways.

3.3.2 Term Frequency (TF)

Frequency is the number of times a word appears in a text compared to the total number of words in the text. Information taken from images is linked to a number that shows how closely each word on a website is related to the text. Images and website text that use the same or similar words will have vectors that are similar. This is what we see when we use the cosine similarity method. Cosine similarity measures the similarity between two term frequency vectors.

3.4 Relevancy Measure of Images in the Web

The main goal of this research is to find out how relevant the information taken from the images is to the text on the web pages. In this research, the Cosine similarity technique (Qurashi et al., 2020) is used because it works better than other similarity techniques like the Jaccard and Euclidean (Huang, 2008) distance techniques and gives more accurate results. This method checks to see if two vectors are related. Text and labels taken from graphical content and web pages can be identified by a number of features, each of which keeps track of how often a certain keyword appears. So, a term frequency vector is a way to describe text that has been extracted from images and web pages. Figure 14(a) and web page text term frequency vectors are:

Table 3.1 Frequency Vectors

	2nd	International	Conferenc e	Business	Management	Social	Science	ICBMASS-22
Figure	2	2	2	2	2	2	2	1
Webpage Text	0	3	2	1	2	3	1	0

Table 3.1 shows the number of times each word appears in the text from Figure 3.7(a) and the text of the web page. For example, the word "International" was found twice in the text taken from Figure 3.7(a), so its term frequency in Figure 3.7(a) is 2, but this word doesn't exist in the text of the web, so its term frequency there is 0. With these vectors, the relevance of this graphic to the text on the website it's on is calculated to be 0.72. After figuring out how relevant each image on a web page is, the average of those relevancies is used to figure out how relevant the page's graphics are as a whole. Figure 3.8 shows the whole process of figuring out how relevant something is.

Suppose A and B are two vectors for assessment. By using the cosine measure as a relevancy function, we have

$$Relevancy\ (A,B) = \frac{A.B}{||A||||B||} \qquad Equation - I$$

Here ||A|| is the Euclidean norm of vector A = $(A_1, A_2, A_3, A_4, \dots\dots\dots A_n)$, characterized as $\sqrt{A_1^2, A_2^2, A_3^2, A_4^2, A_5^2\ \dots\dots\dots A_n^2}$. Conceptually, it is the length of the vector. Correspondingly, ||A|| is the Euclidean norm of vector A. The measurement calculates the cosine of the angle among vectors A and B. For example 3.7(a) figure shows vectors with term frequency of extracted text from graphical content and text of webpage. Relevancy of this graphical content with text of its containing website is computed by using Equation – I in the following way:

$$Relevancy\ (A,B) = \frac{(1*0)+(1*2)+(1*0)+(1*0)+(1*0)+(1*2)+(1*0)+(1*5)+(1*2)+(1*4)+(1*0)}{\sqrt{(1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2)} * \sqrt{(0^2+2^2+0^2+0^2+0^2+2^2+0^2+5^2+2^2+4^2+0^2)}} = 0.62$$

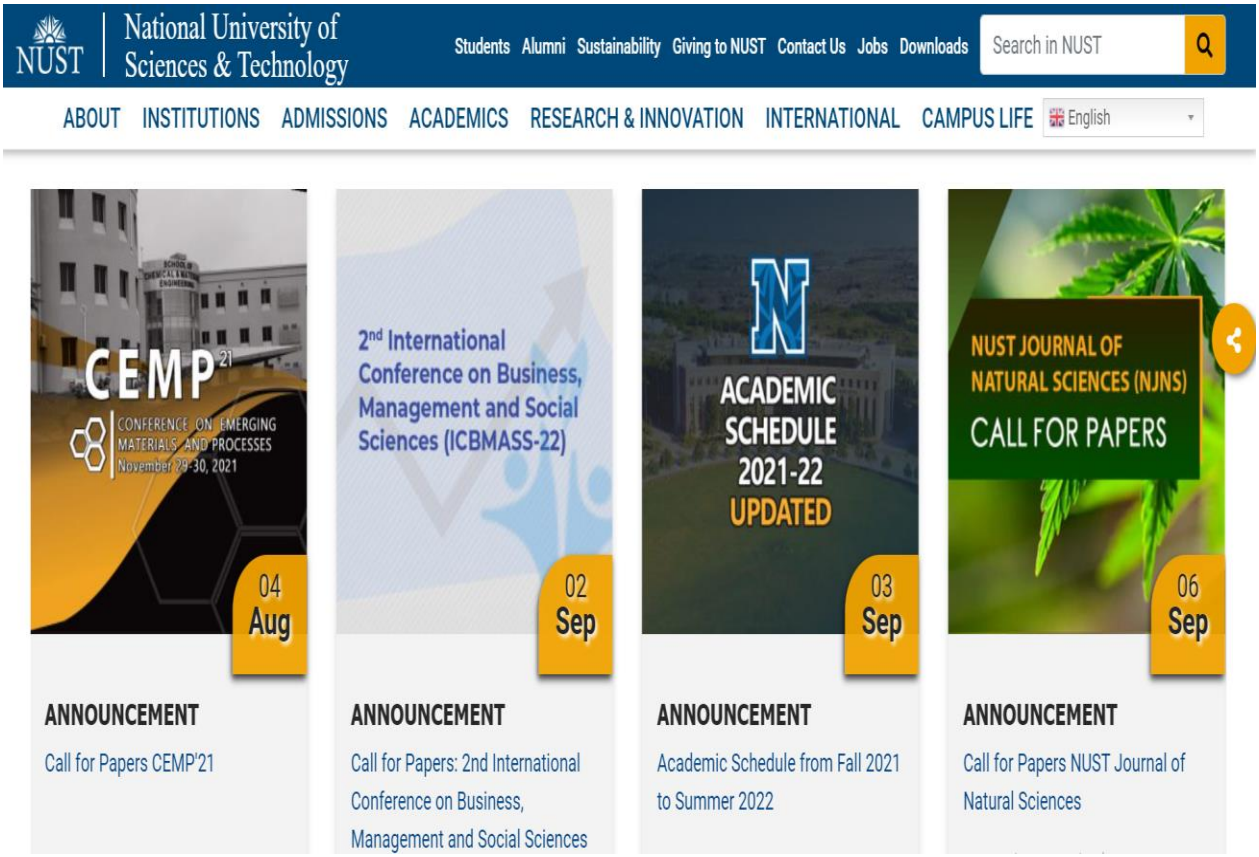


Figure 3.8 Input Images

```
Term Frequency Vector of Graphical Content:
8  CONFERENGE  SONSEMERGING  olen  Novemberg29-30, 2021   AUg  ANNOUNCEMENT  Call  Papers CEMP'21
1      1          1         1      1            1     1      1         1     1      1
Term Frequency Vector of Webpage Text:
0      2          0         0      0            2     0      5         2     4      0
Computed Relevancy: 0.62
```

```
Term Frequency Vector of Graphical Content:
2nd   International   Conference  Business   Management  Social Sciences (ICBMASS-22) ANNOUNCEMENT Call   Papers
2        2             2         2         2         2      2         1          1      1       1
Term Frequency Vector of Webpage Text:
0        3             2         1         2         3      1         0          5      2       4
Computed Relevancy: 0.72
```

```
Term Frequency Vector of Graphical Content:
ACADEMIC N14 9/00  2021-22 ANNOUNCEMENT rom Fall 2021
    1       1   1       1         1            1    1      1
Term Frequency Vector of Webpage Text:
    5       0   0       3         5            0    4      2
Computed Relevancy: 0.75
```

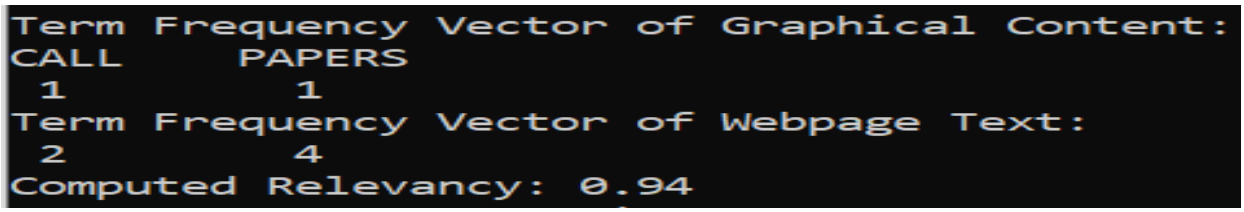


Figure 3.9 Output Images

Same way the relevancies were computed for other three graphical contents in Figure 3.8 as well.

$$Relevance\ (A,B) = \frac{(2*0)+(2*3)+(2*2)+(2*1)+(2*2)+(2*3)+(2*1)+(1*0)+(1*5)+(1*2)+(1*4)}{\sqrt{(2^2+2^2+2^2+2^2+2^2+2^2+2^2+1^2+1^2+1^2+1^2)} * \sqrt{(0^2+3^2+2^2+1^2+2^2+3^2+1^2+0^2+5^2+2^2+4^2)}} = 0.72$$

$$Relevance\ (A,B) = \frac{(1*5)+(1*0)+(1*0)+(1*3)+(1*5)+(1*0)+(1*4)+(1*2)}{\sqrt{(1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2)} * \sqrt{(5^2+0^2+0^2+3^2+5^2+0^2+4^2+2^2)}} = 0.75$$

$$Relevance\ (A,B) = \frac{(1*2)+(1*4)}{\sqrt{(1^2+1^2)} * \sqrt{(2^2+4^2)}} = 0.94$$

So the overall average relevancy of graphical contents with text of web page is:

((0.62+0.72+0.75+0.94)/4)=0.76

Here the average of all relevancy values of individual graphical contents were taken overall relevancy of graphical contents with their container webpage.

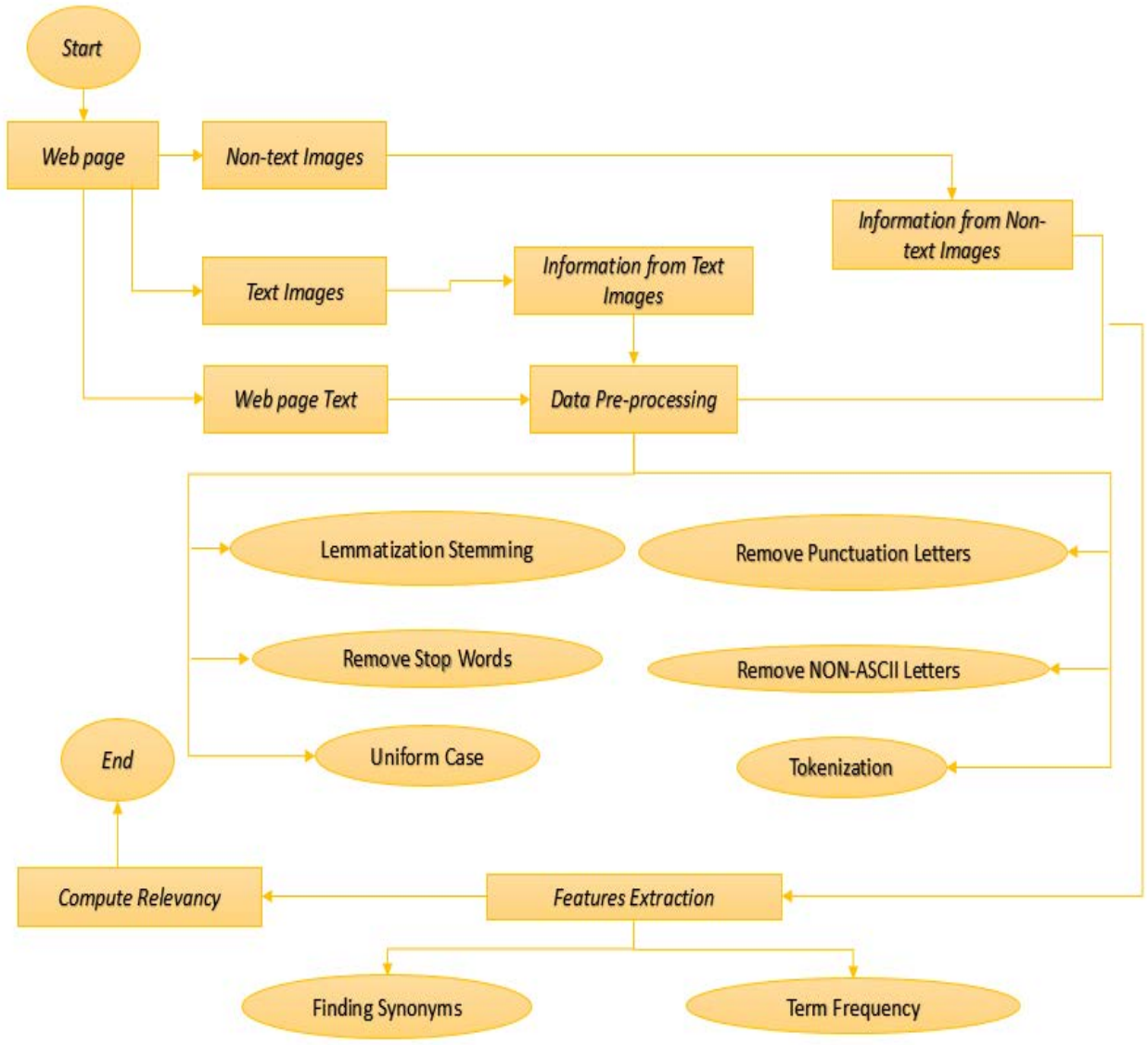


Figure 3.10 Workflow of Relevancy computation

In this chapter, we have proposed methodology that computes the relevancies of the images with the websites. In the next chapter, we will evaluate our proposed methodology and proposal for relevant images that could enhance the readability of the web pages by using a user study.

Chapter 4: Evaluation

In this chapter, first we have evaluated our proposed methodology by using the educational websites in Pakistan then evaluate the proposal that relevant images could enhance the web readability.

