

**WRAP: Classification of Appropriate User's Age for Website Restriction
through Metadata utilizing Data Mining and Naïve Bayes Classification**

A Thesis

Presented to the Faculty of the
College of Computer and Information Sciences
Polytechnic University of the Philippines

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Science in Computer Science

Española, Beverly Dianne D.

Fajiculay, Christian M.

April 2017

AUTHORIZATION

We hereby declare that we are the sole authors of the thesis.

We authorize the Polytechnic University of the Philippines and the College of Computer Management and Information Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

We further authorize the Polytechnic University of the Philippines and the College of Computer Management and Information Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Española, Beverly Dianne D.

Fajiculay, Christian M.

April 2017

AFFIDAVIT OF CO-OWNERSHIP

AFFIDAVIT ON COPYRIGHT CO-OWNERSHIP

REPUBLIC OF THE PHILIPPINES
City of Manila

The undersigned, being duly sworn, do solemnly affirm and say:

1. That they are the creators/authors of the work entitled WRAP: Classification of Appropriate User's Age for Website Restriction through Metadata utilizing Data Mining and Naïve Bayes Classification which is the subject of the application for copyright filed on _____;
2. That they have presented and submitted the said work to the Polytechnic University of the Philippines as part of the requirements of their curricular program: College of Computer and Information Sciences, Bachelor of Science in Computer Science.
3. That upon submission of the work to the Polytechnic University of the Philippines, they acknowledge the University as COPYRIGHT CO-OWNER of the above-mentioned work; and
4. That as copyright co-owner, PUP has the right to reproduce, publish and publicly distribute copies of the said work, in whatever form, electronic or otherwise for instruction purposes, and for any other purposes that promote access to and utilization of intellectual property of the University; provided that the names of the authors shall be acknowledged/cited in all dissemination and utilization of the work.

In witness whereof, the authors/creators hereunto set their signature on this 24 day of April, year 2017, in Manila, Philippines.

Authors'/Creators' Name

Signature:

Española, Beverly Dianne D.
Fajiculay, Christian M.
Montaril, Ranil M., MSECE

[Signature]
[Signature]

Subscribed and sworn to or affirmed before me this _____ day of _____, 2017.

Doc. No. 109
Page No. 2
Book No. IX
Series of 2A

NOTARY PUBLIC

[Signature]
ATTY. DANIEL F. FURAQUE
Notary Public, City of Manila, 4409 Old Sta. Mesa, Manila
PTR No. MLA5991976-12/29/2016 (year 2017)
IBP No. 1005865-11/03/2016 (years 2017-2018)
Rol# No. 50723: Commission No. 2017-059;
MCLE Compliance No. V-0022538-April 14, 2019
My Commission Expires on Dec 31, 2018

CERTIFICATION OF PROOFREAD

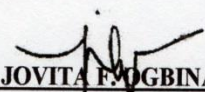
Certification of Proofread

This is to certify that the thesis work entitled **“WRAP: Classification of Appropriate User's Age for Website Restriction through Metadata utilizing Data Mining and Naïve Bayes Classification”** by **Beverly Dianne D. Española and Christian M. Fajiculay** was proofread and edited by the undersigned.

This certification is being issued for whatever legal purpose it may serve.



Signed:


JOVITA P. JOBINAR
 (Signature over Printed Name)

Date:

April 5, 2017

SIGNATURE

The thesis “**WRAP: Classification of Appropriate User’s Age for Website Restriction through Metadata utilizing Data Mining and Naïve Bayes Classification**” submitted and presented by Beverly Dianne D. Española and Christian M. Fajiculay in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science has been

Recommended for Approval and Acceptance:

Date:

Prof. Iluminada Vivien R. Domingo, DBA
Faculty-In-charge

Prof. Ranil M. Montaril, MSECE
Thesis/Technical Advisor

Approved by the Thesis Defense Panel:

Prof. Gisela May A. Albano, MIT

Prof. Mary Jane Magno-Tan, MIT

Prof. Carlo G. Inovero

Accepted for the Department of Computer Science

Prof. Melvin C. Roxas, MSGITS
Department Chairperson

Accepted for the College of Computer and Information Sciences

Prof. Gisela May A. Albano, MIT
Dean of the College

ABSTRACT

The researchers developed a system, WRAP: Classification of Appropriate User's Age for Website Restriction through Metadata utilizing Data Mining and Naïve Bayes Classification, which classifies and analyzes the required age for a website. The researchers come up with the idea of creating the system because of the growing amount people, of all ages, who browses the internet, not realizing the content may be harmful for them. Therefore the researchers thought of an idea in creating a system that will classify a website, crawl its relative links, determine what is the age required and if the age of the user meets that age to access the website, the entered URL will be loaded, otherwise the user will be restricted from accessing the website.

The researchers used Simple HTML Parser as the tool to extract the URL and contents of a website. Once extracted, the URL and the contents will be classified, whether positive or negative. This is through Naive Bayes Classification. The advantage of using Naive Bayes Classification is that, even with less training data, the results are accurate. Right after classifying, the system now, identifies the required age for the website with Fuzzy Logic. From the classification as the input of the Fuzzy Logic System, an output is generated either 18 and above only or for all. If the user's age meets the required age to access a website, he/she can browse through it, otherwise he/she will be restricted to access it.

The study proved that the system was very highly accurate with an 86% of accuracy from the results of the Precision, Recall and the F-Measure from the test data. It also has 100% reliability rate. And lastly, the experiment also proved that, there is no statistical significant difference between the expert's assessment and the system's assessment.

ACKNOWLEDGEMENT

This research will not be possible without the help of these people:

Engr. Ranil M. Montaril for his great dedication in guiding us in the documentation side of the thesis. We also thank him for his undying support and suggestions for the betterment of our research. And also for all the materials he gave us that became one of our great references.

We also thank Dr. Rosa Maria Nancho, for giving us information about child psychology and her support for the importance of the study. And Ms. Jasmin A. Bascos, for helping us in evaluating and assessing our system.

To our esteemed panelists and faculties, we are very thankful for deliberating their comments, assessing both the papers and tool and suggesting on how to improve the study.

To our family, friends and classmates for the undying support, financial needs and for encouragement to do our best to finish this research. To all team Ran 2017, for joining and being a great part for this journey.

And most of all, we thank the guidance of the Almighty God for giving us the strength and courage to finish our research.

B.D.D.E.

C.M.F.

TABLE OF CONTENTS

AUTHORIZATION	ii
AFFIDAVIT OF CO-OWNERSHIP	iii
CERTIFICATION OF PROOFREAD	iv
SIGNATURE	v
ABSTRACT	vi
ACKNOWLEDGEMENT	vii
TABLE OF CONTENTS	viii
LIST OF FIGURES	x
LIST OF TABLES	xi
LIST OF NOTATION.....	xii
CHAPTER 1 The Problem and Its Background	1
1.1 BACKGROUND OF THE STUDY	1
1.2 STATEMENT OF THE PROBLEM.....	4
1.3 HYPOTHESIS.....	4
1.4 THEORETICAL/CONCEPTUAL FRAMEWORK	4
1.4.1THEORETICAL FRAMEWORK	4
1.4.2CONCEPTUAL FRAMEWORK	10
1.5 SIGNIFICANCE OF THE STUDY	11
1.6 SCOPE AND LIMITATION OF THE STUDY	11
1.7 OPERATIONAL TERMS	12
CHAPTER 2 Review of Related Literature	13
2.1. RELATED LITERATURE.....	13
2.2. RELATED STUDIES.....	23
2.3. SYNTHESIS OF THE STUDY.....	34
CHAPTER 3 Research Methodology	35
3.1. RESEARCH DESIGN.....	35
3.2. SOURCES OF DATA	35
• Population	35
• Respondents.....	36
• Sampling Technique	36
3.3. INSTRUMENTATION	36

3.3.1. SOFTWARE/HARDWARE TOOLS	36
3.3.1.1. SYSTEM ARCHITECTURE	36
3.3.1.2. DEVELOPMENT DETAILS	39
3.3.2 RESEARCH INSTRUMENT	39
3.4. DATA GENERATION/GATHERING PROCEDURE	39
3.5. STATISTICAL TREATMENT OF DATA	40
3.5.1. Accuracy	40
3.5.2. Reliability	41
3.5.3. Expert vs. System	42
CHAPTER 4 Presentation, Analysis and Interpretation of Data	44
CHAPTER 5 Summary, Conclusions and Recommendations	50
5.1. Summary of Findings/Results	50
5.2. Conclusions	51
5.3. Recommendation	51
REFERENCES	53
APPENDIX	57
APPENDIX A: SAMPLE RESEARCH INSTRUMENT	57
APPENDIX B: COMMUNICATIONS	62
APPENDIX C: RAW DATA	64
APPENDIX D: SCREENSHOTS.....	70
APPENDIX E: IMPLEMENTATION REPORT	73
APPENDIX F: CURRICULUM VITAE.....	76

LIST OF FIGURES

Figure 1.1: Data mining as a step in the process of knowledge discovery [Han et al. 2012].	5
Figure 1.2: The stages of analysis in processing natural language [Indurkha and Damerau 2010].	6
Figure 1.3: Supervised Text Classification [Natural Language Toolkit 2008].	7
Figure 1.4: Structure of Bayesian Network [Witten et al. 2011].	8
Figure 1.5: Roadmap for the fuzzy inference process [MathWorks 2016].	9
Figure 1.6: Theoretical Framework of WRAP	10
Figure 1.7: Conceptual Framework of WRAP	10
Figure 2.1: List of age restriction for some of the major social sites online [Sheppard 2016].	17
Figure 2.2: Internet Growth in the Philippines [The Rappler 2016].	19
Figure 2.3: Multi-level classification conceptual framework [Amplayo and Occidental 2015].	25
Figure 2.4: Style vs Content: Accuracy from 1975-1988 for Style (Online-Behavior+Lexical-Stylistic) vs Content (BOW) [Rosenthal and McKeown 2015].	30
Figure 2.5: Style and Content: Accuracy from 1975-1988 using BOW, Online Behavior, and Lexical Stylistic features [Rosenthal and McKeown 2015].	30
Figure 2.6: Age Prediction scores per class [van de Loo et al. 2016].	33
Figure 2.7: Recall, Precision and F-measure using different n-gram	34
Figure 3.1: System Architecture of WRAP	36
Figure 3.2: Membership Function for Content.	37
Figure 3.3: Membership Function of URL.	38
Figure 3.4: Membership Function of Age.	38
Figure 3.5: Chi-Square Distribution Table	43
Figure 4.1: Accuracy of WRAP in predicting the appropriate age of 2 age groups.	46
Figure 4.2: Reliability of WRAP in predicting the age of two age group.	47

LIST OF TABLES

Table 2.1: Different methods of Age Verification in UK [Lindley 2015].	14
Table 2.2: Classification of Contents as classified in Germany [Lindley 2015].	15
Table 2.3: Comparison of best performing classifiers using all features and using feature-reduced datasets [Tighe 2016].	24
Table 2.4: Evaluation of Different Classifiers [Amplayo and Occidental 2015].	25
Table 2.5: Results generated from the WEKA classifier using RIPPER algorithm applied to classify Phishing emails [Ferolin R.J. 2011].	27
Table 2.6: Results of Phishing Pages removed after notifications were sent [Ferolin R.J. 2011].	27
Table 2.7: Most important features in the JOINT model with all features (condition 10) [Nguyen 2011].	28
Table 2.8: Feature Accuracy [Rosenthal and McKeown 2015].	31
Table 2.9: Results for data sets [Peersman 2011].	32
Table 3.1: Verbal Interpretation of Percentage of Accuracy	41
Table 3.2: 2x2 Contingency Table	42
Table 4.1: Confusion Table for Determining the Accuracy in Required Age	45
Table 4.2: System's Accuracy rate for Precision, Recall and F-Measure for Respondents with Ages of 17 & below	45
Table 4.3: System's Accuracy rate for Precision, Recall and F-Measure for Respondents with Ages of 18 & above	46
Table 4.4: System's Accuracy, Precision, Recall and F-measure of WRAP	47
Table 4.5: Summary of the Cases for WRAP	48
Table 4.6: Reliability Statistics of WRAP	48
Table 4.7: Result of the System's Output Compared to the Experts Evaluation using Chi-Square	49

LIST OF NOTATION

Symbol	Meaning
d	a document
c	a class; overrated, underrated and neutral
Σ	Summation
$P(d c)$	Probability of d given c
NLP	Natural Language Processing
FL	Fuzzy Logic
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative

LIST OF EQUATIONS

Equation 1: Formula for Precision	40
Equation 2: Formula for Recall.....	40
Equation 3: Formula for F-Measure	40
Equation 4: Formula for Accuracy	41
Equation 5: Equation for Cronbach's Alpha Test.....	41
Equation 6: Chi-Square	42

CHAPTER 1

The Problem and Its Background

1.1 BACKGROUND OF THE STUDY

The early internet was used by computer experts, engineers, scientists, and librarians. But today, everyone knows how to use and access the internet, old, young, boy, girl, and has changed the world we live in. Never before has it been so easy to access information; communicate with people all over the globe; and share articles, videos, photos, and all manner of media. Moreover, this increased popularity is starting to cause a problem, as it is proving a struggle to keep up with demand. The internet uses a lot of a country's power output; like in Britain, wherein it consumes eight percent of the country's power and the demand is only to grow every year, with the worry of it using all of Britain's power supply by 2035 [The Works 2015]. In the Philippines, the median age that's using the internet is 24 and consumes about 150k terabytes of data annually, the millennial who grew up as digital natives. The Filipinos thrive on staying connected with their communities because we need real-time information to make the right choices, especially during time of crisis [The Rappler 2016].

The Internet can be a wonderful resource for kids. They can use it to research school reports, communicate with teachers and other kids, and play interactive games. Kids who are old enough to swipe a screen can have access to the world [KidsHealth]. Researches about child and how the internet can affect them highlighted that; they get into trouble online all the time, even when they aren't looking for it. And could affect their lives, specifically on mental health because of these inappropriate contents of the internet that may occur also in social media, blogs or forums also [Geier 2013] [Internet Matters 2013] [Caroll 2011] [Gonzales 2014]. "The fact is that the growth of Internet as unregulated space has thrown up two major challenges when it comes to protecting our children." as quoted by David Cameron - 22nd July 2013 to the NSPCC in London, UK [Cameron 2013]. The first challenge is illegal material; proliferation and accessibility of child abuse images in the internet. The second challenge is legal; cultural - the fact that many children are viewing online pornography and other damaging material at a very early age.

Even children as young as 4 or 5 years old have the capacity to access information all over the internet as technology becomes more affordable and more accessible. There are specific instances wherein children can be exposed to pornography, and to different degrees, be directly sexually abused over the internet. Access to violent pornography and the more mainstream pornography websites can be encountered by children through a simple search done through any of the popular search engines (Google, Yahoo, etc.) Although offering age based "safeguards" for access, these safeguards are utterly useless, as a 10 year old can easily click on the "I am 18

years old or older” button, pass the validation process and be exposed to all sorts of pornography the web has to offer [Diloy 2013]. The cyberbullying statistics in the Philippines shows that the age group of those who said they are bullied are adult (18+) with 53% and minor (17 and below) with 47%. 57% of this is female and the rest is male. The top most object of attack is attack on reputation. And the nature of attack are spreading photoshopped images [ASKSonnie 2015]. Another concern is that the internet contains adult contents. In a recent survey conducted, the Philippines ranked 1st in the world for time spent watching porn. Further data reveals that majority or 42% of Filipino visitors are aged 18 to 24, followed by 25 to 34 at 31% [Chan 2016]. Still, leaving a 28% of other age group viewers that may be classified as minors. "But both the challenges have something in common; they're about how our collective lack of action on the internet has led to harmful and, in some cases, truly dreadful consequences for children." as continued by Dr. Cameron [Cameron 2013]. The Philippines Cybercrime law, RA 10175 or Cybercrime Prevention Act of 2012 was signed by former Pres. Aquino on September 12, 2012 and agreed by the House of Representatives. This law states that, the government will have a power to monitor and shutdown private internet properties, criminalizes computer crime and impose rules related to online activities [The Summit Express 2012].

There is a growing interest in automatically predicting the gender and age of authors from texts. However, most research so far ignores that language use is related to the social identity of speakers, which may be different from their biological identity [Nguyen 2014]. The web is the huge storage of network-accessible information, and knowledge. As humans, we have a knack for estimating another person's age quite accurately just by glancing at their face. Although age estimation may seem relatively simple to us, computers have a much more difficult time performing the task [Zyga 2011]. Another common approach of uncovering hidden user attributes in social media is to model the writing habits of users by extracting various features from texts they have posted. This approach, however, suffers from the inability of models generated from one genre of social media to be successfully applied to other genres in some cases [Marquardt et al. 2014].

The web pages are continuously increasing in volume and complexity with time so it is going to be difficult to extract the valuable relevant information from internet [Parmar 2016]. And to solve this complexity, data mining could be used. Data Mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness

metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

And in relation with age verification that uses data mining, some systems in different country were invented. The UK Government has started to exert pressure on the ISPs to voluntarily put a basic “shield” in place. When it comes to higher levels of assurance around age, the UK has led the way for some years with the development of online identity and age verification systems. The major providers in the UK have extended their services to cover international markets around the globe. The introduction of the e-ID card in Germany, though, offers a way for age verification systems to be simplified (in Germany), both for the provider and the customer. It has the potential to be the universal approach by 2020 [Lindley 2015].

An interview with Dr. Rosa Maria Nancho, an adolescent medicine doctor of Manila Doctors Hospital was conducted by the researcher. She agreed that there is an importance of implementing this study since, as the internet a source of information can also affect an individual. She also pointed that a child is very impressionable; they don't know what is the reality and artificial world. So what they see, they copy it and may harm them. Moreover, an individual's brain is not fully developed until mid-20 as questioned on what is her insight with the research problem and when does a person's learning in its full development. There are some psychological tests that could determine the mental age of an individual, she added. With this statement, the problem of identifying the users age is now resolved [Nancho 2016].

With the development of age verification systems, data mining, extraction of information is achieved easily nowadays and making the research problem solvable. At first, the user will register and input his/her information such as email, username, password and birthday and login to the system. The user will then input the URL in the box. The system will then determine the polarity of the URL whether it is negative, neutral or positive and assign a score. Simultaneously, the content of the URL will be extracted and analyze using Naïve Bayes Classification to determine whether the content is positive or negative. From the given variables (URL Polarity and Content Polarity) the appropriate age that could access the website can be identified. Then after, the appropriate age to access the website and the user's age is to be determined whether to restrict the user from accessing it or not. This concludes a solution in website restriction through prediction of age in determining the age appropriate of a certain website when accessed.

