Chapter One: INTRODUCTION

World Wide Web is a source of enormous information and such information is increasing exponentially. The basic purpose ofWeb is to provide up-to-date and relevant information to all users. Success of any organization like E-commerce highly depends on the quality of website. Websites are developed by various organizations from very small to large organization with development teams; but small companies develop websites without sufficient resources and professional skills. Thus it is very important to evaluate the quality of website and web development process to improve the quality of websites.

Quality of a website can be viewed in terms of internal and external quality. Internal quality refers to cost effectiveness, maintainability and portability whereas external quality measures from the user's standpoint(Signore, 2005). Despite of various detailed design guidelines and design recommendation provided by various authors it is very difficult to implement them(Nielsen, 1999)(Nielsen, 2000).Web page metrics plays an important role to measure the quality of a website as they can measure various attributes of a web page quantitatively that influence the quality of a web page.

A large number of web metrics have been proposed by different authors that contribute to the goodness or quality of a website. Web metrics covers almost all aspects like page composition, amount of information, presentation, content and size of the websites.

1.1 Motivation of work

Although various guidelines were provided by different authors to design a quality website but these guidelines are not properly defined for the implementation point of view. Thus various developers feel difficulty to design a quality websites by following these guidelines(Nielsen, 1999)(Nielsen, 2000)(Shedroff, 1999)(Friedman, 2008)

Almost for every organization's success; quality of a website is very important regardless of organizations goal whether commerce (Amazon) or content presentation

(Google). But many smaller websites are designed with lack of resources and professional skills, leading to the poor quality of website. Thus it is very important question that how to improve the design of websites. In our research we explore the following issues:

- Relation between web page metrics and quality of websites
- Accuracy and precision of web page metrics to predict the quality of the websites
- Compare performance of different machine learning and logistic regression techniques to predict the quality of website.

1.2 Aim of Work

Aim of our research is to find the relationship between web page metrics and quality and websites and to compare the performance of various machine learning techniques and logistic regression technique to identify the best model to predict the quality of a website.

In this research work WEB METRICS CALCULATOR was developed in ASP.NET language which is used to compute 20 web page metrics like word count, link count, script count etc. Web pages have been collected from various categories of pixel awards to evaluate the quality of a website.

These metrics make a subset of metrics which are related to the quality of web page design. Then different machine learning techniques and logistic regression technique were applied to compare the performance to predict a website into good or bad.

1.3 Organization of Thesis

Remainder part of thesis is organized as follows:

- Chapter 2: Related work

This chapter briefly describes the related work that has been done for evaluating the website.

- Chapter 3: Literature review

This chapter describes the detailed literature about the web metrics and the importance of web metrics.

- Chapter 4: Research methodology

This chapter describes the WEB METRICS TOOLS used for computing web metrics, metrics selected for study and various machine learning techniques in detail which we used in our research.

- Chapter 5: Result Analysis

This chapter discuss the comparative study of results of applying different machine learning techniques on the dataset.

- Chapter 6: Conclusion and Future Work

This chapter discuss the conclusion drawn from the research and scope of the future work.

Chapter Two: Literature Survey

Over past 20 years more than 350 web metrics have been proposed by different authors to improve the quality of web sites and web development process. Bray made the earliest attempt to make global measurements about the web (Bray, May,1996).It basically included the general attributes of web such as page size, site visibility and format distribution.

Many metrics such as no. of hits, click- through rates etc. becomes very popular to quantify the use of web. Pitkow found the problem associated with hit metering as the reliable metric due to the proxy and client caches(Pitkow, 1997).So there is a need of new web metrics that provide the deeper view of the web as a whole and a single web page as a different perspective.

In 2002 on the basis of magnitude and measurement function Dhyani provided a classification of web metrics(Dhyani, Ng, & Bhowmik, 2002).

A lot of existing work has been done on evaluating web page quality, but most quantitativemethods for evaluating web sites focus on statistical analysis of usage patterns in server(Chi, Pirroli, & Pitkow, 2000)(Drott, 1998)(Fuller & Graff, 1996). Traffic-based analysis (e.g., pages-per-visitor or visitors-per-page) and time-based analysis (e.g., click paths and page-view durations) provide data that must be interpreted in order to identify usability problems. The analysis based on such data is quite uncertain since web server logs provide incomplete traces of user behavior, and because timing estimates may be skewed by network latencies.

The above work focuses more on navigation history; explicitly clicked links and thetime spend on a web site. Server logs are problematic because they only track uniquenavigational events (e.g., do not capture use of back button) and thus are hard to understandbecause of caching. Another method for evaluating web pages of user interest automaticallyinvestigates various factors in a user’s browsing behaviour such as number of scrolls, forminput, search text etc.

Another approach that assumed that website evaluation must be rapid and automatic. This approach uses two types of tools and techniques. First approach is

Usability awareness tool (WebSAT), this approach should used by the designers who does not aware of the usability issues and second approach was Web usability tools and techniques (NIST web metric tool) should used by designer to improve the usability of website(Scholtz, Laskowski, & Downey, 1998)

Other approaches were inspection-based that rely on assessing static HTMLaccording to a number of pre-determined guidelines, such as whether all graphics containALT attributes that can be read by screen readers(Velayathan & Yamada, 2006). For example, WebSAT (Web StaticAnalyzer Tool) is used to check the accessibility issues (i.e., support for users withdisabilities), forms use, download speed, maintainability, navigation and readability of Webpages. There are many other techniques that compare quantitative web page attributes – suchas the number of links or graphics – to thresholds(Thimbleby, 1997). However, there are no clearthresholds established for a wider class of quantitative Web page measures.

Simulation has also been used for web site quality evaluation. For example, asimulation approach has been developed for generating navigation paths for a site based oncontent similarity among pages, server log data, and linking structure(Chi, Pirroli, & Pitkow, 2000). The simulationmodels hypothetical users who are traversing the site from described start pages, making useof information “scent” (i.e., common keywords between the user’s goal and linked pagescontent) to make decisions related to navigation. The approach does not consider the impactof various web page attributes, such as the amount of text or layout of links.

Web site effectiveness is also measured in terms of information and service quality. This study uses two instrument WEBQUAL and SERVQUAL instrument. These two instruments are combined in order to capture the interactivity and service retrieval of web(Fink, 2001).

The most closely related work is done in Ivory et.al(Ivory, Sinha, & Hearst, Preliminary findings on quantitative measures for distinguishing highly rated information-centric web pages, 2000)(Ivory, Sinha, & Hearst, 2001) which providespreliminary analysis of collection of web pages and captures various web metrics associatedwith the rated websites, and predicts how the pair-wise correlations are

manifested in thelayout of the rated and unrated sites pages. This work does not apply various machinelearning algorithms to predict the best suited model that can provide high accuracy.

CBR(Case Based Reasoning) and SWR(Step Wise Regression) techniques are used by to proposed size measures and effort predictors for web cost estimation. These methodologies basically tried to estimate the design and authoring effort for Web.(Mendes, Mosley, & Counsell, 2003).

The approach presented by G. Velayathan and S. Yamada(Velayathan & Yamada, 2006) analyzes the user logsmetrics such as number of scrolls, form input, search text etc. and extracts effective rules toevaluate web pages using a machine-learning method known as decision tree. A client sidelogging/analyzing tool GINIS is used to automatically evaluate web pages using theselearned rules. Similarly, M. Zorman et.al(Zorman, Podgorelec, Kokol, & Babic, 1999) has proposed an algorithm to find the good orrelevant websites for keywords provided by the user. They developed an intelligent searchtool which employs TFIDF heuristics for finding term frequency and decision tree machinelearning algorithm for automatically evaluation of the websites.

Another approach was based on applying Ranking SVM(Li & Yamada, Automated Web Site Evaluation – An Approach Based on Ranking SVM, 2009)(Li & Yamada, 2010)which is used toextract evaluation criteria from evaluation data for automated web site evaluation. It choosesthe evaluation criteria which are the discriminant functions learned from a set of rankinginformation and evaluation features such as freshness, accuracy of spelling and grammar, toppage's global link popularity collected automatically by web robots. However, it does notconsider the other algorithms for the website evaluation.

The quality of a website can be defined in terms of functional as well as non-functional properties. K. M. Khan(Khan, 2008) has derived the non-functional attributes such asreliability, usability, efficiency, security and assessed them. The work done in(Khan, 2008) adopts aGoal-Question-Metric (GQM) approach to derive quality metrics. It defines the goals that areneeded to be measured, then it develops the questions derived from goals that are required todetermine if the goals are fulfilled, and finally, their

measurements are the answers of thequestions which are known as metrics. For instance, questions related to the goal failure ratecould be: what is the percentage of incorrect links on the page?

2.1 Importance of Web Metrics

As more and more websites are created day by day, complexity and competition also increases. To check whether we created a good website or not we use web metrics. Web metrics help us to evaluate a web site. Web metrics vary based on nature and purpose of web site.

1. Meta keyword metrics helps us to find what keywords user enters in search engine to locate a particular website. By analyzing the Meta keyword metric we can see on which keywords over website comes in top 10 output of search engine.

2. Out link and in link metrics of a web site helps us to find path where we can enter a website and leave a web site. It also helps to find out any cycle is created or not. If there are more in link then website will have good hit rate.

3. Bound rate metrics helps to find percentage of initial users who bounce away to different website rather than continue to your website. Low bounce rate is good for a website as people are staying more on your website. Identify web pages for a websites which has high bounce rate so that we can modify these pages in order to decreases the bounce rate. For eg in e-commerce sites the major benefit of web analytics may be to find out the average amount of time taken to close an online sale.

4. If your web statistics for example, reveal that 60% of the individuals who watch a demo video also make a purchase, then you'll want to strategize to increase viewership of that video.

5. There are metrics which can show you the percentage of clicks each item on your webpage received. This includes clickable-photos, text links in your copy, downloads and of course, any navigation you may have on the page. Are they clicking the most important items?

6. If you utilize advertising options other than web-based campaigns, your web analytics program can capture performance data if you'll include a mechanism for sending them to your website. Typically, this is a dedicated URL that you include in your advertisement (i.e."www.example.com/offer50") that delivers those visitors to a specific landing page. You now have data on how many responded to that ad by visiting your website.

7. If you are running a banner ad campaign, search engine advertising campaign or even email campaigns, you can measure individual campaign effectiveness by simply using s dedicated URL similar to the offline campaign strategy.

8. Analytics permits you to see where your traffic geographically originates including country, state and city. This can be especially useful if you use geo-targeted campaigns or want to measure your visibility across a region.

9. If you're working to increase visibility, you'll want to study the trends in your New Visitors data. Analytics identifies all visitors as either new or returning.

10. Web traffic generally has peaks at the beginning of the work day, during lunch and toward the end of the work day. It's not unusual however to find strong web traffic entering your website up until the late evening. You can analyze this data to determine when people browse versus buy and also make decisions on what hours you should offer customer service.

Chapter Three: Research Background

This chapter focus on the effect of various web page measures on the quality or goodness of website. Thus it is very important to select web page metrics as independent variable to analyze the website.

3.1 Web Page Metrics

Web page metrics gives the quantitative measurement of different attributes of a website like page size, word count etc. A list of web interface measuresprovided by Ivory(Melody, 2001)based on site architecture, page performance, page formatting, text formatting, link formatting, graphic formatting, text element, link element, graphic element to analyze the quality of a web page by calculating different web page metrics. These web measures can be divided on the basis of efficiency, functionality, maintainability, portability, reliability, and usability quality characteristics.

3.1.1 Efficiency web metrics (Signore, 2005), (Mich, Franch, & Gaio, 2003)

Efficiency metrics as shown in Table 1 are related to size of a web page and the load time of a website/webpage.

Table 1: Efficiency Web metrics list

Metric	Meaning
efficiency_css_size	Css size per page
efficiency_homepage_load_time	Homepage load time
efficiency_image_size	Image size
efficiency_javascript	Script size per page
efficiency_page_load_time	page load time
efficiency_page_size	Page size

3.1.2 Functionality web metrics

Functionality metrics as shown in Table 2 include navigation, forms, identity and other aspects related to the functionality offered by the site.