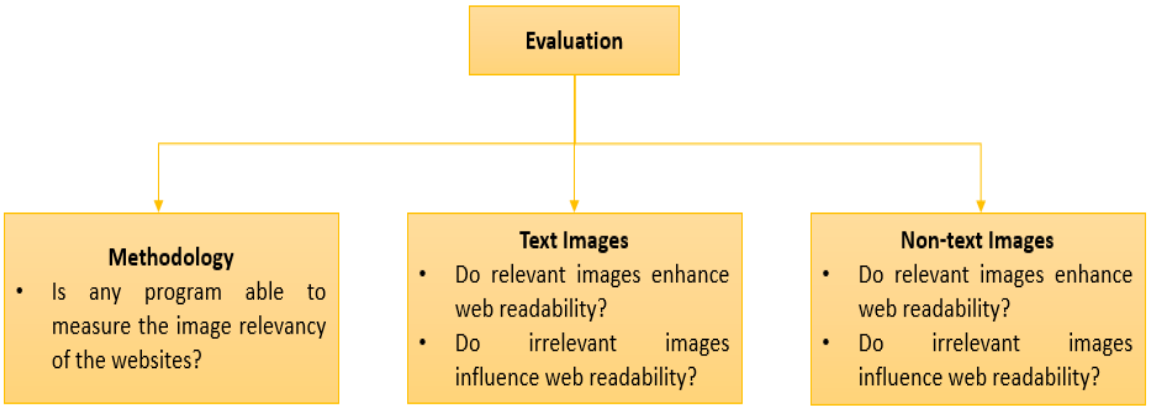


Figure 4.1 Evaluation Process

4.1 Evaluation of the Methodology

For the proposed methodology evaluation we looked at 50 educational websites in Pakistan and rated them. Figure 4.2 shows how the relevance of the websites in the corpus chosen for the evaluation is spread out. Three groups of results were made. Text taken from images on 11 out of 50 websites matched the text on those websites 50–60% of the time. This was true for 60–70% of 24 websites, and 70–80% of the graphical content on 16 other websites was also true for those websites. The results also show that between 70 and 80% of the graphics on about 40% of websites are related to the websites. Figure 4.3 shows each site's score for how relevant it is.

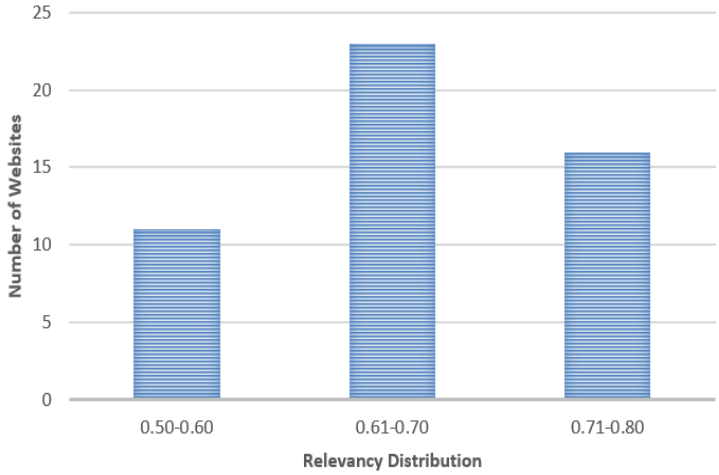


Figure 4.2 Websites Relevancy Distribution with Text Images

On the other hand, information taken from non-text images on 16 out of 50 websites matched the text on those websites 50–60% of the time. This score was between 61 and 70% for 20 websites, while 14 other websites had non-text images that were between 71 and 80% relevant to the websites, as shown in Figure 4.3.

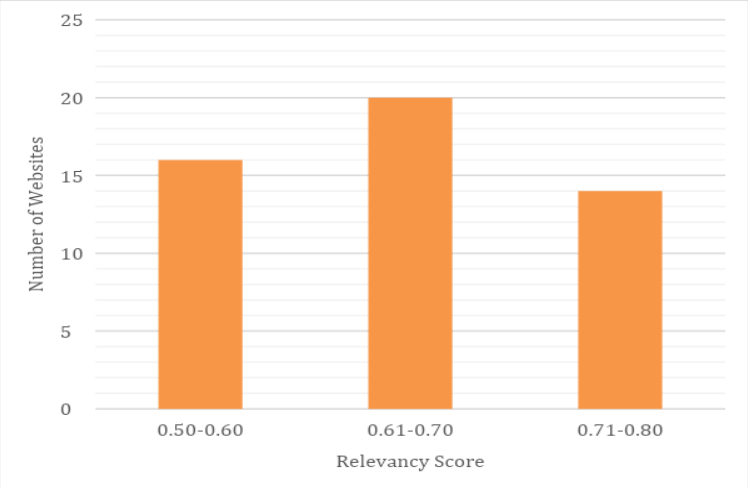


Figure 4.3 Websites Relevancy Distribution with Non-Text Images

So, if you use these websites with 70–80% relevancy as a standard, at least 60% of websites could be better if they made their graphical content more relevant. For example, the National University of Science and Technology (NUST) website has a relevancy score of 0.72. From the screenshot in Figure 4.4, it's clear that the page's highlighted image is a good fit for the page.

Figure 4.4 Example of Better Relevancy due to relevant Content

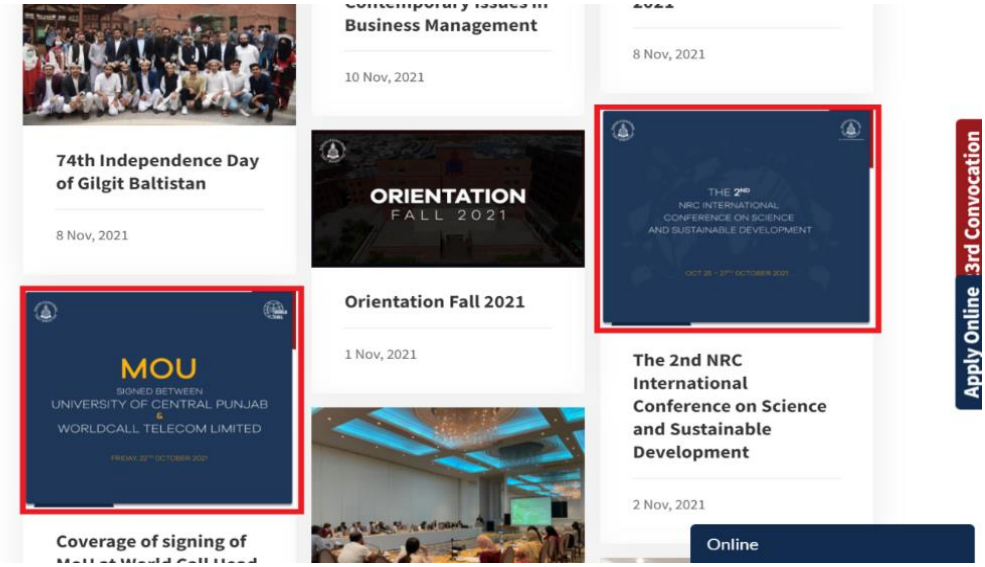


Figure 4.5 Example of Worse Relevancy due to Poor Quality Content

In this case, the confidence values of text extracted from graphics are good because the graphics are well balanced in terms of aspect ratio, colour contrast, resolution, font style, font size, etc. Second, the text in images is the same as the text on a webpage. The University of Central Punjab (UCP), on the other hand, has a relevance score of 0.50. Figure 4.5 is a screenshot that shows that images on web pages are not easy to read. Aspect ratio, font size, colour contrast, etc. are not appropriate. This makes OCR work less well, so there is less confidence in the text that is extracted. Because of this, it was found that text taken from graphics was less relevant to the context of the web page. Figure 4.6 is another example. The image quality is good, and OCR worked well, but the text taken from the image doesn't match the text on the web page very well. Again, this led to a low score for relevance.

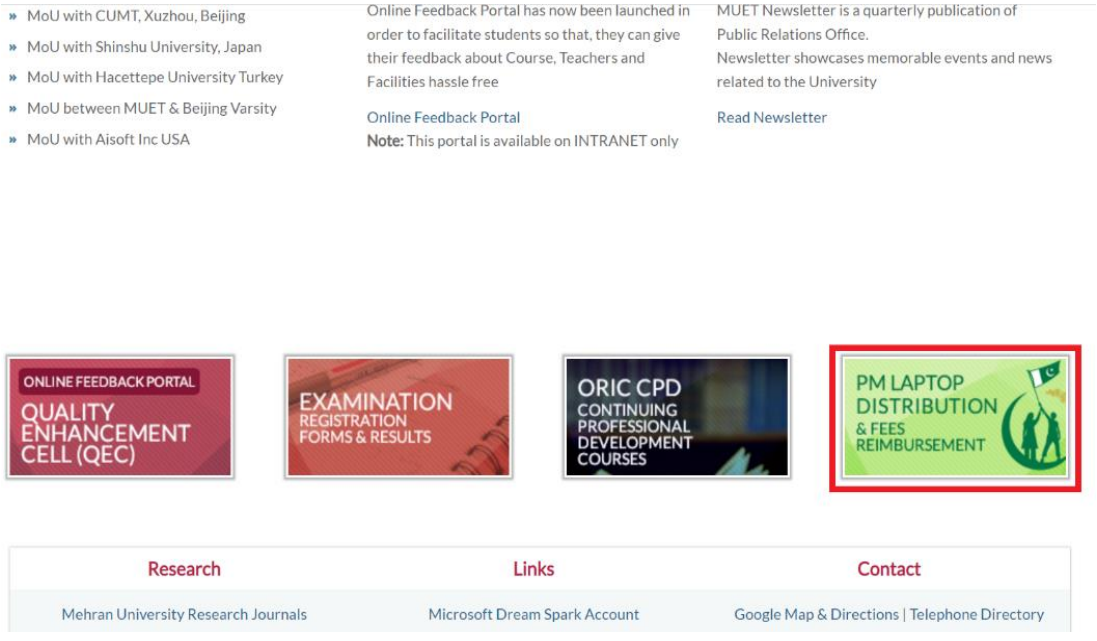


Figure 4.6 Example of Worse Relevancy due to Irrelevant Content

Another example of non-text images, Hajvery University has a relevancy score of 0.48. As it's obvious in Figure 4.7, the image used on the web is out of the context of the webpage. Consequently, extracted information from non-text images has been found to be less relevant to the textual content of the web page.

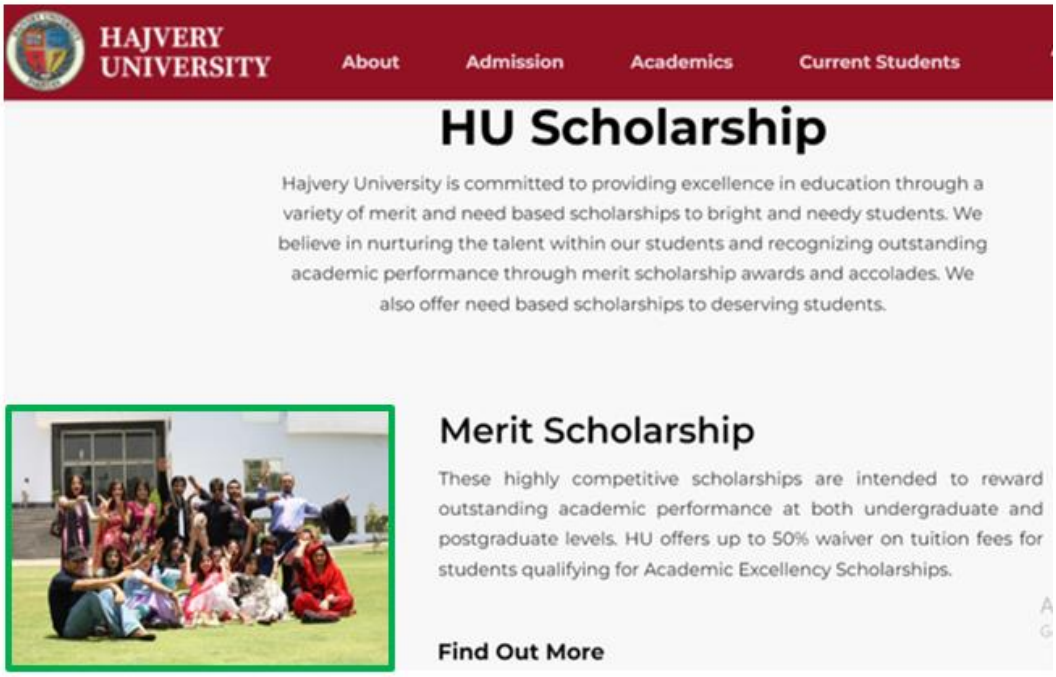


Figure 4.7 Web page with irrelevant non-text image

Table 4.1 Educational websites relevancy scores with text and non-text images

Sr#	Website	Score with Non-Text Images	Score with Text Images	Overall Score
1	http://nu.edu.pk/	63	67	65
2	https://nust.edu.pk/	70	72	71
3	https://lums.edu.pk/	59	58	58.5
4	https://www.giki.edu.pk/	47	61	54
5	https://itu.edu.pk/	65	60	62.5
6	http://www.pu.edu.pk/	59	51	55
7	https://www.aku.edu/	47	49	48
8	https://www.umt.edu.pk/	61	62	61.5
9	https://www.comsats.edu.pk/	67	71	69
10	https://uol.edu.pk/	58	61	59.5
11	https://www.bahria.edu.pk/	64	67	65.5
12	https://www.iba.edu.pk/	61	58	59.5
13	https://www.riphah.edu.pk/	68	70	69
14	https://www.iobm.edu.pk/	46	45	45.5
15	https://uet.edu.pk/	65	71	68
16	https://www.iub.edu.pk/	66	58	62
17	https://qau.edu.pk/	63	67	65

18	https://superior.edu.pk/	58	62	60
19	https://www.ucp.edu.pk/	63	50	56.5
20	https://iiu.edu.pk/	59	62	60.5
21	http://www.uaf.edu.pk/	47	71	59
22	https://www.uok.edu.pk/	49	61	55
23	https://gcuf.edu.pk/	47	67	57
24	https://www.bzu.edu.pk/	58	58	58
25	https://su.edu.pk/	61	69	65
26	https://www.neduet.edu.pk/	58	45	51.5
27	https://www.muet.edu.pk/	47	71	59
28	https://www.fccollege.edu.pk/	61	58	59.5
29	https://www.kmu.edu.pk/	68	67	67.5
30	https://szabist.edu.pk	58	62	60
31	http://www.pieas.edu.pk/	67	66	66.5
32	https://www.uaar.edu.pk/	57	62	59.5
33	https://www.pide.org.pk/	67	71	69
34	https://www.numl.edu.pk/	47	61	54
35	http://www.uop.edu.pk/	67	67	67
36	https://gcu.edu.pk/	59	59	59
37	https://usindh.edu.pk/	47	47	47
38	https://www.hamdard.edu.pk/	61	61	61
39	https://www.uhs.edu.pk/	63	63	63
40	https://uog.edu.pk/main.php	58	58	58
41	https://www.hup.edu.pk/	48	45	46.5
42	https://pafkiet.edu.pk/main/	56	62	59
43	https://fjwu.edu.pk/	51	66	58.5
44	https://iqra.edu.pk/	61	62	61.5
45	https://www.aup.edu.pk/	46	71	58.5
46	https://web.uettaxila.edu.pk/	58	61	59.5
47	https://ssuet.edu.pk/	48	67	57.5
48	https://ndu.edu.pk/	64	59	61.5
49	https://www.au.edu.pk/	67	47	57
50	https://www.ntu.edu.pk/	47	61	54

In order to evaluate the proposal we have proposed following research questions:

i. Do relevant images enhance web readability?
ii. Do irrelevant images influence web readability?

