MAUU5900

MASTER THESIS

in

Universal Design of ICT

November 2020

A Framework for an Automatic Evaluation of Image Description based on an Image Accessibility Guideline

Himmat Kumar Dogra

Supervisor: Dr. Raju Shrestha

Department of Computer Science

Faculty of Technology, Art, and Design

OSLOMET

Abstract

In this digital world, it is mandatory to have equal access for all the people to understand the images. Inaccessible image descriptions can create barriers for some users to access the images. This paper presents a framework to automatically evaluate image descriptions under the consideration of NCAM image accessibility guidelines. Different research papers of image descriptions and image accessibility were studied for the literature review to explore more about the methodology to accomplish the evaluation of image descriptions. Machine learning is used to build a model which predicts how accessible an image description is with respect to NCAM guidelines. Random forest model was trained using Flickr8K dataset, and the dataset was labeled according to the NCAM guidelines. Standard error and Accuracy were used as a metric to calculate the accuracy and performance. Both the quantitative and qualitative evaluations found that the proposed framework is as effective as human evaluation. The framework is believed to be helpful to the web authors to check the accuracy of any image description with NCAM guidelines in order to make accessible images for the web content.

Acknowledgment

I want to say thanks to my supervisor Associate Professor, Dr Raju Shrestha Department of Computer Science Faculty of Technology, Art, and Design. During this thesis period, I attended a number of meetings with my supervisor, and he supported me by giving valuable feedback on my research. Valuable guidance and suggestions from my supervisor pushed me to sharpen my mind to achieve the goal of this research study.

I would like to thank my accessibility experts who provided a lot of useful information about different concepts to build this automatic machine learning framework.

I also want to thanks to my fellow OsloMet University students in the department who participated in my research study and supported me with valuable time and efforts to perform the questionnaire task.

I wish to show my gratitude to all who supported me with their skills, efforts, and time.

Table of Contents

List of Figures

List of Tables

1.Introduction

Accessibility is the key checkpoint for the multimedia contents on the internet. There are various types of web content users and all have their peculiarities to access the internet. Users use different peripheral devices and accessibility options for interacting with the web. Web content might be of different types, such as images, sounds, text, videos, and animations. In this modern world of technology, images might play an important role in representing web content. Some images like graphs, charts, and maps consist of complex information that might create inaccessibility for the people with visual impairments and color blindness.

“Images can be made more accessible with the use of alternative text and image descriptions” (W3C, 2018). Alternative text or descriptions are helpful in this case that conveys the information of images to the people with such disabilities through a screen reader. An alternative text is a text version of an image for providing the user with the same information a sighted person receives (Paciello, 2000). Whereas, image descriptions gives more information as compared to the alternative text (Veroniiiica, 2018).

Image accessibility is the practice of overcoming the inaccessibility of images for better understanding. “Image description should describe the purpose of the image in the context of the web page” (Petrie, Harrison, & Dev, 2005).

“Many images on the internet are either without any alternative text or containing the inaccurate alternative text” (P. Bigham, S. Kaminsky, Ladner, Danielsson, & Hempton, 2006, p. 1). Images having no alternative text or description might become inaccessible for users with disabilities. There are some special guidelines which can be used to write image descriptions for visually impaired people (Veroniiiica, 2018). WCAG gives the standard guidelines with which developers can make their websites more usable and accessible for the wide range of users (W3C, 2018). Cooper Hewitt guidelines for the image descriptions also explains about the basic guidelines and special guidelines for describing people in images (Cooper Hewitt, 2019). NCAM (National Center for Accessible Media) suggest that the descriptions of the images should be written according to the accessibility guidelines (Diagram Center, n.d.). NCAM guidelines provide transparency or clarity on the description of all type of images. General guidelines for all images include style of text, language, formatting, grammar, and layout that apply to all the types of images. The guidelines include the consideration of important elements such as context, audience, and use of suitable tone. So, the developers need to use accessible image descriptions for the people of wide range regardless of their disabilities. It is mandatory to check how image descriptions are correct according to the image guidelines. Evaluation of image descriptions can be carried out either manually or automatically. Manual evaluation is based on experts or web authors. On the other hand, there is a need for the development of a framework that can evaluate the image descriptions automatically. “Machine learning can be defined as a study of algorithms that can make the system able to automatically learn from experience and training data” ("What is Machine Learning," 2017). Machine learning focuses on the building of a model that might be useful for predictions. By providing training data to a machine learning algorithm, a model can be trained according to the specified rules.

This thesis work envisages a machine learning- based framework to evaluate image descriptions automatically based on accessibility guidelines. A machine learning-based model is trained with a labeled dataset to predict the accuracy in compliance with the image accessibility guidelines. In addition to this, a questionnaire study has been conducted with accessibility experts to validate the quality of predictions from the model.

1.1 Problem Statement

Image descriptions are the part of web images that should be accessible enough to convey the contextual information of an image in the form of text on websites. Image describers use image descriptions without following any guidelines that make the web experience inferior for users. Although there is the availability of image accessibility guidelines, most of the websites are not adhering to them (Petrie et al., 2005). Manual evaluation techniques for image descriptions are limited to only a few numbers of images and it is time-consuming for web authors to evaluate thousands of image description manually (Dahal & Shrestha, 2019, p. 4). There is a need for an automated evaluation method that enables quicker evaluation and ensures the quality of image descriptions so that images descriptions convey the same information to all the users.

This thesis work tries to address this by designing and developing an evaluation framework to check the accessibility of image descriptions based on image accessibility guidelines automatically.

1.2 Objective and Research Questions

Based on the problem statement, the main objective of this study is to design and develop a framework to automatically evaluate the accessibility of images descriptions based on NCAM image accessibility guidelines.

To achieve this objective, the major research question has been defined as follows-

- How can we automate the evaluation of the accessibility of image descriptions based on an accessibility guideline using a machine learning-based framework that performs comparably with manual evaluation by experts?

1.3 Research Methodology

There are two major methodologies to evaluate the image descriptions, namely quantitative and qualitative methods. Qualitative research involves collection of the data and analyzing data to find a relevant answer for the research. In contrast, the quantitative approach focuses on explaining a phenomenon and applies statistical techniques (Rhodes, 2014). Another methodology of research is mixed-method design that integrates techniques from quantitative and qualitative methods to answer a research question (Byrne & Humble, 2006). In search of getting answers for my research question, this study has applied a mix method approach.

1.3.1 Mix Methods Research

In mix-method research, both quantitative and qualitative methods are used to strengthen the research study. Mix methodology introduces an alternative to the quantitative and qualitative methods to answer the research question (Subedi, 2016). The research question of this study firstly aims to build the automatic image description evaluation framework based on accessibility guidelines using a machine learning-based framework. The framework predicts the image description accuracy in quantitative nature. The next focus of the research question is to check how this framework performs with manual evaluation by experts. Thus, quantitative data collects first, and then responses from the experts can be used as an explanation for the quantitative results (Creswell, 2012).

Firstly, a quantitative method was used to perform the quantitative evaluation for the automatic evaluation framework. The quantitative results of the framework were further analyzed using different statistical methods. The quantitative research method was used to determine the relationship between framework prediction and ground truth (target results). The ground truth data refers to the most accurate dataset, which was labeled by the expert.

The data collection process was performed with the help of the Microsoft Excel software. The data includes the results from the prediction of the framework and ground truth (target values). Statistical methods were used to find out whether the framework predicts a similar image description accuracy as that of ground truth. Table 4.1 explains the accuracy of the framework in compliance to the ground truth values.

On the other hand, a questionnaire was used to collect the scores from experts, and these scores further were compared with the framework scores to find the association between the scores for each guideline. Using expert judgments, an expert can describe information in statistical data with the help of different types of scales (Iriste & Katane, 2018). Correlation analysis was chosen to figure out how much the framework is efficient in producing results as same as expert results.

1.4 Thesis Organization

This thesis contains five chapters, including this chapter. Chapter 2 presents the background works that include relevant concepts and related research. Chapter 3 presents the design and development of the framework. Chapter 4 explains about experiments and results. Chapter 5 includes discussion, conclusion, future work, and ethical issues of the research study. The reference list and appendixes are placed at the end of the report.

2.Background and Related Works

Web accessibility facilitates accessible web contents to the people with disabilities to develop and maintain their education, employment, social and family life (Lewthwaite, 2014). Web accessibility may rely on different components of web development such as web developers, web tools and web content. Web accessibility guidelines may help the developers to overcome the problems that create barriers for people with special abilities. There are some tools for the developers to create accessible websites such as accessibility evaluation tool, assistive technology and web authoring tools (Yu, 2002). “One of the reasons of inaccessible websites is the scarcity of professionals who are familiar with accessibility evaluation tools” (Abuaddous, Jali, & Basir, 2016, p. 1). Some of the accessibility tools may not support new or changed accessibility guidelines (Trewin, Cragun, Swart, Brezin, & Richards, 2010). “The existing accessibility tools still provide incomplete automation for certain accessibility issues such as calculating alternative text quality (Harper & Chen, 2012, p. 5).

Section 2.1 describes different image accessibility. Section 2.2 explains about different image accessibility guidelines. Section 2.3 explore different evaluation methods used to evaluate the quality of image descriptions.

2.1 Image Accessibility

Images are the relevant part of the web content that describe the information using a chart, graph, table, image of text and maps. Images sometimes contain a lot of information that may create the barriers for people with disabilities. However, people with disabilities have to face difficulties while gaining information from images.