Table 2: Functionality Web metrics list

Metric	Meaning
forms_form_info_request(Olsina & Rossi, 2002),(Group, 2005)	presence of contacts/info form
forms_labels(Americo, 2010)	number of label tags
Identity_auther(Mich, Franch, & Gaio, 2003)	Average presence of author
Identity_logo (Mich, Franch, & Gaio, 2003)	presence of site name in title
Identity_sitename_title(Group, 2005)	Presence of navigation bar
Navigation_bar(Signore, 2005)	Presence of navigation bar
Navigation_bread_crums (Signore, 2005)	Presence of bread_crums(path metric)
Navigation_quality_of_links (Mich, Franch, & Gaio, 2003)	Presence of page title in links

3.1.3 Maintainability web metrics

Maintainability metrics as shown in Table 3 include aspects related to the number of items to maintain (e.g. scripts, styles used, and tables);

Table 3: Maintainability Web metrics list

Metric	Meaning
Maintenance_num_script(Americo, 2010)	Script files no per page
Maintenance_num_styles(Americo, 2010)	Css file number per page
Maintenance_num_tables(Americo, 2010)	Tables number per page

3.1.4 Portability web metrics

Portability metricsas shown in Table 4 include aspects related to page layout, use of html standards, etc.

Table 4: Portability Web metrics list

Metric	Meaning
Page_layout_device_specific (Signore, 2005)	Presence of specific css to device
Page_layout_html_standards(Americo, 2010)	Use of html notations in formatting
Pagelayout_num_divs(Americo, 2010)	Number of divs
Page_layout_num_frames(Calero, Ruiz, & Piattini, 2005)	Number of frames
Pagelayout_num-tables(Americo, 2010)	Number of tables
Pagelayout_num_table_inside_tables (Calero, Ruiz, & Piattini, 2005)	Presence of table inside table

3.1.5 Reliability web metrics

Reliability metrics as shown in Table 5 include aspects related to the validation and links status

Table 5: Reliability Web metrics list

Metric	Meaning
Links_avg_num_words(Calero, Ruiz, & Piattini, 2005)	Average num of words in links
Links_links_titles(Calero, Ruiz, & Piattini, 2005)	Links with title attributes
Links_num_broken_links(Olsina & Rossi, 2002)(Signore, 2005)	Number of broken links
Link_num_extern_links (Signore, 2005)	Number of broken link to another site
Link_num_image_link (Calero, Ruiz, & Piattini, 2005)	Number of link with images
Link_num_intern_broken_link (Signore, 2005)	Number of broken link I the same site
Link_num_intern_links (Signore, 2005)	Number of intern links
Links_num_links(Olsina & Rossi, 2002)(Signore, 2005)	Number of links
Links_num_non_implemented_links(Olsina & Rossi, 2002)	Num of non implemented links
Link_page_withot_link(Olsina & Rossi, 2002)(Signore, 2005)	Pages without links in the site
Links_num_non_implemented_links (Calero, Ruiz, & Piattini, 2005)	Number of non implemented links
Validation errors (Signore, 2005)	Html warning par page

3.1.6 Usability web metrics

Usability web metrics as shown in Table 6 include aspects related to accessibility, multimedia and textual contents.

Table 6: Usability Web metrics list

Metric	Meaning
Accessibility_img_alt (Signore, 2005)	presence of alt attribute in images
accessibility_img_title(Calero, Ruiz, & Piattini, 2005)	presence of title attribute in images
accessibility_validate_access (Signore, 2005),(Mich, Franch, & Gaio, 2003),(Pollilo, 2005)	accessibility issues per page
multimedia_num_img (Signore, 2005)	image number per page
text_font_size_average_em	average of font size in em (percentage) in css
text_font_size_average_px	average font size in css in pixels
text_font_size_max_em	maximum font size in em (percentage) in css
text_font_size_max_px	max font size in pixels
text_font_size_min_em	minimum fonts size in em (percentage) in css
text_font_size_min_px	min font size in pixels
text_heading_len (Signore, 2005)	average heading length
text_heading_reverse_order	number of headings in reverse order
text_italic_text	number of italic text bigger than 20 chars
text_num_diferent_colors	number of different text colors in css
text_num_diferent_fonts (Signore, 2005)	number of different text fonts in css
text_num_sentences_in_paragraph (Signore, 2005)	number of sentences per paragraph
text_num_subheading_heading(Signore, 2005)	number of sub headings per heading
text_num_syllables_in_word(Signore, 2005)	number of syllables per word
text_num_words_in_sentence(Signore, 2005)	number of words per sentence
text_num_words_meta_description(Americo, 2010)	number of words in metatag description

text_num_words_meta_keywords(Americo, 2010)	number of words in metatag keywords
text_paragraph_max_size(Signore, 2005)	maximum size of paragraph
text_paragraph_size(Signore, 2005)	paragraph size
text_subheading_len(Signore, 2005)	sun heading length
text_total_newlines(Signore, 2005)	total number of newlines
text_total_sentences(Signore, 2005)	total sentences
text_total_syllables(Signore, 2005)	total syllables
text_total_words(Signore, 2005)	total words
text_uppercase_text	number of uppercase sentences

3.2 Independent and Dependent variable

This dataset comprises of total 21 variables out of which 20 variables are independent and 1 is dependent variable. Table 7 gives the list of 20 web page measures that we have selected for our study. To compute these web page measures we have developed WEB METRICS CALCULATOR in ASP.NET language. We have used CFS(Correlation based Feature Selection) in WEKA tool to select the subset of independent variables that acts as the best predictors out of all other independent variables(Hall, 1999). This subset is searched through all possible combinations of variables. CFS provides with the good feature subset that are highly correlated with data set.

Table 7: List of Metrics for study

Metrics	Description
Word Count	Total number of words on a Web page
Link Count	Total number of links on a Web page
Graphic Word Count	Total number of words in ALT attribute
Page Size	Size of Web Page
Script Count	Total number of scripts on a Web page
Image Count	Total number of images on a Web page

Inline Element Count	Total number span element count on a Web page
Class Used count	Total number of class used on a Web page
Exclamation Count	Total number of !'s used on a Web page
Load Time	Time to load a Web page
Meta Tag Count	Total number of meta tag on a Web page
Page Title Word Count	Total number of words used for title
List Items	Total number of ordered list on a Web page
Meta Description Length	Total number of words used for meta description
Unordered List Count	Total number of unordered list on a Web page
Division Count	Total number of div tag used on a Web page
Number Of Headings	Total number of headings (H1,H2,H3,H4,H5,H6) used on a Web page
Paragraph Count	Total number of paragraphs on a Web page
Text Link Count	Total number of links that are text.
Image Link Count	Total number of links that are image.

Dependent variable is Category which takes two values either good or bad depending on judgment of pixel awards.

3.3 Empirical Data Collection

Web pages have been selected from pixel awards website. Pixel awards have been given to the websites which shows excellence in design and development and established by Erick and Lisa Laubach in year 2006(www.pixelawards.com). Judging criteria for websites are Innovation, Content, Navigation, Visual Design, and Functionality and Site Experience.

Websites are placed in 24 categories Agency, Animation, Apps, Art, Blogs, Commerce, Community, Experimental, Fashion, Food & Beverage, Games, Geek, Green, Magazines, Movies, Music, Non-Profit, Personal, Sports, Travel, TV and Weird. These websites are judged against judging criteria. There are two types of winners for each

category one of themis People's Champ and another is winner. Dependent variable takes the value good for both of them and bad for other websites in respective category. Thus we have taken 294 websites from these categories and level-1 pages of these websites. Thus 90 websites are nominated in 2010, 109 websites nominated in 2011 and 95 websites are nominated in 2012 year.

3.3.1 Categorization of Websites into Good and Bad

There are 2 awards given in each category, one is chosen by judges as winner, and another is People's Champ Winner. We have considered the winner websites in all the categories as good and all the other nominee websites as bad. In 2010 out of 90 websites 33 websites are categorize in good and 57 websites as bad. In 2011 out of 109 websites 41 websites are categorized in good and 68 websites as bad. Similarly in 2012 out of 95 websites 31 websites are categorized into good and 64 websites as bad. Table 8 shows the website categorization.

Table 8: Categorization of Websites

	Websites 2010	**Websites 2011**	**Websites 2012**
Good	33	41	31
Bad	57	68	64

Chapter Four: Research Methodology

4.1 Methodology

This methodology finds the number of web page metrics like word count, link count etc. and compare the quality of different web pages using these metrics and finally build the models using machine learning and statistical technique to predict the website as good or bad.

Figure 1 shows the basic methodology adapted for this study. Methodology is divided into three Modules where Module 1 is Empirical data collection, Module 2 is Web metrics calculator, and Module 3 is result analysis.

Empirical data collection: First websites were selected from 2010, 2011 and 2012 pixelawards website from different sub-category. Second step is to enter the URL of the website from which we want to calculate different web metrics.

Web metrics calculator: Web metrics calculator is a tool used to compute different web metrics for the input URL of website.

Result analysis: Data computed by web metrics calculator is used for analysis and comparing the different machine leaning algorithm and logistic regression to predict the quality of web page and to compare the prediction accuracy of different machine learning algorithms.

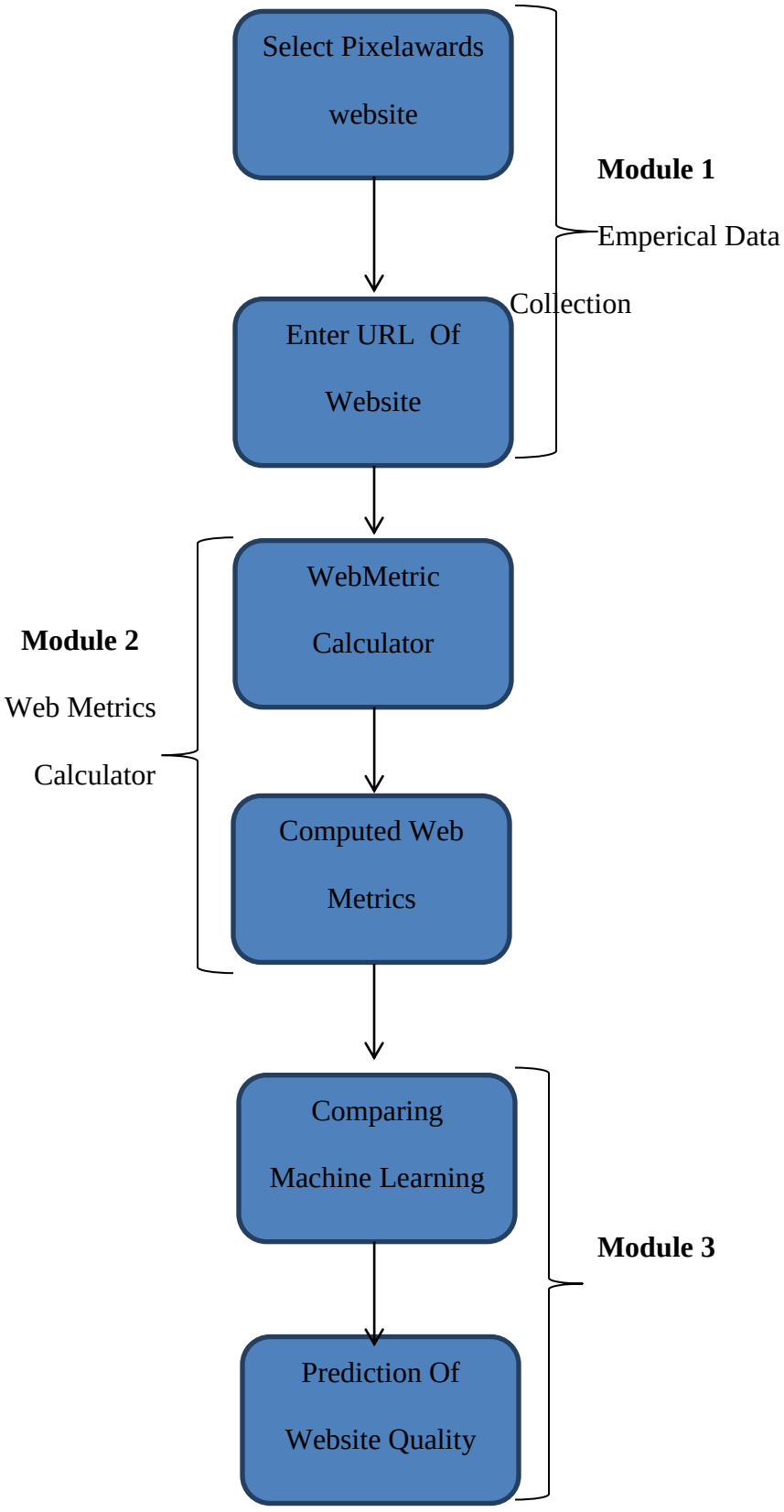


Figure 1: Flow graph of methodology

4.2 Tool description

Web Metrics Calculator was developed in asp.net language that calculates 20 web page metrics

Purpose: The idea is to automatically collect information about the web pages that gives an idea of the flavour of web page document. Web metrics calculated by this tool can be used for analysis of web site quality attributes.