A user study has been conducted to find the answers of above research questions. In a user study, two evaluations, evaluation by final users and evaluation through readability experts, have been performed. Different questions have been asked and readability scores have been computed. Best, average and worst websites have been selected by using the proposed automatic tool. At the end, the relevancy scores of selected websites were compared with readability scores computed in the user study in order to find the research question's answer.

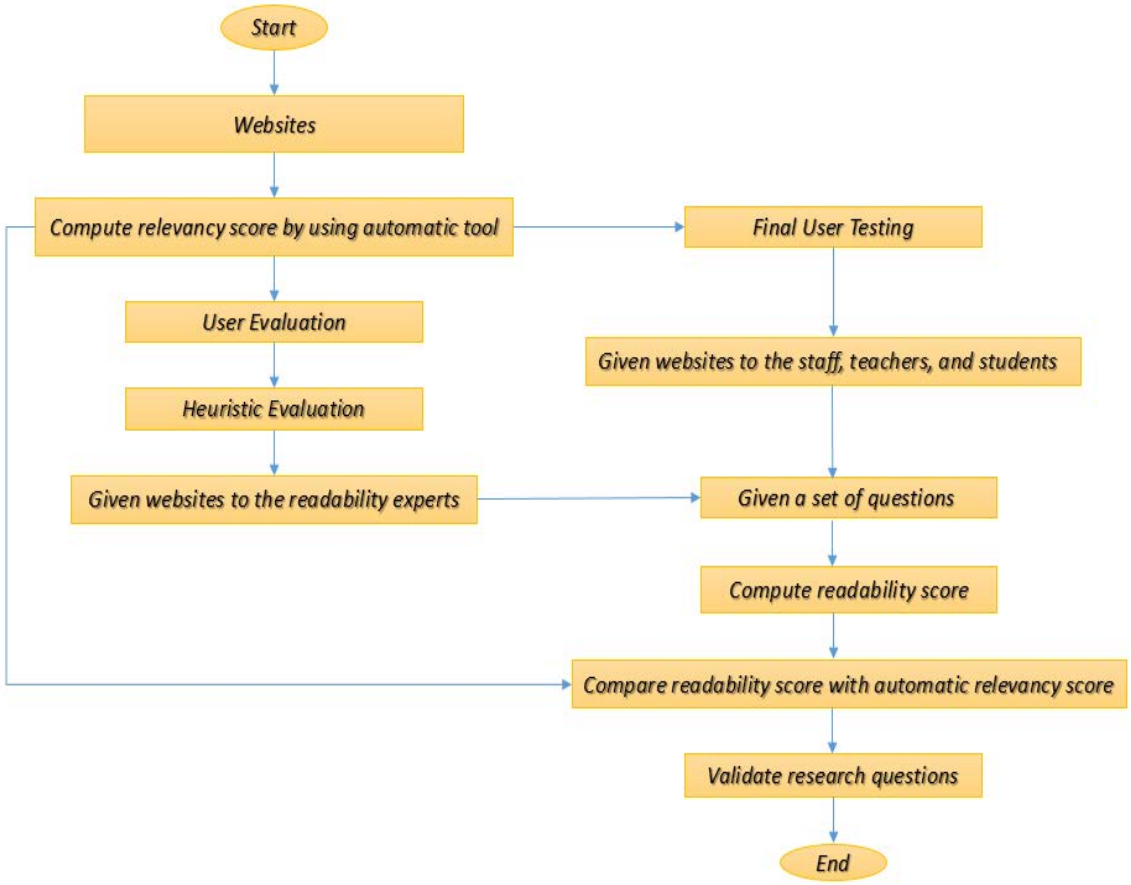


Figure 4.8 User Evaluation

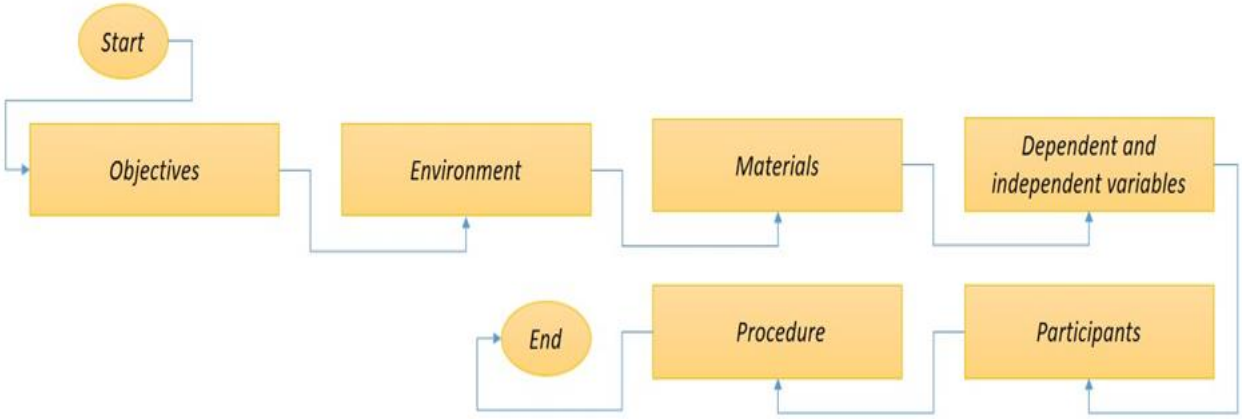


Figure 4.9 User Evaluation Design

4.2 User study For Non-Text Images

In this section, user study has been performed to check that relevant non-text images could enhance web readability. Two evaluations, evaluation by final users in educational institutions and readability experts from different software houses in Pakistan, have been conducted.

4.2.1 Objective

The main goal of this study is to find the answers of research questions, the relevant non-text images could increase web readability for users. In order to achieve objective, the readability score and reading time have been computed through questions that have been asked in the user study.

4.2.2 Environment

An online survey through Google Forms has been conducted. Experts and users have the option to evaluate the web page at any place.

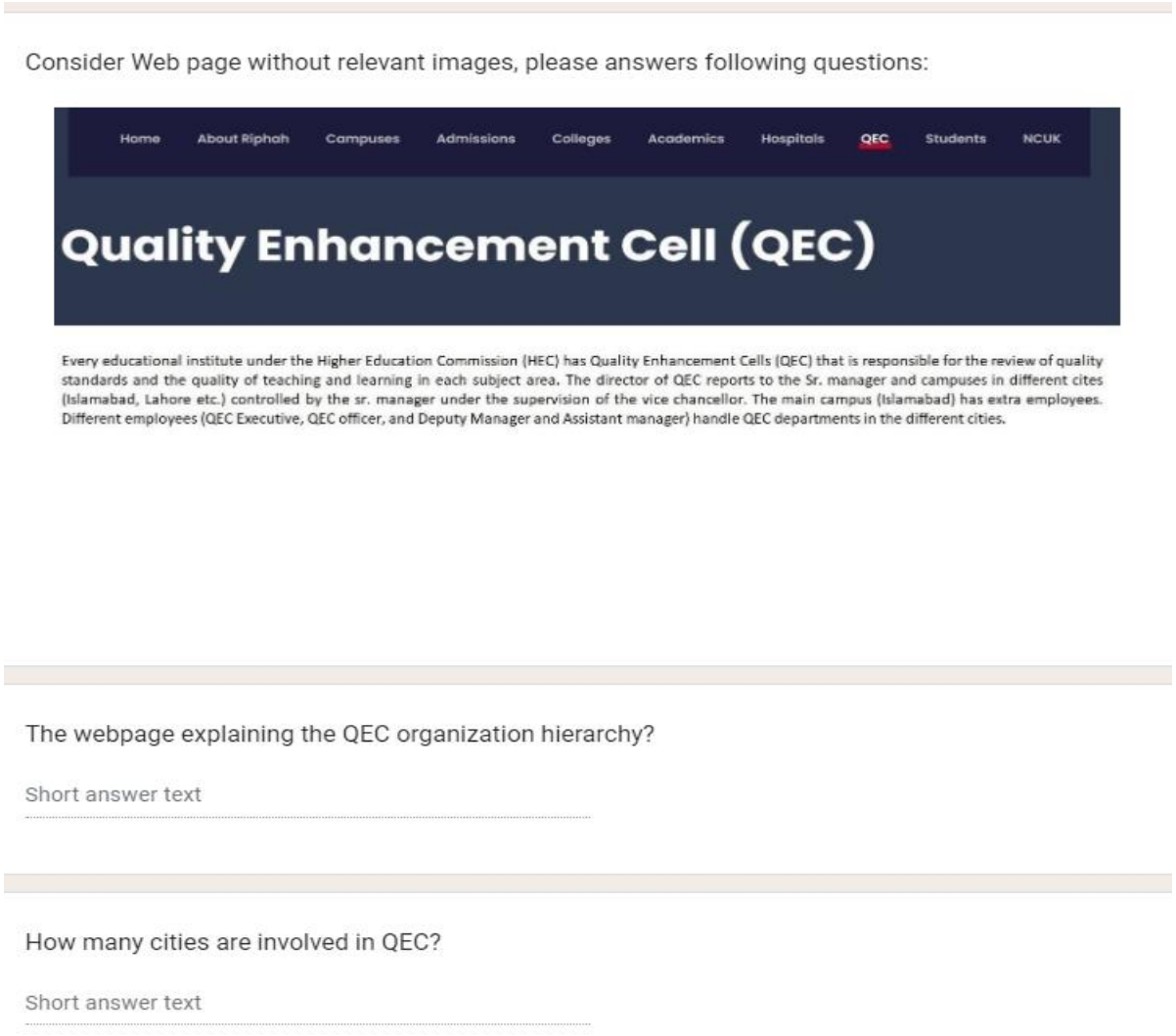


Figure 4.10 Survey Form

4.2.3 Materials

In this study, we have considered fifty educational websites in Pakistan, listed in Table 5. For this research work, the two web pages with a better relevancy score, two web pages with average relevancy score and the two web pages with a worse relevancy score according to the methodology

proposed were selected. Then, a second version of these webpages were constructed by the authors to carry out an A/B Test with the webpages as it is explained in detail in the Procedure. For example, Figure 4.12 shows the screenshot of one of the original web pages with a non-text image, meanwhile a new webpage is specially created for this study for carrying out the A/B test, without the non-text image, shown in Figure 4.11.

4.2.4 Dependent and independent variables

In our case, the things that matter are how well people understand, which can be bad, fair, good, or excellent. This understanding depends on the following factors that are out of our control:

- What kind of picture it is (chart, diagram, flow diagram, or photo)
- The sharpness of pictures
- What non-text images and paragraphs have to do with each other

4.2.5 Participants

A total of 1024 final users (potential readers) were voluntarily enlisted for final user testing (Male =512 and Female=512); and 32 readability experts (Male =16 and Fe-male=16) participated voluntarily in heuristic study, which were developers from different software houses in Pakistan. On the one hand, the users for the potential readers study were users from academic backgrounds. Teachers, staff, and students from the different institutes listed in Table 5 were especially asked to take part. They were hired after an agent test of the client population showed that they fit a certain profile. All of the users met the criteria for inclusion, which was that they weren't readability experts and weren't power users. This means that they hadn't done any web testing yet, but they did have some experience riding the web. Moreover, they were required to graduate and be between 20 to 35 years old.

Table 4.2 Number of Participants

	Total	Male	Female
Final Users	1024	512	512
Experts	32	16	16

On the other hand, the readability specialists were enlisted to perform the heuristic examination. The inclusion criteria for this study was: to have graduate-level coursework in human-PC collaboration, and brutal variables of website architecture, and to have previously been taught and taken an interest in somewhere around one heuristic web assessment project. This is predictable with the thought that master evaluators ought to be utilized for heuristic assessment, as they give better outcomes. The invitation to be enrolled in the studies was shared and advertised using different social media platforms, and also emailed the links to academic users, and to industry people.

4.2.6 Procedure

This experiment has been conducted in two different groups. Firstly, we gave three websites (with non-text images and without non-text images) to half of the experts and users. Another set of three websites (with images and without images) was given to the other half of the experts and users. During the evaluation procedure, experts and users had the opportunity to clarify any doubts or problems. Experts and users checked the relevance of non-text images with the webpage and answers to questions. User feedback has been recorded and this was used to check the relevancy of non-text images with the text of the web page and its readability.

