FilterX: A Web Extension for Moderating Online Hate Speech

A Thesis
Presented to the Faculty of
Information and Communications Technology Program
STI College Balagtas

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

Beatriz G. De Guia
Eman Joseph T. De Leon
Jordan Limwell C. Marcelo
Lois Alysson R. Marquez

November 2024

ENDORSEMENT FORM FOR ORAL DEFENSE

TITLE OF RESEARCH: **FilterX: A Web Extension for Moderating Online Hate Speech**

NAME OF PROPONENTS: Beatriz G. De Guia
Eman Joseph T. De Leon
Jordan Limwell C. Marcelo
Lois Alysson R. Marquez

In Partial Fulfilment of the Requirements
for the degree Bachelor of Science in Computer Science
has been examined and is recommended for Oral Defense.

ENDORSED BY:

Mr. John Irwin T. Vendivil
Thesis Adviser

APPROVED FOR ORAL DEFENSE:

Engr. Regina R. Mape, MIT
Thesis Coordinator

NOTED BY:

Engr. Regina R. Mape, MIT
Program Head

November 2024

APPROVAL SHEET

This thesis titled **FilterX: A Web Extension for Moderating Online Hate Speech**, prepared and submitted by **Beatriz G. De Guia, Eman Joseph T. De Leon, Jordan Limwell C. Marcelo** and **Lois Alysson R. Marquez**, in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science, has been examined and is recommended for acceptance and approval.

Mr. John Irwin T. Vendivil
Thesis Adviser

Accepted and approved by the Thesis Review Panel
in partial fulfillment of the requirements for the degree of
Bachelor of Science in Computer Science

<Panelists' Given Name MI. Family Name>
Panel Member

<Panelists' Given Name MI. Family Name>
Panel Member

<Panelists' Given Name MI. Family Name>
Lead Panelist

APPROVED:

Engr. Regina R. Mape, MIT
Thesis Coordinator

Engr. Regina R. Mape, MIT
Program Head

November 2024

ACKNOWLEDGEMENTS

The researchers would like to thank the following:

Thesis Coordinator, Engr. Regina R. Mape, MIT, for her guidance and assistance throughout the research process, from conceptualization to completion. Her insights and feedback greatly contributed to the quality of the study.

Thesis adviser, Mr. John Irwin T. Vendivil, for his mentorship, expertise, and support. His constructive criticism and encouragement were instrumental in refining the research methodology and analysis.

Thesis Review Panel, <state their contributions to your research> for their thorough evaluation and constructive feedback during the oral defense. Their insights and suggestions helped enhance the validity of the study.

Parents and/ or Guardians, for their support, understanding, and encouragement throughout the journey. Their love and encouragement provided the researchers with the motivation to persevere and excel.

Friends and inspirations, for their moral support, encouragement, and understanding during challenging times. Their presence and encouragement provided the researchers with the motivation to overcome obstacles, strive for excellence and

Others classmates, colleagues, and mentors, for their valuable insights, feedback, and encouragement throughout the research process. Their contributions and support were appreciated and guided us in the successful completion of this thesis.

Lastly, the researchers would like to extend their heartfelt gratitude to all those who, directly or indirectly, contributed to the completion of this thesis. Their support and encouragement have been valuable and appreciated. Thank you all for your support and encouragement.

ABSTRACT

Title of research**:** **FilterX: A Web Extension for Moderating Online Hate Speech**

Researchers: **Beatriz G. De Guia**
Eman Joseph T. De Leon
Jordan Limwell C. Marcelo
Lois Alysson R. Marquez

Degree: **Bachelor of Science in Computer Science**

Date of Completion: **June 2025**

Keywords: **Hate Speech Detection, Text Classification, Web Extension, Natural Language Processing, Machine Learning, Logistic Regression, LightGBM, ELECTRA Embedding, BERT-based Model RoBERTa**

FilterX is an innovative website add-on that addresses the global issue of hate speech on the internet. Hate speech on the internet will be addressed by using sophisticated natural language processing (NLP) tools and machine learning software to monitor and restrict content generated by users across multiple online platforms in real time. The extension takes a multi-step approach to detecting hate speech through analyzing text, context, and user actions. Users can set up their own preferences for moderation as well as specify thresholds to create safe online spaces tailored specifically for them and their needs with an easy-to-navigate UI. This preventative moderation tool also offers educational materials and interventions designed to encourage meaningful discussions among people rather than spreading hateful messages. The document discloses how FilterX was developed, implemented, and appraised, as well as how it has been instrumental in advancing a polite online society characterized by inclusive digital environments.

Table of Contents

LIST OF TABLES

List of Figures

List of Diagram

List of Appendices

Appendix Page

INTRODUCTION

The expansion of social online platforms has rewritten the way individuals engage with digital spaces and created opportunities for interaction, knowledge-sharing, and communication on a global scale. More than 3.5 billion people were active on social media by 2019 or about one-third of the world's population and more than two-thirds of all internet users (Osp, 2019). However, the heightened connectivity has also fueled the spread of harmful conduct, and there are causes for concern about online hate speech. The internet, which operates often with rules established by privately-motivated for-profit entities, continues to be shown to be a fertile ground for hate speech or other dangerous content. This trend imperils healthy discourse, hitting hardest at already marginalized communities, and threatening their well-being and safety.

Hate speech online has emerged as a problem that threatens the very foundation of civil and free communication globally. Soaring cases of discrimination, harassment, cyberbullying, and related matters dictate how online conversations are moderated in real-time. Nave and Lane (2023) in "Countering Online Hate Speech: How Does Human Rights Due Diligence Impact Terms of Service," delved into the complex issue of hate speech management. It underlines the necessity for corporate human rights due diligence on online platforms. As online platforms like Facebook step up efforts to remove 31 million pieces of hate speech between April and June 2021 (Statista, 2023), most of their efforts are still reaction-based in nature and leave much to be desired in proactive moderation techniques. The gap glaringly is in real-time detection and moderation of hate speech online.

Most of the existing approaches rely on user reporting or post-publication moderation, making hateful content visible until the time of action. This delay in response tends to foster hate speech more, hence the need to come up with better tools that are in a position to deal with the issue immediately after it is published. "FilterX" is a browser extension that detects and moderates hate speech in real-time. FilterX uses machine learning and NLP to bridge this gap through an online real-time solution for platforms on the web. The

contribution of this research is in its development of a tool not only practical in solution but also available to solve the increasingly problematic issue of hate speech online.
FilterX fits well with the current trend in content moderation through web extensions. It is an effective solution for protecting users against hate speech. Real-time hate speech detection follows, having some reduction of its prevalence but more importantly, contributes to a safer and more inclusive cyberspace. This extension can also be easily taken over because of its compatibility with Google Chrome, Microsoft Edge, Opera, and Brave major browsers, and access and impact on these users. At a time of such extensive growth and change in communication online, the need for tools such as FilterX becomes increasingly more important. This paper is meant to be both a contribution to academic research in this area and applicable within the everyday digital world, fostering a digital world where respect, inclusion, and safety are rewarded for all users. It will give birth to a concrete automated solution to hate speech; it is going to put firmly in place new standards for content moderation, also opening doors for further innovations that shall buttress the digital world.

Background of the problem

The internet has been very powerful and allows individuals to communicate with the entire world, share ideas, and talk about almost everything. Lately, this increase in connectivity does not come without the challenge of "harmful content," such as hate speech, that needs more regulation. According to the Statista Research Department, Facebook removed 7.2 million pieces of hate speech content, down from 7.4 million in the first quarter of 2024. Between April and June 2021, the social network removed a record number of over 31 million pieces of hate speech. The reality of hate speech reflects a major gap within digital content moderation in its inability to respond to a high increase in harmful language with existing measures. Online hate speech targets minority communities, causing emotional pain social isolation, and even violence offline.

Most social media have embraced policies to reduce the problems but depend on reporting users and also post-moderation rather than intervention in real-time or proactively so that users will be exposed to damaging content before it is identified and

taken down, thus further perpetuating the problem. This thus generates an urgent demand for something much more efficient and time-saving in the context of the growing importance given to online interactions in personal and professional spheres. In pursuit of the objectives of this research, therefore, there is a need to create an internet browser extension that will detect and filter hate speech on a real-time basis.

Since it is intended to monitor the diffusion of harmful content in real-time, the tool hopes to limit drastically the reach of harmful content on internet-based platforms and to make online spaces safer for the user. Compatible with widely used browsers like Google Chrome, Microsoft Edge, Opera, and Brave, this extension is going to apply machine learning algorithms and natural language processing (NLP) techniques to capture the trends in language that have become common indicators of hate speech. Therefore, the novelty of the study lies in its potential to bridge a serious gap in real-time hate speech moderation. In fact, the existing solutions are mainly inefficient because they rely on manual moderation or automated means that cannot keep up with the dynamism of online communication. "FilterX" is therefore proposed as a web extension through which a more robust and scalable solution shall be developed in terms of detecting hate speech as it happens and therefore preventing its spread.

This project will have positive impacts for different stakeholders, including online communities, the users of social media, and even administrators of the online platforms. The tool will ensure that users, particularly those who are in vulnerable groups, are protected from psychological harm associated with hate speech exposure. Secondly, the tool will also assist platform administrators in creating an improved content moderation strategy without having to seek information from user reporting alone.

The implications of the current study go beyond content moderation since it tries to contribute toward a more holistic field of online safety and responsible digital engagement. In that way, it helps build a more respectful and inclusive internet- an addition to a better life for individuals and society at large.

At the same time, digital communication platforms are developing at high rates, and emerging solutions will be needed to deal with hate speech. The suggested browser

extension in this study is a real-time and automated solution to track harmful content online and regulate it accordingly. Since online interactions are increasingly becoming important day by day, the need for such tools in the future will be much more paramount to promote healthier and safer digital environments.

Overview of the current state of the technology

The present landscape of technology in online hate speech detection is intricately intertwined with advancements in data and AI applications, as evidenced by Matt Turck's Data and AI Landscape 2019 chart (FutureLearn, 2022). Some of the critical technologies include natural language processing technology and machine learning algorithms in this area, FutureLearn, 2022. This is logistic regression-the supervised learning technique that distinguishes features on simplicity and effectiveness in identification of hate speech from labeled data. Major social media platforms such as Facebook, YouTube, and Twitter are actively fighting against hate speech through using AI algorithms provisioned with NLP techniques into content moderation strategies. By the same token, their popularity complements broader contemporary research efforts to address problems in the digital space, particularly focusing on the need for algorithmic fairness and transparency within content moderation (FutureLearn, 2022). In such improvements, however, there are still challenges within the field regarding the nuanced and context-dependent nature of hate speech. Some of the challenges encountered include overlapping and competing models in multi-language detection, the lack of availability of labeled datasets, and the problem of being unable to keep up with emerging trends (Kovács et al., 2021; MacAvaney et al., 2019). To find their way out of the problems and continue achieving improvements in precision and fairness, hate speech detection systems would need to respond to these problems. Hence, the application of NLP techniques like sentiment analysis contributes to a better understanding of the complexities associated with hate speech, allowing for real-time moderation and optimization of cross-lingual effectiveness (Boishakhi et al., 2021). Conclusion Online hate speech detection reflects a dynamic interplay among various tools and techniques in its technological state. Although significant developments have been realized, the area remains dynamic, and non-ending challenges have to be continually adapted to overcome them and generate progress that can support a safer and

more inclusive digital environment.