Advantage:

- It automates the extraction of web metrics rather than manually searching the tags or the information in html page.
- Size of the tool is very small (in few Kb).
- Implementation of Sql query will store the result of all web pages in .csv file so there is no need to enter data manually.

Installation:

- Install Asp.net 2008(minimum) on the operating System.
- Install Sql server 2008 on the operating system.

Required operating environment:

- Operating system: Tool can be run on winxp, win7 or in win8 operating system.
- Microsoft .net: Asp.net 2008 and Microsoft .net 3.5(minimum) must be installed on operating system to run this tool.
- CPU: 2.4 Ghz processor and 512Mb (minimum) ram is required.
- Disk space: There must be 40Mb of data must be free to run this tool
- Web connectivity: An active internet connection is required.

Method to calculate Web metrics:

1. Word count: Total number of words that are displayed on the web page. This can be calculated by calculating the number of display word between <body> and </body> tag.

2. Link count: Total number of links that direct to either external page or internal page. This can be calculated by counting <a href> tag in web page.

3. Graphic word count: Total number of words used to save the image file. This can be calculated by calculating number of words between alt" " in web page.

4. Page Size: Total size of the web page(in Bytes).

5. Script count: Total number of scripts used in web page. This can be calculated by counting <script> tag in web page.

6. Image count: Total number of images that exist on web page. This can be calculated by counting total number of ".jpg", ".png" and ".gif" in web page.

7. Inline element count: This can be calculated by counting total number tag used in web page.

8. Class used count: Total number of classes used in web page. This can be calculated by counting total number of class=" " used in web page.

9. Exclamation count: Total number of exclamation(!) used in web page. This can be calculated by counting total number of ! in web page.

10. Load time: Time required to load the web page in web browser. This can be calculated by measuring End time-Start time.

11. Meta tag count: Total number of Meta tag used in web page. This can be calculated total number of <Meta tag used in web page.

12. Page title word count: Total words used in the page title of web page. This can be calculated by counting total number of words between <title> and </title> tag.

13. List Items: Total number of lists used in web page. This can be calculated counting total number of <li> tag in web page.

14. Meta description length: Total number of words used in meta description.
15. Unordered List: Total number of unordered list exists on web page. This can be calculated total number of <ol> tag used in web page.

16. Division count: Total number of div tag used in web page. This can be calculated by counting total number of tag used in web page.

17. Number of headings: Total number of lines that are marked as headings. This can be calculated by counting total number of <h1>,<h2>,<h3>,<h4>,<h5> and <h6> in web page.

18. Paragraph count: Total number of paragraphs used in web page. This can be calculated by counting total number of <p> tag used in web page.

19. Text link count: Total number of links that are text. This can be calculated by counting total number total number of display words between <a> and tag.

20. Image link count: Total number of links that are image. This can be calculated by counting total number of <img between <a> and tag.

Web Metrics Tool works by taking the URL of any web page as input and produce the output of selected web metrics. Basic interface of Web Metric Tool is shown in figure 2.

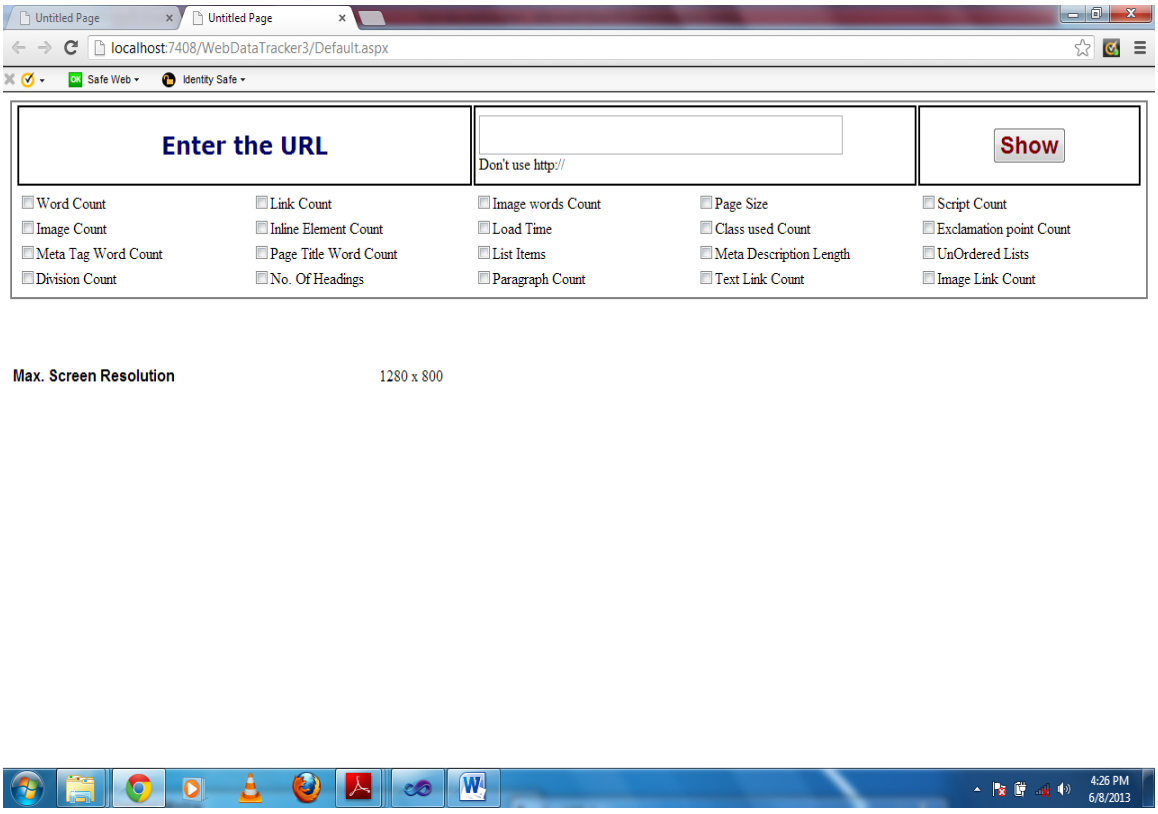


Figure 2: Basic Interface of WEB METRICS CALCULATOR

Web Metrics Calculator stores the source code of the URL as text file temporarily in local directory and then applies parsing techniques to text file to get desired web metrics. “SHOW” button on the Web Metric Tool enables to view the desire web metrics as output.

Figure 3 shows the output window of Web Metrics Calculator of single URL

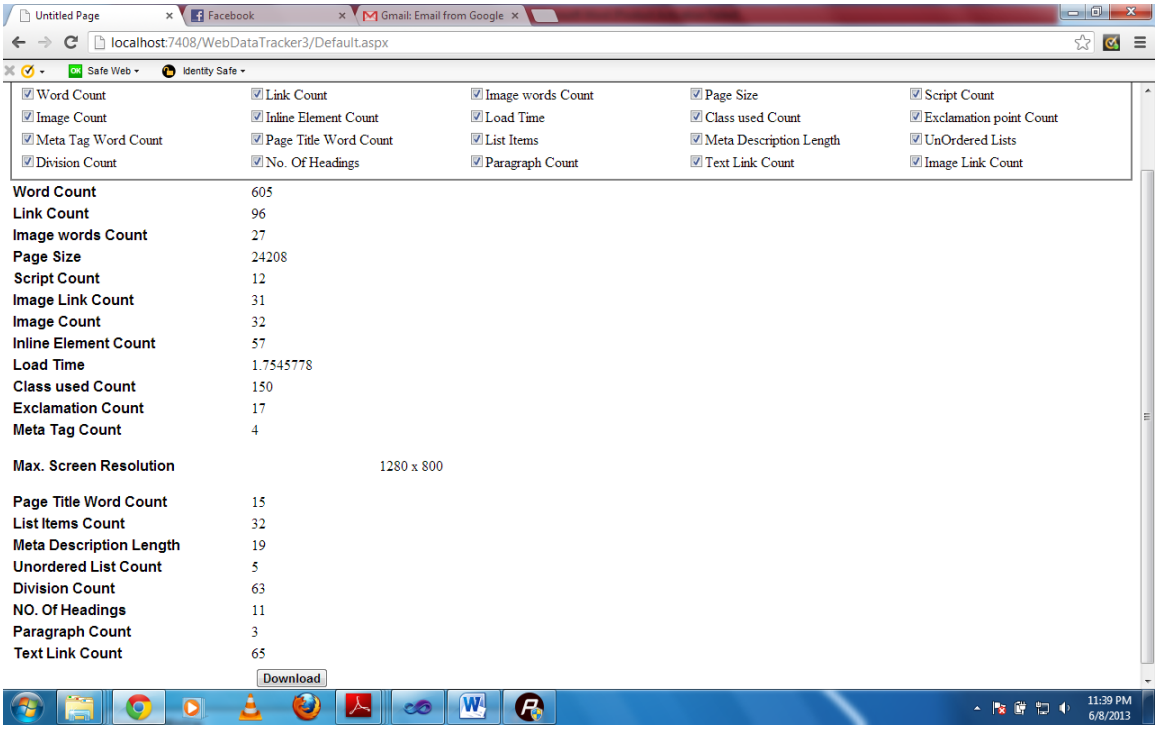
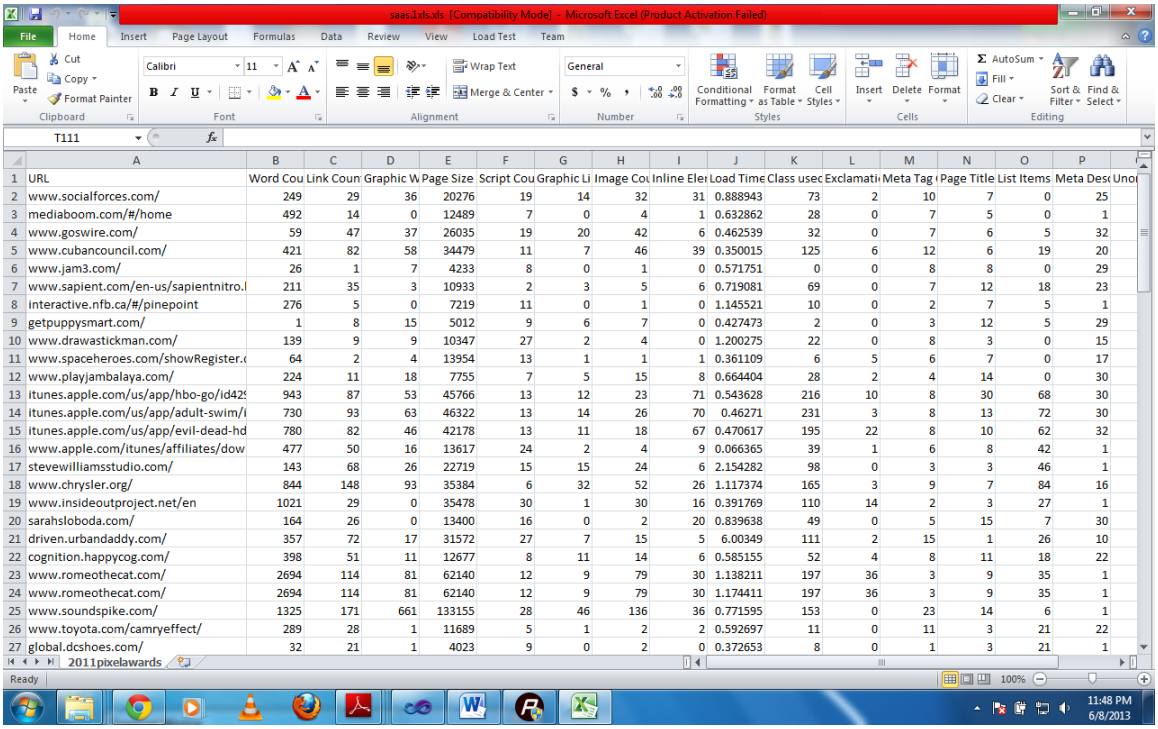