Website 1, Website 2 and Website 3

- Group 1 --> with non-text images / questions / without non-text images / questions about preferences
- Group 2 --> without non-text images / questions / with non-text images / questions about preferences

Website 4, Website 5 and Website 6

- Group 1 --> with non-text images / questions / without non-text images / questions about preferences
- Group 2 --> without non-text images / questions / with non-text images / questions about preferences

4.2.7 Questionnaires

For validation of the hypothesis, different types of questions which consist of control questions, questions related to the user's understanding, and finally, questions relative to the user's feelings have been asked in the user survey. For example, consider the best webpage as shown in Figure 4.11 without relevant non-text images, please answer the following questions.

i. The webpage explains higher educational institutes?
ii. Do you think the educational institute has a clean environment?
iii. Does it consist of male and female students?
iv. Different trees are surrounding the buildings.
v. Do you think the institute has huge buildings?
vi. It has good sports grounds.
vii. Do you think it has a friendly environment?
viii. Time taken to go through the contents and answer these questions was recorded.

Figure 4.11 Webpage without non-text images

On the other hand, the same webpage, as shown in Figure 4.12 with relevant non-text images, please answer the following questions:

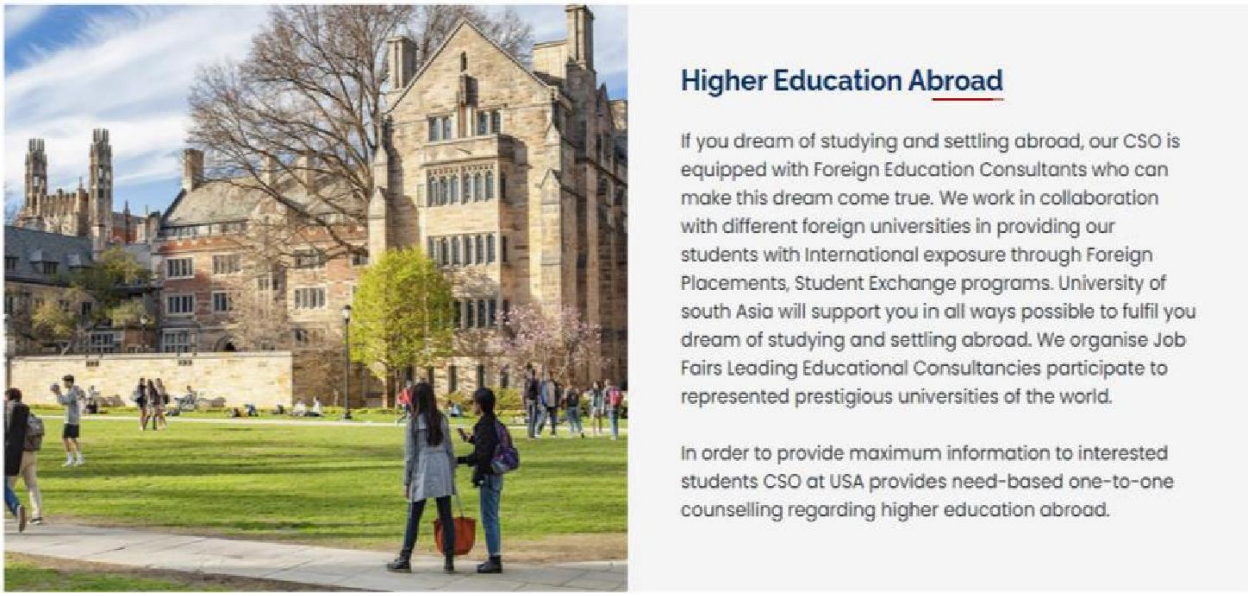


Figure 4.12 Webpage with non-text images

i. The webpage explains higher educational institutes?
ii. Do you think the educational institute has a clean environment?
iii. Does it consist of male and female students?
iv. Different trees are surrounding the buildings.
v. Do you think the institute has huge buildings?
vi. It has good sports grounds.
vii. Do you think it has a friendly environment?

viii. The new image added to the webpage helps me to understand the web content.
ix. Time taken to go through the contents and answer these questions was recorded.
x. I prefer a webpage with relevant non-text images (web page webpage shown in Figure 4.12).

4.2.8 Results

We've looked at six websites and found that two are the best, two are average websites and the other two are the worst of the fifty we were looking at. The results were looked into and put together so that they could be shown in a statistical way. User online results show that web page 1, which has mostly relevant graphics, has a readability score of 53.57% for a page without images and 91.71% for a page with images. The results suggest that relevant, graphical content helped people understand the page best. On the other hand, users had a hard time getting the idea of the same page without images. The same thing holds true for page 3. Without images, Page 3 has a readability score of 52.33%, but with images, it has a score of 89.67%. On the other hand, when websites with irrelevant graphics are served to users without graphics, they are easier to understand than when they were served with graphics. The online results show that page 2 is 49.11% easy to read without any pictures, and 50.01% easy to read with pictures. The same thing happens on page 4. Without images, Page 4 has a readability score of 50.13%, but with images, it has a score of 50.67%. Based on the results, it's clear that negative images that aren't relevant hurt the readability of Figure 4.14. Webpage 5 without images has a readability score of 56.77% while the same webpage with images has 63.66%. Webpage 6 without images has a readability score of 52.11% while the same webpage with images has 59.61%. Figure 4.13 shows that when irrelevant images are taken away, users can see things more quickly and accurately.

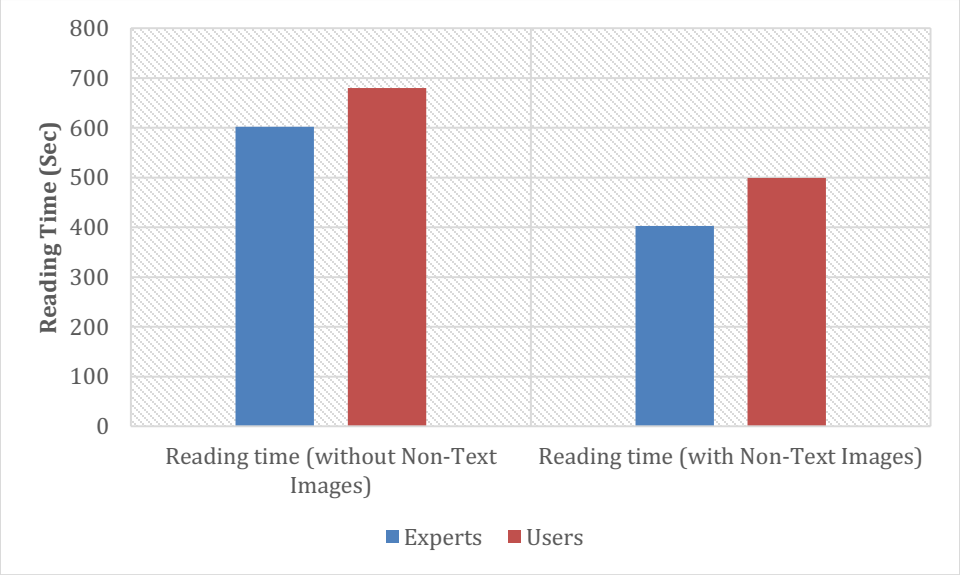


Figure 4.13 Web Readability Time with Non-Text Images

It doesn't change much when experts look at it. Page 1 without images has a readability score of 51.17%, while the same page with images has a readability score of 90.07%. Without images, Page 3 has a readability score of 53.13%, but with images, it has a score of 88.01%.

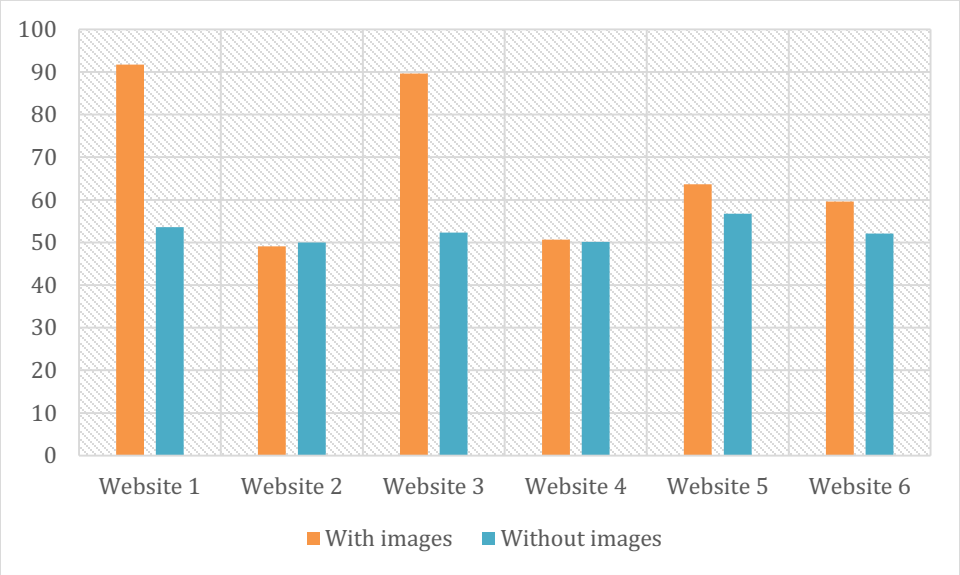


Figure 4.14 User's based readability scores with and without text images

A page with no pictures has a readability score of 50.13 percent, while the same page with pictures has a score of 51.6 percent. Figure 4.15 shows that Page 4's readability score is 51.15% when it doesn't have any irrelevant images, but it drops to 49.63% when it does. Without images, Page 5 has a readability score of 55.77%, but with images, it has a score of 61.66%. Without images, Page 6 has a readability score of 50.7%, but with images, it has a score of 57.66%.

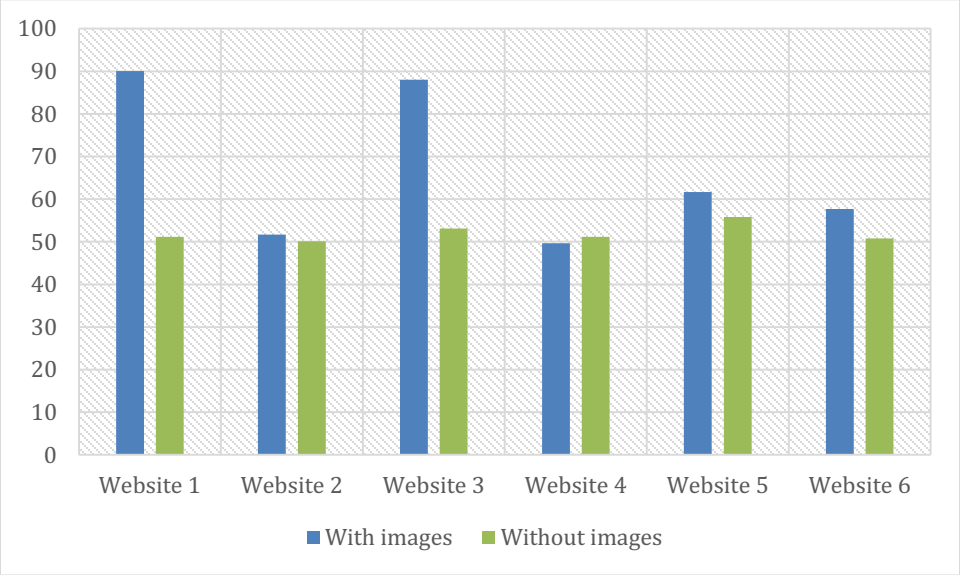


Figure 4.15 Experts-based readability scores with and without text images

In this section, we have evaluated our proposal through research questions that relevant text images could enhance web readability by using user study. Two evaluations, evaluation through final user and evaluation through readability experts, have been conducted in this section. Best, average and worst websites computed through proposed automatic tool have been used in the user evaluation. Different questions related to the user feelings, user understanding etc. were asked, time taken to read the content and answer the questions were recorded and readability scores were computed. We have observed that the results of the final user evaluation are close to the heuristic evaluation, and also observed that the websites have high relevance scores, have high readability scores in the user evaluation that verify our proposed hypothesis and research questions that relevant images could enhance the web readability.

4.3 User study For Text Images

In this section, user study has been performed to check that relevant text images could enhance web readability. Two evaluations, evaluation by final users in educational institutions and readability experts from different software houses in Pakistan, have been conducted.

4.3.1 Objective

The main goal of this study is to find the answers of research questions, the relevant text images could increase web readability for users. In order to achieve objective, the readability score and reading time have been computed through questions that have been asked in the user study.

4.3.2 Environment

An online survey through Google Forms has been conducted. Experts and users have the option to evaluate the web page at any place.

4.3.3 Materials

In this study, we have considered fifty educational websites in Pakistan, listed in Table 5. For this research work, the two web pages with a better relevancy score, two web pages with average relevancy score and the two web pages with a worse relevancy score according to the methodology proposed were selected. Then, a second version of these webpages were constructed by the authors to carry out an A/B Test with the webpages as it is explained in detail in the Procedure. For example, Figure 4.17 shows the screenshot of one of the original web pages with a text image, meanwhile a new webpage is specially created for this study for carrying out the A/B test, without the text image, shown in Figure 4.16.

4.3.4 Dependent and independent variables

In our case, the things that matter are how well people understand, which can be bad, fair, good, or excellent. This understanding depends on the following factors that are out of our control:

- What kind of picture it is (chart, diagram, flow diagram, or photo)

- The sharpness of pictures
- What text images and paragraphs have to do with each other

4.3.5 Participants

A total of 1024 final users (potential readers) were voluntarily enlisted for final user testing (Male =512 and Female=512); and 32 readability experts (Male =16 and Fe-male=16) participated voluntarily in heuristic study, which were developers from different software houses in Pakistan. On the one hand, the users for the potential readers study were users from academic backgrounds. Teachers, staff, and students from the different institutes listed in Table 5 were especially asked to take part. They were hired after an agent test of the client population showed that they fit a certain profile. All of the users met the criteria for inclusion, which was that they weren't readability experts and weren't power users. This means that they hadn't done any web testing yet, but they did have some experience riding the web. Moreover, they were required to graduate and be between 20 to 35 years old.

On the other hand, the readability specialists were enlisted to perform the heuristic examination. The inclusion criteria for this study was: to have graduate-level coursework in human-PC collaboration, and brutal variables of website architecture, and to have previously been taught and taken an interest in somewhere around one heuristic web assessment project. This is predictable with the thought that master evaluators ought to be utilized for heuristic assessment, as they give better outcomes. The invitation to be enrolled in the studies was shared and advertised using different social media platforms, and also emailed the links to academic users, and to industry people.

