**Product Comparison and Analysis using Data Mining**

A Thesis

Presented to the Faculty of the

College of Computer and Information Sciences

Polytechnic University of the Philippines

In Partial Fulfilment

Of the Requirements for the Degree

Bachelor of Science in Computer Science

Averos, Christian M.

Delicano, Jobea Ann F.

//year

AUTHORIZATION

I/We hereby declare that I am/we are the sole author/s of the thesis.

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

I/We further authorize the Polytechnic University of the Philippines and the College of Computer and Information Sciences 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.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The thesis **“SilenSwear: An Automatic English Swearword Detection and Muting Using Keyword Spotting”** submitted and presented by Medrano, Jamie Paul R., and Sefe, Marlon M. in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science has been

Recommended for Approval and Acceptance: Date:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Faculty-In-charge

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Technical Advisor

Approved by the Thesis Defense Panel:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Accepted for the Department of Computer Science

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Department Chairperson

Accepted for the College of Computer and Information Sciences

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dean of the College

# ABSTRACT

**ACKNOWLEDGEMENT**

Table of Contents

[ABSTRACT iii](#_Toc448577050)

[List of Figures vii](#_Toc448577051)

[List of Tables viii](#_Toc448577052)

[Chapter 1: The Problem and Its Background 1](#_Toc448577053)

[1.1 Introduction 1](#_Toc448577054)

[1.2 Background of the Study 1](#_Toc448577055)

[1.3 Conceptual Framework 3](#_Toc448577056)

[1.3.1 Conceptual Framework of the System 3](#_Toc448577057)

[1.3.2 Conceptual Framework of the Study 4](#_Toc448577058)

[1.4 Statement of the Problem 4](#_Toc448577059)

[1.5 Significance of the Study 5](#_Toc448577060)

[1.6 Scope and Limitation 5](#_Toc448577061)

[1.6.1 Scope and Limitation of the System 5](#_Toc448577062)

[1.6.2 Scope and Limitation of the Study 6](#_Toc448577063)

[1.7 Definition of Terms 6](#_Toc448577064)

[Chapter 2: Review of Related Literature and Studies 8](#_Toc448577065)

[2.1 Review of Related Literature 8](#_Toc448577066)

[2.1.1 Impact of Swearword in Movies on A Child’s Mind 8](#_Toc448577067)

[2.1.2 Capacity of a Child in Imitating 9](#_Toc448577068)

[2.2 Review of Related Studies 9](#_Toc448577069)

[2.2.1 Automatic Speech Recognition 9](#_Toc448577070)

[2.2.2 Keyword Spotting 11](#_Toc448577071)

[2.2.3 Pattern-matching Approach 11](#_Toc448577072)

[2.2.4 Speech Recognition Using Hidden Marcov Model 12](#_Toc448577073)

[2.2.5 Keyword Spotting Using Hidden Marcov Model 13](#_Toc448577074)

[2.3 Synthesis of the Study 14](#_Toc448577075)

[Chapter 3: Research Methodology 15](#_Toc448577076)

[3.1 Research Method Used 15](#_Toc448577077)

[3.2 Research Paradigm 15](#_Toc448577078)

[3.3 System Architecture 16](#_Toc448577079)

[3.3.1 Preprocessing 16](#_Toc448577080)

[3.3.2 Main Process 16](#_Toc448577081)

[3.3.3 Video with Muted Swearwords 17](#_Toc448577082)

[3.4 Population Frame and Sample 17](#_Toc448577083)

[3.5 Description of the Respondents 17](#_Toc448577084)

[3.6 Sampling Technique 17](#_Toc448577085)

[3.7 Instrumentation 18](#_Toc448577086)

[3.8 Data Gathering Procedure 19](#_Toc448577087)

[3.9 Statistical Treatment 19](#_Toc448577088)

[3.9.1 Precision, Recall, and F-measure 19](#_Toc448577089)

[3.9.2 Cronbach’s Alpha Test () 20](#_Toc448577090)

[Chapter 4: Presentation, Analysis and Interpretation of data 22](#_Toc448577091)

[4.1 Accuracy of the System 22](#_Toc448577092)

[4.2 Reliability of the System 23](#_Toc448577093)

[4.3 Reliability Statistics 23](#_Toc448577094)

[Chapter 5: Summary of findings, Conclusion and Recommendation 24](#_Toc448577095)

[5.1 Summary of findings 24](#_Toc448577096)

[5.2 Conclusion 24](#_Toc448577097)

[5.3 Recommendation 24](#_Toc448577098)

[5.3.1 Context-Based Approach 25](#_Toc448577099)