Figure 3: Output Window of WEB METRICS CALCULATOR

Figure 4 depicts output of Web Metrics Calculator is automatically saved in a .csv file, when we calculate the web metrics of desired number of URL. By clicking on the "Download" button a .csv file is automatically generated in which columns represent the different web metrics and row represents different URL.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	URL	Word Cou	Link Coun	Graphic W	Page Size	Script Cou	Graphic Li	Image Cou	Inline Ele	Load Time	Class usec	Exclamati	Meta Tag	Page Title	List Items	Meta Des
2	www.socialforces.com/	249	29	36	20276	19	14	32	31	0.888943	73	2	10	7	0	25
3	mediaboom.com/#/home	492	14	0	12489	7	0	4	1	0.632862	28	0	7	5	0	1
4	www.goswire.com/	59	47	37	26035	19	20	42	6	0.462539	32	0	7	6	5	32
5	www.cubancouncil.com/	421	82	58	34479	11	7	46	39	0.350015	125	6	12	6	19	20
6	www.jam3.com/	26	1	7	4233	8	0	1	0	0.571751	0	0	8	8	0	29
7	www.sapient.com/en-us/sapientnitro.l	211	35	3	10933	2	3	5	6	0.719081	69	0	7	12	18	23
8	interactive.nfb.ca/#/pinepoint	276	5	0	7219	11	0	1	0	1.145521	10	0	2	7	5	1
9	getpuppysmart.com/	1	8	15	5012	9	6	7	0	0.427473	2	0	3	12	5	29
10	www.drawastickman.com/	139	9	9	10347	27	2	4	0	1.200275	22	0	8	3	0	15
11	www.spaceheroes.com/showRegister.(	64	2	4	13954	13	1	1	1	0.361109	6	5	6	7	0	17
12	www.playjambalaya.com/	224	11	18	7755	7	5	15	8	0.664404	28	2	4	14	0	30
13	itunes.apple.com/us/app/hbo-go/id42	943	87	53	45766	13	12	23	71	0.543628	216	10	8	30	68	30
14	itunes.apple.com/us/app/adult-swim/i	730	93	63	46322	13	14	26	70	0.46271	231	3	8	13	72	30
15	itunes.apple.com/us/app/evil-dead-hd	780	82	46	42178	13	11	18	67	0.470617	195	22	8	10	62	32
16	www.apple.com/itunes/affiliates/dow	477	50	16	13617	24	2	4	9	0.066365	39	1	6	8	42	1
17	stevewilliamsstudio.com/	143	68	26	22719	15	15	24	6	2.154282	98	0	3	3	46	1
18	www.chrysler.org/	844	148	93	35384	6	32	52	26	1.117374	165	3	9	7	84	16
19	www.insideoutproject.net/en	1021	29	0	35478	30	1	30	16	0.391769	110	14	2	3	27	1
20	sarahsloboda.com/	164	26	0	13400	16	0	2	20	0.839638	49	0	5	15	7	30
21	driven.urbandaddy.com/	357	72	17	31572	27	7	15	5	6.00349	111	2	15	1	26	10
22	cognition.happycog.com/	398	51	11	12677	8	11	14	6	0.585155	52	4	8	11	18	22
23	www.romeothecat.com/	2694	114	81	62140	12	9	79	30	1.138211	197	36	3	9	35	1
24	www.romeothecat.com/	2694	114	81	62140	12	9	79	30	1.174411	197	36	3	9	35	1
25	www.soundspike.com/	1325	171	661	133155	28	46	136	36	0.771595	153	0	23	14	6	1
26	www.toyota.com/camryeffect/	289	28	1	11689	5	1	2	2	0.592697	11	0	11	3	21	22
27	global.dcshoes.com/	32	21	1	4023	9	0	2	0	0.372653	8	0	1	3	21	1

Figure 4: .CSV Data file

4.3 Machine Learning Algorithms

4.3.1 Bayes Net

Bayesian networks pearl (1988) are quite powerful probabilistic representation and that's why they are most often used for classification purpose but unfortunately they perform in a poor way when learned in a standard way(Grossman & Domingos, 2004).Bayes Nets are graphical representation for probabilistic relationships among a set of random variables. Given a finite set $X(x_1, x_2, x_3, \ldots\ldots x_n)$ of discrete random variables where each variable X_i may take values from a finite set, denoted by Val (X_i) (bayes nets, 2007). A Bayes net is an annotated DAG (directed acyclic graph) G that encodes a joint probability distribution over X. In the Bayes networks nodes of the graph correspond to the random variable $X_1, X_2, X_3 \ldots\ldots\ldots X_n$.The links of the graph

correspond to the direct influence from one variable to the other. If there is a directed link from variable X_i to variable X_j, variable X_i will be a parent of variable X_j. Each node is annotated with a conditional probability distribution (CPD) that represents $p(X_i / P_a(X_i))$ z, where $P_a(X_i)$ denotes the parents of X_i in G. The pair (G, CPD) encodes the joint distribution $p(X_i ....., X_n)$.

4.3.2 Naïve Bayes

Naive Bayes classifier is a statistical classifier as well as a supervised learning method which is based on the Bayesian theorem given by Thomas bayes. It predicts class membership probabilities, such as the probability that a given sample is belongs to a particular class or not(Leung, 2007). Given a class variable, a Naive Bayes classifier assumes that the presence of a particular feature of a class is not related to the presence of any other feature. Given the set of variables

$X = \{x_1, x_2, x_3, ...x_n\}$ a probabilistic classifier can be defined as

$p(C \mid x_1, x_2, x_3, ... x_n)$

Where, C is a dependent class variable with a set of possible outcomes conditional on several variables.

Using Bayes Theorem,

$$p(C) \mid p(x_1, x_2, x_3, ...x_n) = \frac{p(C)\, p(x_1, x_2, x_3, ...x_n \mid C)}{p(x_1, x_2, x_3, ...x_n)}$$

Thus, we want to construct the posterior probability of the event C. Thus, the equation can be written as:

$$Posterior = \frac{\Pr ior * \text{likelihood}}{Evidence}$$

Naïve bayes classification provides a very useful approach to understand and evaluate many other learning algorithms. Naive bayes classification is very fast, it calculates explicit probabilities and robust to noises.

4.3.3 Multilayer Perceptron

A Multilayer Perceptron is a feed forward artificial neural network model that maps different input data instances onto a set of appropriate output. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Each node in all the layers is a neuron associated with nonlinear activation function except for the input nodes. MLP utilizes a supervised learning technique called back-propagation for training the network. MLP is a modification of the standard linear perceptron, which can distinguish data that is not linearly separable(Anderson, 2003).

Multilayer perceptron training algorithm

Multilayer perceptron training is done in two phase:

1. Forward phase
2. Backward phase

Weights are fixed in forward phase and input is propagated layer by layer through input layer to output layer.

Error is computed by comparing the actual output and the target response and this error is propagated layer by layer in backward direction through output layer to input layer.

Weight Adjustment in Backward phase

Assume input to input layer is E **,** and the observed output is $o_i(E)$ and target output is $t_i(E)$ and w_{ij} denotes the weight between node I and node j.

- The Error Term for output unit k is

$$\delta_{o_k} = o_k(\mathrm{E})(1-o_k(\mathrm{E}))(\mathrm{t}_k(\mathrm{E})-o_k(\mathrm{E}))$$

- The Error Term for hidden unit k is

$$\delta_{H_k} = h_k(\mathrm{E})(1-h_k(E)))\sum_{i\in outputs} w_{ki}\delta_{o_i}$$

- Now for every weight w_{ij} between node i and node j we have to calculate

$$\Delta_{ij} = \eta \delta_{H_j} x_i$$

η =learning rate

x_i =input to the network through node i

h_k =hidden unit

- Now for every weight w_{ij} between node i and hidden node j we have to calculate

$$\Delta_{ij} = \eta \delta_{o_j} h_i(\mathrm{E})$$

$h_i(\mathrm{E})$ = output from hidden node to E

- Final adjusted weight is

$$\mathrm{w}_{ij} = \mathrm{w}_{ij} + \Delta_{ij}$$

4.3.4 Adaboost

Adaboost is formulated by Freund & Schapire in 1995. Adaboost is an algorithm for constructing a "strong" classifier as linear combination. It used many other learning algorithms to improve their performance. Initially it chooses one learner out of all that classify data correctly as compare to others.

Then data is reweighted so that the "importance" of misclassified classes can be increased. This process continues and weight of each weak learner is identified.

$$f(\mathrm{x}) = \sum_{t=1}^{T} \alpha_t h_t(\mathrm{x})$$

With the help of "weak" and "simple" classifiers $h_t(\mathrm{x})$.

Some interesting properties of adaboost:

- Adaboost is a linear classifier.
- Output of adaboost converges to logarithm.
- Generalization properties are good.
- Adaboost produces sequence of more complex classifiers.
- It is basically a feature selector by minimization of upper bound on an empirical error.

Algorithm(Matas & Sochman)

Given $(x_1, y_1), ..., (x_m, y_m); x_i \in X, y_i \in \{-1,1\}$

Now initialize weights $D_1(i) = 1/m$

For t = 1… T:

- Call weaklearner, and it returns the weak classifier $h_t : X \rightarrow \{-1,1\}$ with minimum error with respect to distribution D_t

  D_t =Given distribution

- Now choose any $\alpha_t \in R$,

- Updating the value of D_{t+1} with respect to D_t

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

  Z_t =Normalization factor chosen for D_{t+1} is a distribution

  D_{t+1} =Output Distribution.

- Final output of the strong classifier is

$$H(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))$$

  $h_t(x)$ = Weak classifier

  $H(x)$ =Strong classifier

4.3.5 Decision Table

Decision table is basically lists cause and effects in matrix form. It is divided into four parts:

- Condition stub: Lists the comparisons and conditions.
- Action stub: which comprehensively lists the action to be taken along the various program branches

- Condition entries: which list in various columns the possible permutations of answers to the question in the condition stub?
- Action entries: which list, in its columns corresponding to the condition entries the action contingent upon the set of answers to question of that column?

4.3.6 Nnge

Nnge stands for Non-nested generalized exemplars. Nnge is one of the instance based machine learning technique. Nnge extends the nearest neighbourhood concept by including the generalized exemplars. Non-nested generalized exemplars theory is first given by Martin in 1995 which used both the simple instances and generalized exemplars. Nnge was implemented in Weka toolkit and proved to be a very competitive and useful technique(Witten & Frank, 1998).

Algorithm(Zaharie, Perian, & Negru, 2011)

- For every example E^j in the training set do:
- Find the hyper rectangle H^k which is closest to E^j
- IF $D(H^k, E^j)$=0 then
- IF $Class(E^j) \neq Class(H^k)$ THEN $Split(H^k, E^j)$
- ELSE $H^{'}$: $Extend(H^k, E^j)$
- IF $H^{'}$ overlaps with conflicting hyper rectangles
- THEN add E^j as non-generalized exemplar
- ELSE $H^k := H^{'}$

Where E^j = training examples

H^k =Generalized exemplars (hyper rectangles)

4.3.7 Part

Part is based on the divide and conquers strategy and basically avoids the global optimization step used in C4.5 rules and Ripper (Witten & Frank, 1998). It provide unrestricted decision list using divide and conquer strategy. It builds a partial C4.5 decision tree for eachiteration and makes the "best" leaf into a rule. Partial decision trees are used to obtain a rule.

4.3.8 Bf-tree

Bf-tree stands for best first decision tree; it is one of the types of decision tree learning. Bf-tree constructs a tree in divide and conquers strategy. In Bf-tree splitting is done at the best node out of given nodes. In bf-tree every nonterminal node tests an attribute whereas terminal nodes are used to assign classification(Haijan, 2007). In construction of a Bf-tree there are 3 important aspects that must be taken care of

- Calculating the best attribute to split.
- Out of all nodes that competing for splitting which should be expanded next.
- Criteria to stop the growing trees.

Selection of Best node is done on the basis of impurity i.e. node having the maximum reduction of impurity.

4.3.9 J-48

J48 is an open implantation c4.8 algorithm by Weka tool in java and this is decision tree based algorithm that builds the tree in the same way as ID3 along with some improvements. Ros Quinlan had developed this algorithm and this is now widely used for the classification purpose now a days. In this algorithm first base cases are checked and then for each attribute normalized information gain are found and the attribute that has the highest information gain is made the root node and this process is done recursively(c4.5 algorithm). J48 is an evolution and refinement of ID3 that accounts for

unavailable values, continuous attribute value ranges, pruning of decision trees, rule derivation, and so on makes it more fruitful.

4.3.10 Random forest

The word Random forest came from “randomized decision forest” which is first proposed by Tin Kam Ho in Bell labs in 1995. Random forest is quite popular and versatile machine learning classification algorithm and it can work on many attributes with large datasets. Beside the class tags it can also provide some other important information about the dataset. It consists of bagging of un-pruned decision tree learners with randomized features at each split .Decision trees are the most commonly method used for the data exploration such as CART and regression trees. The forest consists of randomly selected inputs or combination of inputs at each node to grow each tree. (Montillo, 2009)Random forest is simple and relatively robust to noise and gives quite good result for some data sets with fast learning. Accuracy of Random forest is as good as Adaboost and sometimes it also gives better result than this. One more advantage of this algorithm is that it is relatively faster than the bagging with better strength, variable importance and correlation.

Chapter Five: Result Analysis

In this research following measures are used to evaluate the performance of each predicting model.