Objectives of the study

General Objective

The primary objective of this study is to develop a Browser extension for hate speech detection and filtering compatible with Google Chrome, Microsoft Edge, Opera, and Brave. The goal of this research is to develop an extension that will detect and filter instances of hate speech in real-time. The extension will focus on analyzing text content and identifying trends in language indicative of hate speech. Throughout the development process, machine learning algorithms and natural language processing techniques will be utilized to enhance the accuracy and effectiveness of the extension in detecting hate speech. The ultimate aim of this project is to develop practical and accessible tools to help reduce the prevalence of hate speech on internet-based platforms.

Specific Objectives:

Model

- Train a model that can react in real-time

  Develop a model optimized for real-time data processing, capable of providing immediate predictions to ensure timely intervention against hate speech. This involves using efficient algorithms and leveraging hardware acceleration.

- Train a model that has fast inference speed

  Ensure the model makes predictions quickly by optimizing the architecture, reducing model complexity, and utilizing techniques like quantization and pruning. Fast inference is critical to maintain seamless user experience and quick response times.

- Train a model with a capacity that meets the study's requirements

Design the model to handle the complexity and variety of tasks needed for effective hate speech detection. This includes training on a diverse dataset to learn the necessary features and patterns, and performing regular evaluations to ensure it meets performance benchmarks.

- Train a model that could perform well in general as it would be used in different webpages (it should not be limited to detecting on specific websites only)

  Train the model on a broad and diverse dataset to ensure it can generalize well across various web environments. Employing techniques like cross-validation and transfer learning can enhance its robustness and ensure consistent performance on different websites.

Extension

- Develop an extension that would maximize the usage of the deployed machine-learning model

  Create a browser extension that integrates the trained model, providing features like automatic detection, analysis, and decision-making capabilities directly within the browser. This includes developing a user-friendly interface and ensuring compatibility with multiple browsers while optimizing for performance and resource efficiency.

- Develop a good and efficient system bus for data transferring

  Design a robust communication protocol within the extension to ensure quick and secure data transfer between the browser and the model. Techniques like data compression, asynchronous processing, and efficient serialization/deserialization will help minimize latency and handle high throughput.

- Add a function that would let the user press the button to activate the scan & take action

Implement a user-initiated scan feature in the extension. This function should provide immediate feedback, displaying the model's predictions and actions in a user-friendly manner. Visual indicators and progress bars can enhance the user experience.

- Add a function that would let the user type a specific input and pass it on the model to make a prediction based on that data alone

  Provide a text input feature where users can enter specific data for the model to predict. The extension should clearly and promptly display the prediction result. Ensuring the function handles various input types correctly and delivers meaningful output is crucial.

- Add a report function so that the user can give feedback on the performance of the model

  Include a reporting feature allowing users to rate the model's accuracy, suggest improvements, and report issues. This function should be accessible and user-friendly, offering fields for comments, ratings, and possibly capturing screenshots of incorrect predictions. User feedback will be vital for continuous improvement of the model, guiding further training and fine-tuning efforts.

Scope and limitations of the study

Scope of the Study

This research aims to comprehensively investigate online hate speech within the context of society, focusing on user groups such as teenagers (ages 15 to 18) and the general population of internet users (ages 16 and above) at STI College Balagtas. The study intends to provide insights into the issues and patterns of online hate speech, guide the creation of more efficient moderation solutions, and enhance algorithms like FilterX for hate speech detection. The temporal scope of the study is limited until the first semester of the academic year 2024–2025, ensuring coverage of recent instances of hate speech on the internet at STI College Balagtas.

A quantitative approach will be applied, which includes defining specific research questions, identifying variables, carefully selecting participants, conducting pre-test measurements to establish initial information, and collecting post-test data. This study's scope is to analyze the impact of an individual approach on the dependent variable in a real-world situation while understanding the limitations of non-randomized responsibility and emphasizing a thorough review of any factors that may confuse the methodology.

The study will incorporate the use of the FilterX Chrome extension, which detects hate speech on web pages using machine learning models for both the English and Filipino languages. This extension is usable by all users with desktop PCs and Chromium-based browsers, such as Google Chrome, Microsoft Edge, and Brave. The FilterX system includes components such as content scripts, background scripts, and FastAPI services for inference. The content script embedded in web pages extracts sentences for analysis, while the background script processes these sentences and forwards them to the FastAPI service. The service performs hate speech detection and returns predictions, which are then used to update the web page's content in real-time.

By leveraging this system, the research aims to understand the effectiveness of automated hate speech detection and moderation tools. This will involve evaluating how well the extension performs in identifying and filtering hate speech, and its impact on user experience. The study will also assess user control features, such as adjusting the sensitivity of the filter and the ability to manually scan web pages, to determine their practicality and effectiveness in real-world applications.

Additionally, the findings from this research will contribute to the development of more robust and user-friendly hate speech detection tools. These insights will be valuable for developers, educators, and policymakers in crafting strategies to mitigate the spread of hate speech online, thereby fostering a safer and more inclusive digital environment for all users, particularly within the STI College Balagtas community.

Limitations of the Study

While the FilterX extension offers a promising solution to mitigate online hate speech,

several limitations may influence the study's outcomes:

- Limited Generalizability

The findings may be specific to the selected age group (15–18 years old) and may not be fully generalizable to other age groups or demographics.

- Short-Term User Interaction

Participants may not have sufficient time to fully adapt to the FilterX extension during the experiment, potentially influencing their overall experience and feedback.

- Model Imperfection

The developers acknowledged that no model is perfect, with the possibility of false positives or false negatives in the hate speech detection process.

- Platform-Specific Limitation

The study is bound by the availability of the FilterX extension exclusively on desktop browsers, particularly on the browser. This limitation restricts the generalizability of the findings to users of other browsers and platforms, such as mobile devices.

Review of related literature/studies/systems

Technology, according to Simplilearn (2023), has become a huge part of people's lives and is progressing and growing at a rapid pace. It also changed the way people access resources and how people learn new things. It is also said that people tend to use and rely on technology for everything, including business efficiency and communicating with other people. Internet access is another technology product that has become increasingly important, for it has revolutionized how people connect, enabling people to communicate and share information with others, allowing them to collaborate on projects in real-time and access information quickly and easily.

The Internet is a system architecture that has revolutionized mass communication, mass media, and commerce by allowing various computer networks around the world to

interconnect. The Internet is also referred to as a “network of networks,” the Internet emerged in the United States in the 1970s but did not become visible to the general public until the early 1990s. By 2020, approximately 4.5 billion people, or more than half of the world’s population, were estimated to have access to the Internet. And as of February 2023, Simon Kemp stated that 85.16 million of the 116.5 million population in the Philippines are internet users, which is 72.5 percent of the Philippine population, and that 9.7 percent of the population is between 13 and 17 years old, and 12.6 percent is between 18 and 24 years old. Filipinos according to Manarpiis, N., Cortez, K. M., Cortez M. G., and Bianca Nicole (2021), were seen to have a high average of spending their time on social media, which led the Philippines being known as the “social media capital of the world”. Filipinos, according to them, spend 102,054 hours on social media, that is equivalent to 4,252 days or 11.64 years in total. Internet users’ population is still growing, largely due to the prevalence of “smart” technology and the "Internet of Things," where computer-like devices connect with the Internet or interact via wireless Synthesis networks. These “things” include smartphones, appliances, thermostats, lighting systems, irrigation systems, security cameras, vehicles, even cities (Encyclopædia Britannica, inc., 2023).

As the internet can be used in different ways, the social impact of the internet can be seen. According to TechTarget (2020), both positive and negative societal effects can be attributed to the internet. People believe that the internet increases civic engagement, sociability and the intensity of relationships. On the other hand, other People contend that the internet increases the risk of isolation, alienation, and withdrawal from society that points to the increase of an emotional response, or the fear of missing out. Another example of an internet risk is hate speech. Hate Speech is a speech or an expression that slanders or criticizes a person/s on a basis of allegedly belonging to a certain social group such as, race, gender, ethnicity, sexual orientation, religion, are, physical or mental disability, and so forth. A typical hate speech involves epithets, which according to Oxford Languages pertains to an adjective or descriptive phrase expressing a quality characteristic of the person or thing mentioned. Hate Speech are statements that promote malicious stereotypes, and speech intended to incite hatred or violence against a group. Hate Speech can also include nonverbal depictions and symbols (Encyclopædia

Britannica, inc., 2023). Hate Speech, according to eSafety Commissioner, et al. (2020) is recognized as a growing online issue which can negatively impact a person's mental health, general wellbeing and online engagement. Also, hate speech can lead to harassment and violence offline, in most extreme cases. Critics of hate speech argue not only that it causes psychological harm to its victims, and physical harm when it incites violence, but also that it undermines the social equality of its victims (Encyclopædia Britannica, inc., 2023). According to SELMA (2019), online hate speech may cause direct and indirect effects on individuals' psychological wellbeing, and it can be short and long term depending on the amount of victimization or how the victims cope up with the situation. Some people argue that consequences of hate speech are similar to the effects experienced by victims with traumatic experiences. Victims of online hate speech can have low self esteem, sleeping disorders, anxiety, fear, insecurity. Others can develop social anxiety, and the feeling of needing to be isolated, some feel that their human dignity was violated, no longer seeing themselves as good and appropriate. Another serious effect of experiencing online hate speech is behavioral harm, possibly committing suicide or cutting themselves. A significant number of adolescents and young adults are targeted by online hate speech. The effect of such hateful utterances can involve severe psychological harm, especially for youths who have to master developmental tasks. Many adolescents and young adults are affected by online hate speech. When a person is exposed to statements that offend a group of people they feel they belong to, it can have devastating consequences. This is especially the case with youths, who do not yet have a consolidated personality. (Obermaier, M., & Schmuck, D., 2022). A project named European SELMA found that 57% of their respondents are teens who encountered hate speech online once or several times in three months. It is said that most often, hate speech happens on social media platforms, websites, or apps like Facebook, Twitter, Youtube, and Instagram, and is encountered accidentally. Some victims of online hate speech choose to tell somebody about what they encountered online, some chose to ignore it because they didn't care or they didn't know what to do.