The image description is the key feature of the images to enhance image accessibility. The alternative text consists of few words of information about the image. However, it is recommended to keep the description near about 280 characters, so that screen reader finishes it in an appropriate time (Veroniiiica, 2018). There are different types of images for representing the detail in a different pattern such as informative, decorative, functional, images of text, complex images, group of images and image maps (W3C, 2014). Informative Images. Complex images consist of graphs, charts, maps, flow charts or diagrams

2.2 Image Accessibility Guidelines

WCAG developed some guidelines and standards that explain how to make web content more approachable to diverse users. The primary focus of Web Content Accessibility Guidelines (WCAG) is to enhance the accessibility of web content for all people to give them same services and features.

Image annotations is an important part of the image accessibility and it can be measured using different types of methods. For image descriptions evaluation, there is a need for some standards and guidelines that can be helpful to identify accessibility the image descriptions. WCAG include the accessibility guidelines to enhance the accessibility of web content. WCAG also explains about the guidance on how to create accessible images that meet its guidelines (W3C, 2014).

National Center for Accessible Media (NCAM) explains guidelines for almost all the types of images. These guidelines were developed by the Carl and Ruth Shapiro at NCAM for creating effective and efficient text alternatives to images for the people with disabilities (Diagram Center, n.d.). These guidelines can enhance the overall accessibility of websites which consist of tons of images. The art of writing the alternative text can be more accessible and universal with the implementation of these guidelines in image descriptions. These guidelines give the equivalent web experience to all the people, including people with disabilities and hence decreases the segregation in society. The fourteen NCAM image accessibility guidelines are listed and described below.

1. The description should be succinct.
2. Colors should not be specified unless it is significant.
3. The new concept or terms should not be introduced.
4. The description should be started with a high-level context and drilled down to details to enhance understanding.
5. The active verbs in the present tense should be used.
6. Spelling, grammar, and punctuation should be correct.
7. Symbols should be written out properly.
8. The description vocabulary should be added which adds meaning, for example, “map” instead of an image.
9. The title and axis labels should be provided.
10. The image should be identified as a scatter plot and be focused on the

change of concentration.

11. The central teaching point should be focused to determine if borders, regions, shapes, and bodies of water are important.
12. The description should be organized using number lists and pull the most important information in the beginning.
13. Physical appearance and actions should be explained rather than emotions and possible intentions.
14. The material should not be interpreted or analyzed; instead, the reader should be allowed to form their own opinions.

First 8 guidelines are common to all types of images, while guidelines 9 and 10 are for graph image, guidelines 11 and 12 are for map images and guidelines 13 and 14 are for natural images (Dahal & Shrestha, 2019).

Cooper Hewitt also explains about the accessibility of images with image description guidelines. These guidelines include general, core aspects of description, describing people, enriching description and infographics (Cooper Hewitt, 2019). In the general part, there are guidelines about the repetition of text, avoiding jargons and usable limits of words, color, size, orientation, subject, graph, map, diagrams, and tables.

On the other hand, NPR guidelines are generally based on the persons, locations and time in the captions (FEDERICO, 2016). These guidelines explain the common rules to write proper information about the persons and their names if necessary, so that reader can identify the people in the captions. These guidelines also guide, where a descriptor should use exact date and where a date can be skipped in the captions.

2.3 Evaluation of Image Descriptions

There are two ways for the image description evaluation, manual method and the automatic method. In the case of manual methods, different types of manual techniques are used to solve the problem. BLEU evaluation metric is used to measure the similarity between two sentences (Papineni, Roukos, Ward, & Zhu, 2002). Whereas, ROUGE(recall-Oriented Understudy for Gisting Evaluation) evaluation is carried out via word sequences or word pairs (Lin, 2004). ROUGE works with comparing automatically created summary to the human-created summary. N-grams are simply a combination of adjacent words or letters of a length n of a source text. The basic point of the n-grams is to capture the language structure of a text. But, BLEU does not measure the meaning of a text in a language and it only gives preferences for n-grams that have exact matches (Tatman, 2019). In addition to this, BLEU does also not consider the sentence structure appropriately and can give the same scores if positions of words are changed in a sentence. Eventually, these metrics do not have any specific rule for the specific type of images, and it may be challenging to use these metrics for the work related to the accessibility of the image descriptions (Hodosh, Young, & Hockenmaier, 2013).

In the case of classification, accuracy is the most common type of metric that is used to measure the performance of the model. While regression analysis gives the way to compare the effects of different variables on any scale. There are different types of regression techniques that can be used to make predictions. There are different types of metrics available for regression evaluation. Mean Squared Error(MSE), Mean Absolute Error (MAE) and R Squared are the common metrics for regression evaluation (Drakos, 2018). Regression techniques were used to estimate the relation between the selected dependent variable and independent variables.

R squared may be a good choice to know how good the features explain the variance in the model. More the value of R Squared closer to the 1, better will be the model (Guanga, 2019). RMSE and R Squared were used in the model for the evaluation and other evaluation metrics can also be used in order to completely evaluate the model so that the model will become as much optimized as possible.

2.3.1. Manual Evaluation of Image Description

Accessibility of the Image descriptions can be improved by providing example images with the sample descriptions (Dahal & Shrestha, 2019, p. 3). The study conducted an experiment to evaluate the proposed method in which similar example image with description was given to the user as a sample cue to write the description of another image (Dahal & Shrestha, 2019, p. 3). The description of sample cue was based on NCAM guidelines and written by the image accessibility experts. The participants were asked to write descriptions without giving any cue, then with a random sample with description and finally by providing a similar cue with a description. Similar example cue provided hints to the users for writing accessible descriptions. The outcome of this study showed a lack of availability of descriptive summaries of the images on the web. This literature also indicates that the main reason for this problem may be the lack of professional web authors, the complexity of writing image descriptions and the lack of time to read the image descriptions guidelines. In addition to this, it also explained that image descriptions could be more accessible by providing some hints to the users. However, this research study was limited to a few participants and only five hundred image descriptions were written by the participants. Alahmadi and Drew (2018) investigated the evaluation to measure the accessibility of 120 web images of university web-based system with the help both manual and automated methods. The findings of the study showed that 88% of images were inaccessible and only 14 images had descriptions in case of human evaluation. In comparison, automatic evaluation found that the majority of images requires either a long description or a valid alternative text.

To sum it up, both the manual and automatic evaluation methods are useful to evaluate the image descriptions. But evaluating the image description is a time-consuming process and it can also cost a lot if the higher number of participants will participate. Manual evaluation methods focused on image accessibility in a better way as compared to the automatic evaluation. So, there is a lack of automatic evaluation method to evaluate the image descriptions based on some image accessibility guidelines.

2.3.2. Automatic Evaluation of Image Description

A machine learning basically is a study of teaching the program how to solve the given task after learning from examples. With the help of machine learning, it is possible to promptly and automatically build a model that can analyze big and complex data to produce faster and accurate results. There is a need for a machine learning algorithm to build a model that can learn from the given data. Supervised and unsupervised learning are common types of machine learning algorithms (Heidenreich, 2018).

In Supervised learning, a machine learning model is trained using data consisting of examples and labels. Labels is also called a dependent variable or predictions. Examples can be any type of features or independent variables of the data that help the model to predict the accurate labels. After the complete training of the model, the supervised machine learning model will be able to predict the accurate label for a never-before-seen example. Predicting house prices and classifying spam e-mails are examples of supervised machine learning. Decision Tree, Random Forest, and

Logistic Regression are few examples of supervised learning algorithms. While unsupervised machine learning has no features no labels, and it is used for clustering, grouping, and organizing data.

Existing automatic evaluation tools can also be used to check the image accessibility. This study presented Acrolinx language checker software that can be used as an accessibility evaluation tool (Vázquez & Lehmann, 2015). However, this customized language checker was used to verify the alternative text in the web images, but it has limits to customize the functionalities.

A dataset can be evaluated using sentence based image description on a ranking task that correlates highly with human judgements (Hodosh et al., 2013). The framework used Correlation to find the lexical-based similarity between the words (Hodosh et al., 2013). This study contradicts the use of metrics such as BLEU and ROUGE because of less similarity of these evaluation metrics with human judgments. So, the study indicates to avoid such evaluation metrics for the evaluation of image descriptions.

Accessibility of web pages is also the relevant part of web experience and images in the web content play an important role. The study described a classifier that is capable of evaluating the accessibility of alternative text on the web pages (Bigham, 2007). The inputs for this solution are some URLs with simple rules. Content similarity, word-based similarity and meaningful alternative text are three main types of the criteria were used to calculate the quality of an alternative text. For evaluation, a classifier was trained using labeled examples from a dataset. The classifier performed well with labelled example and the accuracy is near to 86%, but these results are not the same in case of unlabeled examples (Bigham, 2007). Thus, the study indicates that it is essential to have labeled dataset to build an efficient classifier.

3. Design and Development of the Framework

The development of the image description evaluation framework involves four major steps: a selection of a guideline for an accessible image description, creation of a dataset, development of a model, and performance evaluation. Selection of image description describes the reason for choosing the image accessibility guideline and its contributions in the field of accessibility. Whereas the creation of dataset explains the procedure and rules used for the creation of dataset. How the model designed and developed is explained by the development section. Also, performance evaluation includes the different types of evaluation metrics.