1. Sensitivity and Specificity: Sensitivity and Specificity criteria are used to measure the correctness of the models. Sensitivity and Specificity can be defined as follows

$$Sensitivity = \frac{numbers\ of\ \ websites\ correctly\ predicted\ \ good}{no.\ of\ \ websites\ correctly\ predicted\ \ good + no.\ of\ \ websites\ incorrectly\ predicted\ \ bad}$$

$$Specificity = \frac{numbers\ of\ \ websites\ incorrectly\ predicted\ \ good}{no.\ of\ \ websites\ incorrectly\ predicted\ \ good + no.\ of\ \ websites\ correctly\ predicted\ \ bad}$$

Sensitivity is also called as TPR (True Positive Rate) and Specificity is also called as 1FPR (False Positive Rate).

2. ROC(Receiver Operating Characteristics): ROC analysis is used to evaluate the quality and performance of the predicting models. ROC graph is basically a technique for organizing, visualizing and selecting classifiers on the basis of their performance(Fawcett, 2005).ROC curve is a plotted as specificity on the x-coordinate and sensitivity on the y-coordinate. We can select many cut-off points to calculate sensitivity and specificity but the optimal cut-off points give the maximum value of both sensitivity and specificity.

5.1 Descriptive Statistics

Descriptive statistics gives the simple quantitative measure of the dataset. It provide information like “min”, ”max”, ”mean” and “std dev” of the dataset of year 2010, 2011 and 2012. Table 9 describe the descriptive statistics of 2010 year data. Similarly Table 10 and Table 11 describe the descriptive statistics of 2011 and 2012 year data respectively.

Table 9: Descriptive statistics of year 2010 data

	MIN	MAX	MEAN	STD DEV
Word count	0	5343	740.722	1166.13
Link count	0	880	101.022	131.146
Graphic word count	0	436	43.4	84.966
Page size	1438	421596	44832.111	74331.911
Script count	0	72	16.767	14.353
Graphic link count	0	199	14.5	28.298
Image count	0	1917	48.44	206.626
Inline element count	0	807	42.167	108.565
Load time	0.087	2.865	0.592	0.451
Class used count	0	2725	188.367	427.982
Exclamation count	0	85	3.7	10.087
Meta tag count	0	20	5.744	4.289
Page title word count	1	261	9.822	27.117
List item count	0	198	31.344	47.723
Meta description length	1	275	17.233	33.099

Unordered list count	0	252	11.344	28.802
Division count	0	1192	81.667	165.517
Headings count	0	235	11.833	27.491
Paragraph count	0	920	21.289	97.864
Text link count	0	874	69.267	121.201

Table 10: Descriptive statistics of year 2011 data

	MIN	MAX	MEAN	STD DEV
Word count	0	9338	735.444	1167.039
Link count	0	312	76.426	70.491
Graphic word count	0	661	53.361	104.096
Page size	715	327461	41243.139	49870.344
Script count	0	45	15.519	8.403
Graphic link count	0	85	12.889	16.569
Image count	0	346	34.426	50.56
Inline element count	0	527	21.824	66.298
Load time	0.066	2.643	0.652	0.408
Class used count	0	1866	191.343	304.296

Exclamation count	0	36	3.481	6.648
Meta tag count	1	23	7.259	4.879
Page title word count	1	22	7.593	4.561
List item count	0	242	30.444	44.736
Meta description length	1	155	17.13	20.709
Unordered list count	0	74	8.63	11.051
Division count	0	855	97.639	137.781
Headings count	0	71	13.093	15.892
Paragraph count	0	196	19.63	36.486
Text link count	0	288	62.574	64.408

Table 11: Descriptive statistics of year 2012 data

	MIN	MAX	MEAN	STD DEV
Word count	0	5196	623.2	810.205
Link count	0	951	98.463	133.491
Graphic word count	0	803	44.926	98.351
Page size	572	357540	39631.895	58441.994
Script count	0	63	15.832	10.818

Graphic link count	0	96	13.968	19.121
Image count	0	307	31.242	44.509
Inline element count	0	661	52.653	94.663
Load time	0.124	1.457	0.617	0.286
Class used count	0	1433	208.2	274.367
Exclamation count	0	140	4.421	15.636
Meta tag count	0	36	7.916	5.414
Page title word count	1	33	7.368	5.389
List item count	0	344	32.695	54.252
Meta description length	1	136	19.179	21.051
Unordered list count	0	134	12.158	20.535
Division count	0	813	96.8	127.748
Headings count	0	122	14.611	20.617
Paragraph count	0	135	10.726	16.104
Text link count	0	883	84.495	122.264

5.2 Logistic Regression Analysis

Logistic regression is one of the statistical methods of prediction. Table 12 describes the prediction of web pages of all 3 models and Table 13 describes the 10 cross fold validation result of all 3 models.

Observation made from analysis:

- Out of 33 good websites, 20 are correctly predicted and out of 57 bad website, 40 are correctly predicted which gives the sensitivity 76.83 and specificity of 80.45 respectively.
- Out of 41 good websites, 27 are correctly predicted and out of 68 bad website, 54are correctly predicted which gives the sensitivity 79.68 and specificity of 81.43 respectively.
- Out of 31 good websites, 18 are correctly predicted and out of 64 bad website, 49 are correctly predicted which gives the sensitivity 78.34 and specificity of 79.45 respectively.

Table 12: Website prediction of logistic regression for model 1, 2, and 3

PARAMETER	MODEL1	MODEL2	MODEL3
Number of good website correctly predicted	20	27	18
Number of bad website correctly predicted	40	54	49

Table 13: 10-cross fold results using logistic regression for model 1, 2, and 3

	SENSITIVITY	SPECIFICITY	CUTOFF	AUC
Model 1	76.83	80.45	0.356	0.773
Model 2	79.68	81.43	0.114	0.810
Model 3	78.34	79.45	0.874	0.767

5.3 Bayes Net Analysis

Table 14 describes the prediction of web pages of all 3 models and Table 15 describes the 10 cross fold validation result of all 3 models. Figure 5, Figure 6 and Figure 7 shows the ROC curves for Model 1, Model 2 and Model 3 respectively.

Observation made from analysis:

- Out of 33 good websites, 11 are correctly predicted and out of 57 bad website, 53 are correctly predicted which gives the sensitivity 69.70 and specificity of 68.40 respectively.
- Out of 41 good websites, 33 are correctly predicted and out of 68 bad website, 49are correctly predicted which gives the sensitivity 73.20 and specificity of 73.10 respectively.
- Out of 31 good websites, 23 are correctly predicted and out of 64 bad website, 38 are correctly predicted which gives the sensitivity 74.20 and specificity of 73.40 respectively.

Table 14: Website prediction of Bayes net for model 1, 2, and 3

PARAMETER	MODEL 1	MODEL 2	MODEL3
Number of good website correctly predicted	11	33	23
Number of bad website correctly predicted	53	49	54

Table 15: 10-cross fold results using Bayes net for model 1, 2, and 3

	SENSITIVITY	SPECIFICITY	CUTOFF	AUC
Model 1	69.70	68.40	.380	.740
Model 2	73.20	73.10	.703	.770
Model 3	74.20	73.40	.341	.862

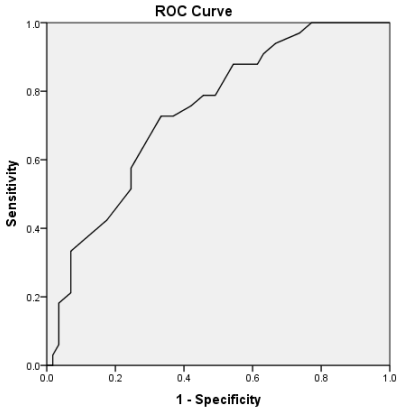


Figure 5: ROC curve of Bayes Net for Model 1

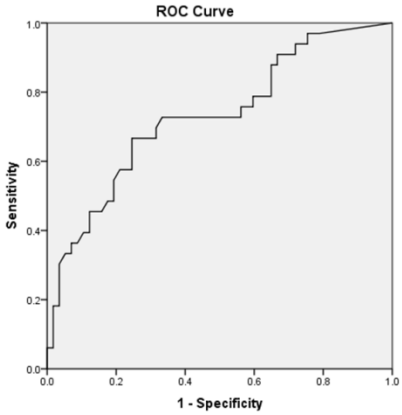


Figure 8: ROC curve of Naïve Bayes for Model 1

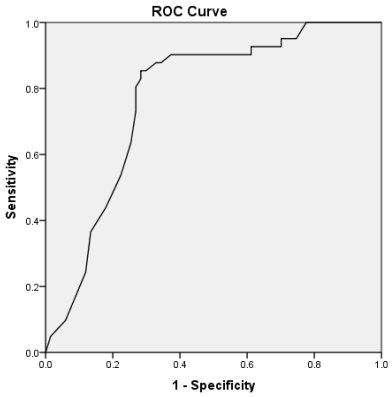


Figure 6: ROC curve of Bayes Net for Model 2

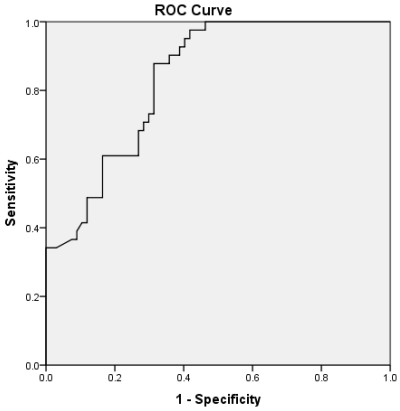


Figure 9: ROC curve of Naïve Bayes for Model 2

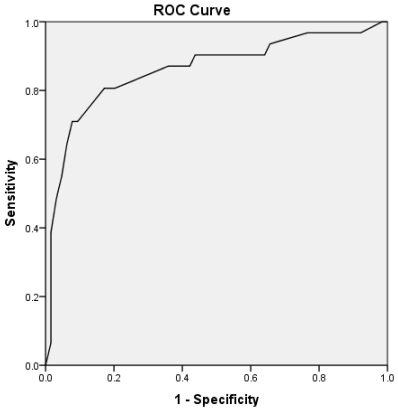


Figure 7: ROC curve of Bayes Net for Model 3

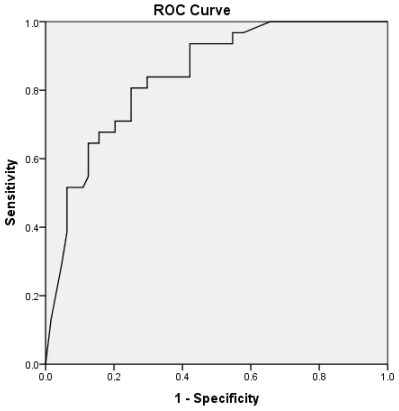


Figure 10: ROC curve of Naïve Bayes for Model

5.4 Naïve Bayes Analysis

Table 16 describes the prediction of web pages of all 3 models and Table 17 describes the 10 cross fold validation result of all 3 models. Figure 8, Figure 9 and Figure 10 shows the ROC curves for Model 1, Model 2 and Model 3 respectively.

Observation made from analysis:

- Out of 33 good websites, 28 are correctly predicted and out of 57 bad website, 20 are correctly predicted which gives the sensitivity 66.70 and specificity of 79.10 respectively.
- Out of 41 good websites, 31 are correctly predicted and out of 68 bad website, 34are correctly predicted which gives the sensitivity 70.70 and specificity of 71.60 respectively.
- Out of 31 good websites, 26 are correctly predicted and out of 64 bad website, 38 are correctly predicted which gives the sensitivity 74.20 and specificity of 75.00 respectively.

Table 16: Website prediction of Naïve bayes for model 1, 2, and 3

PARAMETER	MODEL 1	MODEL 2	MODEL 3
Number of good website correctly predicted	28	31	26
Number of bad website correctly predicted	20	34	38

Table 17: 10-cross fold results using Naïve bayes for model 1, 2, and 3

	SENSITIVITY	SPECIFICITY	CUTOFF	AUC
Model 1	66.70	71.90	.8425	.729
Model 2	70.70	71.60	.876	.836
Model 3	74.20	75.00	.9575	.841

5.5 Multilayer Perceptron Analysis

Table 18 describes the prediction of web pages of all 3 models and Table 19 describes the 10 cross fold validation result of all 3 models. Figure 11, Figure 12 and Figure 13 shows the ROC curves for Model 1, Model 2 and Model 3 respectively.