So in solving the problem, the researchers realized a need to develop “WRAP: Classification of Appropriate User's Age for Website Restriction through Metadata utilizing Data Mining and Naïve Bayes Classification” to provide accurate and reliable results in determining the appropriate age of a website and the user's age when accessed and make a corresponding act; the restriction of the users from access to a website if restrictions are met.

1.2 STATEMENT OF THE PROBLEM

The study WRAP aims to develop a system that restrict a user from a website by determining the user's age and the age appropriate to access a certain website through its metadata by utilizing data mining and Naïve Bayes Classification and further, restricts user if not suitable for them.

Specifically, it aims to answer the following questions:

1. What is the accuracy of the system when getting the predicted age appropriate in accessing the website using Precision, Recall and F-measure?
2. What is the reliability of the system in resulting the predicted age in accessing the website?
3. Is there a significant difference in the expert's assessment on predicting the age appropriate for accessing a website and our system's resulted age in terms of accuracy?

1.3 HYPOTHESIS

Null Hypothesis:

H_0 : There is no significant difference in the expert's assessment on predicting the age appropriate for accessing a website and our system's resulted age in terms of accuracy.

1.4 THEORETICAL/CONCEPTUAL FRAMEWORK

1.4.1 THEORETICAL FRAMEWORK

The study consisted of different principles and concepts that the system utilized especially in terms of functions. These principles supported the study's research. The following are the concepts used in the study:

- **Data Mining**

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. Data mining is a step by step process of knowledge discovery, as shown figuratively in Figure 1.1, which is an iterative sequence of the following steps:

1. Data cleaning – to remove noise and inconsistent data.

2. Data integration – where multiple data sources may be combined.
3. Data selection – where data relevant to the analysis task are retrieved from the database.
4. Data transformation – where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. Data mining – an essential process where intelligent methods are applied to extract data patterns
6. Pattern evaluation – to identify the truly interesting patterns representing knowledge based on interestingness measures
7. Knowledge presentation – where visualization and knowledge representation techniques are used to present mined knowledge to users

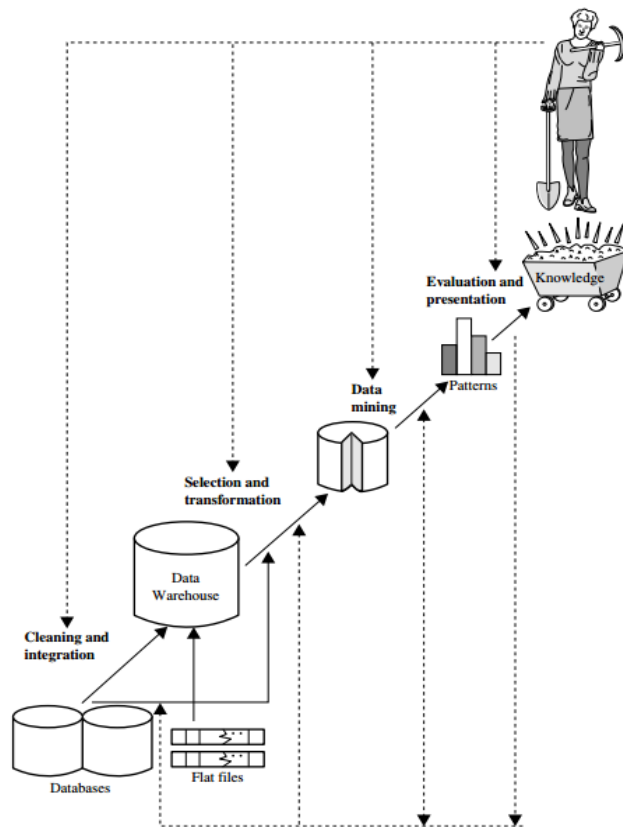


Figure 1.1: Data mining as a step in the process of knowledge discovery [Han et al. 2012].

- **Natural Language Processing**

Natural Language Processing (NLP) has tended to view the process of language analysis as being decomposable into a number of stages, mirroring the theoretical linguistic distinctions drawn between syntax, semantics and pragmatics. Figure 1.2 shows the stages of analysis in processing natural language. The simple view is that the sentences of a text are first analyzed in terms of their syntax; this provides an order and structure that is more amenable to an analysis in terms of semantics, or literal meaning; and this is followed by a stage of pragmatic analysis whereby the meaning of the utterance or text in context is determined [Indurkha and Damerau 2010].

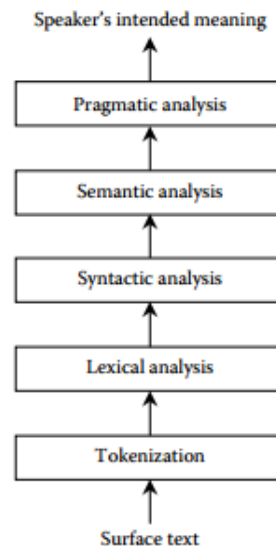


Figure 1.2: The stages of analysis in processing natural language [Indurkha and Damerau 2010].

Text Classification – is the task of choosing the correct class label for a given input. A classifier is called supervised if it is built based on training corpora containing the correct label for each input as shown in the figure below. (a) During training, a feature extractor is used to convert each value to a feature set. Pairs of feature sets and labels are fed into the machine learning algorithm to generate a model. (b) During prediction, the same feature extractor is used to convert unseen inputs to feature sets. These feature sets are then fed into the model, which generates predicted labels. [Natural Language Toolkit 2008]

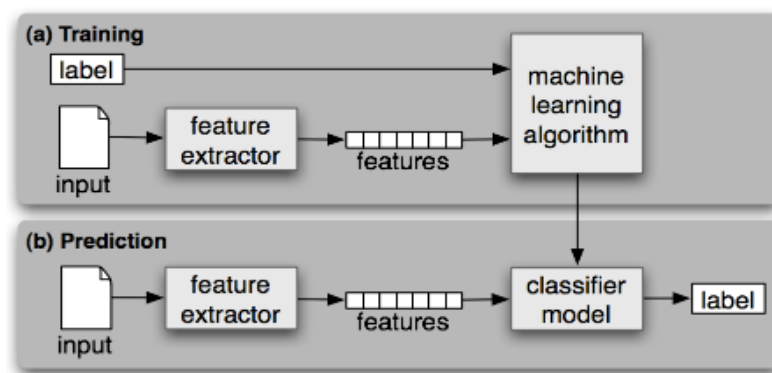


Figure 1.3: Supervised Text Classification [Natural Language Toolkit 2008].

Sociolinguistics – This combines anthropology, statistics and linguistic studies linguistic data in order to answer key questions about the relationship of language and society. Sociolinguists focus on frequency and patterns in linguistic usage, correlations, strength of factors and significance, which together reveal information about the sex, age, education and occupation of speakers/writers but also their history, culture, place of residence, social relationships and affiliations. Sociolinguistics has long investigated the interplay of demographic factors and language use, and it seems likely that the same factors are also present in the data we use to train NLP systems [Tagliamonte 2013].

N-gram – assigns probabilities to sentences and sequences of words. It is a sequence of N words: a 2-gram (or bigram) is a two-word sequence of words and a 3-gram (or trigram) is a three word sequence of words. Features are selected through N-gram. N-grams are used for a variety of different task and it could also be used for developing features for supervised Machine Learning models such as Naïve Bayes, SVM's, MaxEnt models, etc. The idea is to use tokens such as bigrams in the feature space instead of just unigrams [Kavita Ganesan 2011].

- **Bayesian Network**

A theoretically well-founded way of representing probability distributions concisely and comprehensibly in a graphical manner; the structures are called Bayesian networks [Witten et al. 2011]. The Bayesian Network are graphical models for reasoning under uncertainty, where the nodes represent variables (discrete or

continuous) and arcs represent direct connections between them. These direct connections are often causal connections. In addition, BNs model the quantitative strength of the connections between variables, allowing probabilistic beliefs about them to be updated automatically as new information becomes available.

The structure, or topology, of the network should capture qualitative relationships between variables. In particular, two nodes should be connected directly if one affects or causes the other, with the arc indicating the direction of the effect [Korb and Nicholson 2004]. Figure 1.3 is an illustration of the structure of the Bayesian network.

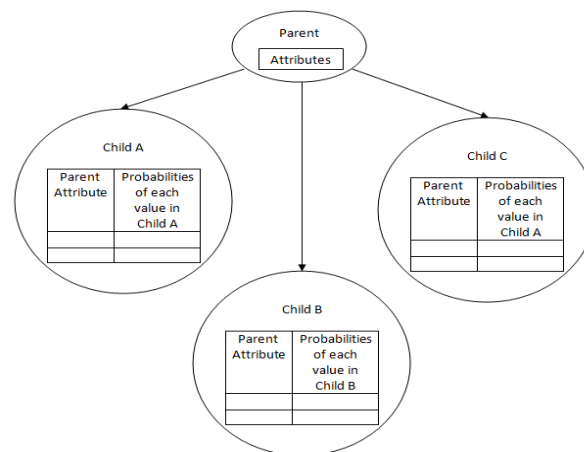


Figure 1.4: Structure of Bayesian Network [Witten et al. 2011].

Naïve Bayes Classification – the Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

- **Fuzzy Logic**

Fuzzy logic is a problem-solving control system methodology that lends itself to implementation in systems ranging from simple, small, embedded micro-controllers

to large, networked, multi-channel PC or workstation-based data acquisition and control systems. It can be implemented in hardware, software, or a combination of both. Fuzzy Logic provides a simple way to arrive at a definite conclusion based upon vague, ambiguous, imprecise, noisy or missing input information. The point of fuzzy logic is to map an input space to an output space, and the primary mechanism for doing this is a list of if-then statements called rules. All rules are evaluated in parallel, and the order of the rules is unimportant. The rules themselves are useful because they refer to variables and the adjectives that describe those variables. Before you can build a system that interprets rules, you must define all the terms you plan on using and the adjectives that describe them. To say that the water is hot, you need to define the range that the water's temperature can be expected to vary as well as what we mean by the word hot. The following diagram provides a roadmap for the fuzzy inference process. It shows the general description of a fuzzy system on the left and a specific fuzzy system on the right.

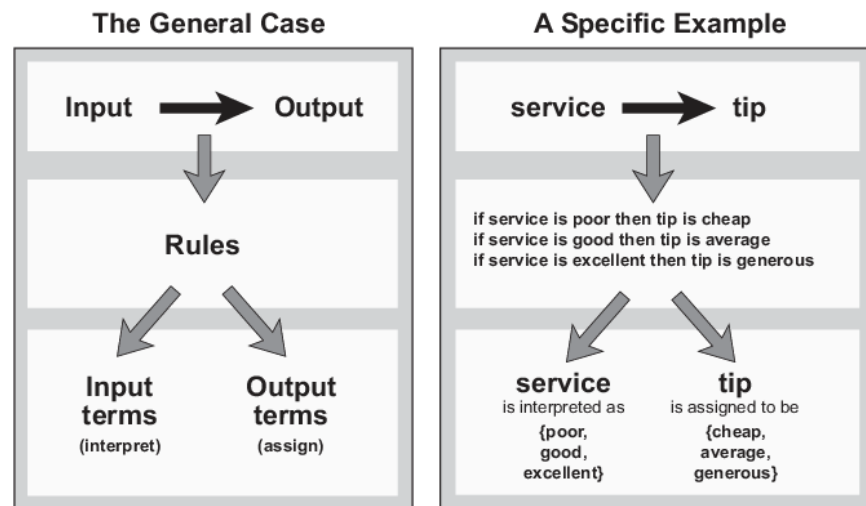


Figure 1.5: Roadmap for the fuzzy inference process [MathWorks 2016].

In the study, the input website undergoes, Data mining, specifically the web. Next phase is Natural Language Processing together with the Character N-gram and Naïve Bayes Classification that includes the pre-processing and classification of the input website. And finally, Fuzzy Logic that predicts the age appropriate in accessing that input website. Figuratively this is shown in Figure 1.6.

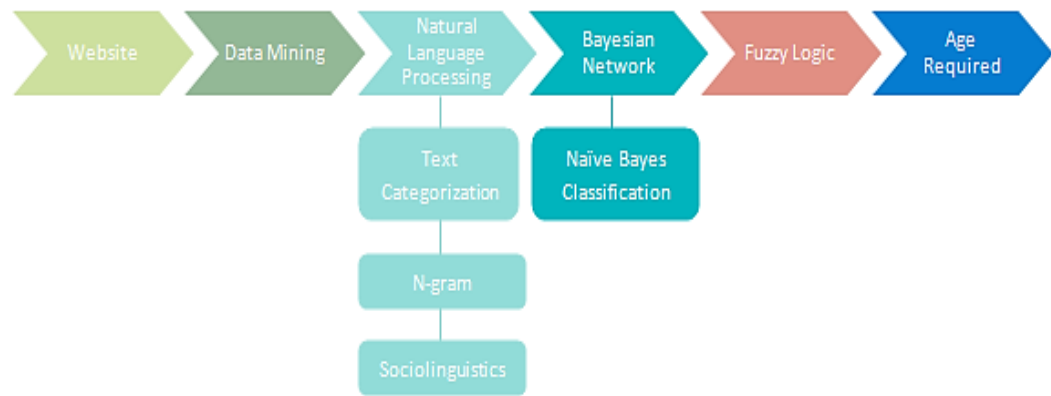


Figure 1.6: Theoretical Framework of WRAP

1.4.2 CONCEPTUAL FRAMEWORK

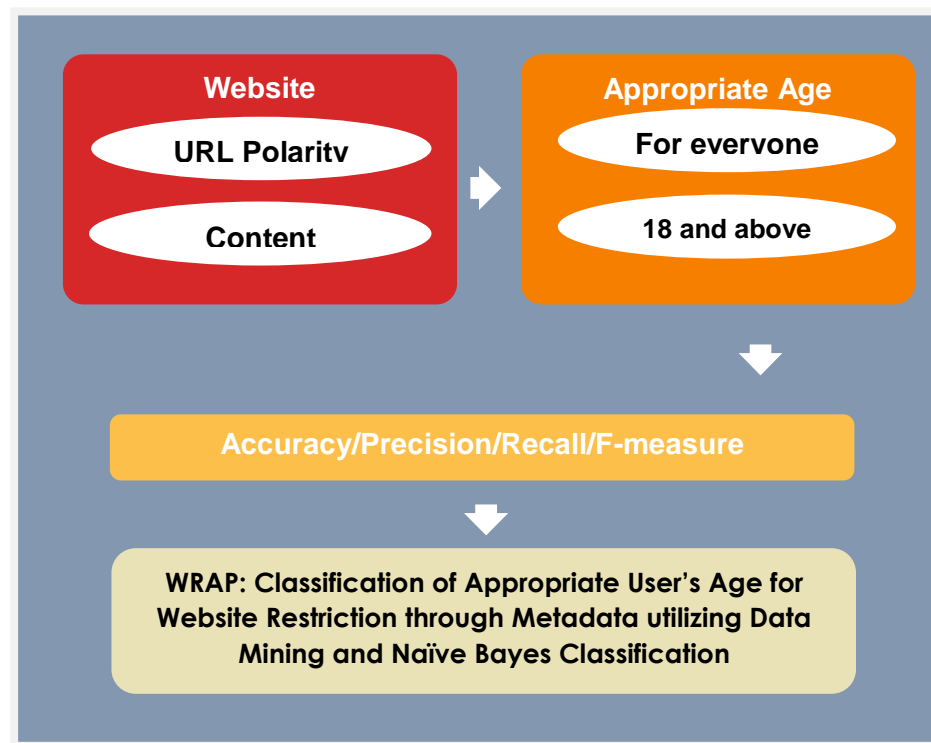


Figure 1.7: Conceptual Framework of WRAP

The Figure 1.7 above illustrated the conceptual framework of WRAP. At first, a website is needed as an input of the system. Then, the Uniform Resource Locator (URL) polarity and Website text contents are gathered. These parameters will serve as the independent variable, and may affect the dependent variable which is the appropriate age that could access a certain

website. Then, the accuracy/precision/recall/f-measure of the system will be evaluated as for the final output of the study, WRAP: Classification of Appropriate User's Age for Website Restriction through Metadata utilizing Data Mining and Naïve Bayes Classification.

1.5 SIGNIFICANCE OF THE STUDY

The study is significantly important in the field of Data Mining since it tackles in acquiring data from the Web and Natural Language Processing for the analysis of the data gathered. And classification of the URL and web contents through the use of character trigram and Naïve Bayes Classification, respectively. This study also tackles Fuzzy Logic theory for identifying the age after classifying.

A brief description of significance of this study will be provided to the following:

Youths: The primary beneficiary of this study. The youths will benefit most from this study since they will be cautioned whether the website they are visiting may contain unwanted or malicious content.

Parents and/or Guardians: Another beneficiary of this study. This study would help them monitor their children's safety against unwanted or malicious sites and will serve as guiding tool for them.

School: Most schools for Secondary and Tertiary provide internet access for students, this will help them to restrict or filter student's access against inappropriate information and contents.

Future Researchers: This will be a big help in understanding more about Natural Language Processing since the study is about Data Mining and Age Prediction. Also, this will become a useful tool for those researchers who will be developing a system where they have the resources in blocking a website depending on age restrictions.

1.6 SCOPE AND LIMITATION OF THE STUDY

The study was conducted in the Philippines specifically within Manila only, as it revolves only in getting the appropriate age for a certain website being accessed by the user. The system WRAP is a web tool based application that has the capability of predicting an age required for a website once accessed through the use of the website's metadata. It was developed using PHP language. The system only accepts Uniform Resource Locator's (URL's) for the input for the

web crawler; then the URL was processed and gave a certain prediction of an age appropriate of accessing a website.

It was designed for computers. A considerable amount of storage was a requisite for the system's database. A fast internet connection was considered. It required an operating system of Windows 7 or Windows 8 in using the system.

It used Data Mining specifically Web Text Content Mining, N-gram and Naïve Bayes Classification for analyzing the websites classification and Fuzzy Logic for getting the predicted Age.

1.7 OPERATIONAL TERMS

Data Mining – the process of extracting and analyzing information from a website to form a new one.

Fuzzy Logic – identification of required age given the URL and content classification score of the website.

Metadata – identified as the set of co-related words in a single class from different documents.

Naïve Bayes Classification – classification of text contents to negative or positive.

Natural Language Processing – analyzing extracted information to determine its syntax, semantics and pragmatics.

N-gram – a character trigram used to classify the URL to negative or positive.

Polarity – state having two opposite or contradictory tendencies such as Negative and Positive.

Website Age – given a certain website, the appropriate age required in accessing the website.

CHAPTER 2

Review of Related Literature

This chapter covers different related literatures that are relevant to the study. It will be used as guidelines and basis throughout the whole research. This chapter also show how the related studies will affect the research made by the researchers.