"With the growth of social media, many children will, unfortunately, come across hate speech online, and we must ensure that young people are aware of what they should do if they come across online hate speech. If children see hate speech online, it is paramount

that they tell someone they trust, be that a parent or a teacher." (Sajda Mughal OBE, 2023). According to Mughal (2023), people, especially young people should know what they should do when they encounter online hate speech, also, according to Mughal, the victim should report the hate speech to the appropriate person, and that it is important to tell an adult to talk through what the child have seen, even though the hate speech doesn't directly pertains or affects the child.

Automated hate speech detection is an important tool in combating the spread of hate speech, particularly in social media. Numerous methods have been developed for the task, including a recent proliferation of deep-learning based approaches (Pang, G,. 2022).

Reducing exposure to hateful speech online by Jack Bowker & Jacques Ophoff (2021), aims to solve people's regular exposure to hateful content online. According to them, regular exposure of hateful content online can reduce levels of empathy in individuals, as well as affect the mental health of targeted groups. It is also said that a significant number of young people fall victim to hateful speech online, but, unfortunately, these contents are poorly managed by online platforms. The research intends to use machine learning and browser extensions to identify hateful content and to assist users in reducing their exposure to hate speech online. They developed a Proof-of-concept extension for Google Chrome web Browser using a local word blocker and a cloud-based model to explore if the browser extension is effective in identifying and managing exposure to hateful speech online. The research concluded that NLP can play a valuable role in hate speech detection, and how browser extensions can be an effective method in managing hate speech online, and it was found that users were interested in this method of reducing their exposure to online hate speech.

Shreyans Jain and Deepali Kamthania (2020), found it unfortunate to experience hate speeches quite often, that is why they developed a Hate Speech Detector called Negator. The Hate Speech Detector Negator was designed to take textual content from a website and send to a server, and using Natural Language processing, the data is cleaned, and with the NLP model all the hate speech, whether it is a statement or a word, are detected. The statement or the word detected then be sent back to the user censored. The Hate

Speech detector, Negator uses NLP by using an Aspect-based Sentiment Analysis (ABSA), wherein, it provides several results for different features mentioned in one statement rather than using one variable to find all the aspects relating to the statement.

The research was concluded by finding that there is no hate speech detector that is directly available for normal users, and that the chrome extension proposed provides an efficient way of detecting and hiding hate speech, for users can set their level of hate speech detection threshold and disable a particular website.

Research by Olawale Onabola, Zhuang Ma, Yang Xie, Benjamin Akera, Abdulrahman Ibraheem, Jia Xue, Dianbo Liu, and Yoshua Bengio (2021), figured that subtle and overt racism is still present both in physical and online communities today and has impacted many lives in different segments of the society, and that people continuously consume information through text, and racially biased content often results in hate speech online. The researchers attempted to automatically detect racially biased content from the web, including comments from online news like Fox News and comments from YouTube videos. They implemented BERT as a base model in automate labeling the dataset with pointers to racial bias. They also implemented a browser extension as a tool to help people identify racially biased content. In conclusion, the researchers decided to make a chrome extension for users to help report and identify racially biased text on the web, and extend to a broader form of bias such as gender. The researchers also plan to reduce the model complexity.

Research entitled, A Machine Learning Model for the Profanity Detection in the Filipino Language, by Beverly F., Crystelle T., Lorenz M., Benedict P., Jr., Ana D., Famela S., Kathleen P., Michael Y., Allen B., and Zareena L. (2022). Aims to solve a prevalent issue in society since the early 1900s. they believed that profanity became a major part of language and speech due to its aid in expression. According to them, profanity can harm a person’s thoughts, and that certain words in a language have been considered profane by certain communities. The researcher believed that online environments are not properly regulated, and that users are prone to the exposure found online. And to be able to avoid this, many applications were developed, but none targeted the Filipino language,

that is why the researchers built a machine-learned model that can classify whether a text contained Filipino profanities or not. The researchers also developed a prototype web application which integrated the predictive model, that can be further improved by integrating semantic and sentiment analysis. The Researcher concluded that the prototype is currently limited to classifying whether a text contains Filipino profanities or not, and that it can still be improved. The researchers also believed that the model can be augmented by implementing a more in-depth classification feature, wherein, profanities can be categorized by severity or by purpose.

Hate Speech in Philippine Election-Related Tweets: Automatic Detection and Classification Using Natural Language Processing, is researched by Mark Edward G. & Charibeth C. (2019), that aims to address the hate speech in the Philippine cyberspace. The researchers thought that research endeavors in the Philippines are limited, and that most of the systems were developed by foreign researchers, and because of that their dataset, and their underlying frameworks and assumptions are not reflective to the Philippine culture and context of hate speech in cyberspace. The researchers seek to develop a model that can automate hate speech detection and classification in Philippine election-related tweets. According to the research, the role of the microblogging site Twitter as a platform for the expression of support and hate during the 2016 Philippine presidential election has been supported in news reports and systematic studies. The research aims to review existing NLP methods employed in hate speech detection and classification, and the techniques that can be utilized to extract features from hate-containing tweets. The researchers also aim to implement the hate speech detection and classifications models following rule-based, machine learning, and deep learning approaches. Based on the research, the researchers found the effectiveness of using simple lexicon-based, language independent-features, specifically term frequency–inverse document frequency, term occurrence, and their combination, in detecting and classifying hate speech in Philippine election-related tweets. According to the researchers, the model was built in a specific domain, and that using the model can be used in other domains but the application of features will be different. According to the researchers, future works may focus on extending the methods of the research to hate speech detection, and that addition of rules may improve the performance of the baseline

rule-based classifier.

Synthesis

According to Encyclopædia Britannica, inc. (2023), approximately 4.5 billion people, that is over half of the world, have access to the internet. For that reason, people mostly get information from the internet. Olawale O., Zhuang M., et al (2021) stated that online communities have impacted many lives in different segments of the society, and that people continuously consume information through text, and racially biased content often results in hate speech online. In addition, Jack B. & Jacques O. (2021) believed that regular exposure of hateful content online can reduce levels of empathy in individuals, as well as affect the mental health of targeted groups. SELMA (2019), also agreed to that statement, and according to them, online hate speech may cause direct and indirect effects on individuals' psychological wellbeing, and it can be short and long-term depending on the amount of victimization or how the victims cope up with the situation. And because of these situations, many researchers have come to a solution. Jack Bowker & Jacques Ophoff (2021), used machine learning and browser extensions to identify hateful content and to assist users in reducing their exposure to hate speech online. Shreyans Jain and Deepali Kamthania (2020), developed a hate speech detection browser extension, Negator, that detects and hides hate speech. The researchers noticed that most of the research about hate speech detection is for English language, however, in the Philippines, both Filipino and English are used as a mode of communication. And that, 85.16 million of the Philippine population, according to Simon Kemp, are internet users which is 73.1 percent of the total population in the Philippines, and Filipinos spend 102,054 hours on social media (Manarpiis, N., Cortez, K. M., Cortez M. G., and Bianca Nicole, 2021).

There is some research that focuses on the Filipino language. Beverly F., Crystelle T., et al (2022), developed a machine-learned model for hate speech detection that can classify whether a text contained Filipino profanities or not, according to them, many applications about hate speech detection were developed, but none targeted the Filipino language. Also, Mark Edward G. & Charibeth C. (2019) thought that research endeavors in the Philippines are limited, and that most of the systems were developed by foreign

researchers, that is why they developed a model that can automate hate speech detection and classification in Philippine election-related tweets. The researchers of FilterX (Filter Extension) aim to develop a hate speech detection browser extension that detects and filters any instances of hate speech in Filipino and English language. Browser Extension is used for the research because as Bowker & Jacques Ophoff (2021) concluded, browser extensions can be an effective method in managing hate speech online, and it was found that users were interested in this method of reducing their exposure to online hate speech. And as what Shreyans Jain and Deepali Kamthania (2020) concluded that there is no hate speech detector that is directly available for normal users, and that the chrome extension they proposed provides an efficient way of detecting and hiding hate speech, for users can set their level of hate speech detection threshold and disable a particular website. Browser extension can be used as a medium, not only in detecting online hate speech, but also filtering hate speech. Although there are different mediums to be used in developing hate speech detection, FilterX intends to scan and filter out active tabs in the browser. Although the hate speech detector, FilterX, has limitations, the researchers believe that there are no models that were made perfect, and that failure is inevitable.

METHODOLOGY

FilterX is a Chrome extension developed to detect and mitigate the presence of hate speech in online content. This tool enables real-time monitoring of harmful language in web browsers, focusing on both English and Tagalog. FilterX is designed with flexibility and user configurability in mind, offering features like real-time detection, batch scanning of page content, and a chat interface for sentence-specific analysis.

To support its multilingual detection capabilities, FilterX utilizes three machine learning models: fastText, an English hate speech detection model, and a custom-built Tagalog hate speech detection model. The extension integrates these models through either a Local API setup, designed for controlled testing and experimentation, or Hugging Face's Serverless Inference API for production. The Local API setup is the primary option for the project's experimental phase to avoid potential cold starts and other latency issues that could impact results. The serverless API, used for production if required, will be reserved only if funding limitations necessitate a cloud-based approach.

Study Methodology

In this study, a two-part questionnaire was designed to assess the impact of a Chrome extension on users' experiences with hate speech on social media. This involved creating a pre-test and a post-test to measure user experiences before and after using the extension. Experiment was conducted on a group of 20 students, who served as participants in evaluating the system's effectiveness.

Data Collection Procedure

1. Pre-Test Administration: Prior to exposure to the extension, each participant completed a pre-test questionnaire. This survey gathered baseline data on participants' experiences with hate speech, comfort levels, and sense of control over content while browsing social media.

2. Intervention: Participants then used the Chrome extension, which was deployed via a local API version, for a designated period. This allowed us to test the system's impact in a controlled environment.
3. Post-Test Administration: After using the extension, participants completed a post-test survey, identical in structure to the pre-test, to measure any changes in their responses.

Data Analysis Approach

To evaluate the impact of the extension, pre-test and post-test results were compared through a series of analytical steps:

- Descriptive Statistics: The mean and standard deviation was calculated for each question in both the pre-test and post-test to summarize and compare the central tendencies and variability of the responses.
- Difference Scores: For each participant, the difference between their pre-test and post-test scores for each question were calculated, to identify any shifts in responses.
- Paired t-test: A paired t-test was conducted for each question to determine if there was a statistically significant difference between pre-test and post-test scores. This test assessed whether any observed changes were likely due to the system rather than random variation.
- Effect Size (Cohen's d): To quantify the impact of the extension, the effect size (Cohen's d) for each question were calculated. This measure helped in understanding the magnitude of the changes observed and assess the practical significance of the extension's impact on user experiences.

Development Methodology

Model Development

This section discusses various experiments conducted to train and fine-tune different machine learning models for optimal predictive performance. The models investigated include Word2Vec Embedding, Tagalog-based ELECTRA, and the BERT-based model

RoBERTa. The goal of these experiments is to evaluate each model's performance and effectiveness in accurately classifying textual data. To support these experiments, a comprehensive data frame was constructed using three distinct datasets, ensuring a diverse and representative sample for evaluation. The results will highlight the comparative strengths and weaknesses of each model in these specific tasks. Performance is measured using several mathematical concepts and metrics, providing a rigorous and comprehensive evaluation. Detailed analysis and findings from these experiments will be presented in the following sections.