4.3.6 Procedure

This experiment has been conducted in two different groups. Firstly, we gave three websites (with text images and without text images) to half of the experts and users. Another set of three websites (with images and without images) was given to the other half of the experts and users. During the evaluation procedure, experts and users had the opportunity to clarify any doubts or problems. Experts and users checked the relevance of text images with the webpage and answers to questions. User feedback has been recorded and this was used to check the relevancy of text images with the text of the web page and its readability.

Website 1, Website 2 and Website 3

- Group 1 --> with text images / questions / without text images / questions about preferences
- Group 2 --> without text images / questions / with text images / questions about preferences

Website 4, Website 5 and Website 6

- Group 1 --> with text images / questions / without text images / questions about preferences

- Group 2 --> without text images / questions / with text images / questions about preferences

4.3.7 Questionnaires

For validation of the hypothesis, different types of questions which consist of control questions, questions related to the user's understanding, and finally, questions relative to the user's feelings have been asked in the user survey. For another example, consider the best webpage as shown in Figure 4.16 without relevant text images, please answers the following questions:

(Control Question)

i. The webpage explaining the QEC organization hierarchy?

(Questions related to understanding)

ii. How many cities are involved in QEC?
iii. Who reports issues to the Vice-Chancellor (VC)?
iv. Who reports hierarchically to the Deputy Manager in Islamabad?
v. Where (city) is the Deputy Manager working?
vi. How many employees are involved in the main campus?
vii. Does the Assistant Manager in Islamabad assist the Deputy Manager or the QEC Officer?
viii. Who can restrict the activities of all employees on different campuses?
ix. Time taken to go through the contents and answer these questions was recorded.

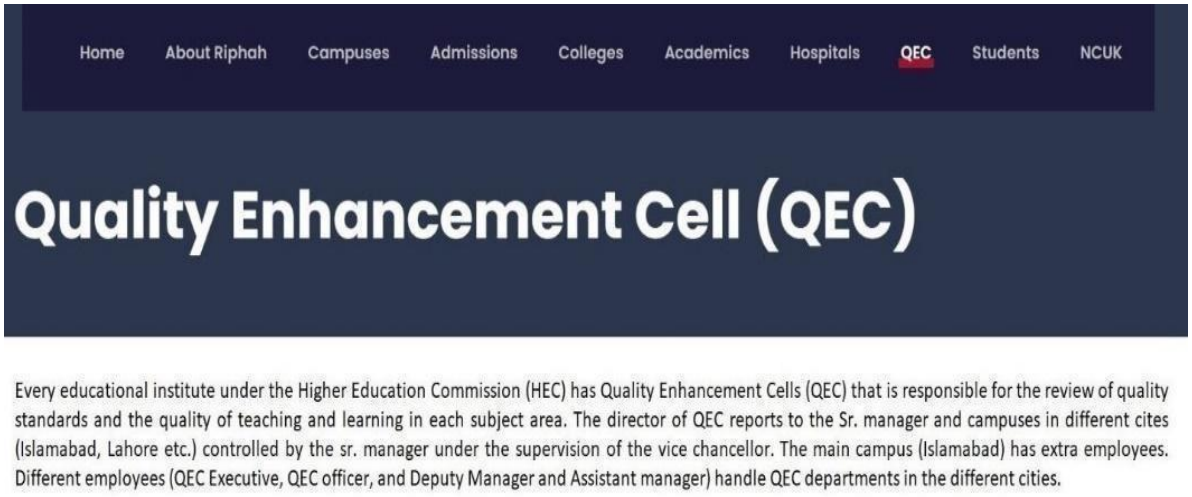


Figure 4.16 Webpage without relevant images

On the other hand, the same webpage, as shown in Figure 4.17 with relevant text images, please answer the following questions:

(Control Question)

i. The webpage explaining the QEC organization hierarchy?

(Questions related to understanding)

ii. How many cities are involved in QEC?
iii. Who reports issues to the Vice-Chancellor (VC)?
iv. Who reports hierarchically to the Deputy Manager in Islamabad?
v. Where (city) is the Deputy Manager working?
vi. How many employees are involved in the main campus?
vii. Does the Assistant Manager in Islamabad assist the Deputy Manager or the QEC Officer?
viii. Who can restrict the activities of all employees on different campuses?
ix. Time taken to go through the contents and answer these questions was recorded.

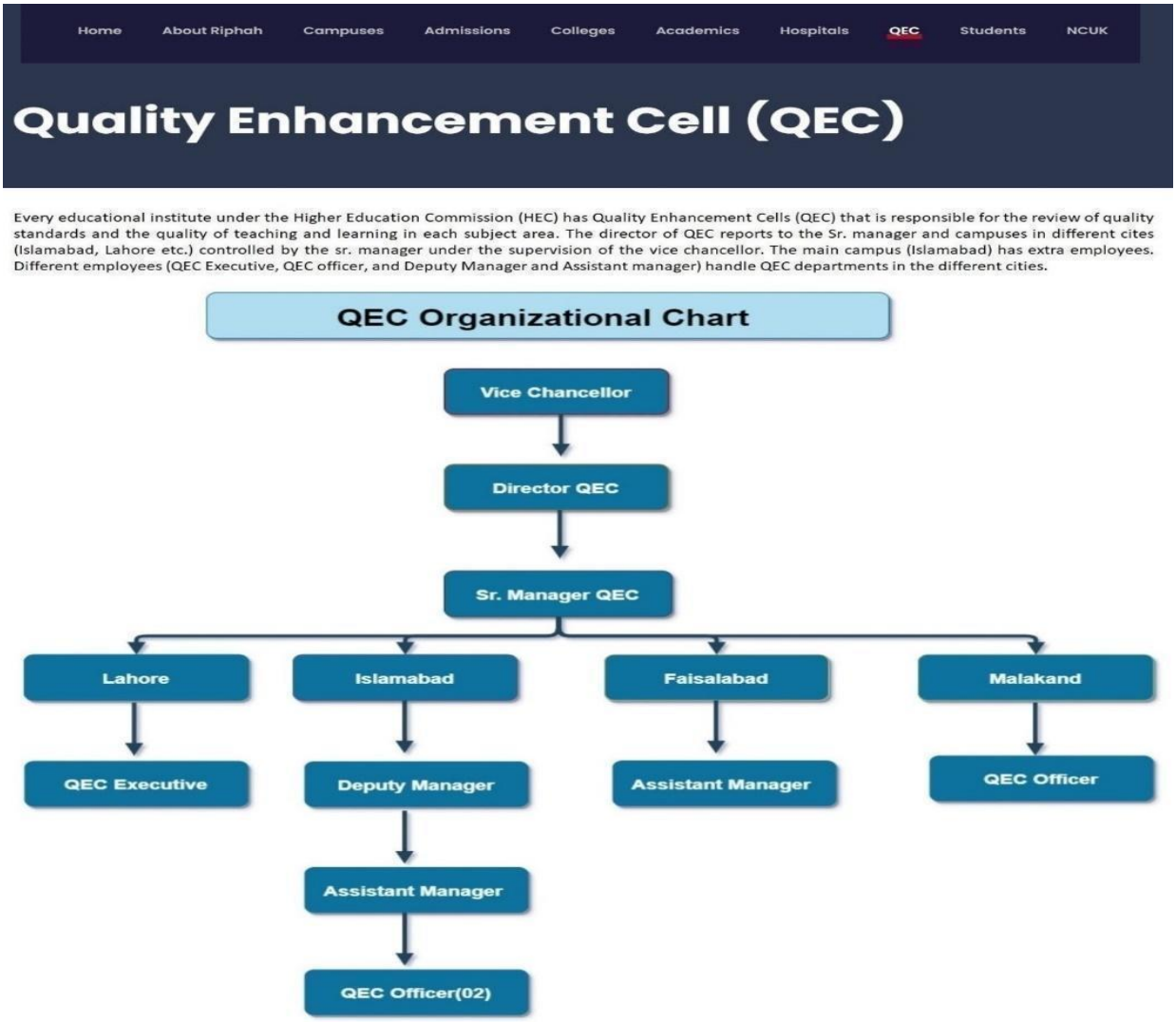


Figure 4.17 Webpage with relevant images

(Questions related to feelings of image relevancy)

i. The new image added to the webpage helps me to understand the web content.
ii. I prefer a webpage with relevant images (Web Page shown in Figure 4.17).

4.3.8 Results

We've looked at six websites and found that two are the best, two are average websites and the other two are the worst of the fifty we were looking at. The results were looked into and put together so that they could be shown in a statistical way. User online results show that web page 1, which has mostly relevant graphics, has a readability score of 52.51% for a page without images and 89.71% for a page with images. The results suggest that relevant, graphical content helped people understand the page best. On the other hand, users had a hard time getting the idea of the same page without images. The same thing holds true for page 3. Without images, Page 3 has a readability score of 50.23%, but with images, it has a score of 87.57%. On the other hand, when websites with irrelevant graphics are served to users without graphics, they are easier to understand than when they were served with graphics. The online results show that page 2 is 49.11% easy to read without any pictures, and 50.01% easy to read with pictures. The same thing happens on page 4. Without images, Page 4 has a readability score of 50.13%, but with images, it has a score of 50.67%. Based on the results, it's clear that negative images that aren't relevant hurt the readability of Figure 4.19. Webpage 5 without images has a readability score of 55.77% while the same webpage with images has 61.66%. Webpage 6 without images has a readability score of 52.11% while the same webpage with images has 59.61%. Figure 4.18 shows that when irrelevant images are taken away, users can see things more quickly and accurately.

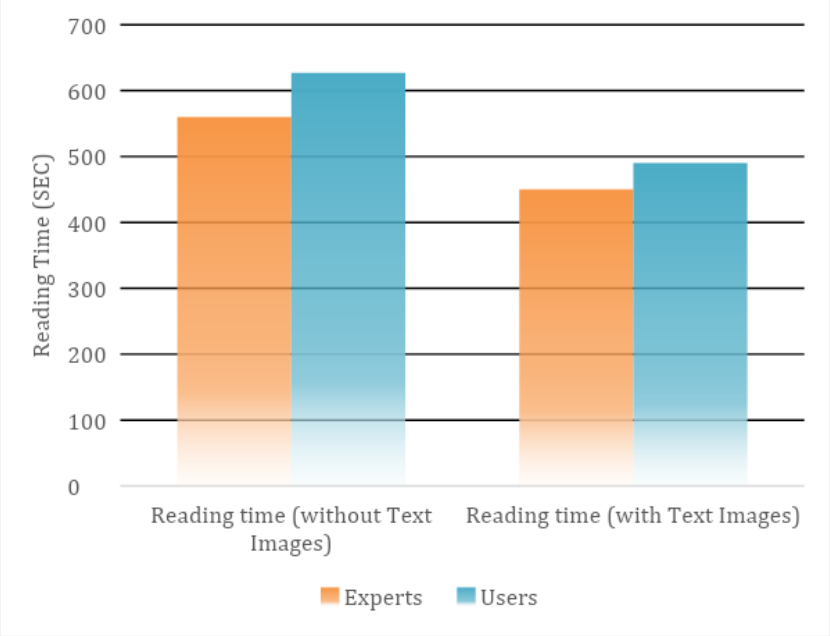


Figure 4.18 Web Readability Time with Text Images

It doesn't change much when experts look at it. Page 1 without images has a readability score of 51.17%, while the same page with images has a readability score of 90.07%. Without images, Page 3 has a readability score of 53.13%, but with images, it has a score of 88.01%.

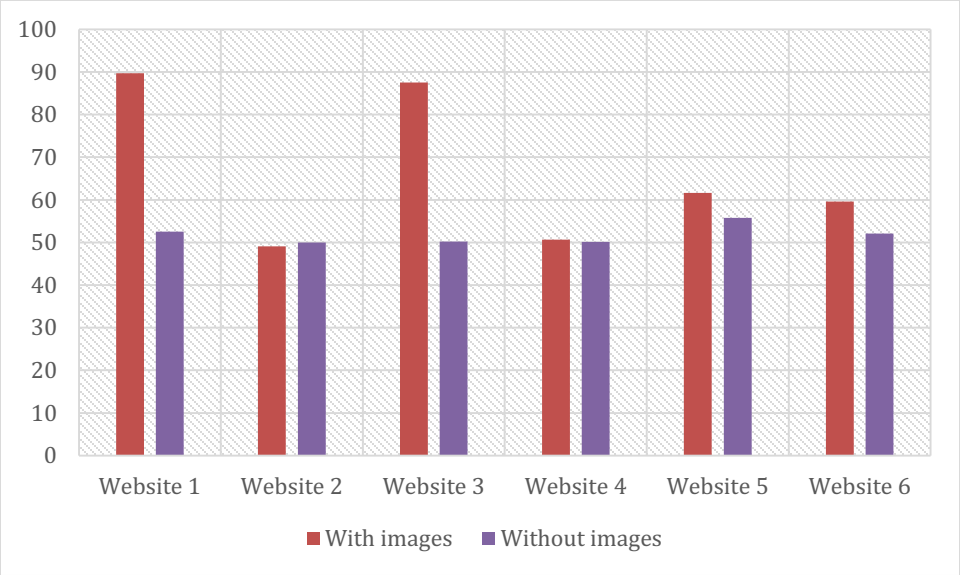


Figure 4.19 User's based readability scores with and without images

A page with no pictures has a readability score of 50.13 percent, while the same page with pictures has a score of 51.6 percent. Figure 4.20 shows that Page 4's readability score is 51.15% when it doesn't have any irrelevant images, but it drops to 49.63% when it does. Without images, Page 5 has a readability score of 55.77%, but with images, it has a score of 61.66%. Without images, Page 6 has a readability score of 50.7%, but with images, it has a score of 57.66%.