Observation made from analysis:

- Out of 33 good websites, 28 are correctly predicted and out of 57 bad website, 20 are correctly predicted which gives the sensitivity 81.80 and specificity of 82.50 respectively.
- Out of 41 good websites, 30 are correctly predicted and out of 68 bad website, 44are correctly predicted which gives the sensitivity 68.30 and specificity of 67.20 respectively.
- Out of 31 good websites, 19 are correctly predicted and out of 64 bad website, 48 are correctly predicted which gives the sensitivity 67.70 and specificity of 67.20 respectively.

Table 18: Website prediction of Multilayer Perceptron for model 1, 2, and 3

PARAMETER	MODEL 1	MODEL 2	MODEL 3
Number of good website correctly predicted	24	30	19
Number of bad website correctly predicted	51	44	48

Table 19: 10-cross fold results using Multilayer Perceptron for model 1, 2, and 3

	SENSITIVITY	SPECIFICITY	CUTOFF	AUC
Model 1	81.80	82.50	.1940	.837
Model 2	68.30	67.20	.5155	.747
Model 3	67.70	67.20	.3705	.749

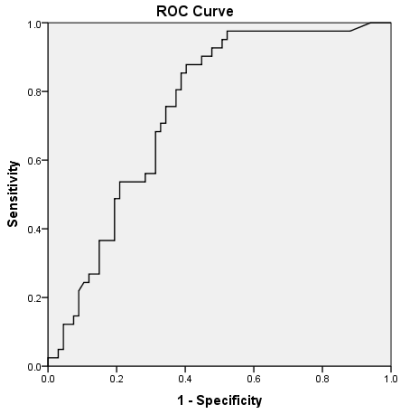


Figure 8: ROC curve of Multilayer perceptron for Model 1

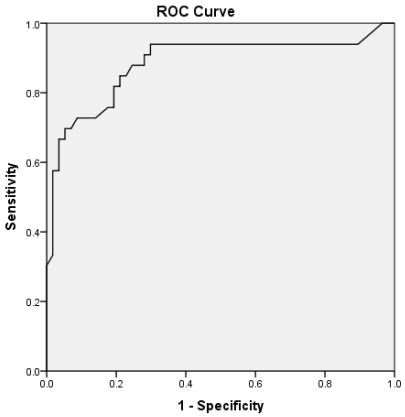


Figure 14: ROC curve of Adaboost for Model 1

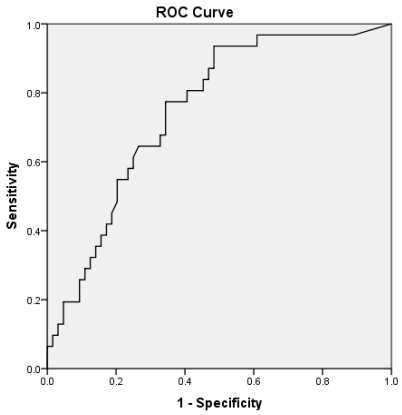


Figure 9: ROC curve of Multilayer perceptron for Model 2

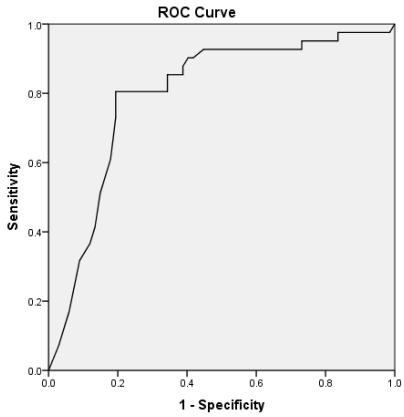


Figure 15: ROC curve of Adaboost for Model 2

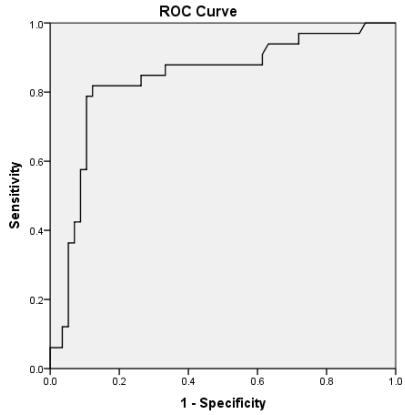


Figure 10: ROC curve of Multilayer perceptron for Model 3

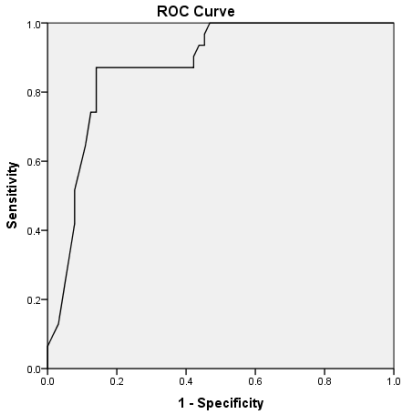


Figure 16: ROC curve of Adaboost for Model 3

5.6 Adaboost Analysis

Table 20 describes the prediction of web pages of all 3 models and Table 21 describes the 10 cross fold validation result of all 3 models. Figure 14, Figure 15 and Figure 16 shows the ROC curves for Model 1, Model 2 and Model 3 respectively.
Observation made from analysis:

- Out of 33 good websites, 26 are correctly predicted and out of 57 bad website, 46 are correctly predicted which gives the sensitivity 81.80 and specificity of 82.50 respectively.
- Out of 41 good websites, 33 are correctly predicted and out of 68 bad website, 54 are correctly predicted which gives the sensitivity 80.50 and specificity of 80.60 respectively.
- Out of 31 good websites, 27 are correctly predicted and out of 64 bad website, 55 are correctly predicted which gives the sensitivity 83.90 and specificity of 85.90 respectively.

Table 20: Website prediction of Adaboost for model 1, 2, and 3

PARAMETER	MODEL 1	MODEL2	MODEL3
Number of good website correctly predicted	26	33	27
Number of bad website correctly predicted	46	54	55

Table 21: 10-cross fold results using Adaboost for model 1, 2, and 3

	SENSITIVITY	SPECIFICITY	CUTOFF	AUC
Model 1	81.80	82.50	.4615	.884
Model 2	80.50	80.60	.5185	.797
Model 3	83.90	85.90	.51	.877

5.7 Decision Table

Table 22 describes the prediction of web pages of all 3 models and Table 23 describes the 10 cross fold validation result of all 3 models. Figure 17, Figure 18 and Figure 19 shows the ROC curves for Model 1, Model 2 and Model 3 respectively.
Observation made from analysis:

- Out of 33 good websites, 24 are correctly predicted and out of 57 bad website, 43 are correctly predicted which gives the sensitivity 72.70 and specificity of 79.10 respectively.
- Out of 41 good websites, 33 are correctly predicted and out of 68 bad website, 54 are correctly predicted which gives the sensitivity 80.50 and specificity of 80.60 respectively.
- Out of 31 good websites, 27 are correctly predicted and out of 64 bad website, 55 are correctly predicted which gives the sensitivity 83.90 and specificity of 85.90 respectively.

Table 22: Website prediction of Decision table for model 1, 2, and 3

PARAMETER	MODEL1	MODEL2	MODEL3
Number of good website correctly predicted	24	29	25
Number of bad website correctly predicted	43	52	53

Table 23: 10-cross fold results using Decision table for model 1, 2, and 3

	SENSITIVITY	SPECIFICITY	CUTOFF	AUC
Model 1	72.70	71.90	.3485	.787
Model 2	70.70	77.60	.603	.745
Model 3	80.60	79.70	.3665	.860

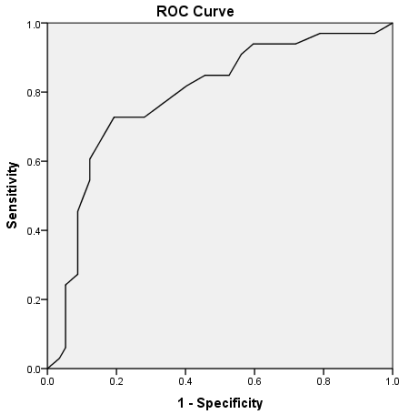


Figure 11: ROC curve of Decision Table for Model 1

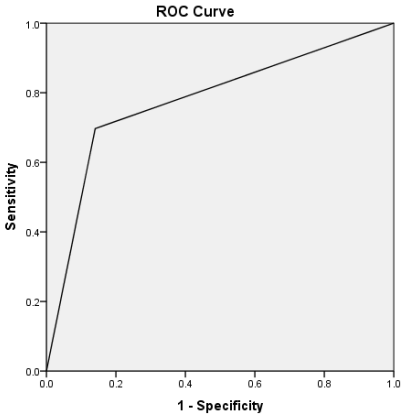


Figure 20: ROC curve of Nnge for Model 1

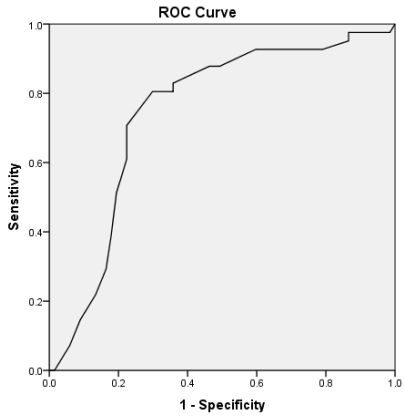


Figure 12: ROC curve of Decision Table for Model 2

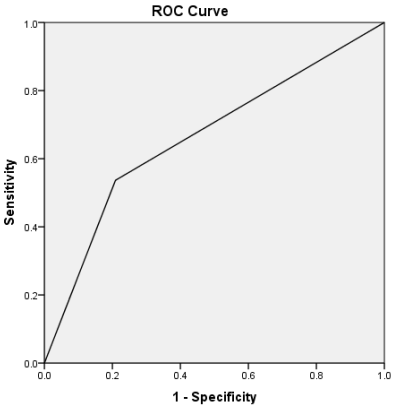


Figure 21: ROC curve of Nnge for Model 2

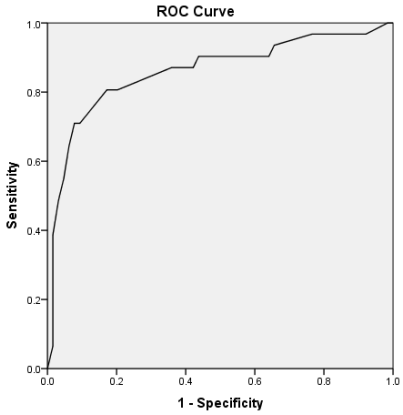


Figure 13: ROC curve of Decision Table for Model 3

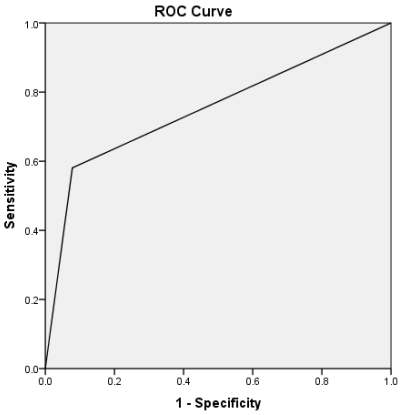


Figure 22: ROC curve of Nnge for Model 3

5.8 Nnge Analysis

Table 24 describes the prediction of web pages of all 3 models and Table 25 describes the 10 cross fold validation result of all 3 models. Figure 20, Figure 21 and Figure 22 shows the ROC curves for Model 1, Model 2 and Model 3 respectively.

Observation made from analysis:

- Out of 33 good websites, 23 are correctly predicted and out of 49 bad website, 53 are correctly predicted which gives the sensitivity 69.70 and specificity of 86.00 respectively.
- Out of 41 good websites, 33 are correctly predicted and out of 68 bad website, 54 are correctly predicted which gives the sensitivity 80.50 and specificity of 80.60 respectively.
- Out of 31 good websites, 27 are correctly predicted and out of 64 bad website, 55 are correctly predicted which gives the sensitivity 83.90 and specificity of 85.90 respectively.

Table 24: Website prediction of Nnge for model 1, 2, and 3

PARAMETER	MODEL1	MODEL2	MODEL3
Number of good website correctly predicted	23	22	18
Number of bad website correctly predicted	49	53	59

Table 25: 10-cross fold results using Nnge for model 1, 2, and 3

	SENSITIVITY	SPECIFICITY	CUTOFF	AUC
Model 1	69.70	86.00	.50	.778
Model 2	53.70	79.10	.50	66.4
Model 3	58.10	92.20	.50	.751

5.9 Part Analysis

Table 26 describes the prediction of web pages of all 3 models and Table 27 describes the 10 cross fold validation result of all 3 models. Figure 23, Figure 24 and Figure 25 shows the ROC curves for Model 1, Model 2 and Model 3 respectively.