I. Data frame Amalgamation, Preparation and Validation

The datasets used are as shown in *Table 1* including each title and their respective authors. Data pre-processing consists of (1) Converting each text to lower case, (2) Replacing newlines with spaces, (3) Removing special characters, (4) Replacing multiple spaces with single space. After pre-processing, there are a total of 21,190 samples, specifically 11,776 Non-hate speech samples and 9,414 hate speech samples with an imbalance ratio of 1.25.

Dataset Title	Author
“hate_speech_filipino”	Jan Christian Cruz
"multilabel-tagalog-hate-speech"	syke9p3
“tagalog-profanity-dataset”	mginoben

Table 1. List Of datasets

After implementing an oversampling technique which is by resampling the data in hate speech dataset and then refitting it onto the main data frame to avoid over-fitting, the data frame is now balanced as shown in *Table 2*.

Label	Count
Hate speech (1)	11,776
Non-Hate speech (0)	11,776

Table 2. Result of Oversampling

The data frame is then validated through investigating its missing values, verifying its data types and checking its class distribution. The result shows that the data frame is now ready for model training. The result is as shown in *Table 3*.

Column Name	Missing Values	Data type
Text	0	Object
Label (1, 0)	0	int64

Table 3. Dataframe Validation Result

II. Model Training

The model training is experimented on different embeddings and tokenization such as Word2vec Embedding, ELECTRA Embedding and RoBERTa Tokenization. Each embedding and tokenization models are experimented on different classifiers such as Logistic Regression, RandomForest and Light Gradient Boosting Machine. Moreover, it is then finalized to use the built dataset above, specifically the text column as an independent variable 'x' trained on a label column known as a dependent variable 'y'.

Term Frequency - Inverse Document Frequency (TF-IDF) is also tried but due to the nature of hate speech where expression changes depending on the context, TF-IDF was marked to have low scalability so no additional experiment attempts were made.

All experiments conducted are summarized and illustrated in *Table 4*. The researchers observed that combining RoBERTa Embedding with traditional classifiers such as Logistic Regression, RandomForest and LightGBM requires many parameters and it was decided that this combination is not suited for real-time inference so no additional attempts were made.

Tokenizer	Classifier
Word2vec Embedding	Logistic Regression
	Random Forest Classifier
	Light Gradient Boosting Machine
Tagalog-Based ELECTRA Embedding	Logistic Regression

	Random Forest Classifier
	Light GBM
BERT-Based Model RoBERTa Fast Tokenizer	RoBERTaSequenceClassification

Table 4. Summary of Experiments

III. Model Evaluation

For model evaluation, the "train-test split" concept is used, where the data frame is divided into 80% for training and then 20% for testing and evaluation as shown in *Figure 1*.

Figure 1. Train-Test Split

Each model's performance is measured by its accuracy score, precision, recall, f-1 score and plotted Receiver Operating Characteristic (ROC) and the Area Under the Curve (AUC). The computation metrics for each evaluation is as shown below.

- Accuracy - Accuracy measures the overall correctness of the model. It is the ratio of correctly predicted observations to the total observations.

Where:

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

- Precision - also called Positive Predictive Value measures the accuracy of the positive predictions.

$$Precision = \frac{TP}{TP+FP}$$

- Recall - (also known as Sensitivity or True Positive Rate) measures the ability of the model to identify all relevant instances.

$$Recall = \frac{TP}{TP+FN}$$

- The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both concerns.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{2TP}{2TP+FP+FN}$$

The summarized results of the machine learning models for hate speech detection show that the RoBERTaSequenceClassifier with RoBERTa Fast Tokenizer outperformed other models, achieving the highest Accuracy of 0.8406, Precision of 0.8341, Recall of 0.8534, and F1 Score of 0.8436. The Tagalog-Based ELECTRA Embedding with Light GBM also showed promising results with an Accuracy of 0.7855, Precision and Recall both at 0.81, and an F1 Score of 0.81. Comparatively, models using Word2vec Embedding had

lower performance metrics, with the Random Forest Classifier achieving an Accuracy of 0.7744, Precision of 0.78, Recall of 0.76, and F1 Score of 0.78. These results indicate that transformer-based models like RoBERTa may be more effective for this application than traditional embedding methods as shown in *Table 5*.

Tokenizer	Classifier	Accuracy	Precision	Recall	F1 Score
Word2vec Embedding	Logistic Regression	74%	74%	73%	74%
	Random Forest Classifier	77%	78%	76%	78%
	Light Gradient Boosting Machine	76%	79%	80%	77%
Tagalog-Based ELECTRA Embedding	Logistic Regression	78%	81%	81%	81%
	Random Forest Classifier	77%	79%	80%	80%
	Light GBM	79%	81%	82%	81%
BERT-Based Model RoBERTa Fast Tokenizer	RoBERTa Sequence Classification	85%	84%	85%	84%

Table 5. Summary of Reports

ROC AUC Plotting

The ROC curve is a plot of the true positive rate (Recall) against the false positive rate (FPR) at various threshold settings. The AUC (Area Under the Curve) of the ROC curve is a single scalar value that summarizes the performance of the model across all threshold levels.The ROC curve is created by plotting the true positive rate (Recall) on the y-axis and the false positive rate (FPR) on the x-axis.

Word2Vec ROC AUC reports are as on the following *Figures*.

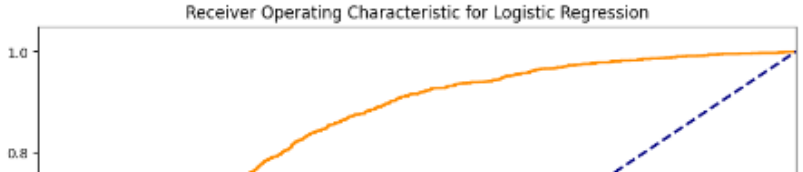


Figure 2. ROC AUC = 82%

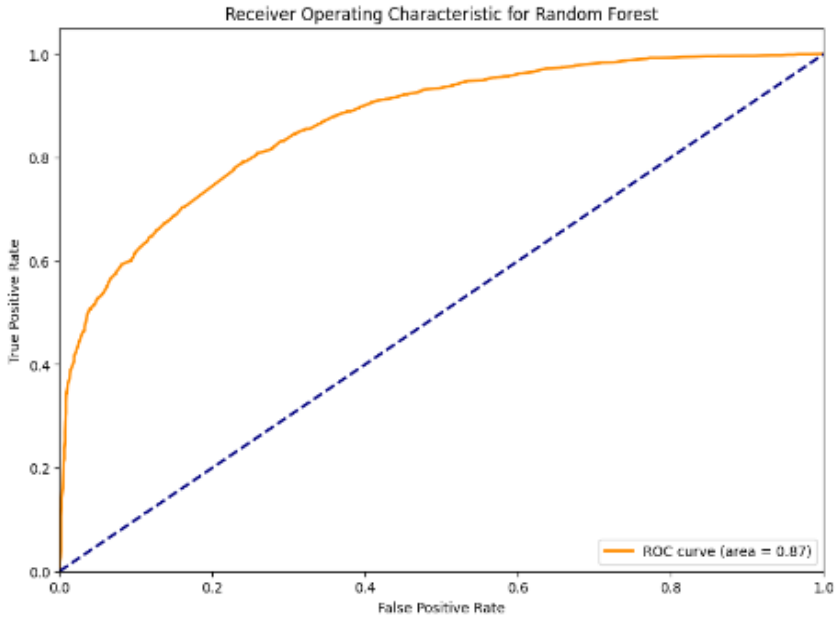


Figure 3. ROC AUC = 87%

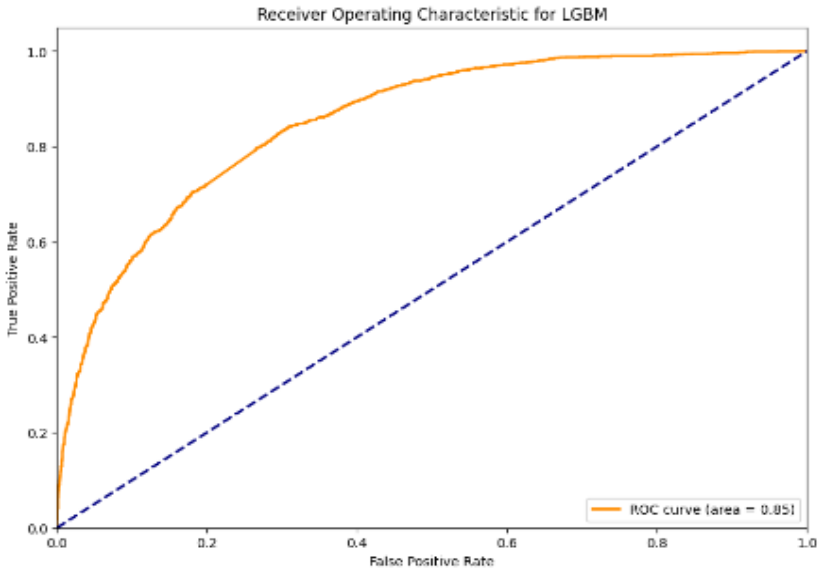


Figure 4. ROC AUC = 85%

ELECTRA ROC AUC reports are as on the following *Figures*.