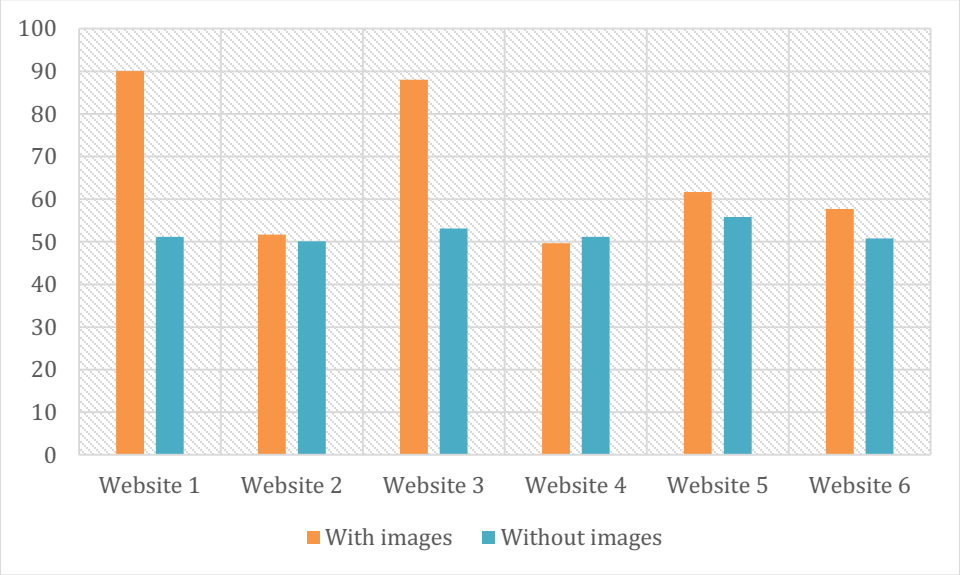


Figure 4.20 Experts-based readability scores with and without images

In this section, we have evaluated our proposal through research questions that relevant text images could enhance web readability by using user study. Two evaluations, evaluation through final user and evaluation through readability experts, have been conducted in this section. Best, average and worst websites computed through proposed automatic tool have been used in the user evaluation. Different questions related to the user feelings, user understanding etc. were asked, time taken to read the content and answer the questions were recorded and readability scores were computed. We have observed that the results of the final user evaluation are close to the heuristic evaluation, and also observed that the websites have high relevance scores, have high readability scores in the user evaluation that verify our proposed hypothesis and research questions that relevant images could enhance the web readability.

Chapter 5: Conclusion and Further Work

5.1 Conclusion

This thesis aims to study the readability problems that users may encounter while interacting with web pages and propose readability guidelines to help developers and designers make web pages easier to read by looking at only the essential parts of a page and not relying on their judgment. The main contributions made in this thesis are analysis of image relevancy measures on websites from a readability perspective, analysis of the factors which could influence the readability of the images and a proposal of an image relevance measure and design a methodology to evaluate the relevancy of images on the websites. The contributions of this thesis are made in light of some limitations that we are considering educational websites having English language limited to the specific country.

Further, we conclude with the work done to achieve the thesis's goals:

First, the analysis of image relevancy measures on websites from a readability perspective was studied, where image relevancy techniques on websites were analyzed in Chapter 2. In addition to conducting a systematic review to find factors that affect readability of images on the web pages. The results from the systematic review were integrated with the proposed methodology to compute image relevancy on the websites, which is presented in Chapter 3. This methodology combines different ways to get information from images by using Cloud Vision API and OCR and reading text from websites to find relevancy between them. Techniques for preprocessing data have been used on the information that has been extracted. NLP technique has been used to determine what images and text on a web page have to do with each other. This tool looks at fifty educational websites' pictures and assesses their relevance. Results show that images that have nothing to do with the page's content and images that aren't very good cause lower relevancy scores. Second, a proposal of an evaluation methodology to assess that relevant images could enhance web readability was introduced in Chapter 4, in which a user study was done to evaluate the proposed methodology based on two evaluations: the evaluation of the end users of the page and the heuristic evaluation, which was done by experts in readability. A user study was done with questions about what the user knows, how they feel, and what they can do. The websites that have high relevancy scores in the proposed automatic tool have higher readability scores in the user study and this validated the hypothesis that the relevant images could enhance web readability and research questions. The results back up the idea that images that are relevant to the page make it easier to read.

5.2 Future work

The present proposal for readability guidelines for web developers is the first attempt in this regard. These guidelines form a basis that can be developed by expanding in specific directions:

Initially, the factors that affect readability of educational websites in English and how relevant images could enhance web readability were studied. In future, we will propose the needed readability guidelines and will update the proposed evaluation methodology for web developers in line with the new studied image relevancy on websites to ensure the compatibility of the proposed methodology for evaluating the readability. The importance of an educational website in Pakistan for students through which they can efficiently study is doubtful. An easy-to-use educational website for college helps to appeal to both students and parents, as it lets them see everything right from their dashboard. Using the organization web portal, they can view their academic progress by logging into their accounts. A website for online classes significantly enhances the learning experience for your students. Taking online courses gives them the freedom to learn when they want. So In recent years, it has become imperative to website design for educational institutions such as universities and colleges. On the other hand, English is an important medium in a number of key educational institutions in Pakistan, is the main language of technology and international business, has a major presence in the media, and is a key means of communication among a national elite. The constitution and the laws of the land are codified in English. In the future, we will consider other domains and countries.

The final users' evaluation of the proposed hypothesis that the relevant images on websites could enhance web readability are being evaluated in the selected educational websites and software houses. The evaluation will involve more users and conduct more sessions, allowing the investigation of additional potential readability barriers. The upcoming users' evaluations could update our proposed guidelines with new guidelines or recommendations. Because this evaluation consists of a limited number of users.

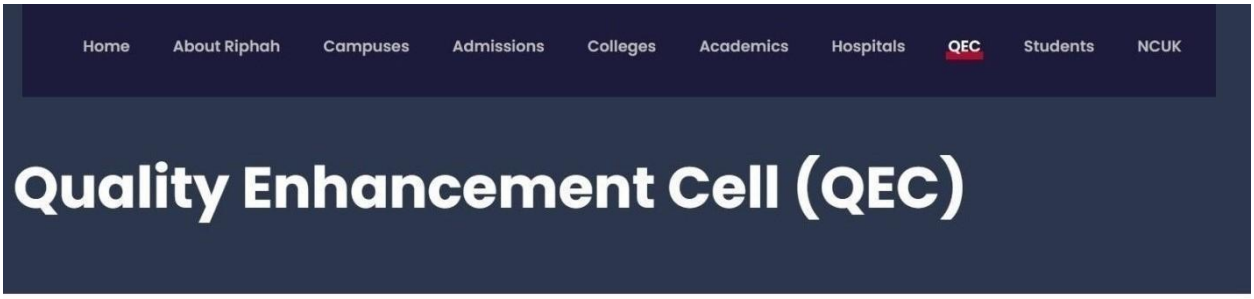
Moreover, the proposed guidelines are presented that images should be relevant to the text of web pages. Irrelevant images on the web could decrease web readability. Lastly, we presented various papers at conferences on relevancy metrics, user surveys, and journal articles on the finished approach and its evaluation. In the not-too-distant future, we hope to make the findings of our doctoral research more widely known, both from a scientific and public relations standpoint, by publishing them in venues such as conferences, journals, and exhibitions. At this time, we are planning to collaborate with various application domains to publish in the following journals:

- Evaluate the user feedback and relevancy of any other applications within the domain
- Evaluate the significance of websites from various countries.

Appendix A

Questionnaire for readability Guidelines Evaluation for Without Text Images

This appendix shows the questionnaire that was used for the User evaluation, which was done with the help of end users and experts on readability. The following questionnaire looks at how easy the new proposed guidelines are to read, how well they work for users, and how happy users are with them.



Every educational institute under the Higher Education Commission (HEC) has Quality Enhancement Cells (QEC) that is responsible for the review of quality standards and the quality of teaching and learning in each subject area. The director of QEC reports to the Sr. manager and campuses in different cites (Islamabad, Lahore etc.) controlled by the sr. manager under the supervision of the vice chancellor. The main campus (Islamabad) has extra employees. Different employees (QEC Executive, QEC officer, and Deputy Manager and Assistant manager) handle QEC departments in the different cities.

i. Please mark your gender:

   a) Male b) Female

ii. The mentioned above webpage explaining the QEC organization hierarchy

iii. How many cities are involved in QEC in the webpage?

iv. Who reports issues to the Vice-Chancellor (VC)?

v. Who reports hierarchically to the Deputy Manager in Islamabad?

v. Where (city) is the Deputy Manager working?

vi. How many employees are involved in the main campus?

vii. Does the Assistant Manager in Islamabad assist the Deputy Manager or the QEC Officer?

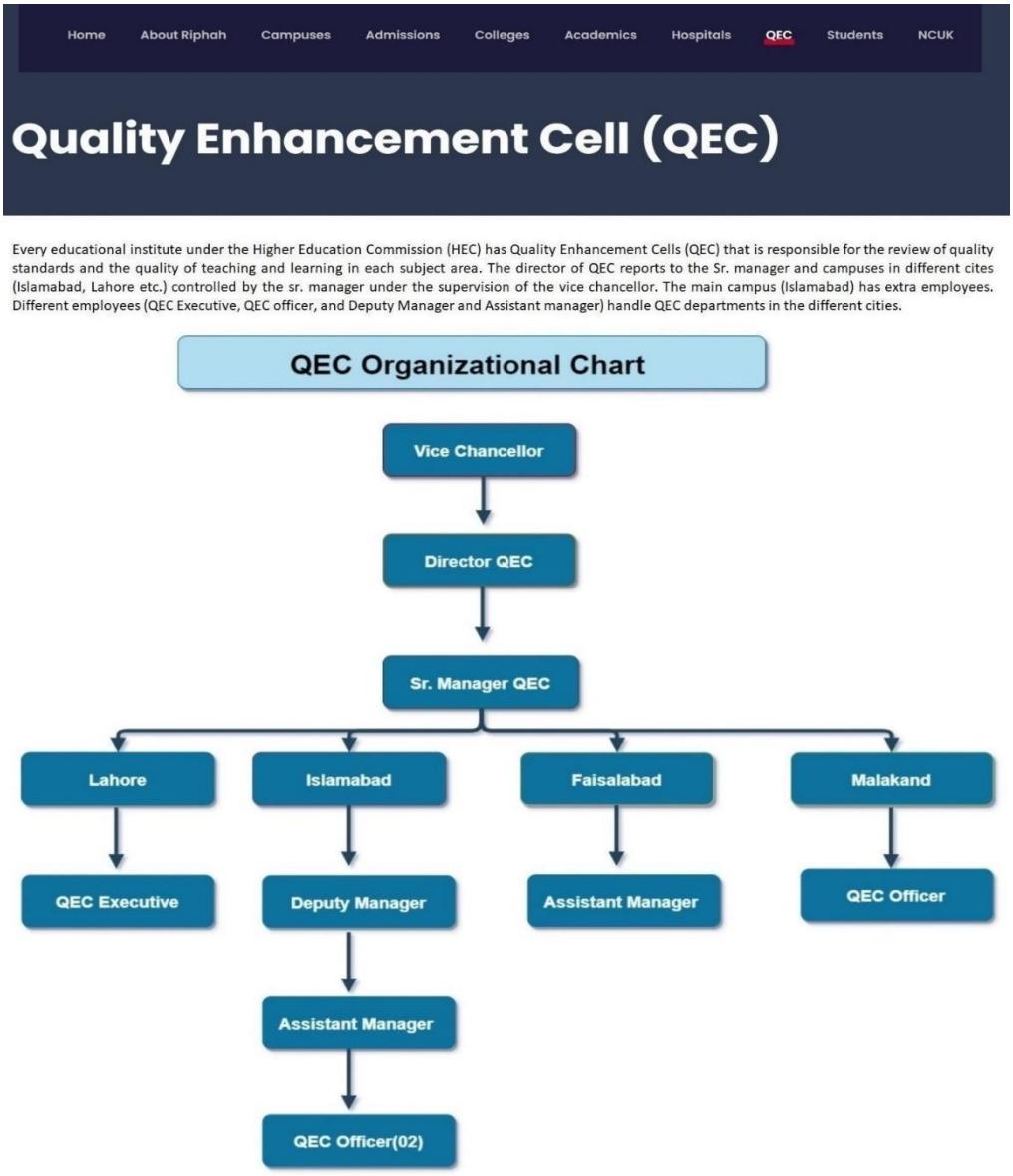
viii. Who can restrict the activities of all employees on different campuses?

ix. Time taken to go through the contents and answer these questions was recorded.

Appendix B

Questionnaire for readability Guidelines Evaluation for With Text Images

This appendix shows the questionnaire that was used for the User evaluation, which was done with the help of end users and experts on readability. The following questionnaire looks at how easy the new proposed guidelines are to read, how well they work for users, and how happy users are with them.



i. Please mark your gender:

a) Male b) Female

ii. The mentioned above webpage explaining the QEC organization hierarchy?

iii. How many cities are involved in QEC in the webpage?

iv. Who reports issues to the Vice-Chancellor (VC)?

v. Who reports hierarchically to the Deputy Manager in Islamabad?

v. Where (city) is the Deputy Manager working?

vi. How many employees are involved in the main campus?

vii. Does the Assistant Manager in Islamabad assist the Deputy Manager or the QEC Officer?

viii. Who can restrict the activities of all employees on different campuses?

ix. The newly added image to the webpage provides clarity on the information presented there.

x. I prefer the above webpage with relevant images.

xi. Time taken to go through the contents and answer these questions was recorded.

Appendix C

Questionnaire for readability Guidelines Evaluation for Without Non-Text Images

This appendix shows the questionnaire that was used for the User evaluation, which was done with the help of end users and experts on readability. The following questionnaire looks at how easy the new proposed guidelines are to read, how well they work for users, and how happy users are with them.

i. Please mark your gender:

a) Male b) Female

ii. The website provides information regarding the higher educational institute?

iii. Do you believe that the educational institution has a healthy and sanitary atmosphere?

iv. Are there both male and female students in it?

v. There are different kinds of trees around the buildings.

vi. Do you think the buildings at the institute are significant?

vii. It has good places to play sports.

viii. Are the people there friendly?

ix. Time taken to go through the contents and answer these questions was recorded.