[5.3.2 Real Time Detection 25](#_Toc448577100)

[REFERENCES 26](#_Toc448577101)

[Appendix A 29](#_Toc448577102)

[Appendix B 41](#_Toc448577103)

[Appendix C 45](#_Toc448577104)

[CURRICULUM VITAE 48](#_Toc448577105)

# List of Figures

[Figure 1.1 Conceptual Framework of the System 3](https://d.docs.live.net/9860edf0ac463815/Documents/Thesispaper_SilenSwear_rev-AJ.docx" \l "_Toc448571741)

[Figure 1.2. Conceptual Framework of the Study 4](https://d.docs.live.net/9860edf0ac463815/Documents/Thesispaper_SilenSwear_rev-AJ.docx" \l "_Toc448571742)

[Figure 3.2 4th Generation Technique Paradigm 15](#_Toc448571743)

[Figure 3.3 System Architecture……………….………………………………………………………………………..16](#_Toc448572809)

# List of Tables

[Table 1.1 List of Swearwords 6](#_Toc448573101)

[Table 4.1 Testing Results for Accuracy 21](#_Toc448398673)

[Table 4.2 Testing Results for Reliability 22](#_Toc448398673)

[Table 4.3 Reliability Statistics 22](#_Toc448398673)

[Table A.1 Experiment Paper 23](#_Toc448398673)

[Table A.2 Experiment Paper with Raw Data 25](#_Toc448398673)

[Table B.1 Time Frame for Implementation 36](#_Toc448398673)

Chapter 1: The Problem and Its Background

1.1 Introduction

In the fast-paced modern world, where technology progresses rampantly, we try to keep up with today's trends as well as needs. With the rise of technology, clinging unto technology makes our lives easier, and some experts say that the internet can now be consider a need along with food and shelter [Pastore, 1998]. Everything is being digitalized as for our communication and work-processes, and with the use of the internet and its rapid development, delivery systems are getting more and more popular alongside of online shopping websites (OSW). With that being said, it has a great affect on us, Filipinos, especially modern Filipina buying products on the internet through OSWs.

Most people all over the Philippines have access to the internet and they are still growing in numbers but there's a problem, how can we diminish the hard work process of search through many OSWs just to get the best deals? As a Filipino, I'm used to seeing my mother go through many store just to get the optimal price that fits in our budget, sometimes, we even come to a point of bargaining or asking the store to lower the price if we're going to buy bulk of their products. Users would search for many websites to compare and analyze what is the best for their searched similar products, weighing in the reviews, quantity, and even quality just to make sure that they're making a good decision upon purchasing such product. Therefore, the researchers decided to develop a system to help with troubled online shoppers (OS).

1.2 Background of the Study

Most Filipino shoppers have a hard time jumping to another store into other store just to find the best deals of their wanted product, may it be a new pair of shoes, shining jeans, or maybe a newly released branded cap. As much as we can, we bargain, bargain, and bargain. Most modern Filipino women would go to mall after mall just to buy off on-sale product which is also true for old-fashion mothers who would prefer to bargain just to get deals from their market suppliers. In today’s technological advantage, many shoppers prefer e-commerce as their mode of shopping [Kitonyi, 2017], primarily because internet is easy to access and buying is just a click away.

Filipinos like to bargain, since the old times, Filipino mothers would go to the market with a fixed money and get much more worth of their money for. Just imagine the hard work of going through and bargaining to many stores just to lessen the pay for a similar product as well as avoiding similar low quality products. This is also true to online shoppers looking for looking for a specific product in an online store, they would search many online store just to get the best deal in their opinion. They would scan for reviews of the said product and to avoid getting scammed or getting less than what they expected. One of the most common problem in OSWs is that there are scammers always present [**Bolido**, 2014] even in the most popular OSWs.

The researchers plan to build a system that caters in the selection of similar products in which the user has to decide to buy or not. Getting the data only from the trusted OSW to minimize cyber market scamming and increasing the satisfaction rate of the users in selecting. Filtering only the 'best' deals in terms of reviews, quantity, and popularity to get the most-likely product to be picked by the user.

1.3 Conceptual Framework

1.3.1 Conceptual Framework of the System

Figure 1.3.1 shows the conceptual framework of the System where the User’s search terms are the input data. The processes are getting the data from OSW and filtering out redundant data. The output data are the collection of best deals.

|  |  |  |
| --- | --- | --- |
| **INPUT** | **PROCESS** | **OUTPUT** |
| * User search term | * Getting data from site * Filtering out redundant data | * Best Deals |

Figure 1.3.1 Conceptual Framework of the System

1.4 Statement of the Problem

This study aims to develop a system that gets the best deal from various online shopping websites based on user’s search terms.