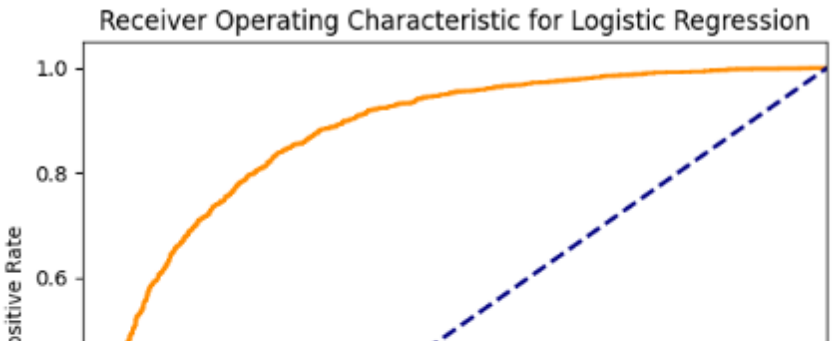


Figure 5. ROC AUC = 87%

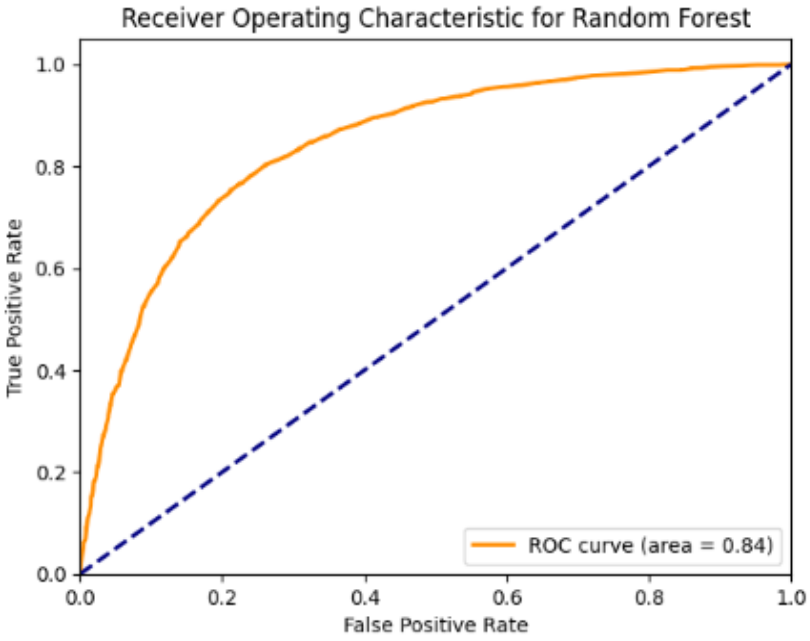


Figure 6. ROC AUC = 84%

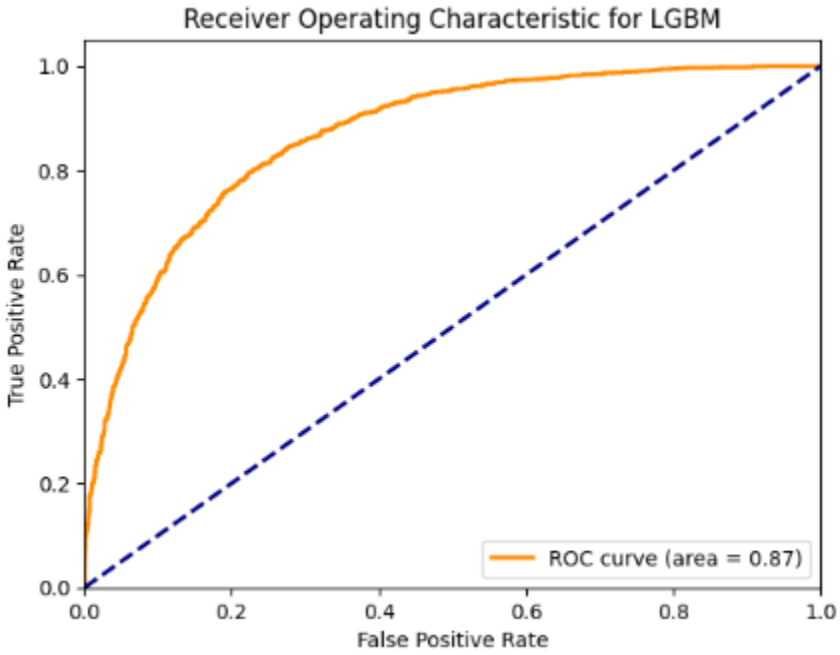


Figure 7. ROC AUC = 87%

RoBERTa ROC AUC report is as follows:

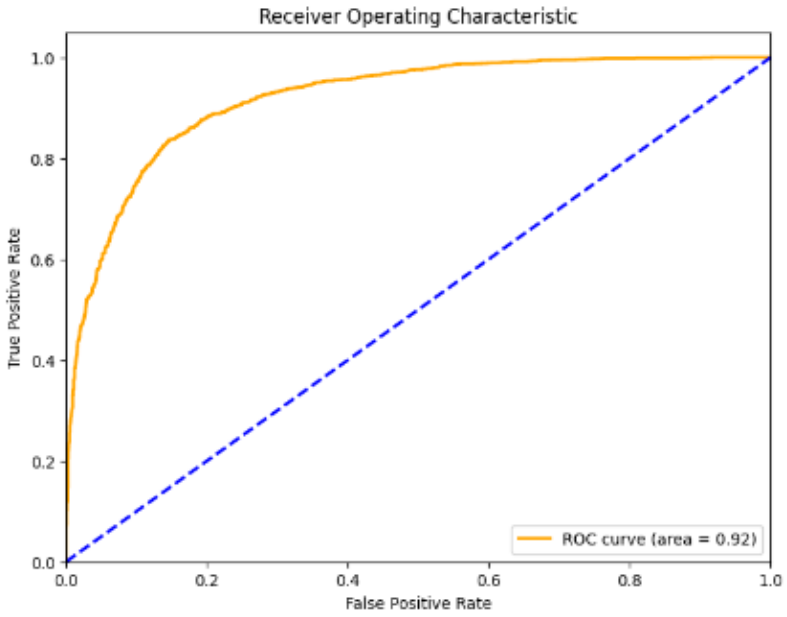


Figure 8. ROC AUC = 92%

Inference Time Test

The inference time test conducted here utilizes the mathematical concept of measuring elapsed time. To initiate the test, a base sample sentence "Ako si Jordan Limwell Marcelo" is generated and replicated 5000 times, creating a sizable dataset for evaluation. This dataset is then subjected to a simulated scenario where 10% of the sentences are systematically removed. This step is crucial as it emulates real-world scenarios where not all data might be processed due to various factors such as noise reduction or computational resource constraints. By reducing the workload in this manner, the test becomes more representative of practical usage scenarios, ensuring that the evaluation accurately reflects the model's performance under realistic conditions. The result is summarized on *Table 6* based on the 95% confidence interval (mean ± 2 standard deviations).

The researchers concluded that ELECTRA embedding requires too many parameters and was pronounced incapable of real time inference. It is considered that further attempts were unnecessary.

Model		Inference Time
Tokenizer	Classifier	
Word2vec	Logistic Regression	0.132 ± 0.032 ms
	Random Forest	5.973 ± 2.794 ms
	LGBM	0.808 ± 0.032 ms
RoBERTa	RoBERTaSequenceClassifier	15.109 ± 0.131 ms

Table 6. Inference Time Report

All the models evaluated in this study successfully met the inference time requirement, which was set at below 50 milliseconds. Specifically, the Word2vec-based Logistic Regression model achieved an inference time of 0.132 ± 0.032 ms, the Random Forest model recorded 5.973 ± 2.794 ms, the LGBM model had an inference time of 0.808 ± 0.032 ms, and the RoBERTaSequenceClassifier model showed an inference time of 15.109 ± 0.131 ms. Given that all models performed well within the acceptable range, the focus was shifted towards deciding which model to use based on overall accuracy. After thorough evaluation, it was concluded that RoBERTa should be used due to its superior accuracy compared to the other models. This decision ensures the selection of the most effective and reliable model for the sequence classification.

IV. Model Visualization

To visualize the actual model validation, researchers employed t-SNE (t-distributed Stochastic Neighbor Embedding) to reduce the dimensionality of the prediction space to two components, which could then be easily plotted. Researchers first obtained model predictions for the validation dataset using the trainer object. t-SNE was then applied to reduce these high-dimensional predictions to two dimensions, preserving data structure and enabling visualization. Utilizing seaborn's scatterplot function, the transformed data was plotted, with each point representing a prediction and colored according to the true validation labels. The "hsv" color palette was used for the binary classification task. The plot was titled per each model. The results for each model are as shown in the following *Figures*.