Appendix D

Questionnaire for readability Guidelines Evaluation for With Non-Text Images

This appendix shows the questionnaire that was used for the User evaluation, which was done with the help of end users and experts on readability. The following questionnaire looks at how easy the new proposed guidelines are to read, how well they work for users, and how happy users are with them.

i. Please mark your gender:

b) Male b) Female

ii. The website provides information regarding the higher educational institute?

iii. Do you believe that the educational institution has a healthy and sanitary atmosphere?

iv. Are there both male and female students in it?

v. There are different kinds of trees around the buildings.

vi. Do you think the buildings at the institute are significant?

vii. It has good places to play sports.

viii. Are the people there friendly?

ix. The newly added image to the webpage provides clarity on the information presented there.

ix. Time taken to go through the contents and answer these questions was recorded.

Appendix E

Questionnaire for readability Guidelines Evaluation for With Text Images

This appendix shows the questionnaire that was used for the User evaluation, which was done with the help of end users and experts on readability. The following questionnaire looks at how easy the new proposed guidelines are to read, how well they work for users, and how happy users are with them.

A relevant graphical content in the web page helps to portray context in a more effective way. Considering this fact, please rate how relevant is the image highlighted inside a rectangle with the web and how effectively playing this role *

1 2 3 4 5 6 7 8 9 10

i. Does the picture help you better understand the website's information?
ii. Is the page attractive to you?
iii. Do you need help understanding the graphics?
iv. Do you think the pictures make the text easier to understand?

v. Taking all of these things into account, please rate the quality of the website as a whole.

GLOSSARY

Guidelines A set of rules to guarantee the success of a certain goal or process (Shekelle, 1999).

Human-Computer Interaction is a multidisciplinary field that aims to solve the real problems of making computer systems easier to use with scientific answers (Gregory, 1991).

Image processing is a set of actions performed on an image to enhance it or extract relevant information from it (Castleman, 1996).

Inclusive design when designers make their products or services with everyone's needs in mind, no matter their age, ability, culture, where they live, or how much money they have (Clarkson, 2010).

Optical Character Recognition is the process of converting an image of text into a format that a machine can read. (Mori, 1999).

Readability states that is the ease with which a reader can understand a document (Klare, 1963).

ACRONYMS

- **AI:** Artificial Intelligence.
- **AICMD:** Annotation by Image-to-Concept Distribution Model.
- **DDC:** Dewey Decimal Characterization
- **HCI:** Human-Computer Interaction.
- **HTML:** Hypertext Markup Language.
- **KDT:** Knowledge Discovery in Texts.
- **LCC:** Library of Congress Grouping.
- **LIS:** Library and Data Science.
- **LSH:** Locality-Sensitive Hashing.
- **NLP:** Natural language processing.
- **OCR:** Optical Character Recognition.
- **OEM:** Object Exchange Model.
- **OWDIG:** Online Webpage Image Downloader and ImageInfo Grabber.
- **TF:** Term Frequency.
- **WCAG:** Web Content Accessibility Guidelines.
- **WCM:** Web Content Mining.
- **WSM:** Web Structure Mining.
- **WUEM:** Web Usability Evaluation Model.
- **WUM:** Web Usage Mining.
- **WWW:** World Wide Web.

BIBLIOGRAPHY

Abedi, M., Torabi, S. A., Norouzi, G. H., & Hamzeh, M. (2012). ELECTRE III: A knowledge-driven method for integration of geophysical data with geological and geochemical data in mineral prospectivity mapping. Journal of applied geophysics, 87, 9-18.

Akerkar, R., & Sajja, P. (2009). Knowledge-based systems. Jones & Bartlett Publishers.

Alemerien, K., & Magel, K. (2014). GUIEvaluator: A Metric-tool for Evaluating the Complexity of Graphical User Interfaces. In SEKE (pp. 13-18).

Alemerien, K., & Magel, K. (2014). GUIEvaluator: A Metric-tool for Evaluating the Complexity of Graphical User Interfaces. In SEKE (pp. 13-18).

Ali, A. Z. M., Wahid, R., Samsudin, K., & Idris, M. Z. (2013). Reading on the Computer Screen: Does Font Type Have Effects on Web Text Readability?. International Education Studies, 6(3), 26-35.

Alqahtani, A., Alhakami, H., Alsubait, T., & Baz, A. (2021). A survey of text matching techniques. Engineering, Technology & Applied Science Research, 11(1), 6656-6661.

Antunes, H., & Lopes, C. T. (2019, June). Readability of web content. In 2019 14th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-4). IEEE.

Arief, R., Mutiara, A. B., & Kusuma, T. M. (2018). Automated extraction of large scale scanned document images using Google vision OCR in apache Hadoop environment. International Journal of Advanced Computer Science and Applications, 9(11).

Asai, T., Abe, K., Kawasoe, S., Sakamoto, H., Arimura, H., & Arikawa, S. (2004). Efficient substructure discovery from large semi-structured data. IEICE TRANSACTIONS on Information and Systems, 87(12), 2754-2763.

Bates, M. J. (1999). The invisible substrate of information science. Journal of the American society for information science, 50(12), 1043-1050.

Belkin, N. J., & Robertson, S. E. (1976). Information science and the phenomenon of information. Journal of the American society for information science, 27(4), 197-204.

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. Educational Psychology Review, 24(1), 63-88.

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. Educational Psychology Review, 24(1), 63-88.

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. Educational Psychology Review, 24(1), 63-88.

Berners-Lee, T., & Hendler, J. (2001). Publishing on the semantic web. Nature, 410(6832), 1023-1024.

Bharanipriya, V., & Prasad, V. K. (2011). Web content mining tools: a comparative study. International Journal of Information Technology and Knowledge Management, 4(1), 211-215.

Biddinika, M. K., Lestari, R. P., Indrawan, B., Yoshikawa, K., Tokimatsu, K., & Takahashi, F. (2016). Measuring the readability of Indonesian biomass websites: The ease of understanding biomass energy information on websites in the Indonesian language. Renewable and Sustainable Energy Reviews, 59, 1349-1357.

Bigelow, C. (2019). Typeface features and legibility research. Vision research, 165, 162-172.

Bilal, D., & Huang, L. M. (2019). Readability and word complexity of serps snippets and web pages on children's search queries: Google vs bing. Aslib Journal of Information Management.

Blackman, M. C., & Funder, D. C. (1998). The effect of information on consensus and accuracy in personality judgment. Journal of Experimental Social Psychology, 34(2), 164-181.

Bothun, L. S., Feeder, S. E., & Poland, G. A. (2021). Poor Readability of COVID-19 Vaccine Information for the General Public: A Lost Opportunity. medRxiv.

Breuel, T. M. (2008, January). The OCRopus open source OCR system. In Document recognition and retrieval XV (Vol. 6815, pp. 120-134). SPIE.

Butler, M., Holloway, L., Marriott, K., & Goncu, C. (2017). Understanding the graphical challenges faced by vision-impaired students in Australian universities. Higher Education Research & Development, 36(1), 59-72.

Caravolas, M., Downing, C., Hadden, C. L., & Wynne, C. (2020). Handwriting legibility and its relationship to spelling ability and age: Evidence from monolingual and bilingual children. Frontiers in Psychology, 11, 1097.

Chen, S. H., & Chen, Y. H. (2017, April). A content-based image retrieval method based on the google cloud vision api and wordnet. In Asian conference on intelligent information and database systems (pp. 651-662). Springer, Cham.

Chitradevi, B., & Srimathi, P. (2014). An overview on image processing techniques. International Journal of Innovative Research in Computer and Communication Engineering, 2(11), 6466-6472.

Christensen, H. D. (2017). Rethinking image indexing?. J. Assoc. Inf. Sci. Technol., 68(7), 1782-1785.

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. ITL-International Journal of Applied Linguistics, 165(2), 97-135.

Collins-Thompson, K., & Callan, J. (2005). Predicting reading difficulty with statistical language models. Journal of the american society for information science and technology, 56(13), 1448-1462.

Conway, M., Meares, K., & Standart, S. (2004). Images and goals. Memory, 12(4), 525-531.

Cook, I. A., Warren, C., Pajot, S. K., Schairer, D., & Leuchter, A. F. (2011). Regional brain activation with advertising images. Journal of Neuroscience, Psychology, and Economics, 4(3), 147.

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. Reading in a foreign language, 23(1), 84-101.

Cyr, D., Head, M., Larios, H., & Pan, B. (2009). Exploring human images in website design: A multi-method approach. MIS quarterly, 539-566.

Da Costa, M. G., & Gong, Z. (2005, June). Web structure mining: an introduction. In 2005 IEEE International Conference on Information Acquisition (pp. 6-pp). IEEE.

De Clercq, O., Hoste, V., Desmet, B., Van Oosten, P., De Cock, M., & Macken, L. (2014). Using the crowd for readability prediction. Natural Language Engineering, 20(3), 293-325.

Del Valle Gastaminza, F. (1999). Documentary analysis of photography. Multimedia documentation notebooks , (8), 26.

Eakins, J. P. (2001). Trademark image retrieval. In Principles of visual information retrieval (pp. 319-350). Springer, London.

Eakins, J. P., & Graham, M. E. (1999). Content-based image retrieval, a report to the JISC Technology Applications programme.

El Desouki, M. I., Gomaa, W. H., & Abdalhakim, H. (2019). A hybrid model for paraphrase detection combines pros of text similarity with deep learning. Int. J. Comput. Appl, 975, 8887.

Elahi, E., Iglesias, A., & Morato, J. (2022). Readability of Non-Text Images on the World Wide Web (WWW). IEEE Access, 10, 116627-116634.

Elahi, E., Iglesias, A., & Morato, J. (2022, June). Readability of Graphical Contents on World Wide Web (WWW). In 2022 17th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-4). IEEE.

Enser, P. (2008). The evolution of visual information retrieval. Journal of Information Science, 34(4), 531-546.

Fang, Q., Xu, C., Sang, J., Hossain, M. S., & Ghoneim, A. (2016). Folksonomy-based visual ontology construction and its applications. IEEE Transactions on Multimedia, 18(4), 702-713.

Feather, J., & Sturges, P. (2003). International encyclopedia of information and library science. Routledge.

Feng, B., Ng, J. H., Heng, J. C. D., & Ng, H. H. (2009). Molecules that promote or enhance reprogramming of somatic cells to induced pluripotent stem cells. Cell stem cell, 4(4), 301-312.

Feng, M. L., & Tan, Y. P. (2004, June). Adaptive binarization method for document image analysis. In 2004 IEEE international conference on multimedia and expo (ICME)(IEEE Cat. No. 04TH8763) (Vol. 1, pp. 339-342). IEEE.

Fernández Huerta, J. (1959). El programa escolar: conclusión. Revista de educación.

Ferrari, A., Pirrotta, L., Bonciani, M., Venturi, G., & Vainieri, M. (2022). Higher readability of institutional websites drives the correct fruition of the abortion pathway: A cross-sectional study. Plos one, 17(11), e0277342.

Ferreira, S., & Okita-Ouma, B. (2012). A proposed framework for short-, medium-and long-term responses by range and consumer states to curb poaching for African rhino horn. Pachyderm, 51, 52-59.

François, T., & Fairon, C. (2012, July). An “AI readability” formula for French as a foreign language. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 466-477).

Grădinaru, M., Negru, A., Boiangiu, C. A., Tarbă, N., Voncilă, M. L., & Deaconescu, R. A. (2022, September). Complete OCR Solution for Image Analysis of World War 2 Documents. In 2022 21st RoEduNet Conference: Networking in Education and Research (RoEduNet) (pp. 1-8). IEEE.

Gradisar, M., Humar, I., & Turk, T. (2006, June). Factors affecting the readability of colored text in computer displays. In 28th International Conference on Information Technology Interfaces, 2006. (pp. 245-250). IEEE.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavior research methods, instruments, & computers, 36(2), 193-202.

Griffin, E., McKenna, K., & Worrall, L. (2004). Stroke education materials on the World Wide Web: An evaluation of their quality and suitability. Topics in Stroke Rehabilitation, 11(3), 29-40.

Gulbrandsen, T. R., Skalitzky, M. K., Shamrock, A. G., Gao, B., Hasan, O., & Miller, B. J. (2022). Web-Based Patient Educational Material on Osteosarcoma: Quantitative Assessment of Readability and Understandability. JMIR cancer, 8(1), e25005.

Guo, J., & Huang, J. (2021). Information literacy education during the pandemic: The cases of academic libraries in Chinese top universities. The Journal of Academic Librarianship, 47(4), 102363.

Haddaway, N. R. (2015). The use of web-scraping software in searching for grey literature. Grey J, 11(3), 186-90.

Hall, R. H., & Hanna, P. (2004). The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. Behaviour & information technology, 23(3), 183-195.

Hazim, R., Saddiki, H., Alhafni, B., Khalil, M. A., & Habash, N. (2022). Arabic Word-level Readability Visualization for Assisted Text Simplification. arXiv preprint arXiv:2210.10672.

He, Y., Li, Y., Xing, L., Qiu, Z., & Zhang, X. (2021). Influence of text luminance, text colour and background luminance of variable-message signs on legibility in urban areas at night. Lighting Research & Technology, 53(3), 263-279.

Heliński, M., Kmieciak, M., & Parkoła, T. (2012). Report on the comparison of Tesseract and ABBYY FineReader OCR engines.

Herrouz, A., Khentout, C., & Djoudi, M. (2013). Overview of web content mining tools. arXiv preprint arXiv:1307.1024.

Hlava, M. M. (2015). Standards and Taxonomies. In The Taxobook (pp. 105-116). Springer, Cham.

Hong, R., Pan, J., Hao, S., Wang, M., Xue, F., & Wu, X. (2014). Image quality assessment based on matching pursuit. Information Sciences, 273, 196-211.

Huang, A. (2008, April). Similarity measures for text document clustering. In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand (Vol. 4, pp. 9-56).

Hussain, W., Sohaib, O., Ahmed, A., & Qasim Khan, M. (2011). Web readability factors affecting users of all ages. Australian Journal of Basic and Applied Sciences.