1. Using the a-priori algorithm, what is the accuracy of getting the relevant data from online shopping websites?
2. How should the algorithm be devised in order to get the best deals out of the listed relevant product?
3. What is the accuracy and reliability of the devised algorithm in getting the best deals?

**1.5 Significance of the Study**

The system will benefit the following people:

Online Shoppers. This study would benefit shoppers who wants to lessen their time on selecting the best deal.

Store Owners. This will benefit them by being able to search how they would price their own products.

Future Researchers. This study will be a help as a guiding reference in making a system related to e-commerce.

1.6 Scope and Limitation

1.6.1 Scope and Limitation of the System

The system will mine data from popular retailer online shopping websites and will not include second hand online shopping websites, specifically:

1. Lazada (lazada.com.ph)
2. Shopee (shopee.ph)
3. Amazon (amazon.com)
4. Zalora (zalora.com.ph)

The scope in assessing the best deals will base on the relevancy, availability, price including the shipping fee, specifications, and numerical reviews regarding the product. The product prices will be displayed in Philippine Peso.

The system will not be affiliated in selling products and will only act as a recommender system for finding the best deals.

**1.6.2 Scope and Limitation of the Study**

This study will focus on the accuracy of the finding the best deals of a product selected by the user at the top online shopping websites.

1.7 Definition of Terms

Data Mining - process of extracting information from a large sets of data.

Online shoppers - product consumers that uses Online Shopping Websites as their medium for purchasing.

Online Shopping Websites - form of e-commerce that allows users to purchase commodities over the internet with the help of a web browser.

Scammers - people who extorts or uses dirty tactics to gain information, power or money.

**References**

Kitonyi, N. (2017, March 07). *E-commerce Is Killing Traditional Retail*. Retrieved from https://www.gurufocus.com/news/490164/ecommerce-is-killing-traditional-retail

Bolido, L. B. (2014, April 23). *Safety In Online Buying and Selling*. Retrieved from http://lifestyle.inquirer.net/157606/safety-in-online-buying-and-selling/

Pastore, M. (1998, December 03). *Internet Becoming Necessity to Users*. Retrieved from https://www.clickz.com/internet-becoming-necessity-to-users/72138/

Chapter 2: Review of Related Literature and Studies

2.1 Review of Related Literature

**2.2 Review of Related Studies**

2.3 Synthesis of the Study

Chapter 3: Research Methodology

3.1 Research Method Used

3.2 Research Paradigm

3.3 System Architecture

Figure 3.3 System Architecture

3.3.1 Preprocessing

3.3.2 Main Process

3.4 Population Frame and Sample

3.5 Description of the Respondents

3.6 Sampling Technique

3.7 Instrumentation

3.8 Data Gathering Procedure

## 3.9 Statistical Treatment

3.9.1

Chapter 4: Presentation, Analysis and Interpretation of data

The purpose of this study was to develop a systems:

4.1 Accuracy of the System

4.2 Reliability of the System

4.3 Reliability Statistics

Chapter 5: Summary of findings, Conclusion and Recommendation

5.1 Summary of findings

5.2 Conclusion

5.3 Recommendation

5.3.1 Context-Based Approach

5.3.2 Real Time Detection

# **REFERENCES**

x

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

x

# Appendix A

Appendix A. Experiment Paper

Table A.1 Experiment Paper

Table A.2 Experiment Paper with Raw Data

Appendix B

Appendix B. Implementation Report

**Introduction**

**Problem Statement**

**Respondents/Subject**

**Time frame**

Table B.1 Time Frame for Implementation

|  |  |
| --- | --- |
| Activity / Tasks | Schedule |
| **Audio Gathering** |  |
| **Testing the Accuracy of the System** |  |
| **Computing the System's Accuracy** |  |
| **Chapter 4** |  |
| **Chapter 5** |  |
| **Finalizing Document** |  |

**Participants:** Researchers **Required Tools and Equipment:** Laptop

**Implementation Procedures**

The developed system is intended to be tested in terms of Accuracy and Reliability.

**The following are done by researchers during system implementation:**

1. **Gathering Data.**
2. **Testing of the System**
3. **Checking of the output**
4. **Determining the Accuracy and Reliability of the system.**

**Illustrations and Pictures**

Photos: Development, Testing and Implementation Period

Photos: Development, Testing and Implementation Period

Photos: Development, Testing and Implementation Period

Photos: Development, Testing and Implementation Period

Appendix C

**Appendix C. Graphical User Interface**

**Graphical User Interface Screenshots**