Section 3.1 explains about the selection of image accessibility guidelines. Section 3.2 explains about the creation image description dataset. Section 3.3 explains about the selection of the machine learning model. Section 3.4 is about the development of the model. Section 3.5 explains the evolution of the model.

3.1 Selection of Image Description Guideline

First of all, there was a need for selecting the appropriate image description guidelines. A number of image description guidelines were studied in the literature study, and all the guidelines have different rules to define the image descriptions. After studying and comparing different image description guidelines, NCAM guidelines were selected for this method. Based on the literature study, NCAM guidelines were found to be better than other guidelines. NCAM and DIAGRAM (Digital Image And Graphic Resources for Accessible Materials) teamed up in 2014 to provide relevant resources for making the images accessible (Diagram Center, n.d.). NCAM guidelines focus not only on STEM (Science, Technology, Engineering and Mathematics) images but also on all types of images. NCAM guidelines, in conjunction with DIAGRAM Center explains the general guidelines for all types of images and the guidelines for a specific type of images. NCAM guidelines for complex images can also be used to extend the research of this project by implementing the specific guidelines on specific type of images. NCAM carries out evaluations for websites, applications and electronic documents for conformance with all levels of WCAG 2.0 (WGBH, n.d.). In addition to this, NCAM also introduced a POET image description tool which can be used to learn and create accessible image descriptions (Poet training tool, n.d.). To sum it up, NCAM guidelines are superior to other guidelines. So, this research is based on the NCAM guidelines, which include general guidelines applies to all and the guidelines for natural images.

3.2 Creation of the Dataset

Machine learning depends on datasets, and it is not possible to train a model without the training data. In this study, the image dataset should be based on image description guidelines. There are various image description datasets available such as MS-COCO, ImageNet, Flickr 8K, Flickr 30K and others which are not based on any accessibility guidelines. Flickr8k was selected initially as an image description dataset for the model building. The Flickr8K dataset has near about 8000 images and each image has five captions. Whereas another dataset like MS-COCO and ImageNet has a huge number of images as compared to Flickr8K. So, there is a need for manual creation of accessible image description dataset to train the model. The dataset should be created according to the selected NCAM guidelines. Manually labeling the dataset with a large number of images takes more time and this is the first reason behind choosing the Flickr8K Secondly, Flickr8K is available for free and it has five captions for each of the guidelines. In addition to this, the model can be trained faster with this dataset as compared to other large datasets (Shinde, 2019). Describing the one image in five different ways will help the model to understand the more accurate image description for that single image. Creation of the dataset includes assumptions of various concepts related to language processing tasks.

The training dataset has been labeled by an image accessibility expert as zero to hundred percentage compliance with each guideline. As the dataset has nearly 8000 images with image descriptions and all the image descriptions has been labeled with image accessibility guidelines. The selected accessibility expert was fully aware of image accessibility and NCAM guidelines.

3.3 Selection of Machine Learning Model

A machine learning library in python programming will be used to perform the coding work of this research project. Sci-kit-Learn is a machine learning python library that provides supervised and unsupervised machine learning algorithms (Seif, 2018). Sci-kit is a python module that provides state-of-the-art implementations of many machine learning algorithms for building the model. Sci-kit is open-source and has many in-built features for data engineering. In addition to this, Sci-Kit provides a user guide consisting of installation instructions, documentation, and tutorials.

Random forest is a type of model that create a collection of weak models and then combine them to give the final output (Sarkar, Bali, & Ghosh, 2018). The proposed model is based on the random forest which can be used to build the model in the field of supervised machine learning. As the multiple decision trees are used to predict the result, the final result is the average of all the results of the decision trees. If a few decision trees predict inaccurate results, still the overall result will go to the majority of decision trees. This technique of using multiple decision trees reduce the overfitting of the model. Kadiyala and Kumar (2018) completed a study on evaluating the performance of different bagging ensemble methods including the random forest by using machine learning libraries on the Windows platform. The random forest regressor which is a type of ensemble technique was used to build the model for the evaluation of image description. Random forest works efficiently in regression problems and gives better results than other ensemble techniques (Kadiyala & Kumar, 2018). The reason for choosing the random forest over the other algorithms is its robustness and adaptability to reduce the overfitting of the model (Garg, 2018). The study revealed that the random forest method performed better than other ensemble methods for prediction (Kadiyala & Kumar, 2018).

3.4 Development of the Machine Learning based Model

Machine Learning Life Cycle is a cyclical process that includes a few steps and it can be used as a development process to implement the data science projects (DataRobot, n.d.). The traditional software development lifecycle can be mapped as a machine learning lifecycle (Ferlitsch, 2019). In planning, developers prepare the plan to build a model that can achieve the goal of the problem. It is mandatory to select the right data for the model building because the model can learn noise from the bad data. The data extraction process is handled by the data engineers, whereas data analysis are handled by the data analysts (Ferlitsch, 2019). In modeling, machine learning-based model is trained to predict the output from the features. So, this suitable approach used as a process to create and implement the machine learning-based models for the prediction of image description based on image description guidelines.

The automatic evaluation framework of image descriptions is based on the random forest method for regression analysis. Number of features with their labels were passed through the model and training was performed in this way to train the model for the better predictions. Total number of trees was 20 on which the model was trained. After checking all the tuning of parameters, this selection gave the best performance of this model.

3.4.1 Features Selection

The main part of a model is to choose the accurate rules to extract the features for each guideline. There are different rules and important points of making the image descriptions accessible and these points helps to extract the relevant features for each guideline. Features are extracted for each guideline as discussed below: -

Guideline-1. The first guideline is that the description should be succinct. The conciseness of the image description allows the readers to read within a short time. Repetition of words, length of the sentence and sentence similarity were considered to calculate the accuracy of the description. There should be no repetition of words in the description to make it concise (Diagram Center, n.d.). Jaccard Similarity calculates how many set of words are similar between the sentences (Sieg, 2018). The similarity of two sentences was checked with Jaccard similarity and if the similarity is higher than 80%, then the model will assign low scores for that description.

Guideline-2. The second guideline is about avoiding the usage of colors in the descriptions. “Information in the image description of a chart, graph or map may become illegible if arbitrary colors will be used in the description” (Diagram Center, n.d.). So, Webcolors library is used to identify the colors in the description of map and chart and the color dictionary has 147 colors (Webcolors, n.d.). If a description has a higher number of colors, then the score will also be less as a greater number of colors reduce the clarity of the image description.

Guideline-3. The third guideline is based that there should not be any new information in the description. So, words which are used to explain the term and concepts are used to identify whether the description includes the new information out of the image context or not.

Guideline-4. This guideline explains which type of information should be added in the description. The starting of the description should include the relevant context of the image and the rest of the information can be included later. There are some pre-defined python packages to solve this sort of problems. The guideline separated into two parts first is to check how an image description is easy to understand and the second is to identify how many complex words in that description. Higher the readability scores, easier will be the image description to understand. A lower number of complex words also help the reader to understand the image description more clearly.

Guideline-5. This guideline is about the type of tenses are used in the image descriptions. According to this guideline, there should be only active verbs in the present tense of the image descriptions. A python library Natural language Toolkit (NLTK) can be used to perform the task for this guideline. Part-of-speech tagging of NLTK were used to identify the past and past participles from the image descriptions and VBD and VBN are the two part-of-speech tags that are used for identifying past and past perfect tenses respectively (Rachiele, 2018). If descriptions include either past tense or a present or past perfect tense, then low score will be assigned by the model.

Guideline-6. This guideline requires that grammar, spellings, and punctuations should be correct. A pre-defined string named as “string. punctuation” which gives all the punctuations. So, punctuations from the image descriptions can be identified with the help of this feature. For the spellings, GloVe vectors were used to identify the incorrect words (Sieg, 2018).

Guideline-7. The seventh guideline is only about symbols. If there is any symbol in the description, then it should be defined in such a way that screen does not face any problem while speaking it out to the users. There are a lot of special symbols that can make the description inaccessible for the user if that user is reading descriptions with a screen reader. For example, if the description has ‘5 cm’ as to describe something in measurement, then image description should include ‘five centimeters’ so that screen will not face any problem (NWEA, 2017).

Guideline-8. This guideline explains that there should be proper description vocabulary for each type of image. If the image description is about a map which is showing three countries, then the description can be written as “map of three countries” instead of “image of three countries.” Rules were created to identify the words that are used in the description of a complex image to implement this guideline.

Guideline-9. The guideline is about the inclusion of actions rather than emotions and intentions. A python library was used that itself calculates the sentiment scores and give accuracy.

Guideline-10. The last guideline says, there should be not the interpretations and analysis of the information in the image. The image descriptions should be in such a way that allows the users to understand according to their own opinions. Also, the image description should not have any controversial and uncomfortable content (Diagram Center, n.d.). Rules were created to extract the words that interpret or analyze the descriptions. The first rule is to check whether the image description writer added his own interpretations. The second rule is to identify any controversial words in the description. The words like sex, politics, race, and gender were also considered as rules in order to identify the inaccessible content in the image descriptions.