2.1. RELATED LITERATURE

2.1.1. The Importance of Social Media Restrictions

Most social media sites have age restrictions for opening an account. To open an account with Facebook, Instagram, Pinterest, Twitter, or many other social media sites, children need to be 13 years old. Some other social media accounts require you to be 17 years old. YouTube requires account holders to be 18 years old, but a 13 year old can sign up for an account at 13 years old with a parent's permission. Many kids are signing up for social media accounts even though age restrictions are in place. Social media websites aren't required to verify that an account holder meets the age requirements. Most social media require you to input a birth date to sign up, but kids can come up with fake birthdays for that. On Facebook, 52% of children between the ages of 8 and 16 have admitted that they didn't care about Facebook's age restrictions. Why are age restrictions for social media so important? The first is because a child's personal information is put at risk. Laws have been put in place that prevents websites from collecting personal information from children under the age of 13. The laws can no longer protect the personal information collected from a child under the age of 13 when they sign up for a social media account. Another reason these age restrictions are important is because we all (children included) make dumb decisions and the internet makes those dumb decisions permanent. Kids can harass or cyber bully other kids and not know the consequences of those actions. Anything they post on social media is forever; even if they delete it, there is still a record of it. While they may understand how to use social media, they may not understand how to use it wisely. Parents can block an underage child's access to social media by using an internet filter and restricting access to social media sites [Net Nanny 2015].

2.1.2. Age Verification within the Internet Infrastructure

In countries around the world there's plenty of information and advice on how to make the Internet a safer place for children. Complex legislation exists that covers different markets, sectors and territories. But in the global virtual world of the Internet where geographical boundaries and legal jurisdictions have limited relevance, enforcement is nigh on impossible and, where it has been successful, raised many issues of surveillance and intrusion into private lives. Age verification and age censorship are defenses that we are all familiar with to protect children in the physical world. Factors that affect the approach taken and its effectiveness are primarily driven by legislation, regulation and enforcement; and to some extent by self-regulation and industry best practice. Corporate social responsibility and reputational risk, particularly with market listed companies, also comes into play as an influencing factor.

Table 2.1: Different methods of Age Verification in UK [Lindley 2015].

Method of Age Verification	Sector	Issues
Self-Affirmation	Alcohol advertising Some adult content	Spoofing
Content Filtering	Adult content, mobile	Parental controls to manage filters in house only Household level Can be circumvented by teenagers
Delivery point validation	Delivery of age restricted physical goods	Driver required to perform check – get signature Not an expert on ID No liability
Credit/Debit Card	Online alcoholic sales Restricted media and content	Cannot differentiate cards held by children, such as pre-paid cards
Electronic checks of age verification databases and ID documents	Online gambling Restricted media and content	70-80% demographic coverage of adult Open to impersonation Cost

Table 2.2: Classification of Contents as classified in Germany [Lindley 2015].

Content Classification	Examples
Illegal	Games sporting excessive levels of gore and violence, or displaying symbols of anti-constitutional organizations like the Nazi swastika or the SS runes.
Endangering minors	Adult only content such as pornography. It is the responsibility of the provider to ensure content can only be accessed by adults. Providers must implement age verification systems within closed user groups.
Harmful to minors	Violent games, chat rooms and communities with a minimum of supervision. Providers have to implement Basic Age Verification systems.

Filtering of content and parental controls are promoted in both countries yet Germany has taken this a stage further to “join-up” the process between the content provider and the in-home filters and controlling software. More emphasis seems to have been placed in Germany on the state setting out the policies and supervision rather than the solutions, with the responsibility residing with both the content providers and the parents and each assuming their responsibilities. The UK Government has started to exert pressure on the ISPs to voluntarily put a basic “shield” in place. When it comes to higher levels of assurance around age, the UK has led the way for some years with the development of online identity and age verification systems. The major providers in the UK have extended their services to cover international markets around the globe. The introduction of the e-ID card in Germany, though, offers a way for age verification systems to be simplified (in Germany), both for the provider and the customer. It has the potential to be the universal approach by 2020 [Lindley 2015].

2.1.3. Are age restrictions even relevant today?

Based from one’s experiences, at some point we have most likely played or watched something we were not supposed to because the age restriction on the package told us we couldn’t, along with strict instructions from our parents. The author was a Cinema Manager, he used to get parents coming with their kids to age restricted

movies where the kids are too young, telling him that they are the parents and they will decide what their kids may or may not watch, and that it isn't my place to tell them what they can or can't do. Now, in dealing with this simple mindedness we must ask ourselves as to the importance that Age restrictions play in this modern society of ours. As time has gone on, he think movies and gaming have pushed the limits in violence and graphical content. So in light of this, standardized authorities have implemented stricter restrictions. Most people I speak to say "The kids see sex and violence everywhere anyway, so what's the difference?" With thoughts like that, it's no wonder things have deteriorated in the entertainment industries, and blame shifted not to parents or government but to the industries making the movies or games. We are all aware that gaming is on the forefront of people's minds in regard to violent behavior in children. The author himself is a parent and can assure us that his son, under his guidance, will never watch or play something that isn't age appropriate. But he cannot watch him 24/7. There are other adults that need to be responsible enough to shelter him from the things that his little eyes should not see. So are parents to blame? No, not entirely. We cannot shift blame to one focus point without having to look at the bigger picture. We have become so accustomed to sex and violence, and yes we see it almost every day, so it becomes almost the norm. But that is us, adults - over 21 years of age (Although from 18 you are pretty much there). We must focus on what we need to do to stop acts, that are attributed to gaming and movies, which needs to start in every aspect of a child's life. We cannot turn around and blame each other, but rather find the solution to where such issues have arisen from. This isn't to say that parents are the main culprits or that the main reason for this behavior is their fault. No, as some kids are born with potential mental issues. So, I ask the question based on what I have said here [SA 2013].

2.1.4. Why age matters in social media?

There is often derision at the age limits on social media sites. It's a good bet that you will know someone whose child has a profile on a site, despite being under the age required in the platform's terms and conditions. And maybe the view of the parent is that they monitor the child's usage and therefore it isn't a problem. But there are a few things you should be aware of before taking such a laid-back approach. First off, most people don't realize the age limits for most of the social sites. The image below shows the listed age restriction for some of the major social sites online at the moment. Most of us are probably aware that Twitter and Facebook stipulate you must

be at least 13 to have a profile, but did you know that you have to be 18 (or 13 with parental consent) to have YouTube, Kik, Flickr or FourSquare account?

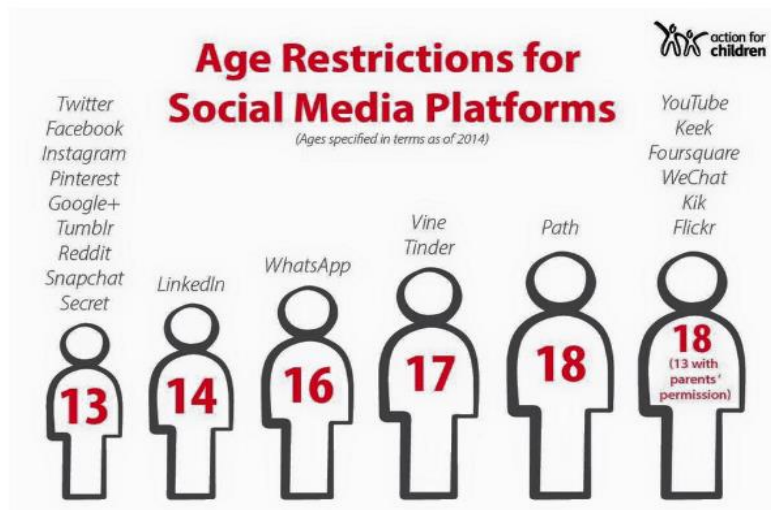


Figure 2.1: List of age restriction for some of the major social sites online [Sheppard 2016].

The BBC recently revealed that private groups on Facebook are being used by pedophile rings to exchange images and to groom young people. A further concern is that when presented with the evidence and some examples of offending content, Facebook decided not to shut down the groups or to remove all of the content. This is in line with their current approach which seems to be not to take too strong a line, presumably at the risk of upsetting and alienating users. This is extremely concerning when back in 2011 the BBC reported that as many as one in five children aged between 9 and 12 had a Facebook profile. This is a situation which is unlikely to have changed. It is quite easy to misrepresent yourself unintentionally on Facebook and therefore attract unwanted attention. And as we all know, with complicated and misleading privacy controls as well, it is easy to also set up your profile to be visible to any manner of people. And whilst laws exist to prosecute and protect us from these criminals, those laws only protect us if the person accused doesn't get off on a technicality. It is also seriously worth thinking about whether you want your children to use sites like Twitter and the like, where content isn't even restricted to a network of friends. It might be cool for their friends to know exactly where they are and what you're doing at every second of the day, but it can also be quite hazardous as well. And at the end of the day, almost all networks absolve themselves from the responsibility of policing their site and safeguarding their users. So it is important that we take seriously our online safety and realize that we might be posting more than we realize when we put something online [Sheppard 2016].

2.1.5. The Importance of Classifications and Age Restrictions

In Australia, the same classification system is used for TV, movies and video/computer games. These classifications are based around age and what content, e.g. violence, language and sexual themes, is appropriate for different age groups. These classifications are G, PG, M15+, MA15+ and R18+. There is more to determining what is suitable for your child than just checking the classification. For example, a game which is rated M15+ is recommended only for people over the age of 15 whereas MA15+ is restricted to only those people aged 15 years or older; what affects the classification is what makes up the game. It is then necessary to look at what the content of that game is, and then assess whether it is appropriate for the child to view. The majority of social media websites are restricted to users aged 13 years and above. Facebook and YouTube require users to confirm that they are 13 or older in order to sign up. This age restriction is there because these sites require a level of maturity to use them safely and responsibly. If a child has lied about their age in order to sign up to one of these accounts, they are sending the wrong message to other internet users. For instance, if a 12 year old lies about their age and pretends to be 18 years old to join Facebook, they could receive communications of an adult nature from other users, or receive adult-oriented targeted advertisements for dating websites or alcohol. Just because they are 13, however, doesn't mean it is appropriate for them to join these sites. Parents need to discuss with their children what is involved in joining these sites and whether their children are mature and responsible enough to use them safely. Check the classifications on the games the child is playing at home and make sure they are age appropriate. Speak to the child about what games they play at friends and relatives houses and make sure that they are also age-appropriate. Don't just take your child's word on what the classification is; do your own research as well. Find out what is involved in the game; do you have to solve problems or kill people? Talk to the child about how they feel during and after the game and whether or not playing the game is impacting on other areas of their life. If children are using sites which aren't age appropriate, parents should discuss with them the reasons why age restrictions are in place. There are mechanisms for reporting underage users to social networking sites and having their profiles removed. However, this may drive their usage underground and the child might simply start up a new account and not let you know [Pearcedale Primary School 2012].

2.1.6. A profile of internet users in the Philippines

The Rappler asked who's using the Internet in the Philippines and presented a snapshot of statistics and insights.

- The median age is 24 – the millennial who grew up as digital natives.
- Filipinos thrive on staying connected with their communities.
- We need real-time information to make the right choices, especially during time of crisis.
- The mobile internet penetration is growing at a rate of 1.5x (or 30 million users) every year.
- We consume about 150k terabytes of data annually.

From a total population of 101 million, 119 million mobile phone subscriptions (117% penetration rate), 95% are prepaid, 55% have a mobile broadband subscription, 10 % have a broadband subscription and 80% are subscribed to the lowest speed tier plans (1-3Mbps). And the time spent online is 3.2 hours on mobile and 5.2 hours on desktop and tablet. The top online activities are social media with 47%, videos 19%, 15% on online mobile games, 13% location-based search and 29% on online shopping. Currently, we have one of the highest digital populations in the world. The Internet audience's growth rate shows no signs of slowing down either as shown in Figure 2.2.

INTERNET GROWTH IN THE PH: ⁹

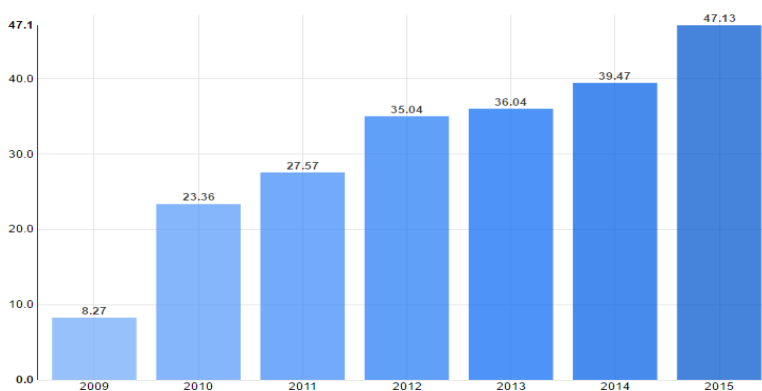


Figure 2.2: Internet Growth in the Philippines [The Rappler 2016].

Information and communications technology (ICT) plays a crucial role towards our nation's development. In the next decade, our country's biggest challenges - education, better transportation, and basic services - can all be solved by Internet access and big data. At the forefront of innovation are mobile solutions, coming primarily from the country's thriving startup market [The Rappler 2016].

2.1.7. Cybercrime Law in the Philippines

RA 10175 or Cybercrime Prevention Act of 2012 has been the hot topic this week especially for the netizens and press people as it will significantly create impact on our freedom of speech and some aspects of the constitution. Imagine if you just tweet, re-tweet or post any issue related to government or any person posting with defamation on Facebook or Twitter and you will be prisoned for at most 12 years in return. This of course is a total failure of Philippine justice. Cybercrime Law was signed by President Aquino on September 12, 2012 and agreed by the House of Representatives. Among the senators that pushed the law were PiaCayetano, Jose “Jinggoy” Estrada, Francis Escudero, Gregorio Honasan, PanfiloLacson, Manuel “Lito” Lapid, Loren Legarda, Ferdinand Marcos Jr., Aquilino Pimentel III, Ralph Recto, Ramon Revilla Jr., Vicente Sotto III and Manuel Villar. Under the new law which was already effective October 3, 2012, the government will have a power to monitor and shutdown private internet properties, criminalizes computer crime and impose rules related to online activities. Among the cybercrime offenses included in the bill are cybersquatting, cybersex, child pornography, identity theft, illegal access to data and libel. The new law received mixed reactions upon its enactment especially on the grounds of freedom of expression, freedom of speech and data security. Several petitions are currently submitted to the Supreme Court of the Philippines questioning the constitutionality of the Act. Being a blogger and certified netizen, I am strongly disagreeing with the cybercrime law being imposed today. The Summit Express still believe that social media and online medium is a mark of our freedom. For the information of all, United Nations declared Internet freedom a basic human right July 8, 2012. The U.N. Human Rights Council passed a resolution that “affirms that the same rights that people have offline must also be protected online, in particular freedom of expression, which is applicable regardless of frontiers and through any media of one’s choice.” [The Summit Express 2012].

2.1.8. Cyberbullying Statistics

Cyberbullying in the Philippines is also at stake. The Age group of those who said they are bullied are adult (18+) with 53% and minor (17 and below) with 47%. 57% of this is female and the rest is male. The top most object of attack is: (1) attack on reputation, (2) attack on appearance and (3) attack against the victims’ opinion. The top 3 nature of attack are: (1) spreading photoshopped imaged, (2) spreading videos that are supposedly private and (3) poser/spreading lies. The cyberbullying

statistics suggests that girls are more susceptible to cyberbullying, is no respecter of age. Filipino cyber bullies appear to be creative since more than words, they use photoshopped images to hurt their victims. Others use supposed private videos as means to harass their victims. Identity theft also plays a big role on cyber harassment. Facebook as the most popular social networking in the Philippines is also the primary platform of bullies. And because we can easily purchase a prepaid sim card, cell phone comes second while blogs come third. It is also interesting to note that Filipino victims are more transparent to their friends than family. Finally, this cyberbullying statistics revealed an interesting point worthy of a separate blog post, bullied because of unpopular opinion [ASKSonnie 2015].

2.1.9. PHP-Proxy and PHP Simple HTML DOM Parser

There have been many other proxy scripts in the past, but all have either perished permanently or has stopped updating them. PHP-Proxy is a web-based proxy script. With PHP-Proxy, complex sites such as YouTube and Facebook can be supported, also it was designed to be fast and easy to customize. This proxy script is intended to replace all other proxy scripts.

PHP-Proxy is a better alternative to Glype, which at that time was extremely lacking in many features that a user want. Most of Glype's site appears to be down. PHP-Proxy is also better than PHP Proxy script from whitefyre because there have been no updates or new features added to it for many years now. It is also hard to customize and breaks on many popular websites such as Facebook and YouTube [MIT 2015].

PHP Simple HTML DOM Parser is written in PHP5+ that lets developer manipulate HTML in a very easy way! It finds tags on an HTML page with selectors just like jQuery. It also extract contents from HTML in a single line. It also solves the problem with hosting, proxies and memory leak [Parser 2000].

2.1.10. A Dose of Business Intelligence: Data Mining

Data mining techniques can be classified into five general areas. First, visual representations techniques are graphical interpretations of complex (and even simple) relationships, which are commonly the “front-end” of other data mining techniques but are also used as “post-hoc” procedures. Data is accessed via specialized views and/or

drill-down processes for deeper analyses. Second, variable/feature selection methods are dimension-reduction techniques to summarize data into “relatively fewer” features, commonly used to identify the “more important” information. These are often conducted as data pre-processing, but are also used for index-derivation objectives. Third, segmentation and clustering techniques are used to find groups of “similar” characteristics based on relevant dimensions. Segments or clusters are made based on different similarity (or dissimilarity) measures, the objective of grouping often for profiling purposes, for “targeting” specific segments, or for classifying (of “new” units). Fourth, association rules are used to look for significant relationships and/or sequences among transactions (or events), with the rules based on frequent patterns. Common applications are collaborative filtering, market basket analysis and sequence analysis. Fifth, predictive modelling looks into developing a “model” based on discovered patterns or trends in the data, with the “model” being used to predict future outcome and/or identify impacts of changes in behaviors or activities. Predictive models are commonly used for robust customer valuation (or scoring) and identification (e.g., customers who are most likely to respond to an offer).

Different sources in the literature and different data mining software provide different frameworks of the data mining process. But somehow, the data mining process (or any analytical procedure for business intelligence, in this case) can be summarized in three stages – (1) objective and/or data setting (2) data processing and/or analysis, and (3) documentation and execution. These can be further classified as follows – under objective and/or data setting, the company must (a) know the business directives and/or identify specific objectives or queries, (b) then translate the business objectives into analytical objectives, and (c) prepare the data and map out the methodology (if data requirements and/or methods do not suffice to meet the objectives, then the objectives must be re-aligned or the data must be gathered and/or methods must be modified); for data processing, activities include (d) extraction, transformation and loading of data, and (e) analytics proper which includes validation and/or assessment procedures; and finally, activities under documentation and execution include (f) report writing and (g) implementation of decisions/actions.