Observation made from analysis:

- Out of 33 good websites, 23 are correctly predicted and out of 57 bad website, 48 are correctly predicted which gives the sensitivity 69.70 and specificity of 71.90 respectively.
- Out of 41 good websites, 33 are correctly predicted and out of 68 bad website, 48 are correctly predicted which gives the sensitivity 70.70 and specificity of 71.60 respectively.
- Out of 31 good websites, 23 are correctly predicted and out of 64 bad website, 52 are correctly predicted which gives the sensitivity 74.20 and specificity of 81.20 respectively.

Table 26: Website prediction of Part for model 1, 2, and 3

PARAMETER	MODEL1	MODEL2	MODEL3
Number of good website correctly predicted	23	33	23
Number of bad website correctly predicted	48	48	52

Table 27: 10-cross fold results using Part for model 1, 2, and 3

	SENSITIVITY	SPECIFICITY	CUTOFF	AUC
Model 1	69.70	71.90	.135	.786
Model 2	70.70	71.60	.6005	.770
Model 3	74.20	81.20	.377	.738

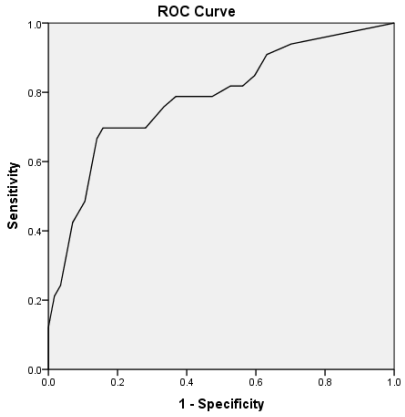


Figure 14: ROC curve of Part for Model 1

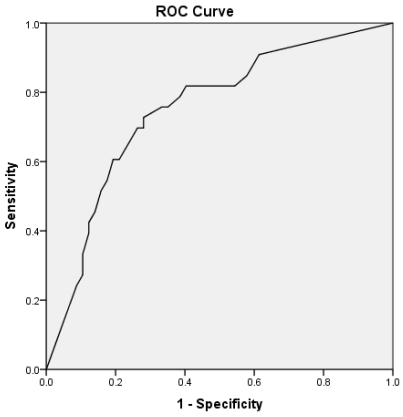


Figure 26: ROC curve of Bf-tree for Model 1

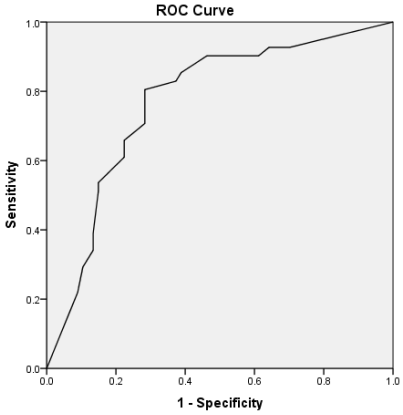


Figure 15: ROC curve of Part for Model 2

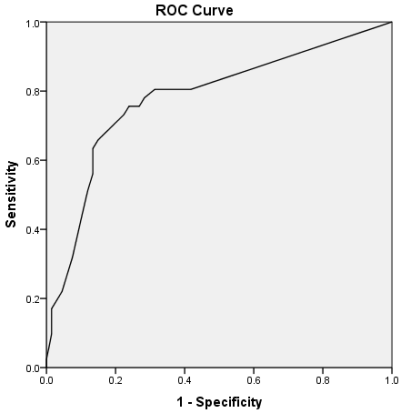


Figure 27: ROC curve of Bf-tree for Model 2

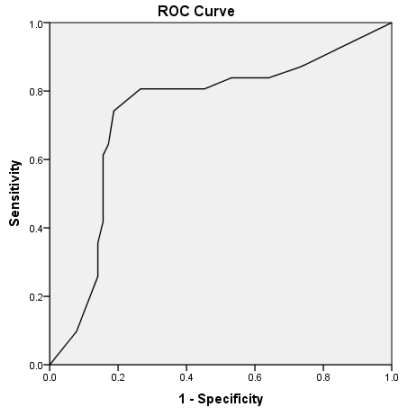


Figure 16: ROC curve of Part for Model 3

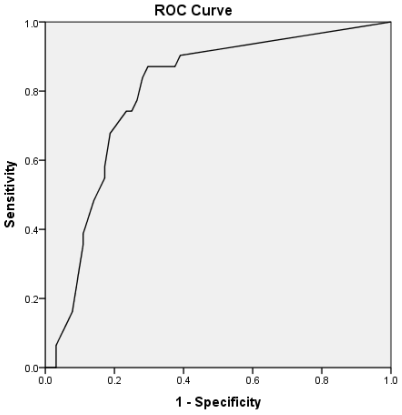


Figure 28: ROC curve of Bf-tree for Model 3

5.10 Bf-tree Analysis

Table 28 describes the prediction of web pages of all 3 models and Table 29 describes the 10 cross fold validation result of all 3 models. Figure 26, Figure 27 and Figure 28 shows the ROC curves for Model 1, Model 2 and Model 3 respectively.
Observation made from analysis:

- Out of 33 good websites, 23 are correctly predicted and out of 57 bad website, 42 are correctly predicted which gives the sensitivity 72.70 and specificity of 71.90 respectively.
- Out of 41 good websites, 30 are correctly predicted and out of 68 bad website, 52 are correctly predicted which gives the sensitivity 75.60 and specificity of 76.10 respectively.
- Out of 31 good websites, 23 are correctly predicted and out of 64 bad website, 48 are correctly predicted which gives the sensitivity 74.20 and specificity of 75.00 respectively.

Table 28: Website prediction of Bf-tree for model 1, 2, and 3

PARAMETER	MODEL1	MODEL2	MODEL3
Number of good website correctly predicted	23	30	23
Number of bad website correctly predicted	42	52	48

Table 29: 10-cross fold results using Bf-tree for model 1, 2, and 3

	SENSITIVITY	SPECIFICITY	CUTOFF	AUC
Model 1	72.70	71.90	.2285	.751
Model 2	75.60	76.10	.2915	.781
Model 3	74.20	75.00	.450	.797

5.11 J-48 Analysis

Table 30 describes the prediction of web pages of all 3 models and Table 31 describes the 10 cross fold validation result of all 3 models. Figure 29, Figure 30 and Figure 31 shows the ROC curves for Model 1, Model 2 and Model 3 respectively.
Observation made from analysis:

- Out of 33 good websites, 23 are correctly predicted and out of 57 bad website, 45 are correctly predicted which gives the sensitivity 72.70 and specificity of 77.20 respectively.
- Out of 41 good websites, 34 are correctly predicted and out of 68 bad website, 51 are correctly predicted which gives the sensitivity 80.50 and specificity of 76.10 respectively.
- Out of 31 good websites, 24 are correctly predicted and out of 64 bad website, 54 are correctly predicted which gives the sensitivity 77.40 and specificity of 78.10 respectively.

Table 30: Website prediction of J-48 for model 1, 2, and 3

PARAMETER	MODEL1	MODEL2	MODEL3
Number of good website correctly predicted	23	34	24
Number of bad website correctly predicted	45	51	54

Table 31: 10-cross fold results using J-48 for model 1, 2, and 3

	SENSITIVITY	SPECIFICITY	CUTOFF	AUC
Model 1	72.70	77.20	.1345	.762
Model 2	80.50	76.10	.7555	.828
Model 3	77.40	78.10	.0665	.802

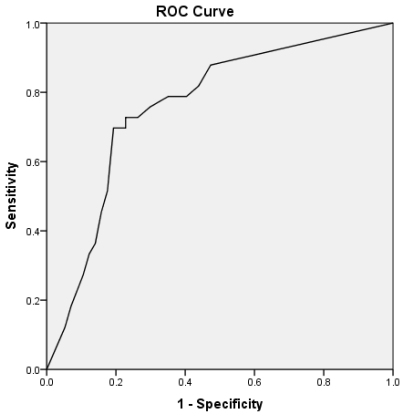


Figure 17: ROC curve of J-48 for Model 1

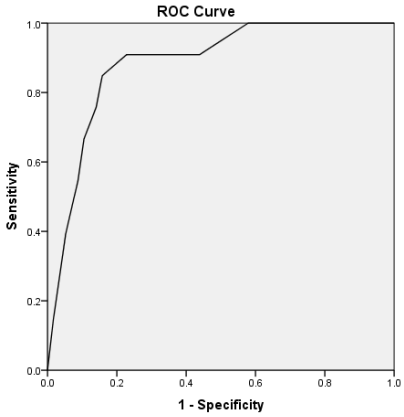


Figure 32: ROC curve of Random Forest for Model 1

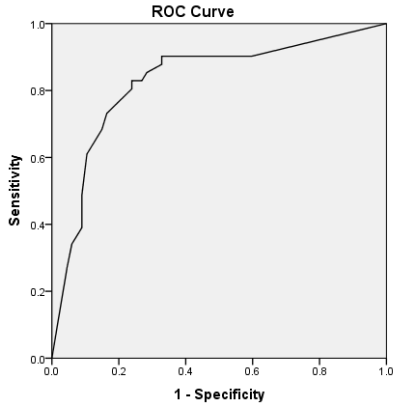


Figure 18: ROC curve of J-48 for Model 2

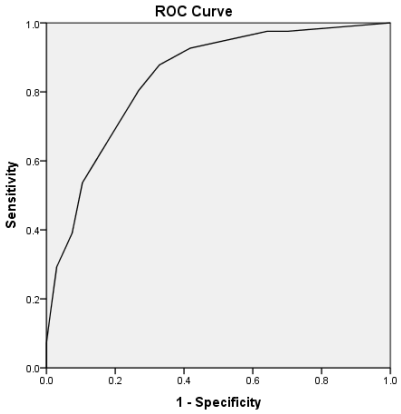


Figure 33: ROC curve of Random Forest for Model 2

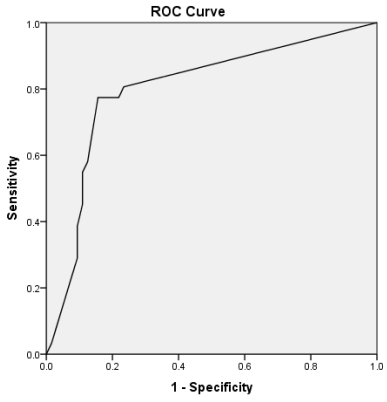


Figure 19: ROC curve of J-48 for Model 3

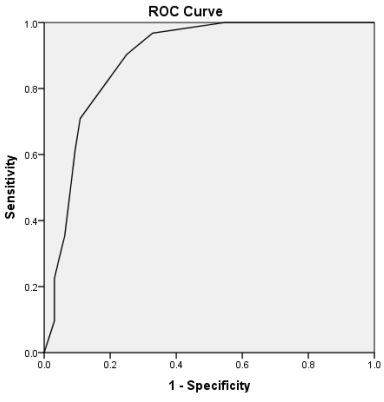


Figure 34: ROC curve of Random Forest for Model 3

5.12 Random Forest Analysis

Table 32 describes the prediction of web pages of all 3 models and Table 33 describes the 10 cross fold validation result of all 3 models. Figure 32, Figure 33 and Figure 34 shows the ROC curves for Model 1, Model 2 and Model 3 respectively.
Observation made from analysis:

- Out of 33 good websites, 28 are correctly predicted and out of 57 bad website, 48 are correctly predicted which gives the sensitivity 84.90 and specificity of 84.20 respectively.
- Out of 41 good websites, 26 are correctly predicted and out of 68 bad website, 56 are correctly predicted which gives the sensitivity 80.50 and specificity of 73.10 respectively.
- Out of 31 good websites, 24 are correctly predicted and out of 64 bad website, 54 are correctly predicted which gives the sensitivity 83.90 and specificity of 79.70 respectively.

Table 32: Website prediction of Random forest for model 1, 2, and 3

PARAMETER	MODEL1	MODEL2	MODEL3
Number of good website correctly predicted	28	26	24
Number of bad website correctly predicted	48	56	54

Table 33: 10-cross fold results using Random forest for model 1, 2, and 3

	SENSITIVITY	SPECIFICITY	CUTOFF	AUC
Model 1	84.90	84.20	.450	.885
Model 2	80.50	73.10	.450	.842
Model 3	83.90	79.70	.350	.891

5.13 Evaluation of model

For dimensionality reduction CFS technique provided in WEKA tool was used, which provide the subset of attributes. When CFS applied to 2010 year data, 21 variables were reduced to 5 variables in which 4 are independent and 1 is dependent. Independent variables in 2010 data are Word Count, Link count, Script Count and List Item Count. Similarly in 2011 year dataset independent variables are Link Count, Script Count, Inline Element Count, Load time, Page title Word Count and Unordered List Count. Similarly in 2012 year dataset independent variables are Word Count, Page size, Script Count, Image Count, Load Time and Paragraph Count.