3.4.2 Training the Model.

Supervised machine learning was used to build and train the model with the image description dataset. Python 3 were installed on a window-based laptop through anaconda machine learning platform. Anaconda is a distribution of packages which is useful for data science tasks and it provides a number of packages and their dependencies (Mathur, 2016). After installation of the anaconda includes the python versions and their packages. Jupyter notebook, an in-built tool of anaconda, was used to write the code for this python project. A package management system ‘conda’ from anaconda distribution were used to create the environment for the project. All the required libraries were installed, and a complete framework consisted of a single model for each guideline. Thus, random forest models were trained for each guideline to predict the accuracy. A python module Sci-Kit was used since it provides implementations of different machine learning algorithms (Pedregosa et al., 2011). So, random forest regressor was used to predict the accuracy of image descriptions.

Furthermore, the image description dataset was separated in training and testing part. The training partition consists of 80% of the dataset, whereas 20% were used as testing purpose. The dataset consists of independent variables (features) which were to train the model. For framework optimization, 10-folds cross-validation were used during the training of the model. In the case of hyperparameter tuning, different types of parameters involved in a random forest model, but the current framework only used two main parameters. Selecting the number of trees (n_estimators) and the maximum number of features used for splitting a node (max_features) were checked to improve the performance. Number of trees mainly used 20 and maximum feature set for the square root of the number of all features. The results of the testing dataset were extracted in an excel file for further analysis.

3.5 Evaluation of the Model

Cross-Validation was used in the method to improve the training of the model as it split the dataset into different parts and use all of them in an appropriate way. K-fold cross-validation was selected for the method where the value of K was 10 in the proposed method. K-fold cross-validation can have 5 or 10 partitions that can be used for computing the error rate of the model and using this strategy with random forest construct the efficient model (Silke & Roman, 2018). After the training part, the model predicts accuracy of image descriptions and gives results in the form of percentage for each guideline. The Flickr8K dataset was small in size but using ten folds of cross validation used all the dataset. Furthermore, cross-validation was used to estimate the prediction error.

Two metrics are used to check the capability of the framework: prediction error and accuracy. Mean of the prediction error is calculated and used to calculate accuracy.

Prediction error: Prediction error is the absolute value of difference between the predicted value of a guideline and the ground truth (target value). This metric is used to calculate how good the model predicts the response variable.

$$error \; = |predicted\ value - target\ value|$$

Accuracy: Accuracy shows the percentage of correctness, and it is the number of accurate predictions over the total predictions multiplied by a hundred.

$$accuracy = \frac{correct\ predictions}{total\ number\ of\ predictions} \times 100\%$$

A prediction is considered as correct if the value is close to the target value within the standard error of the prediction error. Accuracy of testing part of the model is calculated by dividing the correct prediction to the total number of prediction and then multiplied by 100 for a percentage value.

Standard error: The standard error is standard deviation over the square root of the sample size.

$$\text{Standard Error} = \frac{Standard\ Deviation}{\sqrt{Sample\ Size}}$$

The difference between target values and the predicted values were calculated in the form of absolute values. Then, the standard deviation of those values was calculated.

4.Experiments and Results

Two different experiments were conducted, quantitative and qualitative. Performance of the model was evaluated quantitatively. In comparison, qualitative study was conducted to validate the performance of the model. Thus, a hybrid approach was used to achieve the goal of this study. Section 4.1 explains about the experiment for the evaluation of framework and quantitative evaluation results. Section 4.2 explains the experiment for the qualitative evaluation and its results. Section 4.3 shows the interpretation of both the quantitative and qualitative evaluations performed in this research. Whereas sub-section 4.3.1 explains the correlation between framework prediction results and expert evaluation results. Sub-section 4.3.2 explains the comparison of means of framework results and expert results.

4.1 Experiments for the Evaluation of the Framework

The developed framework is based on random forest regressors after trained by the labeled dataset. The experiment performed step by step task to achieve the higher performance of the model. After the successful building of the framework, quantitative evaluation was conducted to identify the accuracy of the model.

The statistical methods, prediction error, accuracy and standard error were applied on the collected data from framework. Thus, quantitative methods were used to obtain a standard error, mean error, and accuracy values, as shown in Table 4.1. The proposed framework give results in the form of accuracy scores from zero to a hundred. The quantitative evaluation of the framework was carried out after training of the model with the labeled dataset. The results from the ground truth (Dataset Result) was compared with the quantitative results obtained from the model. Various types of statistical tests were studied to explore the collected data from participants. Standard deviation means and standard errors might be used to estimate the nature of data.

Mean error was calculated from both of these results, and then standard deviation and standard error were calculated to figure out the accurate prediction of the model.

4.1.1 Results from the Framework

Statistics of the prediction results from the model on each guideline is given in Table 4.1. Mean prediction error along with the standard deviation and standard error, are shown in Figure 4.1. The resulting accuracy is showing graphically in Figure 4.2.

Table 4. 1 Guideline Wise prediction error and accuracy results from the Framework

Guidelines	Mean Error	Standard Deviation	Standard Error	No. of Accurate Predictions (Total=1706)	Accuracy in Percentage (%)
Guideline 1	0.219	3.47	0.084	1692	99.30
Guideline 2	0.188	3.28	0.079	1694	99.53
Guideline 3	0.093	2.47	0.060	1684	98.83
Guideline 4	0.074	2.19	0.053	1657	97.24
Guideline 5	0.056	1.64	0.040	1676	98.36
Guideline 6	0.215	2.99	0.072	1692	99.30
Guideline 7	0.141	2.37	0.057	1686	98.94
Guideline 8	0.047	1.19	0.029	1701	99.82
Guideline 9	0.323	3.00	0.073	1681	98.65
Guideline 10	0.129	2.32	0.056	1695	99.36

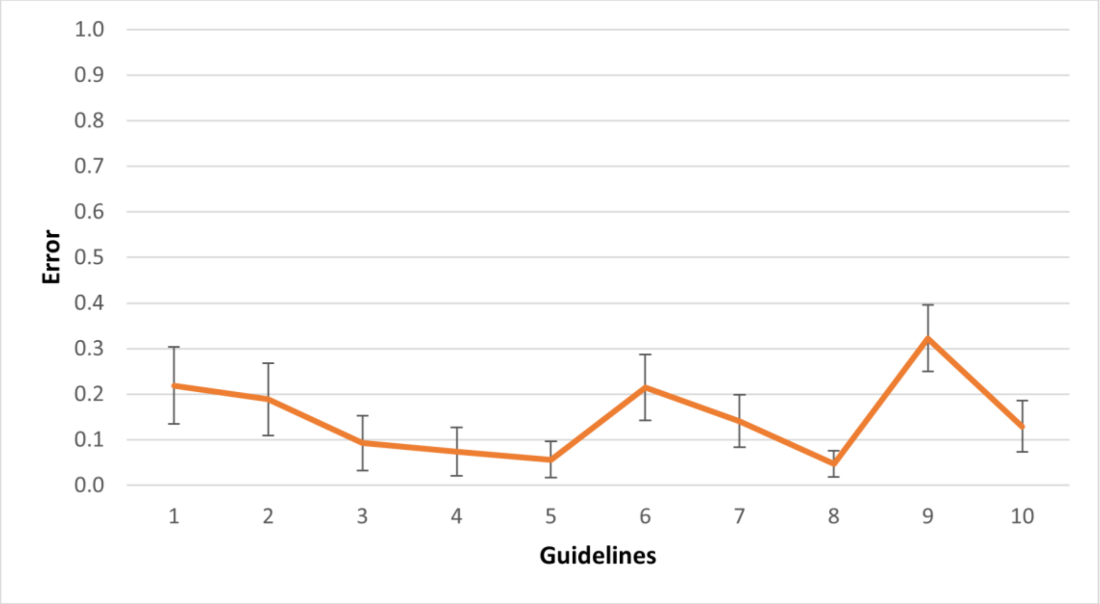

Figure 4. 1 Prediction error plot for each guideline along with the standard error

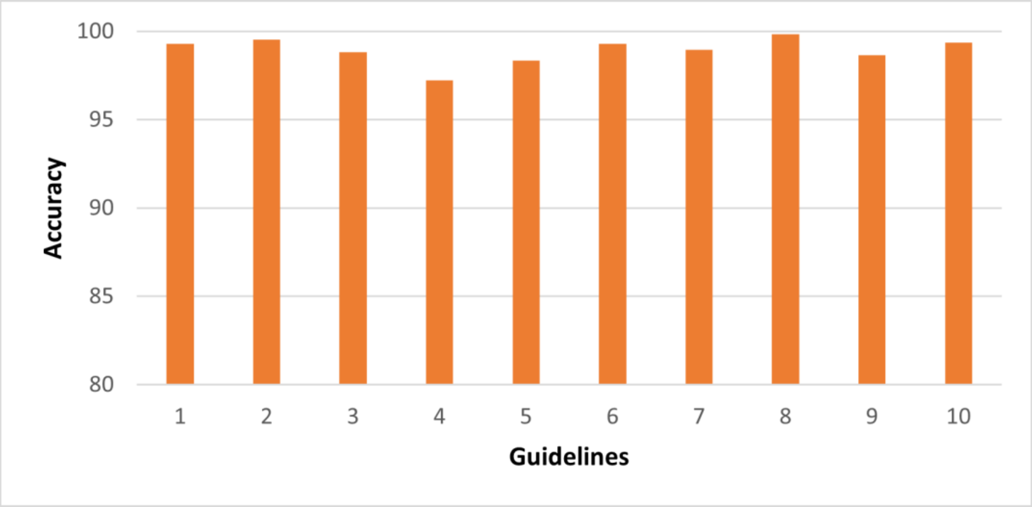

Figure 4. 2 Results of the evaluation framework

The table is representing the different parts of the results for ten guidelines. The column 'Mean error' were calculated as a mean of the difference between the target scores (ground truth scores) and the predicted scores. Third column is the Standard Deviation of the difference between the target and predicted values. Then, standard error column represents a particular standard error value for each guideline. The last column of the table is representing the accuracy of each guideline in percentage. As shown in the table, target scores are very similar to the predicted scores and the model is giving small standard error values. In addition to this, a line chart is representing the mean difference of the target and predicted scores with respect to the standard error.