To best apply the different data mining techniques, one should not only know what technique is appropriate for a given data, but should always be guided by what the business objective/s is/are. Though it seems that data mining is driven by data, what remains fundamental are (1) the company’s motivation or directive – what the company desires to do or needs to address (prior to data mining); and (2) the

company's understanding of the results – how the company reacts with the results (during and/or after data mining) [Lansangan 2011].

2.2. RELATED STUDIES

2.2.1. A New Method for URL-Based Web Page Classification using n-Gram Language Models

There are a number of contexts in which it is important to have an efficient and reliable way to classify a web-page by its Uniform Resource Locators (URLs), without the need to visit the page itself. For example, a social media website may need to quickly identify status updates linking to malicious websites to block them. Additionally, they can use the classification results in marketing researchers to predict users' preferences and interests. Thus, the target of their research is to be able to classify web pages using their URLs only.

Since URLs are normally very concise, and may be composed of concatenated words, classification with only this information is a very challenging task. But in time, much more current research on URL-based classification has achieved reasonable accuracy, but do not scale with large datasets. Thus the researchers applied a solution based on the use of an n-gram language model.

They proved that the n-gram language model is more scalable for large datasets, compared to existing approaches and no feature extraction is required as opposed to some of the existing approaches. It also shows that the classification performance is equivalent to previous successful approaches. Furthermore, it allows for better estimation for unseen sub-sequences in the URLs [Abdallah 2013].

2.2.2. Personality Trait Classification of Essays with the Application of Feature

Feature reduction is also used to identify personality traits. A study was proposed to classify personality traits since it is showing promising results and look to continuously improving the field of psychology by either using new features or by collecting new data from social media; however, a key concept that is not always considered is the use of feature reduction techniques. The research aims to perform feature reduction techniques on linguistic features from essays and classify the author's personality traits based on the reduced feature set.

To extract information from raw text, Linguistic Inquiry and Word Count (LIWC) was utilized. It is a text analysis tool that provides an efficient and effective method for studying emotional, cognitive and structural components present in individuals' written samples. And for the feature reduction, the techniques that are performed are Information Gain; able to characterize the impurity of an arbitrary collection of examples, and Principal Component Analysis; to identify patterns, and highlight the similarities and differences in data. And a 10-fold cross validation was performed on each of datasets in order to evaluate their overall effectiveness and the accuracy, precision and F-measure of all classifiers and the amount of reduction in terms of dataset's feature size.

Table 2.3: Comparison of best performing classifiers using all features and using feature-reduced datasets [Tighe 2016].

	Using All Features				Using Feature-Reduced Datasets			
	Classifier	Accuracy	Precision	F-measure	Classifier	Accuracy	Precision	F-measure
<i>Agreeableness</i>	SimpleLogistic	57.42%	0.572	0.566	SimpleLogistic ^A	57.54%	0.575	0.557
<i>Conscientiousness</i>	SimpleLogistic	54.91%	0.549	0.548	LibSVM ^B	56.04%	0.560	0.560
<i>Extraversion</i>	SMO	53.85%	0.537	0.533	LibSVM ^B	55.75%	0.557	0.556
<i>Neuroticism</i>	SimpleLogistic	57.46%	0.575	0.575	LibSVM ^B	58.31%	0.583	0.583
<i>Openness to Experience</i>	SMO	61.26%	0.613	0.613	SMO ^B	61.95%	0.619	0.619

Classifiers with ^A were trained using Information Gain reduced feature sets and ^B represents classifiers trained using PCA reduced feature sets

2.2.3. Multi-level Classifier for Detection of Insults in Social Media

Data mining was also used to detect insults in social media. The study investigates the use of multi-level classifier to predict insulting content. Negative interactions within a network like social media websites are created through insults. These remarks build up a culture of disrespect in cyberspace and should be prevented. However, current implementations on insult detection using machine learning and natural language processing have very low recall rates.

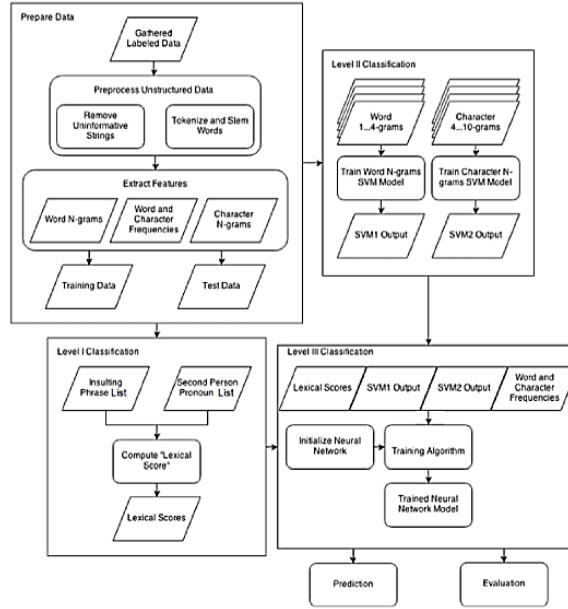


Figure 2.3: Multi-level classification conceptual framework [Amplayo and Occidental 2015].

Figure 2.3 shows the conceptual framework of the multi-level classifier. At its first level, a lexicon-based classifier is used to get the lexical score of the text. Then, at the second level, two Support Vector Machine classifiers are used with word n-grams and character n-grams as input features. These classifiers output two numbers which are two classifications on based on the word and character n-grams. At the final level, the results of the first two classifiers, combined with other input features, which include the number of characters, curse words, second person pronouns, capital letters, and symbols, are used as inputs for the neural network.

Table 2.4: Evaluation of Different Classifiers [Amplayo and Occidental 2015]

Level	Frequency	Character SVM	Word SVM	Lexical Score	Precision	Recall	F1 score	AUC
1				/	100	66.1	79.59	76.64
1	/				76.33	58.82	66.44	80.79
1			/		71.75	65.73	68.61	83.17
1		/			71.36	68.79	70.05	84.39
1		/	/		77.57	65.13	70.81	84.81
2	/			/	70.73	67.71	69.19	83.19
2			/	/	73.9	68.2	70.93	84.7
2	/		/		71.64	69.21	70.41	84.78

2	/	/	/		79.38	65.44	71.74	85.6
2		/		/	77.85	66.8	71.9	85.73
2	/	/			75.54	68.53	71.86	85.87
2		/	/	/	76.1	68.45	72.07	85.97
3	/		/	/	73.95	68.61	71.18	85.18
3	/	/		/	76.78	68.71	72.52	86.19
3	/	/	/	/	75.76	71.22	73.42	86.71

And they concluded the multi-level classifier outperforms other methods in insult detection that use the same dataset and can provide a more reliable way to identify and omit insults in social media, as shown in Table 2.4 [Amplayo and Occidental 2015].

2.2.4. Content Management and Distribution System for Context – Aware Services

We can also consider context data management as a data mining technique since it analyzes the data inside a context. Context data management and distribution is an essential part of context-aware services since huge amounts of up-to-date, reliable, and personalized information are being provided to the growing of end users. Proposed a service provisioning platform, a content management and distribution system, that is generic and configurable enough to support data requests and queries sent by heterogeneous context-aware applications by incorporating different techniques and approaches inherent in existing context-aware services and frameworks[Loyola et al. 2013].

2.2.5. Phishing Attack Detection, Classification and Proactive Prevention Using Fuzzy Logic and Data Mining Algorithm

But it does not end with just data mining. A research study presented a design for removing phishing sites/pages that are hosted probably without the knowledge of the website owner or host server. The system assess and classifies phishing emails using Fuzzy Logic linguistic descriptors that assigns to a range of values for each key phishing characteristic indicators; and the RIPPER Data Mining Algorithm used to characterize the Phishing emails and classify them based both content-based and non-content based characteristics of Phishing emails. Furthermore, the system

proactively gets rid of Phishing site/page by sending a notification to the system administrator of the host server that it is hosting a phishing site which may result in the removal of the site.

Table 2.5: Results generated from the WEKA classifier using RIPPER algorithm applied to classify Phishing emails [Ferolin R.J. 2011].

Validation Mode	10 fold cross validation
Attributes	URL Domain and Entity Criteria
	Email Content Domain
Number of rules	12
Correctly classified	85.4%
Incorrectly classified	14.6%
Number of samples/instances	1000

Table 2.6: Results of Phishing Pages removed after notifications were sent [Ferolin R.J. 2011].

Emails	Traced Server Info	Phishing Page Removed	Removal Success Rate
23	22	18	81.81%

In Tables 2.5 and 2.6, the results showed that the RIPPER algorithm achieved 85.4% for correctly classified Phishing emails and 14.6% for wrongly classified Phishing emails based on publicly available datasets from Phistank. The removal success rate of the identified phishing sites is 81.81% based on the notifications sent to the host of the different phishing pages. And concluded that the study however was able to prove that fuzzy logic and data mining with the use of the RIPPER algorithm is in a way sufficient in assessing the risk of a Phishing email and classifying the email as such, thereby resulting in the issuance of notification to the host server for removal of the Phishing page [Ferolin R.J. 2011].

2.2.6. Author Age Prediction from Text using Linear Regression

Another study was conducted in relation with connection between discourse patterns and personal identification is decades old and came up with two contribution made to the research. First is an investigation of age prediction using multi-corpus

approach wherein they presented results and analysis across three very different corpora: a blog corpus, a transcribed telephone speech corpus and posts from an online forum on breast cancer. With the use of the domain adaptation approach of Daume III, they trained a model on all those corpora together and separate the global features from corpus-specific features that are associated with age. Second is the investigation with age modelled as a continuous variable rather than as a categorical variable. Modelling age as a continuous variable is interesting also for practical benefits of joint modelling of age across corpora since the boundaries for discretizing age into a categorical variable in prior work have been chosen heuristically and in a corpus-dependent way, making it hard to compare performance across different kinds of data. Effective features include both stylistic ones (such as POS patterns) as well as content oriented ones.

Table 2.7: Most important features in the JOINT model with all features (condition 10) [Nguyen 2011].

(a) Features for younger people

Global		Blogs		Fisher		Cancer	
Like	-1.295	You	-0.387	Actually	-0.457	LIWC-Emotic	-0.188
gender-male	-0.539	Went	-0.310	Mean	-0.343	Young	-0.116
LIWC-School	-0.442	Fun	-0.216	Everyone	-0.273	History	-0.092
Just	-0.354	School	-0.192	Definitely	-0.273	Mom	-0.087
LIWC-Anger	-0.303	But	-0.189	Mom	-0.230	ultrasound	-0.083
LIWC-Cause	-0.290	LIWC-Comma	-0.152	Student	-0.182	Kids	0.071
Mom	-0.290	Go	-0.142	Pretty	-0.137	Age	-0.069
So	-0.271	POS-vbpnn	-0.116	POS-lrbcd	-0.315	Mum	-0.069
definitely	-0.263	That's	-0.115	LIWC-Swear	-0.134	POS-symnb	-0.069
LIWC-Negemo	-0.256	Well	-0.112	Huge	-0.126	discharge	-0.069

(b) Features for older people

Global		Blogs		Fisher		Cancer	
Years	0.601	LIWC - Job	0.514	Well	1.644	POS-dt	0.713
POS - dt	0.485	Son	0,257	LIWC – WC	0.855	POS – md vb	0.450
LIWC-Incl	0.483	Kids	0.228	POS - uh prp	0.504	POS - nn	0.369
POS – prpvbp	0.337	Years	0,178	Retired	0.492	LIWC-Negate	0.327
granddaug hter	0,332	Work	0.147	POS-prpvbp	0.430	POS – nnvbd	0.321
grandchild ren	0.293	Wife	0.142	Said	0.404	POS-nnp	0.304
Had	0.277	Husband	0.137	POS – cc fw	0.358	Us	0.287
daughter	0.272	Meds	0.112	Son	0.353	All	0.266
grandson	0.245	Dealing	0.096	Subject	0.319	good	0.248
Ah	0.243	weekend	0.094	POS - cc cc	0.316	POS – cc nn	0.222

Table 2.7 shows features associated with a young age have a negative weight, while features associated with old age have a positive weight. For almost all runs and evaluation metrics the full feature set gives the best performance. Using a linear regression model based on shallow text features; they obtained correlations upto 0.74 and mean absolute errors between 4.1 and 6.8 years. With this data, they concluded that content features and stylistic features to be strong indicators of a person's age [Nguyen 2011].

2.2.7. Age Prediction in Blogs: A Study of Style, Content, Online Behavior in Pre- and Post-Social Media Generations

The same study was conducted to predict age in blogs, based on style, content, and online behavior in pre- and post-social media generations. While features representing writing practices that emerged with social media (e.g., capitalized words, abbreviations, slang) do not significantly impact age prediction on their own, these features have a clear change of value across time, with post-social media bloggers using them more often. They found that the birth years that had a significant change

in writing style corresponded to the birth dates of college-aged students at the time of the creation/popularity of social media technologies, AIM, SMS text messaging, weblogs, Facebook and MySpace.

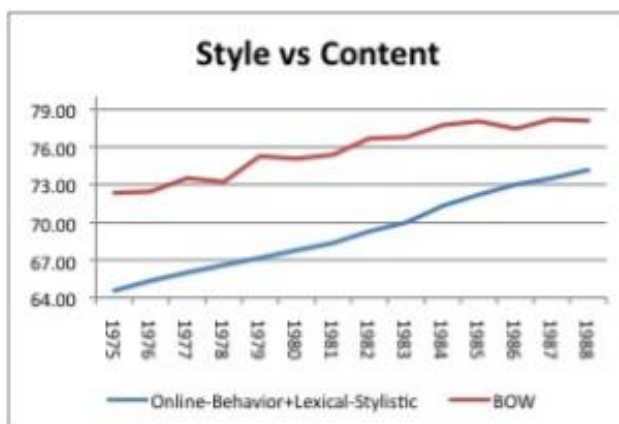


Figure 2.4: Style vs Content: Accuracy from 1975-1988 for Style (Online-Behavior+Lexical-Stylistic) vs Content (BOW) [Rosenthal and McKeown 2015].

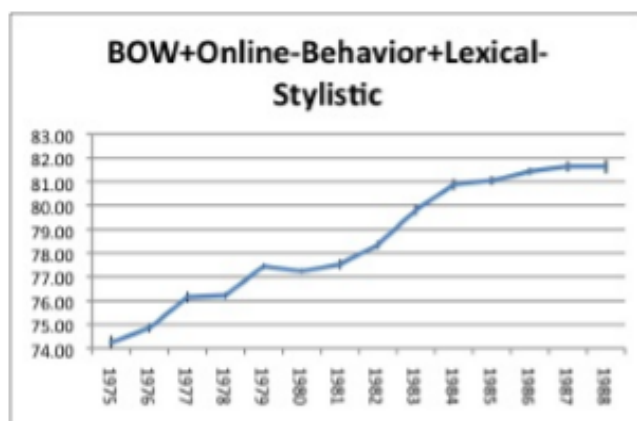


Figure 2.5: Style and Content: Accuracy from 1975-1988 using BOW, Online Behavior, and Lexical Stylistic features [Rosenthal and McKeown 2015].

Through experimentation with a range of years, they found that the birthdates of students in college at the time when social media such as AIM, SMS text messaging, MySpace and Facebook first became popular, enable accurate age prediction. They also show and concluded that internet writing characteristics are important features for age prediction, but that lexical content is also needed to produce significantly more accurate results. The best results allow for 81.57% accuracy as shown in Table 2.8 [Rosenthal and McKeown 2015].

Table 2.8: Feature Accuracy [Rosenthal and McKeown 2015].

Experiment	1979	1984
Online-Behavior	59.66	61.61
Interests	70.22	74.61
Lexical-Stylistic	65.38 ²	67.28 ²
Slang+Emoticons+Acronyms	60.57 ²	62.10 ²
Online-Behavior + Lexical-Stylistic	67.16 ²	71.31 ²
Collocations + Syntax Collocations	53.47 ¹	73.45 ²
POS-Collocations + POS-Syntax Collocations	55.54 ¹	74.00 ²
BOW	75.26	77.76
BOW+Online-Behavior	76.39	79.22
BOW + Online-Behavior + Lexical-Stylistic	77.45	80.88
BOW + Online-Behavior + Lexical-Stylistic + Syntax Collocations	74.8	80.36
BOW + Online-Behavior + Lexical-Stylistic + POS-Collocations + POS Syntax Collocations	74.73	80.54
Online-Behavior + Interests + Lexical-Stylistic	74.39	77.20
BOW + Online-Behavior + Interests + Lexical-Stylistic	79.96	81.57
All Features	71.26	74.07 ²

2.2.8. Predicting Age and Gender in Online Social Networks

While another study about age and prediction was also conducted in online social networks since it is a common characteristic of communication and uses non-standard language variations. These characteristics make this type of text a challenging text genre for natural language processing. Moreover, in these digital communities it is easy to provide a false name, age, gender and location in order to hide one's true identity, providing criminals such as pedophiles with new possibilities to groom their victims. It would therefore be useful if user profiles can be checked on the basis of text analysis, and false profiles flagged for monitoring. Their paper presents an exploratory studying which we apply a text categorization approach for the prediction of age and gender on a corpus of chat texts, which we collected from the Belgian social networking site Netlog. They examined which types of features are most informative for reliable prediction of age and gender on this difficult text type and perform experiments with different data set sizes in order to acquire more insight into the minimum data size requirements for this task.

Table 2.9: Results for data sets [Peersman 2011].

Scores (%)	Age Group	Data set1	Data set 2		
			Exp. 1	Exp. 2	Exp. 3
Precision	Min16	88.5	85.1	61.2	86.5
	Plus25	87.8	90.5	88.3	91.5
Recall	Min16	87.7	80.5	71.5	92.0
	Plus25	88.6	92.9	88.8	85.7
F-score	Min16	88.1	82.7	65.9	89.2
	Plus25	88.2	91.7	88.5	88.5
Accuracy		88.2	88.8	88.5	88.7

In the above table there is an improvement in the result of those on Data set 1 in terms of accuracy, precision, recall and f-scores. After examining these three different approaches of including the metadata for gender in order to investigate their effect on age prediction, for accuracy (88.8%) and recall (92.9%) and f-score (91.7%) of the adult class, the best results were achieved by balancing our data set according to both age and gender: 10,000 instances for both min16 and plus25, including 5000 instances for male and female within each age group. Adding the metadata for gender as an additional feature in each instance produced the best precision score for plus25 (91.5%). All the additional experiments showed improvement to the results of Data set 1, which was only balanced according to age [Peersman 2011].