- Word2vec Embedding

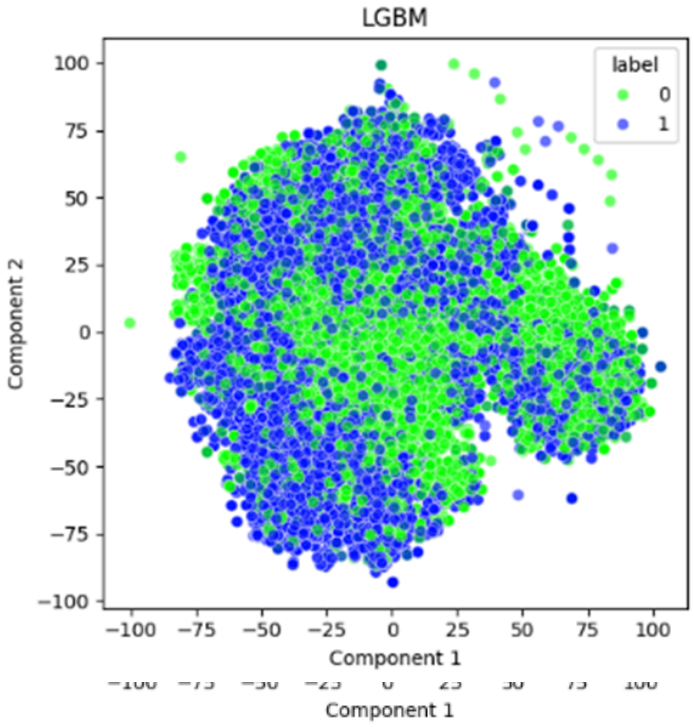


Figure 9. word2vec + Logistic Regression Visualization

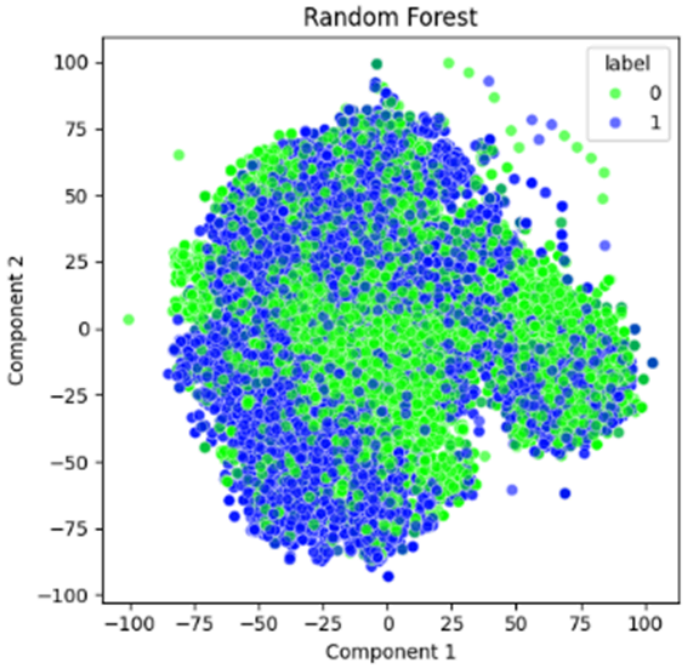


Figure 10. word2vec + RandomForest Visualization

Figure 11. word2vec + LightGBM Visualization

- Tagalog-based BERT model, RoBERTa

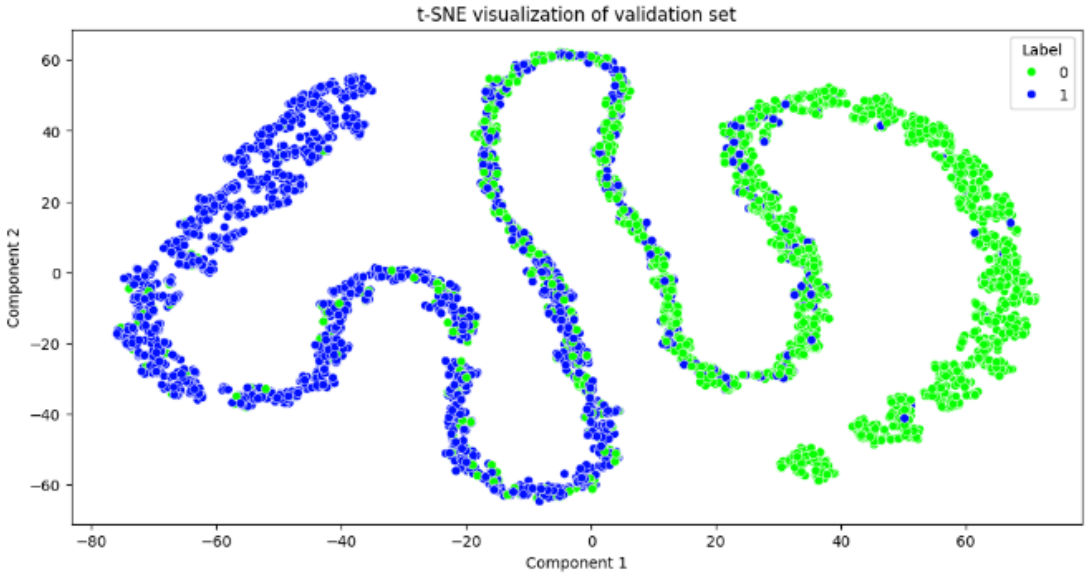


Figure 12. RoBERT Tokenizer + RoBERTaSequenceClassifier Visualization

Among the visualizations compared, the t-SNE visualization of the RoBERTa classifier clearly stands out as the best in terms of class separation. The distinct, non-overlapping clusters in this plot indicate that the dataset is inherently well-structured and the classes are naturally separable.

In contrast, the t-SNE plots of the model outputs (Logistic Regression, Random Forest, and LightGBM) show significant overlap, suggesting that these models are not leveraging the data's inherent structure to achieve effective class separation. This discrepancy highlights a potential gap in the models' ability to generalize and learn the underlying patterns in the data.

Application Development

An Agile approach can be highly beneficial for developing the FilterX due to its iterative nature and focus on user feedback. Here’s a breakdown of the phases of Agile Methodologies:

Phase 1: Planning and Initiation

Purpose: Establish the goals and foundation of the research by defining roles and responsibilities to understand the activities that are needed.

Division of Labor:

- Machine Learning Model Development: Develop an efficient machine leaning model to focus on building hate speech detection models.
- UI Development: Develop and design how the project will look like.
- Backend Development: This includes the API structure and Database connections.
- Testing: This will ensure that all the components of the project will work as it is expected.

Deadlines and Schedules: Ensure steady progress and avoid delays by setting realistic timelines.

Interview: Draft questions for an interview to be able to gather data and insights from users that will help in design and model decisions.

Outcome: Create a structured roadmap with clear roles and timelines to be able to guide the team members through the project.

Phase 2: First Iteration Planning

Purpose: Define technical goals and requirements for the first development iteration.

Model Requirements:

- Aim to achieve at least 70% accuracy and 70% precision in the machine learning model to ensure effective hate speech detection.
- Ensure efficient system performance by maintaining the inference time under 200ms for real-time detection.
- Outcome: These technical benchmarks will guide the model development process.

Phase 3: First Iteration Execution

Purpose: Implement the core functionalities of the system during the first iteration.

Language Detection: LangDetect (Nakatani Shuyo) will be used for language detection to ensure the text is processed in the appropriate language.

CNERG: CNERG will be used to handle English hate speech detection and develop Tagalog hate speech detection model.

Initial User Interface: Develop an initial user interface to gather feedback and improve usability.

Outcome: A functional system with initial UI and core model functionality will be prepared for testing and feedback.

Dive Deep on Tagalog Hate Speech Detection

Purpose: Develop a hate speech detection model for Tagalog text.

RoBERTa-based Model: A Tagalog-specific version of the RoBERTa model will be fine-tuned for text classification.

Tokenizer: To ensure proper sentence parsing, a tokenizer tailored to process Tagalog will be used.

Outcome: A robust Tagalog hate speech detection system will be prepared for evaluation and feedback.

Phase 4: First Iteration Evaluation

Purpose: Evaluate the performance of the Models and User Interface based on the feedback and key performance metrics.

Model Specs: Create a comparison table to evaluate the performance of different models (e.g., accuracy, precision, inference time).

UI Feedback: User feedback will be collected to assess the functionality and design of the initial UI.

Outcome: Areas for improvement in both the model and UI will be identified, guiding the second iteration.

Phase 5: Second Iteration Planning

Purpose: Plan for improvements based on the evaluation from the first iteration.

Enhanced Models: Enhance the machine learning models to improve the accuracy and performance.

Improved UI: Refine the user interface based on the feedback received during the first iteration.

Outcome: Plan for improving the model and UI will be created for the next development cycle.

Phase 6: Second Iteration Execution

Purpose: Implement the improvements planned during the second iteration.

Improved Language Identification: FastText will be used for improved language identification to replace LangDetect.

Update CNERG: Update CNERG for better English hate speech detection, and refine the Tagalog hate speech detection model.

Update UI: Update the user interface to incorporate user feedback and improve user experience.

Outcome: The second iteration will produce enhanced models and an improved UI ready for further evaluation.

Phase 7: Second Iteration Evaluation

Purpose: Evaluate the second iteration's improvements in terms of model performance and user satisfaction.

Model Specs: Generate a comparison table to show the performance of the improved models.

Outcome: Assess whether the improvements meet the set goals and prepare for final deployment.

Phase 8: Release and Deployment

Purpose: Focus on deploying the application to production.

API inference: Hugging Face Hub will be used for API inference to leverage the pre-trained models for real-time hate speech detection.

Local API: Setup a local API using Uvicorn and FastAPI to provide a self-hosted solution.

Outcome: The project will be deployed, making the hate speech detection system available for use in real environments.

Requirement Gathering

Expert Review

In requirement gathering phase, experts in machine learning and web application development were consulted to gather valuable insights and advice on improving the system. These experts provided guidance on optimizing the performance, scalability, and accuracy of the machine learning models, as well as enhancing the overall functionality and user

experience of the web application. The experts' input helped identify potential challenges and areas for improvement, ensuring that the system is both efficient and robust. By incorporating the expert's feedback, the solution will meet the desired standards and effectively address the needs of the target users.

Functional Requirements

Model Requirement	All models shall have an accuracy score of 70% and above.
	All models shall have a precision score of 70% and above.
	All models shall have inference time of below 200 milliseconds (ms).
Chrome Extension Requirement	The extension shall offer multiple methods for hate speech detection.
	The extension shall agree to terms & conditions before the user can access the application.
	The extension shall have a real-time detection feature.
	The extension shall have an alert if an occurrence is detected.
	The extension shall have a report summary of detected hate speeches.
	The extension shall have a report function if a false prediction has possibly occurred.
	The extension shall have a sensitivity control over detection.
	The extension shall have a different censorship type.
	The extension shall have an output to evaluate the exposure of the user in hate speech.
	The extension shall have terms & conditions before diving.
	The extension shall have an information page about the developers and credits.
User Interface Requirement	The user interface shall feature a minimalist theme to avoid interrupting the webpage browsing experience.

	The user interface shall clearly define each feature.

Table 7. Functional Requirements

Non - Functional Requirements

Performance	The extension shall process and respond to user actions (e.g., hate speech detection toggle, censorship changes) within 100 milliseconds.
	The extension should efficiently handle high-traffic web pages, including pages with large amounts of text, without a noticeable slowdown in browsing.
	The extension shall consume minimal memory, with peak memory usage staying under 100 MB during real-time detection operations.
Usability	The extension's features shall be intuitive, requiring no more than two actions for common tasks (e.g., enabling detection, viewing reports).
Reliability	The extension shall gracefully handle API call failures, with user-friendly error messages and automatic retry mechanisms.
	Detected hate speech data and reports shall remain consistent and accurate across sessions.

Table 8. Non-Functional

System Design Specification

Model

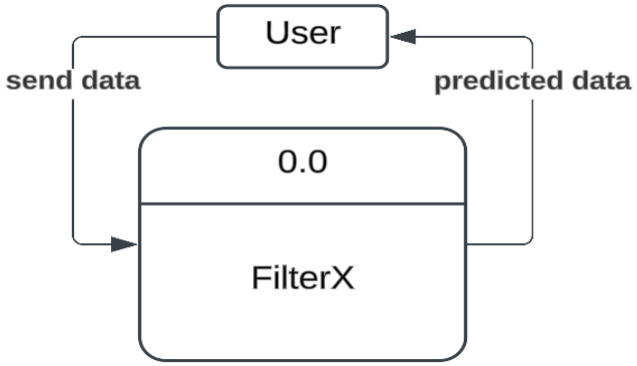
BERT-based RoBERTa:

- ROC AUC: 0.92
- F1-Score: 0.853639
- Inference Time: 15.109 ± 0.131 ms
- The RoBERTa model exhibited the highest ROC AUC among all models, suggesting superior capability in distinguishing hate speech from non-hate speech. The F1-Score and inference time further corroborate its efficiency and reliability in real-time applications.

FilterX

- Data Flow Diagram (DFD): A DFD illustrates the flow of data within the system.
- A Level 0 Context Flow Diagram provides an overview of the FilterX, representing it as a single process interacting with external entities. This

high-level diagram illustrates how the system connects and exchanges



information with users.

Diagram 1. Level 0 Context Flow Diagram

- The Level 1 DFD builds on the single process shown in the Context Diagram, breaking it down into the primary sub-processes of the student monitoring system. This expanded diagram outlines the main functional components and their interactions within the system.