As shown in the line chart, Guideline 5 and Guideline 6 has higher standard error values as compared to other guidelines. On the other hand, Guideline 1, 2, 3 and 4 has low values of the standard errors followed by guidelines 7, 8 and 10. To be concluded, guideline 5 is based on the NLTK library which was used to identify past and past participles. The performance of this guideline can be improved with the improvement in the dataset by giving some more samples. In case of guideline 6, GloVe vectors were used to check the spellings and grammatical mistakes that might be improved by substituting the 50-dimensional vector to any other size of the vectors (Pennington, Socher, & Manning, 2014).

4.2 Experiments for the Qualitative Evaluation

Firstly, a consent form was created that includes the confirmation of the voluntary participation of participants in the web-based experiment. The consent form briefed about the motive of the research as shown in appendix A. An option was provided to the participants if they agree to contribute for the research study. Furthermore, questionnaires were created to collect the data from participants as shown in appendix B. The reason for choosing the online questionnaire was the different location of participants in the countries and also the pandemic situation. The questionnaire was made user friendly for the participants in order to build up their interest. Questionnaires were interactive in nature and one could easily give answers even on a smartphone. Each questionnaire has ten image descriptions. A total of 35 participants participated in this experiment with a unique questionnaire. There were 35 questionnaires which separately distributed to each participant. Each participant has one questionnaire which further consisted of a few questions. The first question was related to the consent form for the voluntarily participation in this research. Participants were asked with an option to proceed further for the experiment. A slider scale was used to receive the response from the participants. In this expert evaluation, participants were used as experts to evaluate the image descriptions in compliance with NCAM image accessibility guidelines.

4.2.1 Questionnaire Design

There were several steps involved in developing online questionnaires. Prior to the main questions, participants were provided with brief introduction of the research and instructions for the questionnaire. Furthermore, each questionnaire included ten image descriptions and participants were asked whether these image descriptions were accessible or not with respect to NCAM guidelines. Chronology of the questions remained the same for all the questionnaires. However, image descriptions used in questionnaires were selected randomly from the Flickr 8K dataset with the help of a python program.

The slider scale range from zero to hundreds which represent the percentage of the image description in compliance with the NCAM image accessibility guidelines. Whereas the value zero corresponds to strongly disagree and the value hundred corresponds to strongly agree. While comparing the results of radio buttons with the slider scales, no statistically significant difference was calculated (Roster, Lucianetti, & Albaum, 2015). So, an online service were used to build a web-based slider scale questionnaire and the complete web-based experiment was anonymously conducted. Unlike Likert scale, visual analog scales helps to avoid statistical challenges and problems due to the ordinal scales (Voutilainen, Pitkäaho, Kvist, & Vehviläinen-

Julkunen, 2016).

4.2.2 Participants Selection

The selection of the participants was required to have sufficient knowledge for the image accessibility and image accessibility guidelines. Participants were selected from different locations but with the same requirement. Participants who are studying or studied in universal design and image accessibility were selected for participation. The educational background or profession of participants must be related to universal design and accessibility. Since participants with expertise in accessibility may evaluate better than participants without expertise. Thus, participants selection were conducted to perform expert evaluation. Majority of the participants were postgraduate students from Oslo Metropolitan University. The researcher contacted each candidate on a social media platform and briefed them about the motive of the research. Participants were instructed appropriately to take part in the anonymous questionnaire process.

4.2.3 Tools Used

Questionnaires were created with the help of an online survey providing website. Various types of online questionnaires services were checked to explore more about the privacy and eventually one was selected for the experiment. The questionnaires were anonymous so that no data could be collected by the researcher. The researcher collected only quantitative scores from the participants. The main equipment used for this experiment was a Dell laptop, spreadsheet software (MS Excel) and Google Chrome web browser. The questionnaires were interactive and also tested on the mobile version of the google chrome so that participants may also be able to complete this questionnaire on a smartphone.

Before administering the questionnaire to all, a pilot study was performed on a web-based questionnaire on five participants as explained in section 4.2.6. Participants were invited online from different places and a web link was shared with them as an invitation. As the experiment was anonymous, so information related to geodata, IP address, names, address, age, date of birth and profession were not collected.

4.2.4 Data Collection

A web-bases questionnaire provided an appropriate environment to the participants for contributing for the experiment. This experiment was specifically designed for observing the participants responses based on their prowess in image accessibility and NCAM guidelines. Participants read the image descriptions and gave scores for all the ten guidelines from zero to hundreds with the help of a slider scale. On the left, there was a mark of zero score and on the right, there was a hundred score mark. Participant can see the actual score when scrolling left or right the between the zero and hundred. Moreover, zero scores were described as completely disagree, while a hundred were completely agreed. Thus, participants scored to all ten guidelines for all of the ten image descriptions.

Secondly, the researcher encouraged participants to take part independently in this experiment as per their understanding. Different participants took time to complete this task and answers were collected in the form of quantitative data. Thus, data from all participants were gathered and exported to excel for analysis.

4.2.5 Data Analysis

A suitable statistical analysis required to measure the relationship between the framework produced scores and the participants scores. These types of research objectives can be quantitatively addressed with the help of correlation analysis (Schober, Boer, & Schwarte, 2018). There may be two commonly used correlation coefficients, the Pearson coefficient and Spearman coefficient. Pearson correlation could be performed if the data is normally distributed with a monotonic relationship between two continuous variables. Also, there should be no relevant outliers in the data according to the assumptions of pearson correlation. On the other hand, if collected data violates the assumptions of Pearson correlation, then spearman correlation can be used as an alternative correlation method.

The collected data were analyzed using the software MS excel. Scores from all participants were together arranged in a single column for each guideline. There were ten columns and each column of the guideline consisted of all participant scores. According to the data, none of the participants gave straight-line answers.

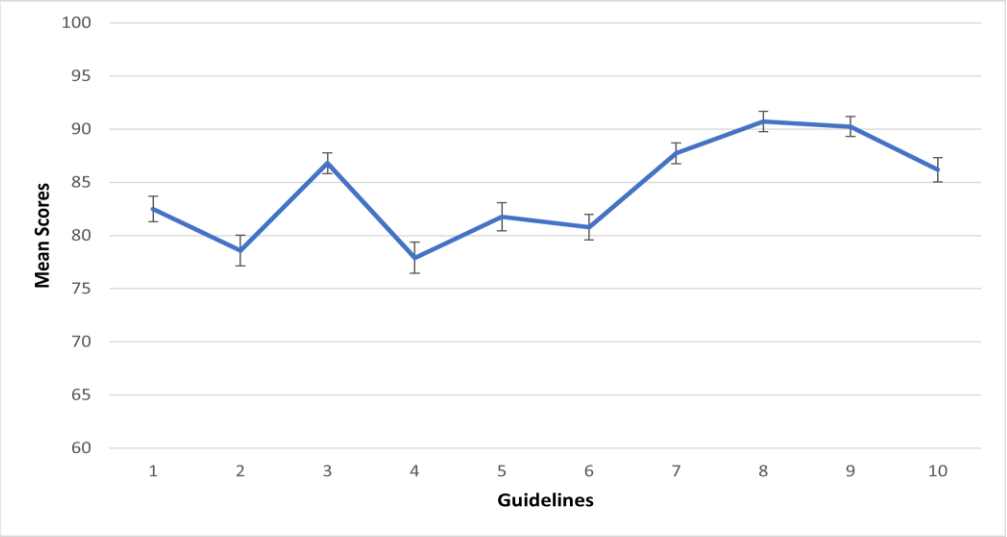

Figure 4. 3 Mean Scores of Participant and Standard Errors

As shown in Figure 4.3, a line chart representing the mean scores for individual guideline with standard errors as error bars. The figure indicates a very low value of the standard error and has a high value of mean for all guidelines. Higher mean values of the guidelines are showing that participants considered the majority of the image descriptions as accessible with respect to the NCAM guidelines. Whereas guideline second and guideline fourth had mean value lower than eighty. The guideline second and fourth is about the inclusion of any color name in the image description and lack of readability in the image descriptions, respectively. So, participants might observe colors and readability easily in the image descriptions.

4.2.6 Pilot Study before Actual Experiment

First of all, the study tried to find out the different accessibility issues related to web experiment tool. The input fields were checked with different window browsers including the mobile browsers. Color contrast, font size, headings and layout of the questionnaire were checked.

Pilot study was conducted with five participants to address the bias on the actual data. One questionnaire was given to the five participants twice on a different time. Test-retest process was conducted to check the reliability of the questionnaire. Two responses from a single participant were collected for each questionnaire.

During the pilot study, several potential biases were identified. Participants demanded for more instructions before answering the questionnaire. Few participants were not satisfied with the color combination used in the questionnaire and some participants had problems with the font size of the questions.