Observation made from evaluation of models after applying CFS technique

- Script Count is very significant metric in all three year dataset, so it should be consider by designers for the good design of website.
- Word Count is common in 2010 and 2012 year dataset and Link Count is common in 2010 and 2011 year dataset.
- Number of significant metrics either same or increases by the time

Cut-off point of all the models is computed using ROC analysis which maintains a balance between predicted website as good and bad. Area under curve (AUC) of ROC is a measure of combination of sensitivity and specificity and ROC curve is plotted between sensitivity and 1-specificty. So Area under ROC curve is used for computing the accuracy of prediction model.

Table 34 describes the prediction result of 10 machine learning techniques for model 1.Table 35 describes the prediction result of 10 machine learning techniques for model 2.Table 36 describes the prediction result of 10 machine learning techniques for model 3.

Table 34: Prediction results of 10 machine learning techniques of model 1

MACHINE LEARNING TECHNIQUE	SENSITIVITY	SPECIFICITY	CUT-OFF	AUC
Bayes Net	69.70	68.40	.380	.740
Naïve Bayes	66.70	71.90	.8425	.729
Multilayer Perceptron	81.80	82.50	.194	.837
Adaboost	81.80	80.70	.4615	.884
Decision Table	72.70	71.90	.3485	.787
Nnge	69.70	86.00	.50	.778
Part	69.70	71.90	.135	.786
Bf-tree	72.70	71.90	.2285	.751
J-48	72.70	77.20	.1345	.762
Random Forest	84.90	84.20	.450	.885

Table 35: Prediction results of 10 machine learning techniques of model 2

MACHINE LEARNING TECHNIQUE	SENSITIVITY	SPECIFICITY	CUT-OFF	AUC
Bayes Net	73.20	73.10	.703	.770
Naïve Bayes	70.70	71.60	.876	.836
Multilayer Perceptron	68.30	67.20	.5155	.747
Adaboost	80.50	80.60	.5185	.797
Decision Table	70.70	71.60	.6005	.745
Nnge	53.70	79.10	.50	.664
Part	70.70	71.60	.6005	.770
Bf-tree	75.60	76.10	.2915	.781
J-48	80.50	76.10	.7555	.828
Random Forest	80.50	73.10	.45	.842

Table 36: Prediction results of 10 machine learning techniques of model 3

MACHINE LEARNING TECHNIQUE	SENSITIVITY	SPECIFICITY	CUT-OFF	AUC
Bayes Net	74.20	73.40	.341	.862
Naïve Bayes	74.20	75.00	.9575	.841
Multilayer Perceptron	67.70	67.20	.3705	.749
Adaboost	83.00	85.90	.510	.877
Decision Table	80.60	79.70	.3665	86.00
Nnge	58.10	92.20	.500	.751
Part	74.20	81.20	.377	.738
Bf-tree	74.20	75.00	.450	.797
J-48	77.40	78.10	.665	.802
Random Forest	83.90	79.70	.350	.891

Logistic regression and machine learning techniques have been employed to evaluate their performance for predicting the quality of the websites. The AUC of all the models predicted using Random Forest technique is greater than the AUC of all the other models predicted using the logistic regression as well as other machine learning techniques (Bayes Net, Naïve Bayes, Multilayer Perceptron, Adaboost, Decision Table, Nnge, Part, Bf-tree, J-48, and Random Forest).

Model 1 with respect to dataset of 2010 has an AUC of 0.885 using Random Forest technique which is greater than that using other techniques and same trend is seen for the models with respect to dataset of year 2011 and 2012 with the AUC of 0.842 and 0.891 respectively. All the models performed best with Random Forest classifier, which is reflected in their AUC values.

Both the sensitivity and specificity should be high to predict good and bad websites. The models predicted with the Random Forest technique have higher prediction performance in terms of sensitivity and specificity. For Model 1, Random Forest classifier provides the sensitivity of 84.90 and specificity of 84.20. Model 2 has

sensitivity of 80.50 and specificity of 73.10. For Model 2, Random Forest provides the sensitivity and specificity of 83.90 and 79.70, respectively.

Thus, on overall basis in terms of sensitivity, specificity and area under ROC curve, the best model suitable for predicting the class of websites as good or bad is determined to be Random Forest Model. It is said that Random Forest outperforms more sophisticated classifiers on many datasets, achieving impressive results.

Chapter Six: Conclusion and Future Work

Basic goal of this research is to categorize the websites into good and bad on the basis of the web page metrics. Further different machine learning algorithms and logistic regression techniques have been employed to classify websites into good and bad and finally compare the performance of different machine learning algorithms.

So we can finally summarize this work into three sub-parts:

1. 294 websites and their level-1 pages from various category from the pixel awards website of year 2010, 2011 and 2012 have been collected.
2. WEB METRICS CALCULATOR which was developed in ASP.NET used to compute 20 web page metrics for these webpages.
3. Logistic regression and 10 machine learning(Bayes net, Naïve Bayes, Multilayer Perceptron, Adaboost, Decision Table, Nnge, Part, Bf-tree, J-48, Random forest) techniques were applied to classify the website and compare the accuracy of logistic regression and different machine learning techniques.

Result of this report can be summarized as follows:

1. Script Count is very significant metric in all three year dataset, so it should be consider by designers for the good design of website.
2. Most significant metrics in 2010 Word count, Link count, Script count and List item count. In 2011 most significant metrics are Link count, Script count, Inline element count, Load time, Page title word count, unordered list count. In 2012 most significant metrics are Word count, Page size, Script count, Image count, Load time and paragraph count.
3. Performance of Random Forest technique is better than all other machine learning techniques and logistic regression under ROC analysis. Range of Area Under Curve of Random Forest is .842-.891.

6.1 FUTURE WORK

Although this research work is conducted on three year dataset and computed 20 web page metric. Analysis should be done on larger and different datasets as well as with more number of web page metrics to generalize our result. Further this research work should extend for all level web pages instead of only 1-level pages and define new web page metrics.

REFERENCES

(n.d.). Retrieved october 14, 2012, from www.pixelawards.com.

Americo, R. (2010). Websites Quality: Does It depend on the application Domain ? *International Conference on the quality of Information and Communications Technology.*

Anderson, J. A. (2003). *An Introduction to Neural Networks.* Prentice Hall.

Bray, T. (May,1996). Measuring the web. *5th International World Wide Web Conference.* Paris,France.

c4.5 algorithm. (n.d.). Retrieved may 24, 2013, from e.wikipidea.org: http://en.wikipedia.org/wiki/C4.5_algorithm

Calero, C., Ruiz, J., & Piattini, M. (2005). Classifying web metrics using the web quality model. *Emerald Group Publishing*, 227- 248.

Chi, E. H., Pirroli, P., & Pitkow, J. (2000). The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a web site. *ACM CHI 00 Conference on Conference on Human Factors in Computing Systems.*

Dhyani, D., Ng, W., & Bhowmik, S. (2002). A survey of web metrics. *ACM Computing Surveys*.

Drott, M. C. (1998). Using web server logs to improve site design. *16th International Conference on Systems Documentation*, (pp. 43-50).

Fawcett, T. (2005, december 19). *An Introduction to ROC analsis.* Retrieved june 06, 2013, from http://people.inf.elte.hu/: http://people.inf.elte.hu/kiss/13dwhdm/roc.pdf

Fink, D. (2001). Web Site Effectiveness: A Measure of Information and Service Quality. *IRMA international conference.*

Friedman, V. (2008, january 31). *10 Principles Of Effective Web Design*. Retrieved june 25, 2013, from http://uxdesign.smashingmagazine.com/: http://uxdesign.smashingmagazine.com/2008/01/31/10-principles-of-effective-web-design/

Fuller, R., & Graff, J. D. (1996). Measuring user motivation from server log files. *Human Factors and the Web 2 Conference.*

Grossman, D., & Domingos, P. (2004). *mlc04.* Retrieved june 02, 2013, from homes.cs.washington.edu: http://homes.cs.washington.edu/~pedrod/papers/mlc04.pdf

Group, M. W. (2005). *qualitycommentary050314final.* Retrieved from http://www.minervaeurope.org/: http://www.minervaeurope.org/publications/qualitycommentary/qualitycommentary050314final.pdf

Haijan, S. (2007). Best-first Decision Tree Learning. New Zealand.

Hall, M. A. (1999). Correlation-based Feature Subset Selection for Machine Learning. New Zealand.

Ivory, M. Y., Sinha, R., & Hearst, M. (2000). Preliminary findings on quantitative measures for distinguishing highly rated information-centric web pages. *6th Conference on Human Factors and the Web.*

Ivory, M. Y., Sinha, R., & Hearst, M. A. (2001). Empirically Validated Web Page Design Metrics. *SIGCHI conference on Human factors in computing systems*, (pp. 53-60). Washington.

Khan, K. M. (2008). Assessing Quality of Web Based System. *IEEE/ACS International Conference on Computer Systems and Applications* (pp. 763-769). AICCSA.

Leung, K. M. (2007). *naiveBayesianClassifier.pdf.* Retrieved 06 04, 2013, from naiveBayesianClassifier.pdf: http://cis.poly.edu/~mleung/FRE7851/f07/naiveBayesianClassifier.pdf

Li, P., & Yamada, S. (2009). Automated Web Site Evaluation – An Approach Based on Ranking SVM. *International Joint Conferences on Web Intelligence and Intelligent Agent Technologies.* IEEE/WIC/ACM.

Li, P., & Yamada, S. (2010). Extraction of Web Site Evaluation Criteria and Automatic Evaluation. *Journal of Advanced Computational Intelligence and Intelligent Evaluation*.

Matas, J., & Sochman, J. (n.d.). *Adaboost_matas.* Retrieved 06 06, 2013, from http://www.robots.ox.ac.uk/: http://www.robots.ox.ac.uk/~az/lectures/cv/adaboost_matas.pdf

Melody, I. y. (2001). An empirical foundation for automated web interface evaluation. *ACM digital library*.

Mendes, E., Mosley, N., & Counsell, S. (2003). Early web size measures and effort prediction for web costimation. *IEEE International Software Metrics Symposium(METRICS'2003)* (pp. 18-29). Sydney: IEEE CS Press.

Mich, L., Franch, M., & Gaio, L. (2003). Evaluating and Designing the Quality of Web Sites. *IEEE Multimedia* (pp. 34-43). Ieee computer society.

Montillo, A. A. (2009, february 04). *Montillo_RandomForests_*. Retrieved may 15, 2013, from http://www.dabi.temple.edu/: http://www.dabi.temple.edu/~hbling/8590.002/Montillo_RandomForests_4-2-2009.pdf

Nielsen, J. (1999). User interface directions for the Web. *Communications of ACM*, 65-72.

Nielsen, J. (2000). *Designing Web Usability: The Practice of Simplicity.* Indianapolis: New Riders Publishing.

Olsina, L., & Rossi, G. (2002). Measuring Web Application Quality with WebQEM. *Ieee computer society*, 20-29.

Pitkow, J. (1997). In search of reliable usage data on the WWW. *6th World Wide Web Conference.*

Pollilo, R. (2005). Un modello di qualità per i siti web. *AICA*, 32-44.

Scholtz, J., Laskowski, S., & Downey, L. (1998). Developing Usability Tools and Techniques for Designing and Testing Web Sites. *4th Conference on Human Factors & the Web.*

Shedroff, N. (1999). *Recipe for a successful web site*. Retrieved may 14, 2013, from www.nathan.com: http://www.nathan.com/thoughts/recipe

Signore, O. (2005). A comprehensive model for Web sites quality. *Seventh IEEE International Symposium on Web Site Evolution*, (pp. 30-38). Budapest.

Thimbleby, H. (1997). Gentler: A tool for systematic web authoring. *International Journal of Human-Computer Studies*, 139-168.

Velayathan, G., & Yamada, S. (2006). Behavior-Based Web Page Evaluation. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. (WI-IAT Workshops.

Witten, I. H., & Frank, E. (1998). Generating Accurate Rule Sets Without Global Optimization. In *Generating Accurate Rule Sets Without Global Optimization.* Department of Computer Science, University of Waikato, 1998.

Zaharie, D., Perian, L., & Negru, V. (2011). A VIEW INSIDE THE CLASSIFICATION WITH NON-NESTED GENERALIZED EXEMPLARS. *IADIS European Conference Data Mining.*

Zorman, M., Podgorelec, V., Kokol, P., & Babic, S. H. (1999). Using machine learning techniques for automatic evaluation of Websites. *Third International Conference on Computational Intelligence and Multimedia Applications ICCIMA* (pp. 169-173). New Delhi: IEEE Computer Society Press.