2.2.9. Text-Based Age and Gender Prediction for Online Safety Monitoring

The study explores the capabilities of text-based age and gender prediction geared towards the application of detecting harmful content and conduct on social media. It was focused more specifically on detecting sexual predators that uses children and gave false age and gender information in their profiles. The researcher performed age and gender classification experiments on a dataset of nearly 380,000 Dutch chat posts from a social network. This study was evaluated into three prediction task which is age prediction, gender prediction, and combined. Five-fold-cross-validation experiment was performed for each task and the results if both balanced and full unbalanced datasets were compared.

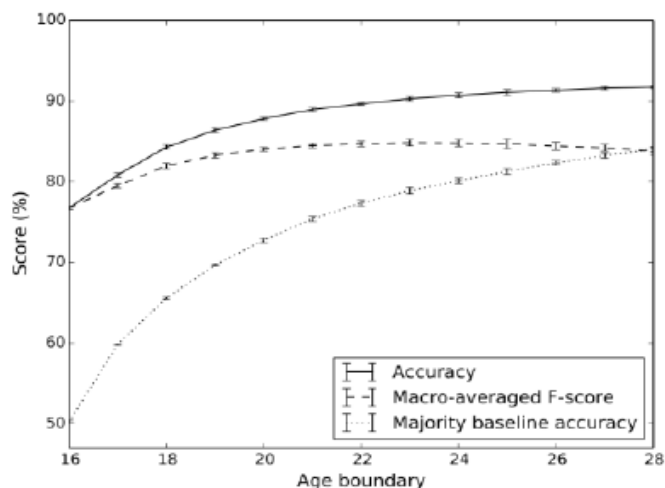


Figure 2.6: Age Prediction scores per class [van de Loo et al. 2016].

The accuracy rises from 76.7% with age boundary 16 to 91.7% with age boundary 28. The curve is quite steep in the beginning and starts to level off towards the end. The macro-averaged F-score reaches a maximum of 84.8% at age boundary 23 and slowly decreases after that. The rise in the accuracy score is mainly due to increased precision and re-calls scores for the younger class. This is caused partially by the growing class imbalance: as the age boundary rises, the portion of instances in the younger class grows, which has a positive effect on the scores for this class.

The paper shows the use-case applicable performance levels can be achieved for the classification of minors versus adults, where it serves a useful component in a cyber security monitoring tool for social network moderators [van de Loo et al. 2016].

2.2.10. Classification of documents based on contents using the n-gram method of MNB model

Classification of large number of documents by content is important and is difficult in some terms like the language origin e.g. Arabic documents. So a study was conducted and aims to apply classification technique for files management to raise the level of organization and retrieval of file. In their study, they created an enhanced Multinomial Naïve Bayes model by using n-gram. Document data was selected as consecutive pairs of keywords which called the bigram, or as three consecutive keywords called trigram and so on, by used of the n-grams the classification performance was increased.

The following figures will show the comparison of Recall and Precision for the MNB classifier with (a) bigrams, (b) trigram and (c) 4-gram classifier:

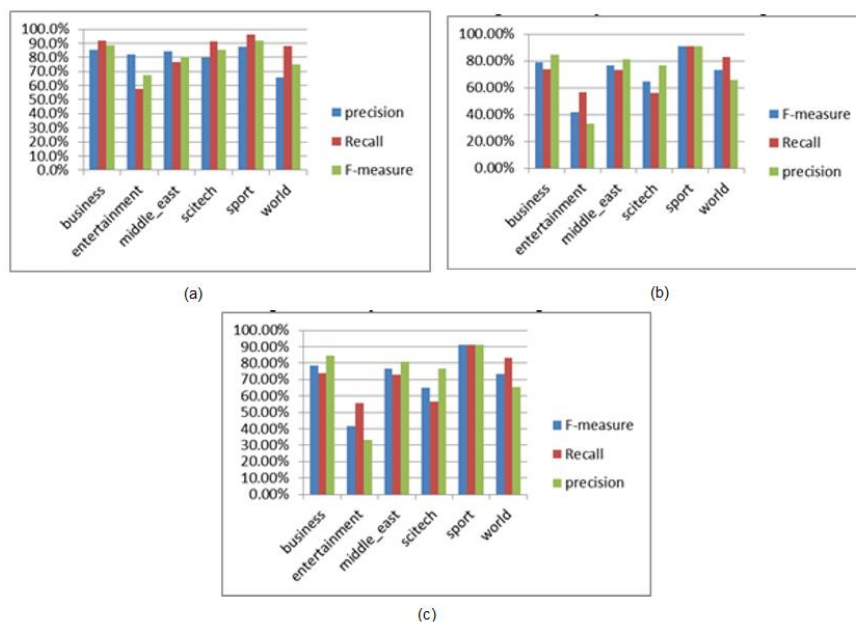


Figure 2.7: Recall, Precision and F-measure using different n-gram

In Figure 2.7, it shows that the most efficient process in the Multinomial Naïve Bayes model was that of bigrams.

2.3. SYNTHESIS OF THE STUDY

Most of the studies stated above were about prediction and classification of age, gender, personality and trait and etc. using websites text contents. This study has an identical output with other studies which will have a resulted age, but have different concept. It is focused on analyzing the websites contents to get the appropriate age in accessing that specific website.

Awareness for the youth's online safety is important because it has a great effect on the individual itself. Information acquired from the internet might result to bad influences to their mind and personality too. In the Philippines, more and more children/youth are able to access the internet without the guidance of the parent because of this the proponents decided to determine the appropriate age for the website based on its contents. In this study it will help, especially children/youth and parents, to be safe from internet, it help protects from accessing website that could possibly harm them mentally or physically and in improving the method of restriction of websites. Locally, this would be unique and the first to be studied in the country.

CHAPTER 3

Research Methodology

3.1. RESEARCH DESIGN

The researchers used experimental research design as its research methodology in testing the relationship of the independent and dependent variable to reach a valid conclusion. And also to control over all factors that may affect the result of an experiment. After-only with Control design was used since the researcher attempts to control for all variables.

The dependent variable in the study is the appropriate predicted age for the website and the independent variable is website itself. The controlling variable on the other hand is the other links connected to the website.

In this study, the researchers applied Natural Language Processing techniques, N-gram and use Naive-Bayes for classification. And also, Fuzzy Logic for predicting age appropriate for accessing a website. For the identification of the user's age, it is required to register his/her information such as email, username, password and birthday. After the development of the system, it will be tested. A URL served as an input and tabulating its performance in terms of scores in accuracy and precision in restricting the user if requirements did not met. Then, different respondents have to perform the same testing and tabulate same as the system to identify if there is a significant difference in the predicting the age appropriate in accessing the website in terms of accuracy and reliability.

This experimental design proves that there is no significant difference in the predicting the age appropriate in accessing the website in terms of accuracy.

3.2. SOURCES OF DATA

- **Population**

In the paper of [Griffin and Hauser 2011] they found out that "20-30 in-depth interviews are necessary to uncover 90-95% of all customer needs for the product categories". Thus, the authors determined that a sample size of 50 websites would provide a reasonable starting point. And a 30 sample websites were evaluated by the expert, on the other hand, each respondent tested 5 websites.

- **Respondents**

The target respondents can be any 18 year old and above, below 18 year old and an expert to increase the accuracy and reliability of the evaluation of the results identified by the system's output. These include psychiatrist, counselor and students.

- **Sampling Technique**

The sampling technique that the researchers used in this study is a Non-Probabilistic approach and will use Quota Sampling technique. Wherein, a sampling frame is not available and subsets are chosen and then either convenience or judgment sampling is used to choose people from each subset. The researcher decides how many of each category is selected.

3.3. INSTRUMENTATION

3.3.1 SOFTWARE/HARDWARE TOOLS

3.3.1.1 SYSTEM ARCHITECTURE

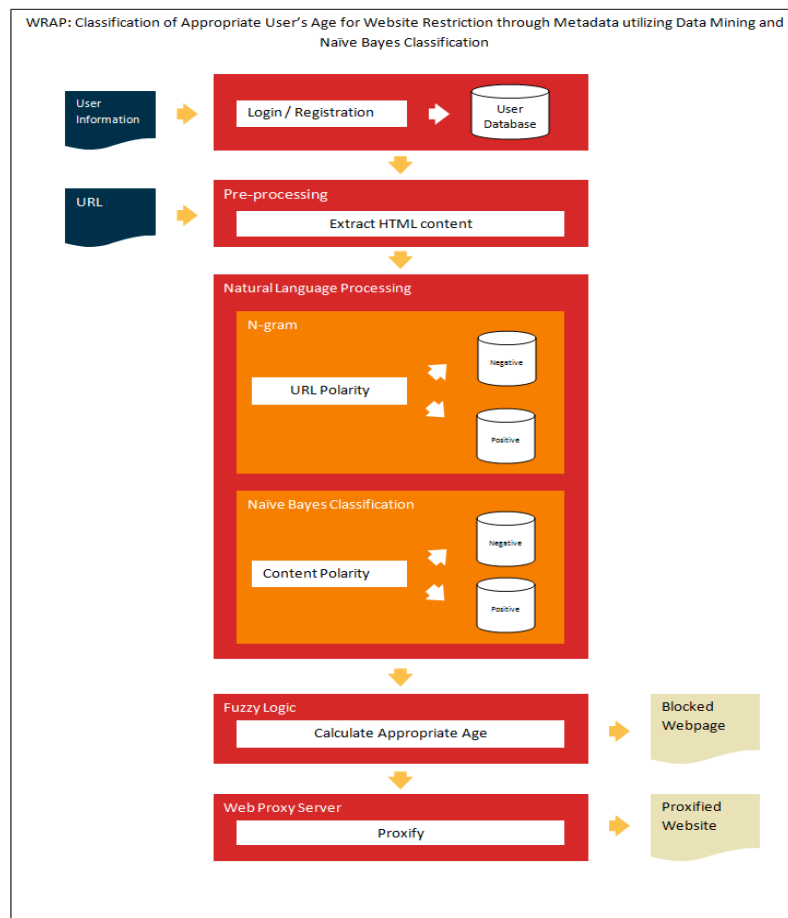


Figure 3.1: System Architecture of WRAP

The user created a user profile that holds the information about his/her email address, username, password and birthday; and it is stored at the database. From the given birthday, the user age is calculated.

A URL serves as the main input of WRAP. In which it undergoes web crawling and pre-processing using Simple HTML DOM Parser, for extracting the text content within the URL.

A dataset of positive and negative URL is provided which is composed of 500 URL's each. From this datasets the URL inputted were classified into two classes; Positive and Negative using character trigram and score accordingly. Based on another dataset of positive and negative contents; which is composed of titles/captions, sentences, paragraphs, short stories, etc., from different website that are classified as negative and positive. Then, the content of the webpage was classified according to the dataset into Positive and Negative contents using the Naïve Bayes Classifier.

For the age identification, the URL Polarity and the Content Polarity were applied with Fuzzy Logic concept. Fuzzy Logic is based on natural language because fuzzy logic is built on the structures of qualitative description used in everyday language; fuzzy logic is easy to use.

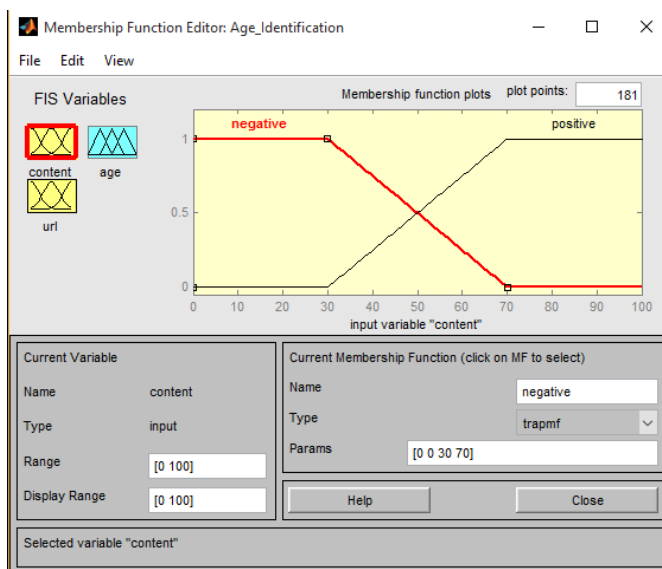


Figure 3.2: Membership Function for Content.

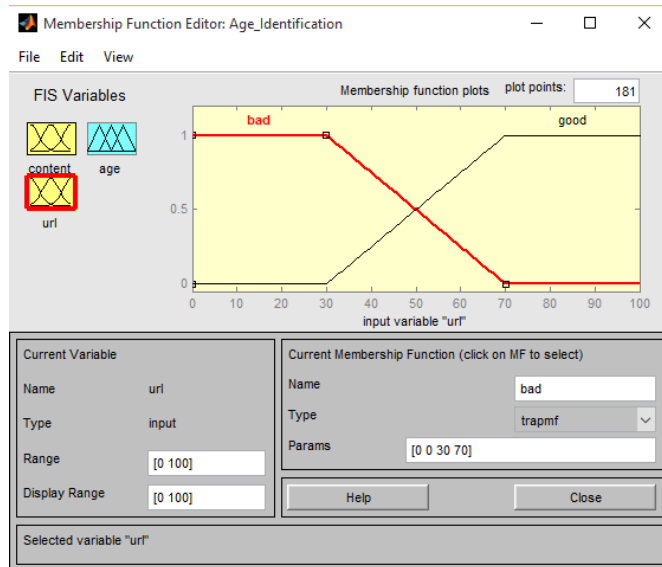


Figure 3.3: Membership Function of URL.

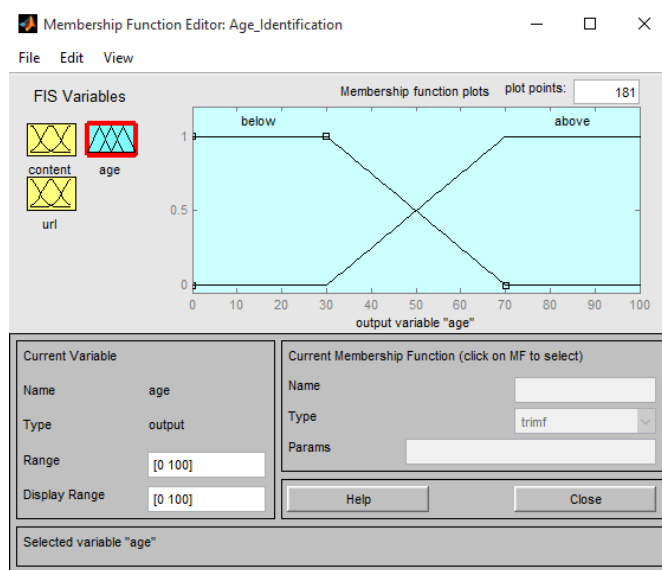


Figure 3.4: Membership Function of Age.

Figures 3.2 and 3.3 shows the input variables for the membership function used which are “content” and “URL” while Figure 3.4 shows the output variable which is “age” in MATLAB Simulation. The degrees of membership function as shown in the figures are the same since they are poles, meaning two things that contradict each other.

Lastly, if the system’s assessment of the website is suitable for the user, the URL entered will be proxified; otherwise a forbidden page will pop out.

3.3.1.2 DEVELOPMENT DETAILS

- **CSS, HTML, JavaScript and PHP** – The researchers had utilized HTML and CSS for front-end programming, JavaScript for client-side communication and PHP Programming for server-side communication in developing the software.
- **Google Chrome Browser** – used as the supported browser by XAMPP for web browsing.
- **IBM SPSS Statistics** – used program for statistical analysis.
- **MATLAB** – Used for simulation process throughout the development of the system.
- **Notepad++** – A text editor used in developing the system.
- **PHP Proxy** – A web proxy server used to access some websites.
- **Simple HTML DOM Parser** – A PHP class that could extract or scrape HTML DOM properties.
- **XAMPP** – Serves as the web server.

3.3.2 RESEARCH INSTRUMENT

Clerical tool was used, since the study is focused on people and gathers data on their judgements of the subject. The clerical tool used is experiment paper.

- **Experiment paper**

The researchers prepared an experimentation paper for the respondents to gather data that were used in getting the accuracy and reliability of the system in generating the output with regards to the age appropriate for the website the user is accessing.

3.4. DATA GENERRATION/GATHERING PROCEDURE

The data used by the researcher were gathered through the help of our experts and respondents.

Step 1: The researchers gathered 50 different website to be the samples for the expert and respondents.

Step 2: Each expert and respondent were given a set of websites to test to.

Step 3: Then the testing was performed through:

Step 3.1: Examining the samples by being tested several times.

Step 3.2: Examining the samples by checking if the inputted sample is accessed or blocked.

Step 3.3: Examining the samples if it is applicable for 18 and above, below 18 or applicable for all.

Step 4: Given an experiment paper, the respondents and expert gave a feedback regarding the systems output that was used in getting its accuracy and reliability.

Step 5: The researchers then analyzed the experiment paper results.

3.5. STATISTICAL TREATMENT OF DATA

3.5.1. Accuracy

In order to evaluate the accuracy of the system in predicting the age appropriate for a website the following formulas used:

- **Precision** - looks at the ratio of correct positive observations

$$precision = \frac{TP}{TP + FP} \quad \text{Equation 1: Formula for Precision}$$

Where:

TP (True Positive) = if the system correctly identified the required age.

FP (False Positive) = if the system incorrectly identified the required age.

- **Recall** - the ratio of correctly predicted positive events

$$recall = \frac{TP}{TP + FN} \quad \text{Equation 2: Formula for Recall}$$

Where:

TP (True Positive) = if the system correctly identified the required age.

FN (False Negative) = if the system incorrectly rejected the required age.

- **F-Measure** - the weighted average of Precision and Recall.

$$F - score = 2 * \frac{precision * recall}{precision + recall} \quad \text{Equation 3: Formula for F-Measure}$$

- **Accuracy** – the ratio of correctly predicted observations

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \quad \text{Equation 4: Formula for Accuracy}$$

Where:

TP (True Positive) = if the system correctly identified the required age.

FP (False Positive) = if the system incorrectly identified the required age.

FN (False Negative) = if the system incorrectly rejected the required age.

TN (True Negative) = if the system correctly rejected the age required.

Verbal Interpretation

Table 3.1: Verbal Interpretation of Percentage of Accuracy

Rating	Level of Accuracy in terms of Precision/Recall/ F-Measure and Total Accuracy
75.01 – 100%	Very High Accuracy
50.01 – 75%	High Accuracy
25.01 – 50%	Low Accuracy
0 – 25%	Very Low Accuracy

After the proponents determined the total accuracy in terms of Precision, Recall, F-measures and Total Accuracy, Table 3.1 was used to verbally interpret the computed values.