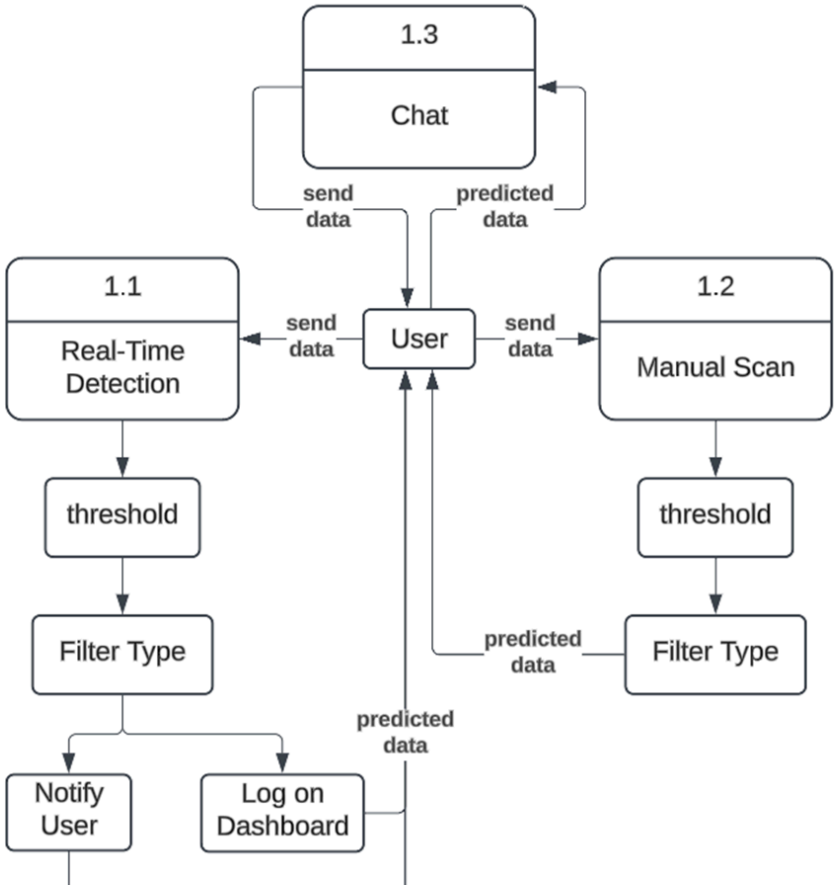


Diagram 2. Level 1 DFD

- Use Case Diagram: Use case diagram referred as a Behavior model or diagram. It simply describes and displays the relation or interaction between the users or customers and providers of application service or the system. It describes different actions that a system performs in collaboration to achieve something with one or more users of the system. Use case diagrams are used a lot nowadays to manage the system.

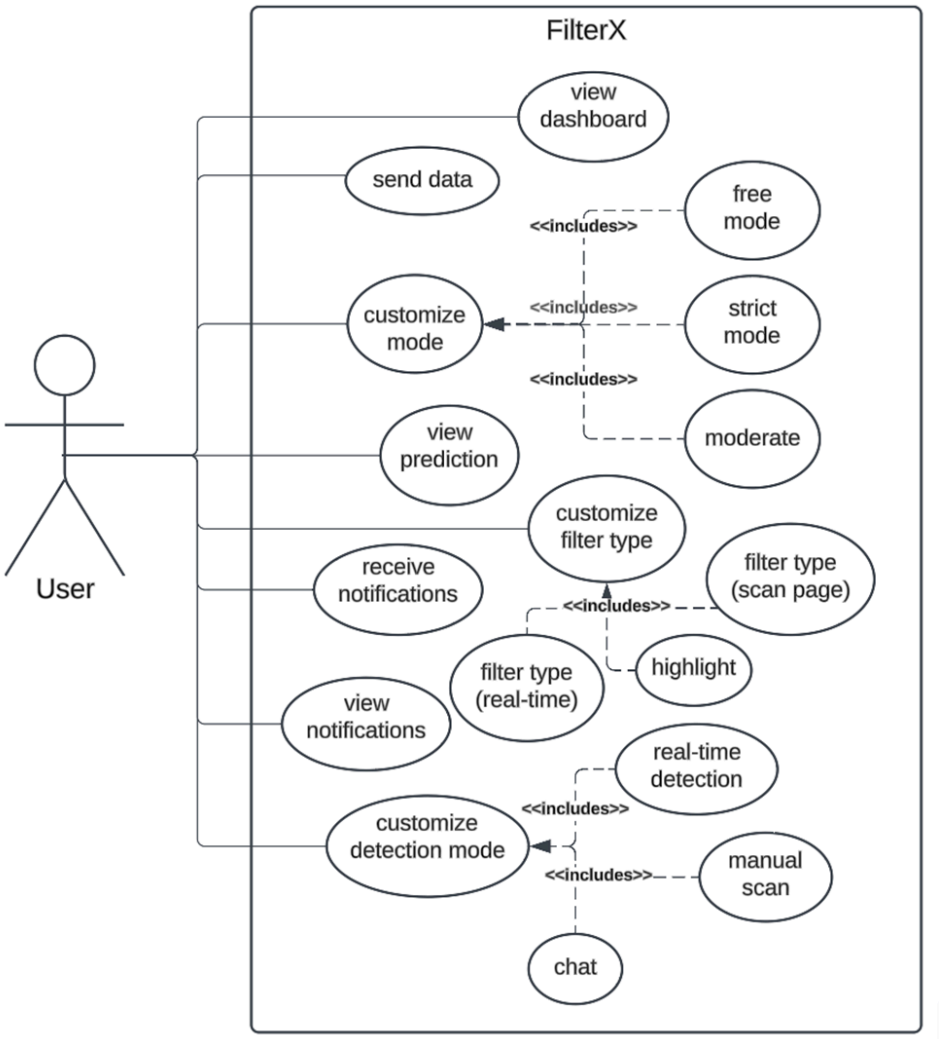


Diagram 3. Use Case Diagram

Features

Multilingual Support: The system incorporates separate machine learning models for English and Filipino languages, enabling effective hate speech detection in multilingual web content.

Real-Time Detection: Efficient content script processing and asynchronous API calls ensure timely hate speech detection as users browse web pages.

User Control: User-friendly options in the extension's popup provide users with control over the hate speech filter and manual scanning of web pages, enhancing the browsing experience.

Highlighting and Censorship: The extension offers two display modes (highlighting and censoring) for detected hate speech, catering to users' preferences and ensuring a customizable browsing environment.

Terms & Conditions: This allows users to be aware of the set guidelines and to enhance users' browsing experience.

Adjustable Potency: The extension allows the user to adjust the potency based on their own preference, just somewhat resolving the contextual dependence of hate speech.

- Dashboard: Contains Chart to display the summary of session in real time detection
- Notification: Notify users immediately when hate speech is found real time.

Deployment and Implementation

- Hugging face Inference API

  The API inference setup utilized the Hugging Face Hub to leverage pre-trained models for real-time hate speech detection. By doing so, it ensured access to state-of-the-art models, streamlining the process of identifying and classifying hate speech efficiently. This approach allowed for rapid, scalable detection with minimal need for additional model training.

- Local API

  Additionally, a local API was set up using Uvicorn and FastAPI to offer a self-hosted solution. This local deployment provided enhanced control, security, and flexibility, allowing users to integrate the hate speech detection capabilities within their own infrastructure while reducing dependency on external services. This combination of remote and local solutions facilitated both broad accessibility and tailored control for the application.

Results and Discussions

To evaluate the impact of the Chrome extension on participants' experiences with hate speech on social media, the data was analyzed from the pre-test and post-test using several statistical methods. This analysis provided insights into whether the extension effectively improved user experiences in terms of reducing exposure to hate speech and enhancing comfort and control over content. *Table 9* below summarizes the results, including means, standard deviations, t-statistics, p-values, and Cohen's d effect sizes for each question.

Question	Pre-Test Mean	Pre-Test Std Dev	Post-Test Mean	Post-Test Std Dev	t-statistic	p-value	Cohen's d (Effect Size)
Q1	**3.95**	**0.999**	**4.75**	**0.444**	**-3.24**	**0.004**	**0.80 (large)**
Q2	**3.50**	**1.000**	**4.30**	**0.657**	**-3.11**	**0.006**	**0.80 (large)**
Q3	**3.95**	**1.050**	**4.70**	**0.470**	**-2.88**	**0.010**	**0.71 (large)**
Q4	**4.05**	**0.826**	**4.70**	**0.571**	**-3.32**	**0.004**	**0.79 (large)**
Q5	**3.35**	**1.496**	**4.45**	**0.686**	**-3.24**	**0.004**	**0.74 (large)**
Q6	**3.45**	**1.099**	**4.60**	**0.503**	**-4.20**	**0.000**	**1.05 (very large)**
Q7	**4.05**	**1.050**	**4.45**	**0.605**	**-1.45**	**0.163**	**0.38 (small)**
Q8	**3.95**	**1.050**	**5.00**	**0.000**	**-4.47**	**0.000**	**1.00 (very large)**
Q9	**3.90**	**1.021**	**4.60**	**0.598**	**-2.67**	**0.015**	**0.69 (large)**
Q10	**3.75**	**1.164**	**4.80**	**0.410**	**-3.80**	**0.001**	**0.90 (large)**

Table 9: Summary of Pre-Test and Post-Test Results

1. Descriptive Statistics:

   As shown in *Table 9*, the mean scores for most questions increased from pre-test to post-test, indicating an improvement in users' experiences after using the extension. For instance, the mean for Question 1 (regarding exposure to hate speech) increased from 3.95 to 4.75, and similar positive changes were observed across other questions.

2. Difference Scores:

   The difference scores for each question, also highlighted in *Table 9*, demonstrate consistent positive shifts, suggesting that participants

generally reported a better experience in the post-test. This pattern reinforces the extension's effectiveness.

3. Paired t-test:

   The paired t-tests (*Table 9*) revealed significant differences ($p < 0.05$) between pre-test and post-test scores for most questions, particularly for questions related to browsing comfort and content control. These significant p-values suggest that the observed improvements were likely a result of using the Chrome extension rather than random chance.

4. Effect Size (Cohen's d):

   To understand the practical significance of these improvements, the effect sizes were calculated, shown in the last column of *Table 9*. Large effect sizes (above 0.8) were observed for several questions, especially Q6 (comfort while browsing) and Q8 (quality of online communities). These large effect sizes highlight that the extension had a substantial impact on user experiences, aligning with our hypothesis.

Discussion

The results support the hypothesis that the Chrome extension would have a positive impact on users' experiences with hate speech on social media. The consistent improvements in post-test scores across various questions indicate that participants felt more comfortable, less exposed to hate speech, and more in control of the content they encountered. The statistical significance of the t-tests, coupled with the large effect sizes, suggests that the extension provided both measurable and meaningful benefits to users, confirming its effectiveness in enhancing user experience on social media platforms.

Conclusion and Recommendations

Conclusions

The findings from this evaluation indicate that the Chrome extension had a significant positive impact on participants' experiences with hate speech on social media platforms. Consistent increases in post-test scores across multiple measures highlight substantial improvements in users' perceived exposure to hate speech, comfort while browsing, and their sense of control over online content. The results of paired t-tests confirmed that these changes were statistically significant, suggesting that the improvements were not due to random chance. Furthermore, large effect sizes observed for key questions underscore the practical importance of these enhancements, demonstrating that the extension delivered meaningful benefits for users.

Overall, these results support the hypothesis that targeted interventions, such as this Chrome extension, can effectively improve online experiences by reducing negative exposure and fostering more positive interactions within social media spaces.