This study resolved the issues that came out from the pilot study and made the web questionnaire ready for the actual experiment. Also, color and font sizes were corrected, and the questionnaire were made more user friendly to use regardless of any device and web browsers.

4.2.7 Reliability and Validity

Another challenging part of an experimental study is to make the questionnaire reliable and valid for the research. Two important qualities for an acceptable questionnaire are reliability and validity.(Oden, n.d.). There are some errors that may appear during the experiment. Thus, pilot study was conducted for the experiment to point out the systematic errors. Five major sources of systematic errors are; measurement instrument, experimental procedure, behavior, participants and environment (Lazar, Feng, & Hochheiser, 2017).

There might be chances of bias caused by measurement instruments if the researcher does not study the functioning of instruments. Firstly, the study ensured if the provided image descriptions were written with correct spellings and font size so the participants could read them easily. Furthermore, the study confirmed that image descriptions given in the web-based questionnaire were capable of showing the same information regardless of the participants' devices. Questionnaire was provided to the participants with clear and understandable instructions to avoid the errors caused by experimental procedure. Participants were able to read the instructions before answering the questionnaire and the chances of errors were less.

However, bias may also be caused by participants. To overcome this sort of error, participants were allowed to answer the questionnaire any time without any intervention by the researcher. In addition to this, participants were motivated to take part in this experiment when they were alone or having free time.

4.3 Comparative Analysis of Quantitative and Qualitative Results

As the objective of this research is to have the framework whose performance is comparable to manual expert evaluation, this section compares the quantitative results from the framework and the qualitative results from the expert evaluation.

4.3.1 Correlation Analysis

Correlation analysis can be used to find out how the values of one variable relate to the values of another variable (Puth, Neuhäuser, & Ruxton, 2015). While looking at the histogram and outliers of the data, non-normally distributions were identified in the data. The collected data violated the Pearson assumptions, so spearman rank-order correlation was selected after examining its assumptions to find the strength of association between framework scores and participants from each guideline. So, spearman correlation was used to find the association between framework results and expert results. Spearman’s rank-order correlation is a non-parametric technique which is also known for estimating the strength of the relationship between two variables whenever the collected data has some outliers and data do not satisfy the usual assumptions of normality (A & Nwankwo, 2014). In spearman assumptions, a relationship should either monotonic or non-monotonic. Unlike pearson, spearman coefficient correlation is based on ranks rather than the actual scores and all observations are assigned with ranks separately. The significance level of spearman correlation is checked with a value named as ‘p’ whereas, p<0.05 or p<.01 both might be considered valid to prove a correlation statistically significant.

The spearman rank-order correlation was calculated to find out the relation between the participants scores and the framework scores. Spearman correlation pointed out the type of relationship between two different observed values in this research. Furthermore, the correlation coefficient (r) values vary between -1 to +1, whereas -1 indicate the strong negative relationship and +1 represent a strong positive relationship. Higher the positive values of correlation coefficient, stronger will be the positive relationship and higher the negative correlation value, stronger will be the negative relationship between the framework and participant scores.

In this research, a null hypothesis for a spearman correlation is:

H_0 : There is no monotonic or non-monotonic relationship between framework scores and participant scores.

H_1 : There is a relationship between framework scores and participant scores.

Table 4. 2 Quantitative Results of Correlation Between Evaluation Framework and Participant Scores

Guidelines	1	2	3	4	5	6	7	8	9	10
Correlation Coefficient (r_s)	0.640	0.766	0.739	0.733	0.781	0.796	0.762	0.581	0.591	0.611

Note: $p < .001$, two-tailed. *N =350.*

Scores from framework and participants were correlated using the Spearman's Correlation method and Table 4.2 is showing the correlation coefficient between both the scores for each guideline. Before calculating the correlation, all the assumptions were tested, and it was found that all guidelines are statistically significant(p<.001). The correlation coefficient (r_s) were used to describe the strength of a relationship between the participant scores and framework scores. The correlation coefficient (r_s) values of guideline first, eighth, ninth, and tenth has a moderate positive relationship between the scores from framework and participants. Whereas the rest of the guidelines has higher correlation coefficient (r_s) values that signified the strong positive monotonic relationship. It is cleared from the correlation values given in Table 4.2; positive correlation coefficient indicates when participants' scores increases, framework scores increase also increases, thus there is a monotonic relationship between framework scores and participant scores for all the guidelines. In addition to this, results clear that all the guidelines have positive association between the framework scores and the expert scores. The results in Table 4.2 suggested to reject the null hypothesis. So, the null hypothesis was accepted that revealed, there is a relationship between framework scores and participants' scores.

4.3.2 Comparison of Means

Quantitative results were obtained from framework and participants respectively and the mean of these scores were calculated for each guideline. Mean comparison helps to understand the variation of scores of each guideline in both the cases of framework and participants. Standard error is used to indicate the uncertainty around mean estimation (Altman & Bland, 2005). Also, standard error was calculated to identify the overlapping of two different means for the same guideline.

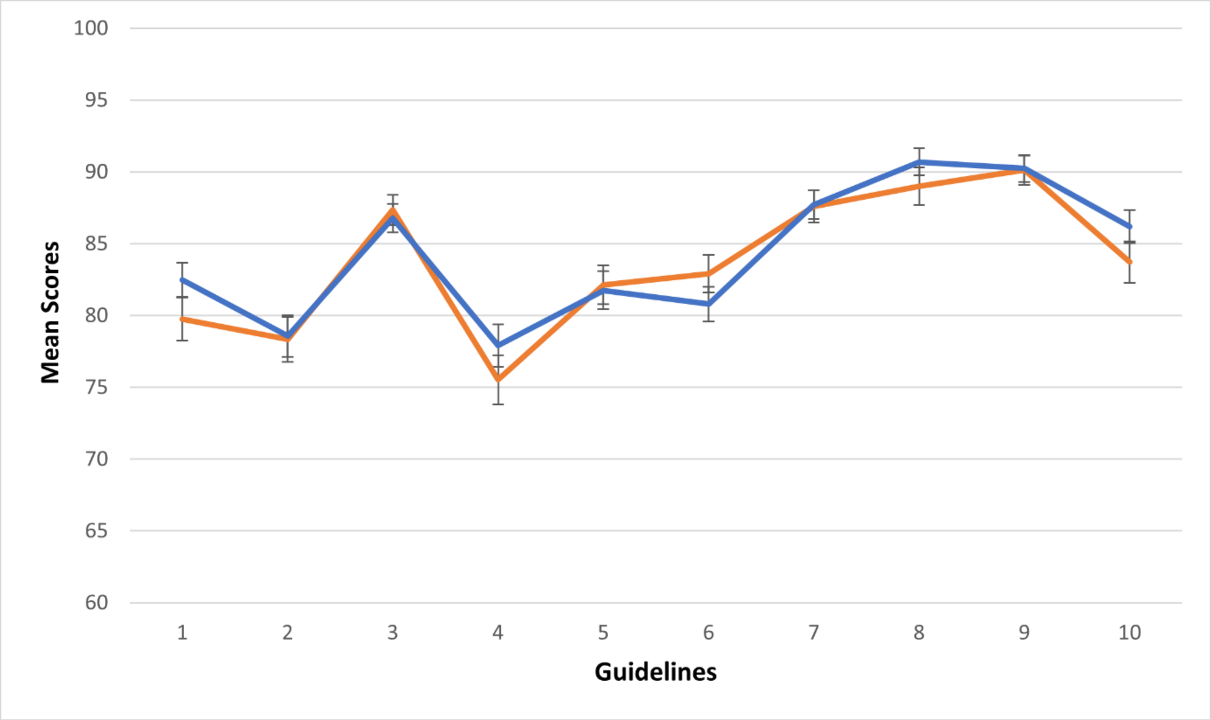

Figure 4. 4 Guideline wise mean scores from the quantitative and qualitative evaluation

Figure 4.4 shows means for each guideline and standard error on the error bars is showing the overlapping of means. When standard error bars on the error bars of two groups overlap, and the sample size is same, then difference between the two means can be considered non-significant (GraphPad, n.d.). The mean value in orange belongs to participant scores whereas the mean value in blue belongs to the framework scores. The x-axis representing the number of guidelines from one to ten and y-axis representing the mean scores. The mean scores of both the cases of framework and participants scores in guideline one and ten lie in between 80 to 85 with a higher value of standard errors than the other guidelines. Also, Guideline four has a lowest mean score for framework and participants scores. While means of framework scores almost overlapped with the means of participants scores in guidelines two, three, five, seven and nine. When the standard error bars overlap then it indicates that the two means are not statistically significant (Motulsky, 2015). In the case of Figure 4.4, standard error bars for each guideline are overlapping with each other. Thus, results showed that the difference between quantitative and qualitative is not statistically significant.

5.Discussion, Conclusion and Future Perspective

This chapter elaborates the discussion part of this study and ethical issues during this research. Section 5.1 explains the discussion of this study and section 5.2 explain the conclusion of this study. Section 5.3 explain what can be done in future from this research study and section 5.2 explains ethical challenges faces in this research study from designing to reporting.

5.1 Discussion

The findings of this study are representing that the framework is capable of automatically predicting the accuracy based on NCAM guidelines. Two different results were represented in this research; the first was checking the accuracy of the model while comparing the proposed framework scores with the ground truth (dataset scores) scores, the second was finding the relationship between expert scores and framework scores.