3.5.2. Reliability

On the other hand, in order to evaluate the reliability of the system in predicting the age appropriate for a website the following formula used:

- **Reliability Analysis** - test that finds how tool produces stable, consistent and repeatable the results.
 - **Cronbach's Alpha Test**

$$\alpha = \frac{n}{n-1} \left(1 - \frac{\sum V_i}{V_{test}} \right) \quad \text{Equation 5: Equation for Cronbach's Alpha Test}$$

Where:

n = Number of Trials

V_i = Variance of scores on each Trials

V_{test} = Total variance of overall scores on the entire Test

3.5.3. Expert vs. System

In order to determine the significant difference between the expert's assessments versus the system's assessment of predicting the required age in terms of accuracy the following formulas used:

- **Chi-Square** – compares the counts of categorical responses between the expert and system. Table 3.2 shows the variables needed.

Table 3.2: 2x2 Contingency Table

	For All	18 & above only	Total
Expert Assessment	a	b	a+b
System Assessment	c	d	c+d
Total	a+c	b+d	a+b+c+d

$$x^2 = \frac{(ad - bc)^2(a + b + c + d)}{(a + b)(c + d)(b + d)(a + c)} \quad \text{Equation 6: Chi-Square}$$

Where:

a = count of Expert Assessment on For All Category

b = count of Expert Assessment on 18 & above only Category

c = count of System Assessment on For All Category

d = count of System Assessment on 18 & above only Category

- **Degrees of Freedom** – are a measure the amount of variability involved.

$$DF = (\text{Number of columns} - 1) * (\text{Number of Rows} - 1)$$

- **Significance Level** – is a measure of how certain the results are. In this study the significance level is 0.05.

- **Chi-square Distribution Table** – to approximate the p-value.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Figure 3.5: Chi-Square Distribution Table

CHAPTER 4

Presentation, Analysis and Interpretation of Data

This chapter presents the analysis of the developed **WRAP: Classification of Appropriate User's Age for Website Restriction through Metadata utilizing Data Mining and Naïve Bayes Classification**. It also shows the interpretation of the results of data gathered to respond to the problems raised within the study. The researchers underwent thorough investigation, analysis, and systematic procedure to come up with the following results. In order to simplify the discussion, the researchers provided tables that summarize the collective data of the respondents.

As stated on the researcher's data gathering procedure, in this problem statement, the researchers themselves evaluated, tested, and analyzed the developed system gathered from the experts and respondents being known as the users of the system. The proponents conducted tests to assess the Accuracy rate of the system's output in terms of Precision, Recall, and F-measure based on the tallied TP (the website is positive and the required age is for everyone), FP (the website is positive and the required age is 18 & above only), FN (the website is negative and the required age is for everyone) and FN (the website is negative but the required age is 18 & above only). Researchers conducted several trials to assess the Reliability of the system's output. Also the researchers identify the significant difference between the expert's assessments in comparison with the system's results. The researchers guided the users in using the system and answering the experiment paper.

Based on the problems stated in Chapter 1, below are the gathered data that can answer the said issues:

1. What is the accuracy of the system when getting the predicted age appropriate in accessing the website using Precision, Recall and F-measure?

Evaluation of the System's Accuracy in Terms of Precision, Recall and F-Measure

The confusion table for determining the accuracy of the system is shown in Table 4.1. It shows the relationship between the classification of a website (positive or negative) and identification of age required for a certain website (18 and above or 17 and below).

Table 4.1: Confusion Table for Determining the Accuracy in Required Age

	WRAP Required Age (For All)	WRAP Required Age (18 & above only)
Actual Classification of Website (Positive)	TP	FP
Actual Classification of Website (Negative)	FN	TN

The research study consisted of two age groups: 17 & below and 18 & above. These two age groups differ from each other, thus in getting the accuracy, precision, recall and f-measure, it is as well separated. On the other hand, the overall accuracy consists of the mean of the two age group's accuracy.

Table 4.2: System's Accuracy rate for Precision, Recall and F-Measure for Respondents with Ages of 17 & below

No. of Websites	Evaluation of System's Answer				Accuracy (%)	Precision (%)	Recall (%)	F- measure (%)	Verbal Interpretation
	TP	FP	TN	FN					
25	10	3	9	3	76	76.92	76.92	76.92	Very High Accuracy

Table 4.2 shows that the respondents with age 17 & below were given 25 random websites. Thirteen (13) websites were classified as positive, 10 of which were accessed by the user with the ages 17 and below, and 3 websites were blocked. The rest of the 25 websites were negative, 9 of which were correctly classified as negative but not accessed by the user, and 3 were classified as negative but the user accessed it. From these, WRAP resulted with a 76% of total accuracy, 76.92% of precision, recall and f-measure which was verbally interpreted as very high accuracy.

Table 4.3: System's Accuracy rate for Precision, Recall and F-Measure for Respondents with Ages of 18 & above

No. of Websites	Evaluation of System's Answer				Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)	Verbal Interpretation
	TP	FP	TN	FN					
25	12	0	1	12	96	100	92.30	96	Very High Accuracy

On the other hand, Table 4.3 shows the system's accuracy for respondents with ages 18 & above. They were given 25 random websites, 13 positive websites and 12 negative websites. All 12 negative websites with an age requirement of 18 & above were accessed by the users. On contrast, out of 13 positive websites, 12 of which were accessed by the users and one website was not accessed but the age of the user was appropriate for the website. From these, the system acquired a 96% of total accuracy, 100% of precision rate, 92.30% of recall rate, and 96% of f-measure rate which is verbally interpreted as very high accuracy and an error rate of 4%.

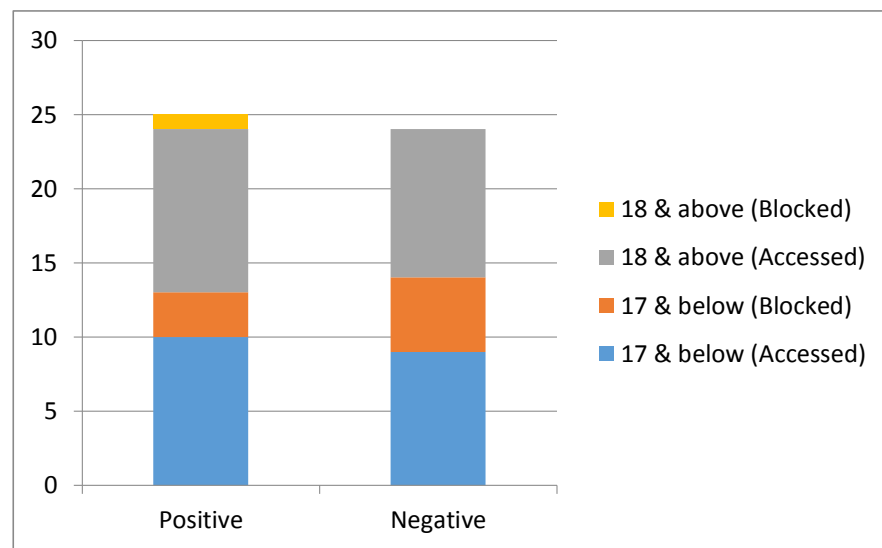


Figure 4.1: Accuracy of WRAP in predicting the appropriate age of 2 age groups

There were 26 positive websites and 24 negative websites, in total of 50 websites. In these positive web pages, 22 of them were accessed by both age groups and 3 respondents of 17 & below and one respondent, in total of 4 respondents were blocked from accessing the

web pages. On the other hand, 15 web pages classified as negative, were accessed, 5 of them came from 17 & below age group and 10 from 18 & above. And 9 of websites were accessed by the users of ages 17 & below. The graph presented in Figure 4.1 shows the presentation of data.

Evaluation of the System's Overall Accuracy

The evaluation of the system's overall accuracy presented below in Table 4.4 shows the accuracy of 76% and 96% of ages 17 and below and ages 18 and above, respectively. As an overall accuracy it resulted with 86% and has a verbal interpretation of a Very High Accuracy.

Table 4.4: System's Accuracy, Precision, Recall and F-measure of WRAP

Evaluation of System's Answer		Overall Accuracy	Verbal Interpretation
Accuracy (17 and below)	Accuracy (18 and above)		
76%	96%	86%	Very High Accuracy

2. What is the reliability of the system in resulting the predicted age in accessing the website?

Evaluation of the System's Reliability using Cronbach's Alpha Test

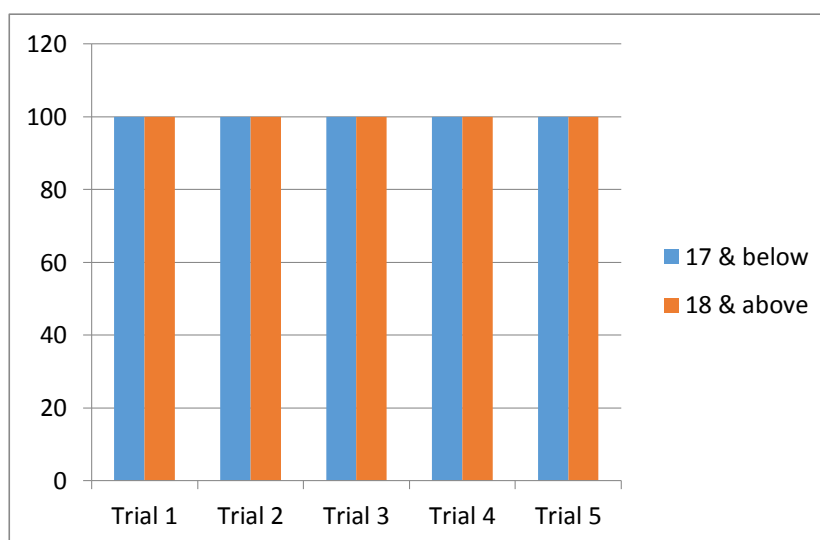


Figure 4.2: Reliability of WRAP in predicting the age of two age group.

As shown in Figure 4.2, predicting the appropriate age required for either 17 & below or 18 & above, after 5 trials, there is a consistency in the output. This means that the system's result does not change at all from the start until the last trial for each two different age groups out of 50 websites.

Table 4.5: Summary of the Cases for WRAP

		N	Percentage (%)
Cases of Websites	Valid	50	100
	Excluded	0	0
	Total	50	100

Table 4.6: Reliability Statistics of WRAP

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Trials
1.000	1.000	5

Shown in the Table 4.5, using the statistical treatment that was mentioned on Chapter 3 in getting the reliability of the system, it produced a total of 100% of the case processing that resulted of the reliability statistics shown in Table 4.6, where it resulted an alpha of 1.0, thus it is said that it has an excellent reliability result.

- What is the significant difference in the expert's assessment on predicting the age appropriate for accessing a website and our system's resulted age in terms of accuracy?

Evaluation of the System's Output Compared to the Experts Evaluation using Chi-Square

As shown at Table 4.7, the expert classified 30 websites: 8 websites as For All ages and 22 websites for 18 and above only. On the other hand, the system's assessment consists of 12 websites classified as For All, and 18 websites for 18 and above only with a total of 30

websites. The total of the category For All is 20 and for 18 & above only, 40 websites were classified. It has a Grand Total of 60. Applying the statistical treatment discussed in Chapter 3, it resulted a 1.2 chi square statistics.

Table 4.7: Result of the System's Output Compared to the Experts Evaluation using Chi- Square

	For All	18 & above only	Total	Chi-Square
Expert's Assessment	8	22	30	1.2
System's Assessment	12	18	30	
Total	20	40	60	

CHAPTER 5

Summary, Conclusions and Recommendations

The aim of this chapter is to summarize the study that was conducted. Included in this summary are the following; a review of the purpose of the study, the restatement of the research questions, the research methodology used, and the summary of the study results, conclusions and discussion. Recommendations for further research and possible studies conclude this chapter.

5.1. Summary of Findings/Results

The accuracy of the system regarding on its performance was tested and evaluated by the researchers. The system has an overall performance in terms of Accuracy rate of Precision, Recall and F-score. Negative websites should not be accessed by any user with ages 17 & below and on the other hand, positive websites should be accessed by both age groups. Out of 24 negative websites, 5 of which websites were accessed by 3 respondents of ages 17 & below thus the system gathered a 76% accuracy rate. On the contrary, there was a 96% of accuracy rate gathered for age group 18 & above. Out of 26 websites, a respondent was blocked from accessing a certain website. As a result, the overall accuracy of the system garnered a total of 86% from the accuracy of age group 17 & below which is 76% and age group 18 & above which is 96%, obtained from the test data presented in Chapter 4.

The reliability of the system in resulting the predicted age for a website was tested and analyzed by the researchers. Based on the data presented in Chapter 4, the system has a reliability rate of 100% wherein a 1.0 alpha was acquired in predicting the age required for a website after 5 trials. This indicates that, there is no change at all in the output of the system in each trial conducted regardless if it is correctly classified for a user to access a website or not.

Lastly, the expert's assessment on predicting the age of a website versus the system's assessment in terms of accuracy was answered. Based on the data presented, the chi-square statistics as calculated given the following variables discussed in Chapter 3, it resulted with 1.2 chi-square value. The expert's evaluation and the system's evaluation for the chi-square test have a 1.2 chi-square statistic. The corresponding probability is between the 0.9 and 0.1 probability levels. That means that the p-value is below 0.05. A 0.27 value was obtained from the one-tailed probability value for a chi-square test. Since a p-value below 0.05 which is 0.27 and is lesser than the conventionally accepted significance level of 0.05 (i.e. $p < 0.05$) we accept the null hypothesis. In other words, there is no statistically significant difference in the expert's evaluation in predicting the age versus the system's evaluation regarding its accuracy.

5.2. Conclusions

The researchers have arrived on the following conclusions. Regarding on the accuracy rate in predicting the required age for a website is highly accurate. An excellent reliability rate was also concluded in determining the reliability of the system in predicting the required age for a website. Thus the system WRAP is a suitable tool in identifying the appropriate age required in accessing a website through its metadata by utilizing Data Mining and Naïve Bayes Classification in terms of its accuracy and reliability.

Lastly, from the results, it can be inferred that referring to an expert and using the system WRAP is both an acceptable way in identifying the appropriate age in accessing a website for users, since the system's evaluation has no significant difference with the expert.

5.3. Recommendation

This research can be improved in many aspects. One of these is improving the classification of the URL and content. In terms of the URL classification, the use of n-gram in classifying the URL's polarity is somehow not satisfactory. In this study character tri-gram was used thus the distance between the characters has larger distances that can affect the accuracy of the algorithm. To solve the problem, the use of character bigram is recommended. On the other hand, implementing new approach in classifying opinions, suggestions and comments are highly recommended, like the use of Sentiment Analysis which is better in identifying the polarity especially with sentiments. This would help in the accuracy of predicting the required age for a website. And also by adding more data in both the URL dataset and Content dataset; it would affect the performance in terms of accuracy of the system. The researchers also recommends, adding Tagalog or Tagalog-English corpus so that when accessing websites, with a Tagalog or Tagalog-English languages like social media or information websites, it can be classified also.

Future researchers may deploy the tool in the web to have a server and accessible without using web server application. The researchers used XAMPP as a web server. Also the researchers did not do this since it involves a lot of financial budget required, because free web server does not support other features of PHP.

Moreover, changing the platform of the system is recommended also. In PHP Proxy there are styles in how HTML tags are displayed. Also in developing the system there are also styles as developed. When accessing the website, if the name of the variable in the stylesheet of a website and the variable in the developed system, it will be changed. Thus the HTML tags in the website are modified. Also in this research, the system was coded in PHP, a server-side programming language, used as a web proxy-server to proxify links and load them. In this way, the webpage is loaded right after it is analyzed but modifies some attributes of the webpage. By

this we mean, the HTML tags of the website and the HTML tag of the WRAP is the same, thus modifying the attribute of the tag of the website. If possible, try creating a Google extension as the platform or a desktop application.

In the field of networking and security this study can be used too. Nowadays, security professionals can detect malicious contents with some techniques. So try other algorithm that specializes about data security.

In technical side, a better and faster internet connection can affect the crawling of the website. Also the use of high end hardware is also recommended, to handle the crawling of websites. Crawling more links may improve the classification than just crawling a specific link or small amount of links.

REFERENCES

- [Abdallah 2013] Abdallah, Tarek Amr. A New Method for URL-Based Web Page Classification using N-Gram Language Models (2013).
- [Amplayo and Occidental 2015] Amplayo, Reinald Kim and Occidental, Jason. Multi-level classifier for detection of insults in social media. *15th Philippine Computing Science Congress* (2015).
- [ASKSonnie 2015] ASKSONNIE. 2015. Cyberbullying Statistics. from *ASKSonnie*: <http://asksonnie.info/cyberbullying-statistics/>
- [Cameron 2013] Cameron, David. *Protecting Our Children*. NSPCC, London UK, 2013.
- [Caroll 2011] Caroll, J.A., Kirkpatrick, R.L. Impact of social media on adolescent behavioral health in california. *Oakland, CA: California Adolescent Health Collaborative*. (2011).
- [Creswell 2012] Creswell, John W. *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*. Pearson Education, Inc., USA, 2012.
- [Diloy 2013] Diloy, Ysrael C. Online Risks Filipino Children Face Today. *Stairway Foundation Inc. 2013* (2013), 1-25.
- [Ferolin R.J. 2011] Ferolin R.J., Gerardo, B.D., Byun, Y-C., Kang, C-U. Phishing Attack Detection, Classification and Proactive Prevention Using Fuzzy Logic and Data Mining Algorithm (2011).
- [Geier 2013] Geier, Eric. 2013. How to child-proof the Internet. from *PC World from IDG*: <http://www.pcworld.com/article/2042233/how-to-child-proof-the-internet.html>
- [Gonzales 2014] Gonzales, R. Social Media as a Channel and its Implications on Cyber Bullying. *DLSU Research Congress* (2014), 1-7.
- [Griffin and Hauser 2011] Griffin, Abbie and Hauser, John. *Voice of the Customer*. USA, 2011.
- [Han 2013] Han, H., Otto, C., Jain, A.K. Age Estimation from Face Images: Human vs. Machine Performance. *The 6th IAPR International Conference on Biometrics (ICB)*. (2013).
- [Han et al. 2012] Han, Jiawei, Kamber, Micheline, and Pei, Jian. *Data Mining Concepts and Techniques*. Elsevier Inc., USA, 2012.