Im, D. H., & Park, G. D. (2015). Linked tag: image annotation using semantic relationships between image tags. Multimedia Tools and Applications, 74(7), 2273-2287.

Isheawy, N. A. M., & Hasan, H. (2015). Optical character recognition (OCR) system. IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN, 2278-0661.

Ismail, A., & Kuppusamy, K. S. (2018). Accessibility of Indian universities' homepages: An exploratory study. Journal of King Saud University-Computer and Information Sciences, 30(2), 268-278.

Jaimes, A., & Chang, S. F. (1999, December). Conceptual framework for indexing visual information at multiple levels. In Internet Imaging (Vol. 3964, pp. 2-15). SPIE.

Jain, P., Taneja, K., & Taneja, H. (2021). Which OCR toolset is good and why: A comparative study. Kuwait Journal of Science, 48(2).

Jones, M. G., & Broadwell, B. (2008). Visualization without vision: students with visual. In Visualization: Theory and practice in science education (pp. 283-294). Springer, Dordrecht.

Kadayat, B. B., & Eika, E. (2020, July). Impact of sentence length on the readability of web for screen reader users. In International Conference on Human-Computer Interaction (pp. 261-271). Springer, Cham.

Kadupitiya, J. C. S., Ranathunga, S., & Dias, G. (2016, December). Sinhala short sentence similarity calculation using corpus-based and knowledge-based similarity measures. In Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016) (pp. 44-53).

Kesorn, K., & Poslad, S. (2011). An enhanced bag-of-visual word vector space model to represent visual content in athletics images. IEEE Transactions on Multimedia, 14(1), 211-222.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. Psychological review, 85(5), 363.

Kosslyn, S. M. (1975). Information representation in visual images. Cognitive psychology, 7(3), 341-370.

Krejcar, O. (2012). Smart implementation of text recognition (OCR) for smart mobile devices. In INTELLI, The First International Conference on Intelligent Systems and Applications (Vol. 19, p. 24).

Kuric, E., & Bielikova, M. (2015). ANNOR: Efficient image annotation based on combining local and global features. Computers & Graphics, 47, 1-15.

Lara-Clares, A., Lastra-Díaz, J. J., & Garcia-Serrano, A. (2022). A reproducible experimental survey on biomedical sentence similarity: a string-based method sets the state of the art. arXiv preprint arXiv:2205.08740.

Lau, T. P., & King, I. (2006, May). Bilingual web page and site readability assessment. In Proceedings of the 15th international conference on World Wide Web (pp. 993-994).

Li, J., & Wang, J. Z. (2003). Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Transactions on pattern analysis and machine intelligence, 25(9), 1075-1088.

Li, S. H., Yen, D. C., Lu, W. H., & Lin, T. L. (2012). Migrating from WCAG 1.0 to WCAG 2.0–A comparative study based on Web Content Accessibility Guidelines in Taiwan. Computers in Human Behavior, 28(1), 87-96.

Lindley, J., Akmal, H. A., & Coulton, P. (2020). Design research and Object-oriented ontology. Open Philosophy, 3(1), 11-41.

Lu, Z., & Wang, L. (2015). Learning descriptive visual representation for image classification and annotation. Pattern Recognition, 48(2), 498-508.

Lund, W. B., Ringger, E. K., & Walker, D. D. (2014, March). How well does multiple OCR error correction generalize?. In Document Recognition and Retrieval XXI (Vol. 9021, pp. 76-88). SPIE.

Magnini, V. P., Crotts, J. C., & Zehrer, A. (2011). Understanding customer delight: An application of travel blog analysis. Journal of Travel Research, 50(5), 535-545.

Makvana, K., Jay, P., Shah, P., & Thakkar, A. (2016, March). An Approach to identify semantic relations between user's queries in text retrieval. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (pp. 1-6).

Man, A., & van Ballegooie, C. (2022). Assessment of the Readability of Web-Based Patient Education Material From Major Canadian Pediatric Associations: Cross-sectional Study. JMIR Pediatrics and Parenting, 5(1), e31820.

Mannan, H., MacLachlan, M., McVeigh, J., & EquitAble Consortium. (2012). Core concepts of human rights and inclusion of vulnerable groups in the United Nations Convention on the rights of persons with disabilities. Alter, 6(3), 159-177.

Manzoor, M., & Hussain, W. (2012). A web usability evaluation model for higher education providing Universities of Asia. Science, Technology and Development.

Markwood, I., Shen, D., Liu, Y., & Lu, Z. (2017). Mirage: Content Masking Attack Against {Information-Based} Online Services. In 26th USENIX Security Symposium (USENIX Security 17) (pp. 833-847).

Marzan, L. R. (2022). Readability Level Analysis and the Usage of Complex Words on Grade 8 Students' Argumentative Essay. LADU: Journal of Languages and Education, 2(5), 169-175.

Meade, M. J., & Dreyer, C. W. (2020). Web-based information on orthodontic clear aligners: a qualitative and readability assessment. Australian dental journal, 65(3), 225-232.

Metzger, N., & Weil, S. (2019). DFG-Projekt: Optimierter Einsatz von OCR-Verfahren–Tesseract als Komponente im OCR-D-Workflow.

Metzler, D., Dumais, S., & Meek, C. (2007, April). Similarity measures for short segments of text. In European conference on information retrieval (pp. 16-27). Springer, Berlin, Heidelberg.

Mihalcea, R., Corley, C., & Strapparava, C. (2006, July). Corpus-based and knowledge-based measures of text semantic similarity. In Aaai (Vol. 6, No. 2006, pp. 775-780).

Miniukovich, A., Scaltritti, M., Sulpizio, S., & De Angeli, A. (2019, May). Guideline-based evaluation of web readability. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-12).

Morgado, M. A. (1993). Animal trademark emblems on fashion apparel: A semiotic interpretation: Part I. Interpretive strategy. Clothing and Textiles Research Journal, 11(2), 16-20.

Mori, Y., Takahashi, H., & Oka, R. (1999, October). Image-to-word transformation based on dividing and vector quantizing images with words. In First international workshop on multimedia intelligent storage and retrieval management (pp. 1-9).

Neudecker, C., Baierer, K., Gerber, M., Clausner, C., Antonacopoulos, A., & Pletschacher, S. (2021, September). A survey of OCR evaluation tools and metrics. In The 6th International Workshop on Historical Document Imaging and Processing (pp. 13-18).

Nietzio, A., Naber, D., & Bühler, C. (2014). Towards techniques for easy-to-read web content. Procedia Computer Science, 27, 343-349.

Nova, A., Sansalone, S., Robinson, R., & Mirza-Babaei, P. (2022, September). Charting the Uncharted with GUR: How AI Playtesting Can Supplement Expert Evaluation. In FDG'22:

Proceedings of the 17th International Conference on the Foundations of Digital Games (pp. 1-12).

Ojha, P. K., Ismail, A., & Srinivasan, K. K. (2021). Perusal of readability with focus on web content understandability. Journal of King Saud University-Computer and Information Sciences, 33(1), 1-10.

Oliveira, D., Bruno, R., Madeiral, F., & Castor, F. (2020, October). Evaluating code readability and legibility: An examination of human-centric studies. In 2020 IEEE International Conference on Software Maintenance and Evolution (ICSME) (pp. 348-359). IEEE.

Oydanich, M., Kuklinski, E., & Asbell, P. A. (2022). Assessing the quality, reliability, and readability of online information on dry eye disease. Cornea, 41(8), 1023.

Palotti, J., Goeuriot, L., Zuccon, G., & Hanbury, A. (2016, July). Ranking health web pages with relevance and understandability. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (pp. 965-968).

Pantula, M., & Kuppusamy, K. S. (2022). A machine learning-based model to evaluate readability and assess grade level for the web pages. The Computer Journal, 65(4), 831-842.

Patel, A. J., Kloosterboer, A., Yannuzzi, N. A., Venkateswaran, N., & Sridhar, J. (2021, August). Evaluation of the content, quality, and readability of patient accessible online resources regarding cataracts. In Seminars in ophthalmology (Vol. 36, No. 5-6, pp. 384-391). Taylor & Francis.

Peng, Y. H., Wu, J., Bigham, J., & Pavel, A. (2022, October). Diffscriber: Describing Visual Design Changes to Support Mixed-Ability Collaborative Presentation Authoring. In Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (pp. 1-13).

Petrovic, V., & Cootes, T. (2006, July). Information representation for image fusion evaluation. In 2006 9th International Conference on Information Fusion (pp. 1-7). IEEE.

Pinto Molina, M., Alonso Berrocal, JL, Cordón García, JA, Fernández Marcial, V., García Figuerola, C., García Marco, J., Gómez Camarero, C., Zazo, Á.F. and Doucet, AV, 2004. Qualitative analysis of the visibility of the research of Spanish universities through their web pages.

Qurashi, A. W., Holmes, V., & Johnson, A. P. (2020, August). Document processing: Methods for semantic text similarity analysis. In 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA) (pp. 1-6). IEEE.

Rajpathak, D. G., & Singh, S. (2013). An ontology-based text mining method to develop D-matrix from unstructured text. IEEE transactions on systems, man, and cybernetics: systems, 44(7), 966-977.

Rayner, K. (1986). Eye movements and the perceptual span in beginning and skilled readers. Journal of experimental child psychology, 41(2), 211-236.

Reitz, J. M. (2004). Dictionary for library and information science. Libraries Unlimited.

Roca, J., Insa, B., & Tejero, P. (2018). Legibility of text and pictograms in variable message signs: can single-word messages outperform pictograms?. Human factors, 60(3), 384-396.

Romanov, M., Miller, M. T., Savant, S. B., & Kiessling, B. (2017). Important new developments in arabographic optical character recognition (ocr). arXiv preprint arXiv:1703.09550.

Rughani, G., Hanlon, P., Corcoran, N., & Mair, F. S. (2021). The readability of general practice websites: a cross-sectional analysis of all general practice websites in Scotland. British Journal of General Practice, 71(706), e391-e398.

Sakai, Y. (2011). Improvement and evaluation of readability of Japanese health information texts: an experiment on the ease of reading and understanding written texts on disease. Library and Information Science, (65), 1-35.

Sharma, A. K., & Gupta, P. C. (2012). Study & analysis of web content mining tools to improve techniques of web data mining. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 1(8).

Sharma, B., & Rao, A. K. (2020). OCR related technology methods. International Journal, 9(3).

Sheats, M. K., Royal, K., & Kedrowicz, A. (2019). Using readability software to enhance the health literacy of equine veterinary clients: An analysis of 17 American Association of Equine Practitioners' newsletter and website articles. Equine veterinary journal, 51(4), 552-555.

Si, L., & Callan, J. (2001, October). A statistical model for scientific readability. In Proceedings of the tenth international conference on Information and knowledge management (pp. 574-576).

Sinha, T., Li, N., Jermann, P., & Dillenbourg, P. (2014). Capturing" attrition intensifying" structural traits from didactic interaction sequences of MOOC learners. arXiv preprint arXiv:1409.5887.

Sirisuriya, D. S. (2015). A comparative study on web scraping.

Sleiman, H. A., & Corchuelo, R. (2012, June). A reference architecture to devise web information extractors. In International Conference on Advanced Information Systems Engineering (pp. 235-248). Springer, Berlin, Heidelberg.

Smith, R. (2007, September). An overview of the Tesseract OCR engine. In Ninth international conference on document analysis and recognition (ICDAR 2007) (Vol. 2, pp. 629-633). IEEE.

Stenner, A. J. (2023). Measuring reading comprehension with the Lexile framework. In Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement (pp. 63-88). Springer, Singapore.

Su, J. H., Chou, C. L., Lin, C. Y., & Tseng, V. S. (2011). Effective semantic annotation by image-to-concept distribution model. IEEE Transactions on Multimedia, 13(3), 530-538.

Tafti, A. P., Baghaie, A., Assefi, M., Arabnia, H. R., Yu, Z., & Peissig, P. (2016, December). OCR as a service: an experimental evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym. In International Symposium on Visual Computing (pp. 735-746). Springer, Cham.

Tan, W. S., Liu, D., & Bishu, R. (2009). Web evaluation: Heuristic evaluation vs. user testing. International Journal of Industrial Ergonomics, 39(4), 621-627.

Vijaymeena, M. K., & Kavitha, K. (2016). A survey on similarity measures in text mining. Machine Learning and Applications: An International Journal, 3(2), 19-28.

Weninger, T., Palacios, R., Crescenzi, V., Gottron, T., & Merialdo, P. (2016). Web content extraction: a metaanalysis of its past and thoughts on its future. ACM SIGKDD Explorations Newsletter, 17(2), 17-23.

Xia, Z., Peng, J., Feng, X., & Fan, J. (2014). Automatic abstract tag detection for social image tag refinement and enrichment. Journal of Signal Processing Systems, 74(1), 5-18.

Xu, H., Ma, Y., Liu, H. C., Deb, D., Liu, H., Tang, J. L., & Jain, A. K. (2020). Adversarial attacks and defenses in images, graphs and text: A review. International Journal of Automation and Computing, 17(2), 151-178.

Yamasaki, T., & Tokiwa, K. I. (2014). A method of readability assessment for web documents using text features and html structures. Electronics and Communications in Japan, 97(10), 1-10.

Yu, C. H., & Miller, R. C. (2010, April). Enhancing web page readability for non-native readers. In Proceedings of the sIGCHI conference on human factors in computing systems (pp. 2523-2532).

Yuan, Z., Xu, C., Sang, J., Yan, S., & Hossain, M. S. (2015). Learning feature hierarchies: A layer-wise tag-embedded approach. IEEE Transactions on Multimedia, 17(6), 816-827.

Zhang, S., Tian, Q., Hua, G., Huang, Q., & Gao, W. (2014). ObjectPatchNet: Towards scalable and semantic image annotation and retrieval. Computer Vision and Image Understanding, 118, 16-29.

Zhang, S., Tian, Q., Huang, Q., Gao, W., & Rui, Y. (2014). USB: Ultrashort binary descriptor for fast visual matching and retrieval. IEEE Transactions on Image Processing, 23(8), 3671-3683.