The first results were obtained after the training and testing of the machine learning model. The first results showed the value of mean, standard error, standard deviation, and accuracy of each guideline. During the statistical calculation, this study calculated the accuracy of nearly 20% of the image description dataset and resulted in Table 4.1. Standard error and standard deviation had lower values and accuracy of each guideline were above ninety-five. High accuracy in prediction of each guideline suggested that the proposed framework is efficient to predict an image description with respect to NCAM guidelines. However, Flickr8K dataset was used efficiently to train the model but Flickr30K could have been used to build the framework on a large set of images. Results obtained from statistical analyses showed the role of image description guidelines to improve the image accessibility. Thus, results do support the claim that it is possible to build an automatic evaluation framework based on the image accessibility guidelines.

In the case of qualitative evaluation, all guidelines have a good relationship between participant scores and framework scores, and some guidelines had moderate to strong relationship to reject the null hypothesis. In addition to this, each participant gave scores according to their knowledge and experience in image accessibility, but framework results were still adequate enough to match with the overall results of participants.

The outcome of this research supports the work conducted by (Dahal & Shrestha, 2019) in which participants were given images with or without cues to write the image descriptions. However, it may require a large amount of time to manually evaluate the image descriptions and specifically when there is multiple number of image descriptions, but the manual evaluation can be overcome with the help of an automated evaluation method. Furthermore, Vázquez and Lehmann (2015) presented an automated evaluation tool 'acrolinx', designed to verify the correctness of alternative text of images. But the 'acrolinx' was not based on any accessibility guidelines. The inclusion of image accessibility guideline plays an important role when there is a need of evaluating different types of image descriptions, such as image descriptions for natural images, map, chart, or a graph.

Furthermore, Bigham (2007) described a classifier to measure the quality of the alternative text with content similarity, word similarity and automatic labeling that performed nearly 86% accurate after training with few labeled examples. However, the study concluded that checking the quality of the alternative text is possible but did not mention whether the model followed any image accessibility guidelines or not.

Participants having different background and countries with proper knowledge of image accessibility made the results more generalizable. However, this experiment was web-based and there was a high chance for the multiple responses for the same questionnaire, but one questionnaire was given to one participant only to avoid this sort of problem. Since the questionnaire consisted of ten image descriptions, participants were encouraged to choose personally comfortable participation time. A total of 350 image descriptions were distributed among different participants, where each questionnaire had ten image description. So, there were 35 questionnaires for 35 participants to reduce the multiple responses from the same participant.

This study has points that may be considered as weak aspects of this research. It can also be argued that there might be more than one accessibility expert for the evaluation of image description dataset. The number of participants were limited to 35 to in this experimental study since there was a need for selecting the participants only with proper knowledge of image accessibility and image accessibility guidelines. Also, this study is limited to predict the accuracy of image descriptions without any consideration of visual information of images.

5.2 Conclusion

A machine learning-based automatic evaluation framework has been developed using a random forest model. A single random forest was used for each guideline that reduced the complexity in predicting the image descriptions. Five-folds cross-validation was used to train the model and thus random forest performed efficiently to predict the accuracy of the image description with respect to the NCAM image description accessibility guidelines. The results from the experiments show high performance from the framework. Qualitative evaluation of the framework from the experts validates the performance and effectiveness of the framework being comparable to manual evaluation by accessibility experts.

5.3 Future Work

As the proposed model is based on the ensemble method of machine learning. Deep learning can also be used to build the prediction model. Keras is a neural network API that runs on TensorFlow whereas TensorFlow, is an open-source machine learning library (Mwiti, 2019). The environment setup needs a high power of CPU and GPU to build and run the deep learning model with a huge dataset with the inclusion of images with image descriptions. In addition to this, deep learning will also help to avoid the manual selection of features to build a model. User evaluation can be carried out in the future which helps the study to find out the robustness of the framework in the real world. However, images with image descriptions could be used to train the model and thus model could be enhanced image accessibility.

5.4 Ethical issues

This research is aimed to improve the image accessibility of image description and ethical issues were considered at every stage of this research study. Proper references and citations have been given in the report. A consent form was prepared that distributed with the questionnaire. The questionnaire for the participants only included the relevant information related to this research and no personal and private data was involved. Comfortability of participants was the first priority for the researcher. All the data related to this research were stored in the researcher’s laptop and was backed up in a password protected drive.

There is no risk involved in this study that may cause stress on the participants. An informed consent formed briefly explain the role of participants and their confirmation of participation in the study. Furthermore, pilot testing was conducted to make sure the appropriate execution of this experiment. It helped the researcher to explore any uncomfortable and irrelevance questions for the participants.

The main objective of the research during this experiment was to allow the participants to give their feedback without giving any personal information. Thus, an anonymous web-based questionnaire was used so that, no information would be collected except the slider scale scores for the image descriptions. However, a web-based anonymous questionnaire was used in this experiment, but participants were informed that the study is independent and have no connection with any software company.

References

A, O., & Nwankwo, C. (2014). Ties Adjusted Rank Correlation Coefficient. *IOSR Journal of Mathematics, 10*, 09-17. doi:10.9790/5728-10530917

Abuaddous, H. Y., Jali, M. Z., & Basir, N. (2016). Web accessibility challenges. *International Journal of Advanced Computer Science and Applications, 7*(10), 172-181.

Alahmadi, T., & Drew, S. (2018). *Evaluation of image accessibility for visually impaired users* (Vol. 8).

Altman, D. G., & Bland, J. M. (2005). Standard deviations and standard errors. *Bmj, 331*(7521), 903.

Bigham, J. P. (2007). *Increasing web accessibility by automatically judging alternative text quality.* Paper presented at the Proceedings of the 12th international conference on Intelligent user interfaces.

Byrne, J., & Humble, A. (2006). An Introduction to Mixed Method Research.

Cooper Hewitt. (2019). COOPER HEWITT GUIDELINES FOR IMAGE DESCRIPTION. Retrieved from https://www.cooperhewitt.org/cooper-hewitt-guidelines-for-image-description/

Creswell, J. W. (2012). *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*: Pearson.

Dahal, D., & Shrestha, R. (2019). *Accessible Image Description Using Sample Example Cues*. Paper presented at the The Fourth International Conference on Universal Accessibility in the Internet of Things and Smart Environments, Athens,Greece. http://www.thinkmind.org/download_full.php?instance=SMART+ACCESSIBILITY+2019

DataRobot. (n.d.). Machine Learning Life Cycle. Retrieved from https://www.datarobot.com/wiki/machine-learning-life-cycle/

Diagram Center. (n.d.). Image Description Guidelines

Retrieved from http://diagramcenter.org/general-guidelines-final-draft.html#3

Drakos, G. (2018). How to select the Right Evaluation Metric for Machine Learning Models: Part 1 Regression Metrics. Retrieved from https://towardsdatascience.com/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0

FEDERICO, S. (2016). These are NPR's photo caption guidelines. Retrieved from https://training.npr.org/2016/01/12/these-are-nprs-photo-caption-guidelines/

Ferlitsch, A. (2019). Making the machine: the machine learning lifecycle. Retrieved from https://cloud.google.com/blog/products/ai-machine-learning/making-the-machine-the-machine-learning-lifecycle

Garg, R. (2018). A PRIMER TO ENSEMBLE LEARNING – BAGGING AND BOOSTING Retrieved from https://analyticsindiamag.com/primer-ensemble-learning-bagging-boosting/

GraphPad. (n.d.). Advice: Don't pay much attention to whether error bars overlap. Retrieved from https://www.graphpad.com/guides/prism/7/statistics/stat_relationship_between_significa.htm

Guanga, A. (2019). Understand Regression Performance Metrics. Retrieved from https://becominghuman.ai/understand-regression-performance-metrics-bdb0e7fcc1b3

Harper, S., & Chen, A. Q. (2012). Web accessibility guidelines. *World Wide Web, 15*(1), 61-88. doi:10.1007/s11280-011-0130-8

Heidenreich, H. (2018). What are the types of machine learning? Retrieved from https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f

Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research, 47*, 853-899.

Iriste, S., & Katane, I. (2018). Expertise as a research method in education. *Rural Environment. Education. Personality (REEP)*, 74-80.

Kadiyala, A., & Kumar, A. (2018). Applications of Python to Evaluate the Performance of Bagging Methods. *Environmental Progress & Sustainable Energy, n/a*(n/a). doi:10.1002/ep.13016

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research methods in human-computer interaction*: Morgan Kaufmann.

Lewthwaite, S. (2014). Web accessibility standards and disability: developing critical perspectives on accessibility. *Disability and Rehabilitation, 36*(16), 1375-1383. doi:10.3109/09638288.2014.938178

Lin, C.-Y. (2004). *Rouge: A package for automatic evaluation of summaries.* Paper presented at the Text summarization branches out.