- [Indurkhya and Damerau 2010] Indurkhya, Nitin and Damerau, Fred J. *Handbook of Natural Language Processing*. Taylor and Francis Group, LLC, 6000 Broken Sound Parkway NW, Suite 300, 2010.
- [Internet Matters 2013] INTERNET MATTERS. 2013. Inappropriate Content. from *Internet Matters*: <https://www.internetmatters.org/issues/inappropriate-content/>
- [Kavita Ganesan 2011] KAVITA GANESAN. 2011. What are N-grams? from *Text Mining, Analysis & More*: <http://text-analytics101.rxnlp.com/2014/11/what-are-n-grams.html>
- [KidsHealth] KIDSHEALTH. Internet Safety. from *KidsHealth*: <http://kidshealth.org/en/parents/net-safety.html>
- [Korb and Nicholson 2004] Korb, Kevin B and Nicholson, Ann E. *Bayesian Artificial Intelligence*. Chapman & Hall/CRC, Washington D.C, 2004.
- [Lansangan 2011] Lansangan, J.R.G. A Dose of Business Intelligence: Data Mining. *The Philippine Statistician Vol 60* (2011), 125-128.
- [Lindley 2015] Lindley, E., Green, I., Laurence, R. AGE VERIFICATION WITHIN THE INTERNET INFRASTRUCTURE. *THE INTERNATIONAL FOUNDATION FOR ONLINE RESPONSIBILITY* (2015).
- [Loyola et al. 2013] Loyola, Elizabeth, Dino, Arvin, and Morales, Noelyn Joyce. Content Management and Distribution System for Context-Aware Services. *Computing Society of the Philippines* (2013), 144-156.
- [Marquardt et al. 2014] Marquardt, James, Farnadi, Golnoosh, Vasudevan, Gayathri, Moens, Marie-Francine, Davalos, Sergio, Teredesai, Ankur, and De Cock, Martine. Age and Gender Identification in Social Media. *PAN-AP-14 corpus - Author Profiling Shared Task* (2014), 1-8.
- [MathWorks 2016] MATHWORKS. 2016. Foundations of Fuzzy Logic. Retrieved September 15, 2016 from *MathWorks*: <http://www.mathworks.com/help/fuzzy/foundations-of-fuzzy-logic.html>
- [MIT 2015] MIT. 2015. PHP-Proxy. from *PHP-Proxy*: www.php-proxy.com
- [Nancho 2016] Nancho, Rosa Maria. Manila, 2016.
- [Natural Language Toolkit 2008] NATURAL LANGUAGE TOOLKIT. 2008. Supervised Text Classification. Retrieved September 9, 2016 from *Natural Language Toolkit*: <http://www.nltk.org/book/ch06.html>

- [Net Nanny 2015] NET NANNY. 2015. Net Nanny. from *The Importance of Social Media Age Restrictions*: <https://www.netnanny.com/blog/the-importance-of-social-media-age-restrictions/>
- [Nguyen 2011] Nguyen, D., Smith, N.A., Rose, C.P. Author Age Prediction from Text using Linear Regression. *Language Technologies Institute* (2011).
- [Nguyen 2014] Nguyen, D., Trieschnigg, D., Dogruoz, A.S., Gravel, R., Theune, M., Meder, T., de Jong, F. Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowd sourcing Experiment. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (2014), 1950-1961.
- [Parmar 2016] Parmar, N., Richhariya, V., Maurya, J.P. An Exploratory Review of Web Content Mining Techniques and Methods. *International Journal of Advanced Research in Computer and Communication Engineering* 5(5), (2016), 144-148.
- [Parser 2000] Parser, Simple HTML DOM. 2000. Simple HTML DOM Parser. from *Simple HTML DOM Parser*: www.simplehtmldomparser.com
- [Pearcedale Primary School 2012] PEARCEDALE PRIMARY SCHOOL. 2012. The Importance of Classifications. from *Pearcedale Primary School*: <http://pearcedaleschool.com.au/wp-content/uploads/2012/10/classifications-and-restrictions.pdf>
- [Peersman 2011] Peersman, C., Daelemans, W., Vaerenbergh, L.V. Predicting Age and Gender in Online Social Networks. *e 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (2011), 115-123.
- [Rosenthal and McKeown 2015] Rosenthal, S and McKeown, K. Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in Pre- and Post-Social Media Generations. *17th International Conference, SPECOM 2015, Athens, Greece* (2015), 113-120.
- [SA 2013] SA, Gabe. 2013. Are age restrictions even relevant today? from *MWeb*: <http://www.mweb.co.za/games/view/tabid/4210/article/7981/are-age-restrictions-even-relevant-today.aspx>
- [Sheppard 2016] Sheppard, P. 2016. Why Age Matters in Social Media. from *SiteSet Digital*: http://siteset.digital/blog/why_age_matters_in_social_media
- [Survey Monkey 2016] SURVEY MONKEY. 2016. Calculating the Number of Respondents You Need. from *Survey Monkey*:

- http://help.surveymonkey.com/articles/en_US/kb/How-many-respondents-do-I-need
- [Tagliamonte 2013] Tagliamonte, Sali. Sociolinguistics for Computational Social Science (2013).
- [The Rappler 2016] THE RAPPLER. 2016. Rappler. from *A Profile of Internet Users in the Philippines*: <http://www.rappler.com/brandrap/profile-internet-users-ph>
- [The Summit Express 2012] THE SUMMIT EXPRESS. 2012. Cybercrime Law in the Philippines review. from *The Summit Express*: http://www.thesummitexpress.com/2012/10/cybercrime-law-in-philippines-review_5.html
- [The Works 2015] THE WORKS. 2015. How the digital world has changed Britain. Retrieved August 26, 2016 from *We are the Works*: <http://wearetheworks.com/latest/2015/may/how-the-digital-world-has-changed-britain/>
- [Tighe 2016] Tighe, E.P., Ureta, J.C., Pollo, B.A.L., Cheng, C.K., de Dios Bulos, R. Personality Trait Classification of Essays with the Application of Feature. *4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016* (2016), 22-28.
- [van de Loo et al. 2016] van de Loo, Janneke, De Pauw, Guy, and Daelemans, Walter. Text-Based Age and Gender Prediction for Online Safety Monitoring. *International Journal of Cyber-Security and Digital Forensics (IJCSDF)*, 5, 1 (2016), 46-60.
- [Walkowiak 2015] Walkowiak, T., Piasecki M. Web-based Natural Language Processing Workflows for the Research Infrastructure in Humanities. *JADH2015* (2015).
- [Witten et al. 2011] Witten, Ian H, Frank, Eibe, and Hall, Mark A. Data Mining Practical Machine Learning Tools and Techniques. In *Witten, Ian H; Frank, Eibe; Hall, Mark A*. Elsevier Inc., USA, 2011.
- [Zyga 2011] Zyga, L. 2011. PhysOrg. from *Researchers hope to build universal human age estimator*: <http://phys.org/news/2011-12-universal-human-age.html>

APPENDIX

APPENDIX A: SAMPLE RESEARCH INSTRUMENT

Sample experiment paper that will be given to the respondents

Experiment Paper of WRAP: Classification of Appropriate User's Age for Website Restriction through Metadata utilizing Data Mining and Naïve Bayes Classification

Objectives:

- To determine the accuracy in accessing and blocking a website based on the output of the system.
- To test the reliability in accessing and blocking a website based on the output with several trials of the system.
- To determine the significant difference between the expert's assessment in predicting the age appropriate for accessing a website and our system's resulted age in terms of accuracy.
- The data gathered are envisioned to provide information and guidelines to improve the application for future researchers.

Materials/Equipment:

- ✓ Laptop / PC with at least 2 GB RAM
- ✓ Pocket Wi-Fi /any source of Internet Connection
- ✓ Experiment paper
- ✓ Pencil or pen

Procedures:

1. Prepare the experiment paper to be answered by the expert and respondents and the websites to be presented.
2. Each respondent will conduct 5 trials in accessing the websites presented to them. On contrast, the expert will be given 30 set of website to assess if the website is positive or negative.
3. Guide both the respondents and the expert in answering the experiment paper.

List of Website Samples:

Website #	Website URL
1	http://thoughtcatalog.com/jim-goad/2014/06/8-cases-of-extreme-animal-cruelty-that-you-should-absolutely-not-read-if-you-love-animals/
2	http://www.mirror.co.uk/news/world-news/gang-kids-kick-flamingo-death-
3	http://www.creepypasta.org/creepypasta/the-magicians-game#read
4	http://faq.ph/visit-rizal-park-in-manila/
5	http://www.medicalnewstoday.com/info/cancer-oncology
6	http://theunboundedspirit.com/the-negative-effects-of-religion-on-society/
7	https://www.playstation.com/en-us/games/until-dawn-ps4/
8	http://www.mensjournal.com/expert-advice/the-18-best-tequilas-in-the-world-20131217/siete-leguas
9	http://www.lazada.com.ph/
10	http://tagaloglang.com/religion-in-the-philippines/
11	http://www.the-line-up.com/6-creepy-crimes-read-bed-tonight/
12	http://9gag.com/nsfw
13	http://www.youporn.com
14	http://www.marcandangel.com/2010/06/21/18-things-i-wish-someone-told-me-when-i-was-18/
15	https://www.thefactsite.com/2011/07/top-100-random-funny-facts.html
16	https://www.common sense media.org/blog/10-most-violent-video-games-of-2015-and-what-to-play-instead
17	https://www.scoopwhoop.com/inothernews/weird-death-facts/#.y3rbyfu4j
18	http://www.scaryforkids.com/
19	http://sports.inquirer.net/
20	http://www.biblestudytools.com/bible-verse-of-the-day/
21	https://www.illuminatioofficial.org/the-official-website-for-the-illuminati/
22	https://www.vaporfi.com/electronic-cigarettes/
23	http://www.advocatesforyouth.org/component/content/article/450-effective-sex-education
24	https://betobaccofree.hhs.gov/about-tobacco/facts-figures/

25	https://www.lonelyplanet.com/philippines
26	http://www.joysporn.com
27	http://www.thejokeyard.com/funny_insults/insult_jokes.html
28	http://www.wikihow.com/Hack
29	http://www.biography.com/people/jos%C3%A9-rizal-39486
30	http://news.abs-cbn.com/sports
31	http://www.grubstreet.com/2013/10/kitchen-horror-stories-2013.html
32	http://www.theliquorbarn.com/
33	http://www.lifehack.org/articles/featured/learn-something-new-every-day.html
34	http://countrystudies.us/philippines/53.htm
35	http://www.storystar.com/php/list.php?sub_category_id=1
36	https://www.compellingtruth.org/Luciferianism.html
37	http://spectangles.net/vikings-ww.php?utm_term=34374&utm_content=2128026&click_id=UWZGWTdSLTNmeDFLUnVNWnU0VXo1cDZWV1RSY2NhQ2I4SIRRUVBhVTdHQld6VnV5dWhPVVBkODc0eS1wTkNKUI8xNDg5ODkzMzEz
38	http://startupguide.com/world/greatest-innovators/
39	http://drugfree.org/drug/marijuana/
40	http://www.freedommag.org/issue/201412-expansion/l-ron-hubbard/religious-influence-in-society.html
41	https://www.livejasmin.com/en/
42	https://www.youtube.com/results?search_query=horror+stories
43	http://www.gov.ph/about/gov/
44	https://www.w3schools.com/
45	http://www.wannafact.com
46	http://www.cosmopolitan.com/lifestyle/advice/a6504/female-genital-mutilation-survivor-stories/
47	www.playboyenterprises.com/
48	http://www.wheninmanila.com/category/travel-adventure/
49	http://edition.cnn.com/2016/05/24/foodanddrink/50-delicious-philippines-dishes/
50	https://nuts.com/chocolatessweets/

Guidelines:

- For each item, complete the table based on the output displayed by the system.
Refer to the following interpretation of values below.

- For determining the Accuracy and Reliability of Predicting Age Appropriate for Accessing a Website.

Respondent # _____

Name: _____ (Nickname) Course/Year/Section: _____

Birthday: _____ Age: _____ Gender: ☐ F ☐ M

Instruction: Write

✓ - If the website is accessed.

X - If the website is blocked.

INPUT	Trials					Average
	1	2	3	4	5	
Website 1						
Website 2						
Website 3						
Website 4						
Website 5						

- For determining the Significant Difference between the expert's assessment and the our systems in terms of accuracy
 - Record the total accuracy assessment by the expert for each website.
 - Record the total accuracy assessment by the system of Trial 1 for each website.

Expert # _____

Name: _____ (Nickname) Course/Year/Section: _____

Birthday: _____ Age: _____ Gender: ☐ F ☐ M

Instruction: Write / if the website is applicable for everyone otherwise, **X** if the website is applicable for 18 years old and above only.

INPUT	Expert Assessment	System's Assessment
Website 1		
Website 2		
Website 3		
Website 4		
Website 5		
Website 6		
Website 7		
Website 8		
Website 9		
Website 10		
Website 11		
Website 12		
Website 13		
Website 14		
Website 15		
Website 16		
Website 17		
Website 18		
Website 19		
Website 20		
Website 21		
Website 22		
Website 23		
Website 24		
Website 25		
Website 26		
Website 27		
Website 28		
Website 29		
Website 30		

APPENDIX B: COMMUNICATIONS

Request Letter for Interview



Polytechnic University of the Philippines
 College of Computer and Information Sciences
 Department of Computer Science
 Sta. Mesa, Manila



September 6 2016

Rosa Maria H. Nancho, MD
 Pediatrics, Adolescent Medicine
 Manila Doctors Hospital

Dear Madam:

We are students of Computer Science in the Department of Computer Science of Polytechnic University of the Philippines currently researching for our thesis, WRAP: Website Restriction of Predicted Age through Metadata utilizing Data Mining and Sentiment Analysis. We are in awe of your expertise on the field of child and adolescent psychiatry, and we are taking the opportunity to ask for advice regarding child's thinking and behavior with regards the internet.

We would like to schedule a brief informational interview with you in person and if you would let us, we are humbly asking to record the conversation for educational purposes. We hope that you will be available on Wednesday, 7 Sep 2016 around 10:30 am.

Here is the list of our agenda:

- Children Behavior and the Internet.
- Determination of Child's age through interrogative response.
- Ways to Protect a Child from the Internet.

We have at least 6 questions to ask and will probably take 20 minutes of your time. We appreciate your consideration of our request.

We will contact you on 6 Sep 2016 to see if you are available or you may leave a message for us at 09484377269 (SMART). Thank you so much and we are looking forward to meet with you.

Sincerely,

Beverly Dianne D. Española (sgd)
 Researcher

Christian M. Fajiculay (sgd)
 Researcher

Ranil M. Montaril, MSECE (sgd)
 Faculty/Thesis Adviser

Noted by:

Michael B. dela Fuente, MSIGTS (sgd)
 Chairperson, DCS

Request Letter for Data Gathering



Polytechnic University of the Philippines
 College of Computer and Information Sciences
 Department of Computer Science
 Sta. Mesa, Manila



February 15 2017

Bascos, Jasmin A.
 Psychologist/Counseling
 Polytechnic University of the Philippines

Dear Madam:

We are students of Computer Science in the Department of Computer Science of Polytechnic University of the Philippines currently researching for our thesis, WRAP: Classification of Appropriate User's Age for Website Restriction through Metadata utilizing Data Mining and Naïve Bayes Classification. We are in awe of your expertise on the field psychology, and we are taking the opportunity to ask you for testing our thesis tool in determining the appropriate age for websites.

We would like to schedule a brief informational interview with you in person and if you would let us, we are humbly asking to record the conversation for educational purposes. We hope that you will be available on Thursday, 16 Feb 2017 around 10:30 am.

Here is the list of our agenda:

- Assessment on different websites.

We have at least 30 websites to be assessed and will probably take 30 minutes of your time. We appreciate your consideration of our request.

You may leave a message for us at 09484377269 (SMART), 09334572394 (SUN) or 09061638426 (GLOBE). Thank you so much and we are looking forward to meet with you.

Sincerely,

Beverly Dianne D. Española (sgd)
 Researcher

Christian M. Fajiculay (sgd)
 Researcher

Ranil M. Montaril, MSECE (sgd)
 Faculty/Thesis Adviser

APPENDIX C: RAW DATA

- a. For determining the Accuracy and Reliability of Predicting Age Appropriate for Accessing a Website.

Respondent # 1Name: Jeri (Nickname) Course/Year/Section: _____Birthday: February 21, 2000 Age: 16 Gender: F ☒ M ☐

Instruction: Write

✓ - If the website is accessed.

X - If the website is blocked.

INPUT	Trials					Average
	1	2	3	4	5	
Website 1	X	X	X	X	X	0/5
Website 2	/	/	/	/	/	5/5
Website 3	X	X	X	X	X	0/5
Website 4	/	/	/	/	/	5/5
Website 5	/	/	/	/	/	5/5

Respondent # 2Name: Alyanna (Nickname) Course/Year/Section: _____Birthday: June 23, 1999 Age: 17 Gender: F ☒ M ☐

Instruction: Write

✓ - If the website is accessed.

X - If the website is blocked.

INPUT	Trials					Average
	1	2	3	4	5	
Website 6	/	/	/	/	/	5/5
Website 7	X	X	X	X	X	0/5
Website 8	X	X	X	X	X	0/5
Website 9	/	/	/	/	/	5/5
Website 10	/	/	/	/	/	5/5

Respondent # 3Name: Emil (Nickname) Course/Year/Section: BSCS 4-1NBirthday: September 27, 1997 Age: 19 Gender: F ☐ M ☒

Instruction: Write

✓ - If the website is accessed.

X - If the website is blocked.

INPUT	Trials					Average
	1	2	3	4	5	
Website 11	/	/	/	/	/	5/5
Website 12	/	/	/	/	/	5/5
Website 13	/	/	/	/	/	5/5
Website 14	/	/	/	/	/	5/5
Website 15	/	/	/	/	/	5/5

Respondent # 4Name: Emerald (Nickname) Course/Year/Section: BSCS 4-1NBirthday: March 23, 1997 Age: 19 Gender: F ☒ M ☐

Instruction: Write

✓ - If the website is accessed.

X - If the website is blocked.

INPUT	Trials					Average
	1	2	3	4	5	
Website 16	/	/	/	/	/	5/5
Website 17	/	/	/	/	/	5/5
Website 18	/	/	/	/	/	5/5
Website 19	/	/	/	/	/	5/5
Website 20	/	/	/	/	/	5/5

Respondent # 5Name: Veinn (Nickname) Course/Year/Section: BSCS 4-3Birthday: October 28, 1997 Age: 19 Gender: F ☐ M ☐

Instruction: Write

✓ - If the website is accessed.

X - If the website is blocked.

INPUT	Trials					Average
	1	2	3	4	5	
Website 21	/	/	/	/	/	5/5
Website 22	/	/	/	/	/	5/5
Website 23	X	X	X	X	X	0/5
Website 24	/	/	/	/	/	5/5
Website 25	/	/	/	/	/	5/5

Respondent # 6Name: Vincent (Nickname) Course/Year/Section: BSCS 4-3Birthday: December 27, 1996 Age: 21 Gender: F ☐ M ☒

Instruction: Write

✓ - If the website is accessed.

X - If the website is blocked.