Mathur, P. (2016). What is Anaconda and Why should I bother about it? Retrieved from https://medium.com/pankajmathur/what-is-anaconda-and-why-should-i-bother-about-it-4744915bf3e6

Motulsky, H. (2015). The link between error bars and statistical significance. *GraphPad Software,[Online]. Available: https://egret. psychol. cam. ac. uk/statistics/local_copies_of_sources_Cardinal_and_Ai tken_ANOVA/errorbars. htm.[Accessed 22 4 2016].*

Mwiti, D. (2019). How To Build a Deep Learning Model to Predict Employee Retention Using Keras and TensorFlow. Retrieved from https://www.digitalocean.com/community/tutorials/how-to-build-a-deep-learning-model-to-predict-employee-retention-using-keras-and-tensorflow

NWEA. (2017). NWEA Image Description

Guidelines for Assessments. Retrieved from https://www.nwea.org/content/uploads/2017/06/Image-Description-Guidelines-for-Assessments-2017.pdf

Oden, C. (n.d.). Validity and Reliability of Questionnaires: How to Check. Retrieved from https://www.projecttopics.org/validity-and-reliability-of-questionnaires-how-to-check.html

P. Bigham, J., S. Kaminsky, R., Ladner, R. E., Danielsson, O., & Hempton, G. (2006). *WebInSight: : making web images accessible* (Vol. 2006).

Paciello, M. (2000). *Web accessibility for people with disabilities*: CRC Press.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: a method for automatic evaluation of machine translation.* Paper presented at the Proceedings of the 40th annual meeting on association for computational linguistics.

Pedregosa, F., Ga, #235, Varoquaux, l., Gramfort, A., Michel, V., . . . Duchesnay, d. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res., 12*, 2825-2830.

Pennington, J., Socher, R., & Manning, C. (2014). *Glove: Global vectors for word representation.* Paper presented at the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).

Petrie, H., Harrison, C., & Dev, S. (2005). Describing images on the web: a survey of current practice and prospects for the future. *Proceedings of Human Computer Interaction International (HCII), 71*.

Poet training tool. (n.d.). How to describe images. Retrieved from https://poet.diagramcenter.org/how.html#general-guidelines

Puth, M.-T., Neuhäuser, M., & Ruxton, G. D. (2015). Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits. *Animal Behaviour, 102*, 77-84.

Rachiele, G. (2018). Tokenization and Parts of Speech(POS) Tagging in Python’s NLTK library. Retrieved from https://medium.com/@gianpaul.r/tokenization-and-parts-of-speech-pos-tagging-in-pythons-nltk-library-2d30f70af13b

Rhodes, J. (2014). On Methods: What’s the difference between qualitative and quantitative approaches? Retrieved from https://www.evidencebasedmentoring.org/on-methods-whats-the-difference-between-qualitative-and-quantitative-approaches/

Roster, C. A., Lucianetti, L., & Albaum, G. (2015). Exploring slider vs. categorical response formats in web-based surveys. *Journal of Research Practice, 11*(1), D1-D1.

Sarkar, D., Bali, R., & Ghosh, T. (2018). *Hands-On Transfer Learning with Python: Implement advanced deep learning and neural network models using TensorFlow and Keras*: Packt Publishing Ltd.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia, 126*(5), 1763-1768.

Seif, G. (2018). An Introduction to Scikit Learn: The Gold Standard of Python Machine Learning. Retrieved from https://towardsdatascience.com/an-introduction-to-scikit-learn-the-gold-standard-of-python-machine-learning-e2b9238a98ab

Shinde, R. (2019). Image Captioning With Flickr8k Dataset & BLEU. Retrieved from https://medium.com/@raman.shinde15/image-captioning-with-flickr8k-dataset-bleu-4bcba0b52926

Sieg, A. (2018). Text Similarities : Estimate the degree of similarity between two texts. Retrieved from https://medium.com/@adriensieg/text-similarities-da019229c894

Silke, J., & Roman, H. (2018). *On the overestimation of random forest's out-of-bag error*.

Subedi, D. (2016). Explanatory Sequential Mixed Method Design as the Third Research Community of Knowledge Claim. *American Journal of Educational Research, 4*, 570-577. doi:10.12691/education-4-7-10

Tatman, R. (2019). Evaluating Text Output in NLP: BLEU at your own risk. Retrieved from https://towardsdatascience.com/evaluating-text-output-in-nlp-bleu-at-your-own-risk-e8609665a213

Trewin, S., Cragun, B., Swart, C., Brezin, J., & Richards, J. (2010). *Accessibility challenges and tool features: an IBM Web developer perspective.* Paper presented at the Proceedings of the 2010 international cross disciplinary conference on web accessibility (W4A).

Vázquez, S. R., & Lehmann, S. (2015). *Acrolinx: a controlled-language checker turned into an accessibility evaluation tool for image text alternatives.* Paper presented at the Proceedings of the 12th Web for All Conference.

Veroniiiica. (2018). How to Write Alt Text and Image Descriptions for the visually impaired. Retrieved from https://www.perkinselearning.org/technology/blog/how-write-alt-text-and-image-descriptions-visually-impaired

Voutilainen, A., Pitkäaho, T., Kvist, T., & Vehviläinen-Julkunen, K. (2016). How to ask about patient satisfaction? The visual analogue scale is less vulnerable to confounding factors and ceiling effect than a symmetric Likert scale. *Journal of advanced nursing, 72*(4), 946-957.

W3C. (2014). Images Concepts. Retrieved from https://www.w3.org/WAI/tutorials/images/

W3C. (2018). Web Content Accessibility Guidelines (WCAG) 2.1. Retrieved from https://www.w3.org/TR/WCAG21/#text-alternatives

Webcolors. (n.d.). Module contents. Retrieved from https://webcolors.readthedocs.io/en/1.5/contents.html

WGBH. (n.d.). Evaluations. Retrieved from https://www.wgbh.org/foundation/what-we-do/ncam/evaluations

What is Machine Learning. (2017). Retrieved from https://www.expertsystem.com/machine-learning-definition/

Yu, H. (2002). Web accessibility and the law: recommendations for implementation. *Library Hi Tech*, *20*(4), 406-419. doi:10.1108/07378830210452613

Appendices

Appendix A is representing the information related to the researcher and his supervisor. It also includes the informed consent form which distributed to each participant before performing the experiment study. Appendix B constitutes a sample of the questionnaire for one image description. Whereas ten of these types of questionnaires were given to each participant with different image descriptions.

Appendix A: Informed Consent

TITLE OF STUDY

Image Description Evaluation Framework based on Image Accessibility Guidelines.

PRINCIPAL INVESTIGATOR

Himmat Kumar Dogra
Department of Information Technology at Oslo Metropolitan University
+47 90017067
S329917@oslomet.no

Collective Data for the Evaluation of Framework

Informed Consent for participation in the research questionnaire.

Thank you for your interest to participate in the study about "An evaluation framework for image description based on an accessibility guideline". This study is conducted by Himmat Kumar Dogra in Oslo, Norway. The purpose of this study is to understand the knowledge and practice of making image description accessible. This web-based questionnaire process is a part of the Master thesis for program Universal Design of ICT with Oslo Metropolitan University. The duration of the questionnaire is 30 minutes involving 10 questions. This will be an anonymous study, in which answers will be obtained in the form of scores and responses will be noted by the researcher. Participants have the right not to answer the question. If a participant feels uncomfortable in answering the questions, the participant has the right to withdraw. This research study will be used to analyze in the evaluation of the framework for the image descriptions in the master thesis. The researcher has been given the explicit guarantee that the researcher will not identify the name, location and other personal information of participants and collected data will remain secure. The researcher will make notes from the questionnaire that will be used only for analysis, from which you would not be personally identified.

○ Select this option to participate voluntarily in this research study.

Appendix B: Questionnaire

Instructions

1. This form has ten different slider questions that have a minimum value of 0 and a maximum value of 100.
2. A value 0 corresponds to Strongly Disagree and value 100 corresponds to Strongly Agree.
3. The participants can give answers according to their knowledge.
4. This Form includes 10 different image descriptions.
5. A participant is given 10 different questions to answer for each image description.
6. These 10 questions are based on the following 10 NCAM image description guidelines: -

   a. The description should be succinct.
   b. Colors should not be specified unless it is significant.
   c. The new concept or terms should not be introduced.
   d. The description should be started with a high-level context and drilled down to details to enhance understanding.
   e. The active verbs in the present tense should be used.
   f. Spelling, grammar, and punctuation should be correct.
   g. Symbols should be written out properly.
   h. The description vocabulary should be added which adds meaning, for example, "map" instead of an image.
   I. Physical appearance and actions should be explained rather than emotions and possible intentions.
   j. The material should not be interpreted or analyzed, instead the readers should
   be allowed to form their own opinions.

7. Participants can evaluate and answer the image descriptions with respect to the given questions.
8. All questions are mandatory.
9. Participants have no limit of time to answer these questions.
10. A Submit button can be used to submit all answers at the end of this form

Image Description : Three angry dogs who are brown, white, and black perhaps engaged with one another in a dirt field.

1. There is no repetitive and unnecessary word in the description.

0 100

51

2. There is no colour name or complex image with the colour name in the description.

0 100

44

3. There is no new concept or terms introduced in the description.

0 100

62

4. The description is easy to read and it does not have difficult words.

0 100

28

5.The description used only present tense.

0 100

39

6. The description used Spelling, grammar, and punctuation correctly.

0 100

39

7. Symbols are written out properly to ensure proper pronunciation by screed reader.

0 100

60

8. If the description is for a map, chart or a graph, the description then uses proper vocabulary.

0 100

99

9. The description has no words related to emotions and possible intentions.

0 100

49

10. There is no analysis, interpretation or Opinions include in the description.

0 100

43