INPUT	Trials					Average
	1	2	3	4	5	
Website 26	/	/	/	/	/	5/5
Website 27	/	/	/	/	/	5/5
Website 28	/	/	/	/	/	5/5
Website 29	/	/	/	/	/	5/5
Website 30	/	/	/	/	/	5/5

Respondent # 7Name: Clowie (Nickname) Course/Year/Section: BSCS 4-1Birthday: August 13, 1996 Age: 21 Gender: F ☐ M ☒

Instruction: Write

✓ - If the website is accessed.

X - If the website is blocked.

INPUT	Trials					Average
	1	2	3	4	5	
Website 31	/	/	/	/	/	5/5
Website 32	/	/	/	/	/	5/5
Website 33	/	/	/	/	/	5/5
Website 34	/	/	/	/	/	5/5
Website 35	/	/	/	/	/	5/5

Respondent # 8Name: Joshua (Nickname) Course/Year/Section: _____Birthday: June 10, 2003 Age: 13 Gender: F ☐ M ☒

Instruction: Write

✓ - If the website is accessed.

X - If the website is blocked.

INPUT	Trials					Average
	1	2	3	4	5	
Website 36	X	X	X	X	X	0/5
Website 37	/	/	/	/	/	5/5
Website 38	X	X	X	X	X	0/5
Website 39	X	X	X	X	X	0/5
Website 40	/	/	/	/	/	5/5

Respondent # 9Name: Chesca (Nickname) Course/Year/Section: _____Birthday: July 20, 1999Age: 17Gender: F ☒ M ☐

Instruction: Write

✓ - If the website is accessed.

X - If the website is blocked.

INPUT	Trials					Average
	1	2	3	4	5	
Website 41	X	X	X	X	X	0/5
Website 42	X	X	X	X	X	0/5
Website 43	X	X	X	X	X	0/5
Website 44	/	/	/	/	/	5/5
Website 45	/	/	/	/	/	5/5

Respondent # 10Name: Rochelle (Nickname) Course/Year/Section: _____Birthday: April 29, 1999Age: 17Gender: F ☒ M ☐

Instruction: Write

✓ - If the website is accessed.

X - If the website is blocked.

INPUT	Trials					Average
	1	2	3	4	5	
Website 46	X	X	X	X	X	0/5
Website 47	X	X	X	X	X	0/5
Website 48	/	/	/	/	/	5/5
Website 49	/	/	/	/	/	5/5
Website 50	/	/	/	/	/	5/5

- b. For determining the Significant Difference between the expert's assessment and the our systems in terms of accuracy

Name: Jasmin Bascos

Age: 38

Gender: F ☒ M ☐

Position: Guidance Councilor CCIS PUP

Instruction: Write / if the website is applicable for everyone otherwise, X if the website is applicable for 18 years old and above only.

INPUT	Expert Assessment	System's Assessment
Website 1	/	/
Website 2	X	X
Website 3	X	X
Website 4	X	/
Website 5	X	X
Website 6	X	/
Website 7	X	/
Website 8	/	/
Website 9	/	/
Website 10	/	/
Website 11	/	/
Website 12	X	X
Website 13	/	/
Website 14	X	/
Website 15	/	/
Website 16	X	X
Website 17	X	X
Website 18	/	/
Website 19	X	X
Website 20	X	X
Website 21	X	X
Website 22	X	X
Website 23	X	X
Website 24	X	X
Website 25	X	X
Website 26	X	X
Website 27	X	X
Website 28	X	X
Website 29	X	X
Website 30	X	X

APPENDIX D: SCREENSHOTS

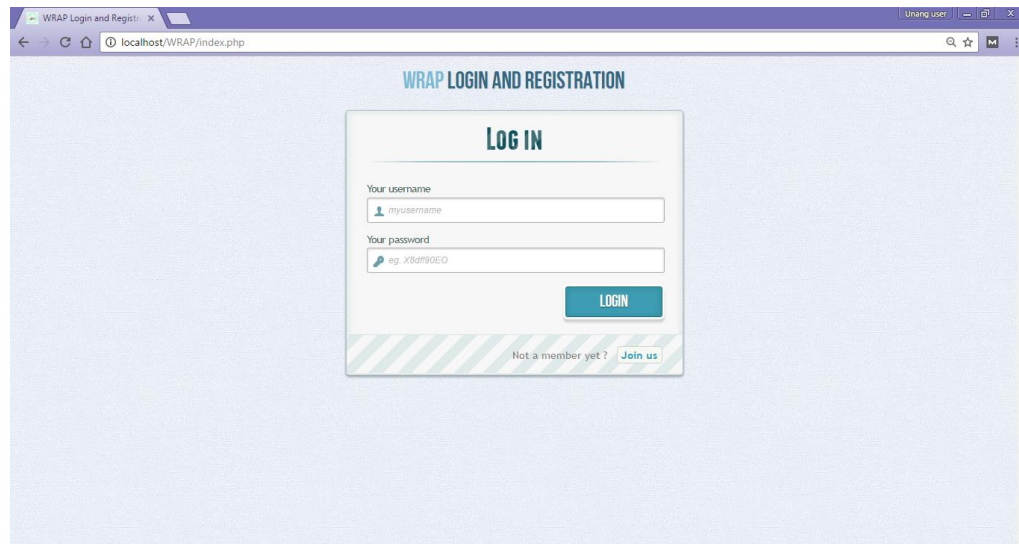


Figure D.1: Log-in Page for Existing Users.

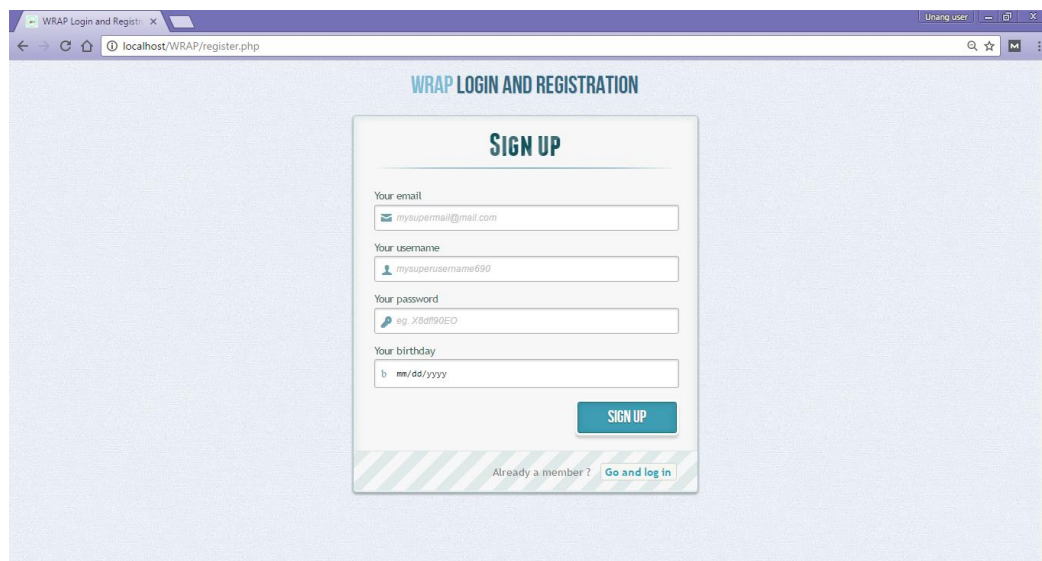


Figure D.2: Sign-up Page for New Users.



Figure D.3: Searching Bar for the system WRAP.

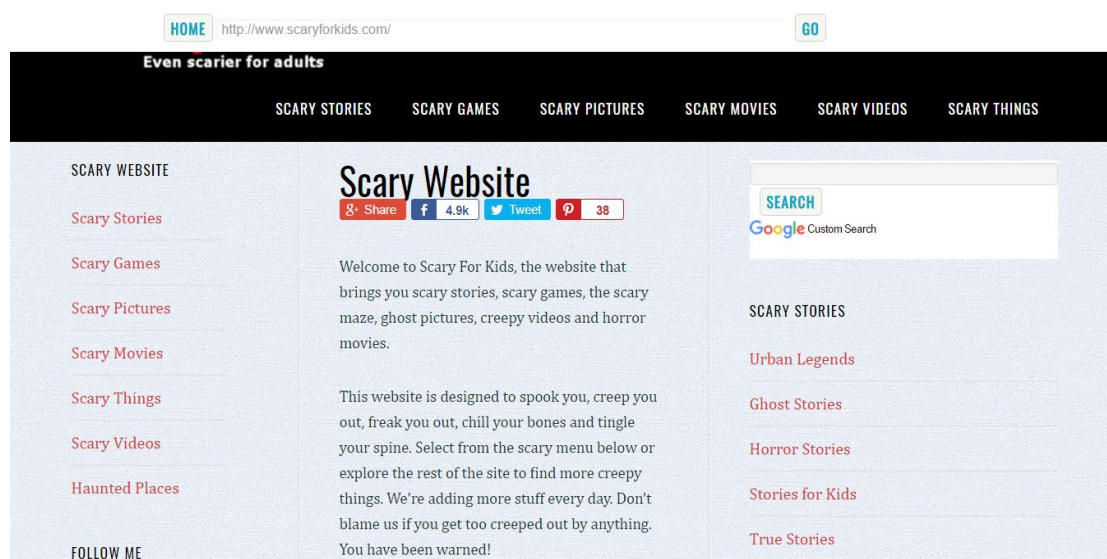


Figure D.4: Appearance of the System when the Website is accessed.

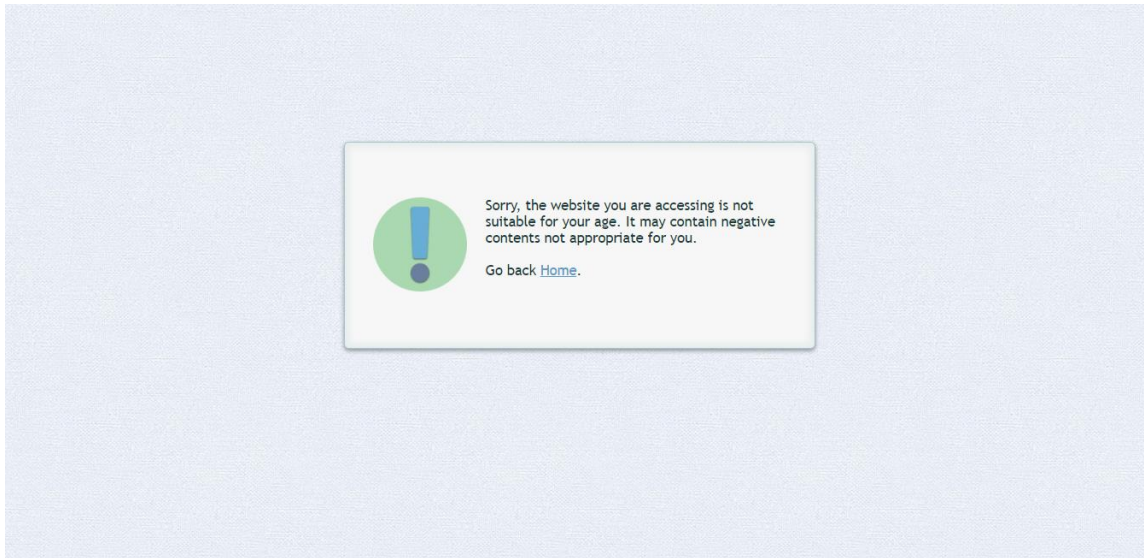


Figure D.5: Appearance of the System when the Website is blocked.

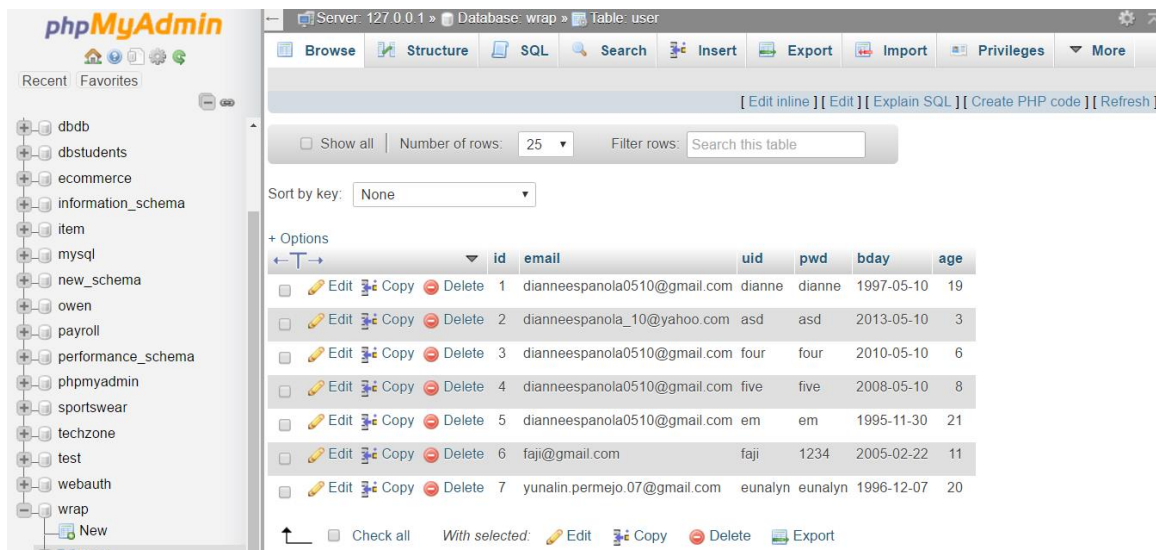


Figure D.6: Database for the System WRAP using PHP MyAdmin.

APPENDIX E: IMPLEMENTATION REPORT

Implementation Report

Introduction

WRAP: Classification of Appropriate User's Age for Website Restriction through Metadata utilizing Data Mining and Naïve Bayes Classification is a tool that classifies a websites and analyzes the required age for a user. It then determines whether the user can access or be restricted depending on the age of the user and the required age

Problem Statement

The study WRAP aims to develop a system that restrict a user from a website by determining the user's age and the age appropriate to access a certain website through its metadata by utilizing data mining and Naïve Bayes Classification and further, restricts user if not suitable for them.

The following problems are to be solved after the implementation:

1. What is the accuracy of the system when getting the predicted age appropriate in accessing the website using Precision, Recall and F-measure?
2. What is the reliability of the system in resulting the predicted age in accessing the website?
3. What is the significant difference in the expert's assessment on predicting the age appropriate for accessing a website and our system's resulted age in terms of accuracy?

Respondents

The respondents consisted of two groups with different age groups. The first age group is 17 and below, with 5 respondents of students of Polytechnic University of the Philippines. The other age group is 18 and above, with also 5 respondents of students of the said campus. In total, the number of respondents is 10. A Guidance Counselor is also considered as a respondent.

Table E.1: Number Respondents Table

Respondents	Number of Respondents
Expert (Guidance Counselor)	1
Student with an age of Below 18	5
Student with an age of 18 and Above	5

Time Frame

Table E.2: Time Frame for the Implementation of WRAP

Activity	February 2017			Remarks
	Week 1	Week 2	Week 3	
Collection of Data				DONE
Assessment of Data				DONE
Tabulation and Analysis				DONE
Calculation of Data				DONE
Conclusion				DONE

Implementation Procedures

The researchers gathered random test websites in the internet. After selecting 50 websites through random selection, the researchers checked if it's positive or negative website. The sites were tested individually and the results were tallied by comparing the answer of the system in the answer of the Guidance Counselor in the same sites. The researchers then evaluated the system through correctly identified the required age (TP), incorrectly identified the required age (FP), incorrectly rejected the required age (FN), correctly rejected the age required (TN).

These are some **Images taken during the Implementation** at Polytechnic University of the Philippines, Sta. Mesa Manila. The implementation was conducted by the group with Ms. Jasmin A. Bascos and the respondents.

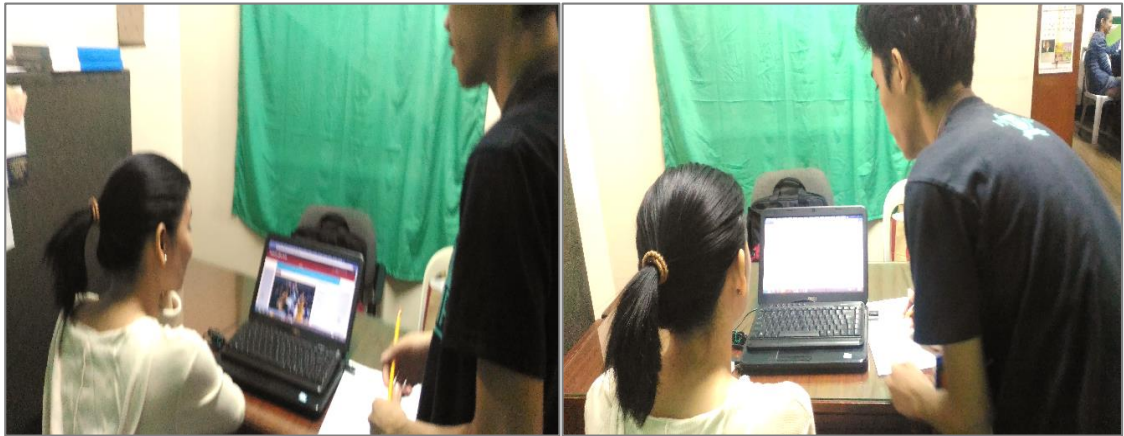


Figure E.1: Ms. Bascos assessing the 30 sample Websites.



Figure E.2: Researchers with Ms. Jasmin A. Bascos



Figure E.3: The Researcher with one of the Respondent

Issues and Concerns

The following are the issues and concerns that were encountered during implementation:

1. Slow internet connection.
2. Low end hardware.

APPENDIX F: CURRICULUM VITAE**Espanola, Beverly Dianne D.**

Email: dianneespanola0510@gmail.com

Contact: 09334572394



Personal Details:

Date of Birth:	May 10, 1997
Age:	19
Gender:	Female
Nationality:	Filipino
Language:	English and Filipino
Address:	54 B. M. Yulo Street Mandaluyong City

Skills:

- Programming Languages: Java, Visual Basic, C
- Web Development Skills: HTML, CSS and PHP
- Database Knowledge: MySQL, Microsoft Access
- Knowledgeable in Microsoft Office applications

Fajiculay, Christian M.

Email: chrisfajiculay@gmail.com

Contact: 09484377269



Personal Details:

Date of Birth:	September 28, 1997
Age:	19
Gender:	Male
Nationality:	Filipino
Language:	English, Filipino and Visaya
Address:	4461 Old Sta. Mesa St. Sta. Mesa, Manila

Skills:

- Programming Languages: Java, VB.Net, C
- Web Development Skills: HTML, CSS, PHP and JS
- Database Knowledge: MySQL, SQLite, Microsoft Access
- Knowledgeable in Microsoft Office applications