Recommendations

1. **Broader Deployment and Further Testing:** Given the demonstrated positive impact, it is recommended to scale up the deployment of the Chrome extension to a wider audience. Conducting additional studies with larger and more diverse participant groups can provide further validation and help generalize the findings.

2. **Feature Enhancements:** To build on the observed effectiveness, it is recommended to enhance the functionality of the extension. Features such as user customization options, integration with different social media platforms, and real-time reporting of harmful content could further empower users and amplify positive outcomes.

3. **Longitudinal Studies:** Future research should include longitudinal studies to evaluate the sustained impact of the extension over extended periods.

This can help assess whether the positive changes are maintained over time and identify any areas for improvement.

4. **Collaboration with Social Media Platforms:** Engaging directly with social media platforms to integrate similar features or leverage the extension's capabilities could further enhance its impact. Collaborative efforts may lead to more comprehensive solutions for mitigating hate speech online.

5. **User Education and Support:** Providing users with educational resources and support on navigating hate speech, using the extension's features, and promoting positive digital interactions can strengthen its overall effectiveness.

References

Nave, E., & Lane, L. (2023). Countering online hate speech: How does human rights due diligence impact terms of service? *Computer Law & Security Review*, 51, 105884. https://doi.org/10.1016/j.clsr.2023.105884

Jaki, S., & De Smedt, T. (2019). Right-wing German Hate Speech on Twitter: Analysis and Automatic Detection. *arXiv*:1910.07518 [cs.CL]. https://doi.org/10.48550/arXiv.1910.07518

Khan, S., Fazil, M., Sejwal, V. K., Alshara, M. A., Alotaibi, R. M., Kamal, A., & Baig, A. R. (2022). BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection. Journal of King Saud University - *Computer and Information Sciences*, 34(7), 4335-4344. https://doi.org/10.1016/j.jksuci.2022.05.006

Berglind, T., Pelzer, B., & Kaati, L. (2020). Levels of hate in online environments. ASONAM '19: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 842–847. https://doi.org/10.1145/3341161.3343521

Balayn, A., Yang, J., Szlavik, Z., & Bozzon, A. (2021). Harmful, Aggressive, Abusive, and Offensive Language on the Web: A Survey of Technical Biases Informed by Psychology Literature. *ACM Transactions on Social Computing*, 4(3), Article 11, 1–56. https://doi.org/10.1145/3479158

Razali F., Jusoh Z., Salleh@Omar A. and Azizan N.(2021). *Implementation of Anti-Profanity Words in Mobile Application Platform*. https://iopscience.iop.org/article/10.1088/1757-899X/1062/1/012026/meta

Vanessa H., Dana R., Thomas K., Dietrich K. (2021) *Modeling Profanity and Speech in Social Media with Semantic Subspaces* https://arxiv.org/abs/2106.07505

Gonçalves, J., Weber, I., Masullo, G. M., Torres da Silva, M., & Hofhuis, J. (2023). *Common sense or censorship: How algorithmic moderators and message type influence*

perceptions of online content deletion New Media & Society, 25(10), 2595–2617. https://doi.org/10.1177/14614448211032310

Mandl, T., Modha, S., Shahi, G., Madhu, H., Satapara, S., Majumder, P., Schaefer, J., Ranasinghe, T., Zampieri, M., Nandini, D., & Jaiswal, A. (2021). *Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages Computing Research Repository*, 2023 (2112) https://arxiv.org/abs/2112.09301

Brave Software. (2023, July 17). *What are browser extensions, and are they safe?* Last updated: July 17, 2023. https://brave.com/browser-extensions-safety/

Bahador, B. (2020, November 17). Classifying and Identifying the Intensity of Hate Speech. *Social Science Research Council*. https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/classifying-and-identifying-the-intensity-of-hate-speech/

Laub, Z. (2019, June 7). Hate speech on social media: Global comparisons. *Council on Foreign Relations*. https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons

Bahador, B. (2020, November 17). Classifying and Identifying the Intensity of Hate Speech. *Social Science Research Council*. https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/classifying-and-identifying-the-intensity-of-hate-speech/

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. *PLOS ONE*, 15(8), e0237861. https://doi.org/10.1371/journal.pone.0237861

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8), e0221152. https://doi.org/10.1371/journal.pone.0221152

Cinelli, M., Pelicon, A., Mozetič, I., Quattrociocchi, W., Novak, P. K., & Zollo, F. (2021). Dynamics of online hate and misinformation. *Scientific Reports*, 11(1), 22083. https://www.nature.com/articles/s41598-021-01487-w#Sec7

Nascimento, F. R. S., Cavalcanti, G. D. C., & Da Costa-Abreu, M. (2023). Exploring automatic hate speech detection on social media: A focus on content-based analysis. *Social Media + Society*, 9(1), 21582440231181311. https://doi.org/10.1177/21582440231181311

Omran, E., Al Tararwah, E., & Al Qundus, J. (2023). A comparative analysis of machine learning algorithms for hate speech detection in social media. *Online Journal of Communication and Media Technologies*, 13(4), e202348. https://doi.org/10.30935/ojcmt/13603

Singh Pasi, P. (2024). SVM based Hate Speech Detection [GitHub repository]. https://github.com/piyushsinghpasi/SVM-based-Hate-speech-detection

Saksesi, A. S., Nasrun, M., & Setianingsih, C. (2018). Analysis of Hate Speech Detection Using Recurrent Neural Network. In 2018 International Conference on Control, Electronics, *Renewable Energy and Communications (ICCEREC)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICCEREC.2018.8712104

Boishakhi, F. T., Shill, P. C., & Alam, M. G. R. (2023). Multi-modal Hate Speech Detection using Machine Learning. *arXiv preprint arXiv*:2307.11519. https://arxiv.org/abs/2307.11519

MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019c). Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8), e0221152. https://doi.org/10.1371/journal.pone.0221152

Tenny, S. (2022, September 18). Qualitative study. *StatPearls - NCBI Bookshelf*. https://www.ncbi.nlm.nih.gov/books/NBK470395/

Statista Research Department. (2023, December 18). *Facebook: hate speech content removal as of Q3 2023*. https://www.statista.com/statistics/1021691/facebook-hate-speech-content-removal-quarter/

Simplilearn. (2023, August 18). Has technology improved our lives?: Simplilearn. *Simplilearn.com*. https://www.simplilearn.com/how-has-technology-improved-our-lives-article

Simplilearn. (2022, December 20). What are the top products of Technology?: Simplilearn. *Simplilearn.com*. https://www.simplilearn.com/products-of-technology-article#:~:text=Internet%20access%20is%20another%20essential,share%20information%20with%20people%20worldwide.

Encyclopædia Britannica, inc. (2023, November 13). Internet. *Encyclopædia Britannica*. https://www.britannica.com/technology/Internet

Contributor, T. (2020, December 21). What is the internet? definition from whatis.com. *WhatIs.com*. https://www.techtarget.com/whatis/definition/Internet

Encyclopædia Britannica, inc. (2023b, December 1). Hate speech. *Encyclopædia Britannica*. https://www.britannica.com/topic/hate-speech

Australian Government, eSafety Commissioner, Netsafe, & UK Safer Internet Centre. (2020) *Online Hate Speech: Findings from Australia, New Zealand, and Europe* (p.3). https://www.esafety.gov.au/sites/default/files/2020-01/Hate%20speech-Report.pdf

Obermaier, M., & Schmuck, D. (2022, July 22). Youths as targets: factors of online hate speech victimization among adolescents and young adults. *Academic.oup.com*. https://academic.oup.com/jcmc/article/27/4/zmac012/6648458

Australian Government, eSafety Commissioner, Netsafe, & UK Safer Internet Centre. (2020) *Online Hate Speech: Findings from Australia, New Zealand, and Europe* (p.23). https://www.esafety.gov.au/sites/default/files/2020-01/Hate%20speech-Report.pdf

What is the real-world impact of online hate speech on young people?. *Internet Matters*. (2023, October 31). https://www.internetmatters.org/hub/question/what-is-the-real-world-impact-of-online-hate-speech-on-young-people/

Designed and Developed by South West Grid for Learning (https://swgfl.org.uk/). (2019, April 11). The consequences of online hate speech – a teenager's perspective. *SELMA - Hacking Hate.* https://hackinghate.eu/news/the-consequences-of-online-hate-speech-a-teenager-s-perspective/

Pang, G. (2022, March 7). Deep learning for hate speech detection: A large-scale empirical evaluation. *Medium*. https://towardsdatascience.com/deep-learning-for-hate-speech-detection-a-large-scale-empirical-evaluation-92831ded6bb6

Bowker, J., & Ophoff, J. (2022). Reducing exposure to hateful speech online. In K. Arai (Ed.), Intelligent Computing: Proceedings of the 2022 Computing Conference (Vol. 3, pp. 630-645). *Springer, Cham*. http://dx.doi.org/10.1007/978-3-031-10467-1_38

Jain, S., & Kamthania, D. (2020). Hate Speech Detector: Negator. *Proceedings of the International Conference on Innovative Computing & Communications (ICICC)* 2020, 4. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3563563

Onabola, O., Ma, Z., Xie, Y., Akera, B., Ibraheem, A., Xue, J., Liu, D., & Bengio, Y. (2023). hBert + BiasCorp - Fighting Racism on the Web. Montreal Institute for Learning Algorithms (Mila), Carnegie Mellon University, Independent Researcher, University of Toronto, *SCIFAR Program Co-director*. https://arxiv.org/pdf/2104.02242.pdf

Jestec.taylors.edu.my. (n.d.). https://jestec.taylors.edu.my/
(PDF) hate speech in Philippine election-related tweets: Automatic ... (n.d.-b).

Cabasag, N. V., Chan, V. R., Lim, S. C., Gonzales, M. E. (2019). Hate Speech in Philippine Election-Related Tweets: Automatic Detection and Classification Using Natural Language Processing. *Philippine Computing Journal*, XIV(1), 1-14.https://www.researchgate.net/publication/375911232_Hate_Speech_in_Philippine_Election-Related_Tweets_Automatic_Detection_and_Classification_Using_Natural_Language_Processing

Kemp, S. (2023, February 8). Digital 2023: The Philippines - datareportal – global digital insights. *DataReporta*l. https://datareportal.com/reports/digital-2023-philippines#:~:text=Internet%20use%20in%20the%20Philippines,at%20the%20start%20of%202023.

Manarpiis, N., Cortez, K. M., Cortez, M. G., & Bianca Nicole. (2021, December). *Online hate speech and the personal experiences of young adult Filipinos*. https://www.researchgate.net/profile/Noel-Manarpiis/publication/357117277_Online_Hate_Speech_and_the_Personal_Experiences_of_Young_Adult_Filipinos/links/61bc511b63bbd932429c5d5a/Online-Hate-Speech-and-the-Personal-Experiences-of-Young-Adult-Filipinos.pdf

Chinn, A. (2022, July 22). What's the System Usability Scale (SUS) & How Can You Use It? *HubSpot.* https://blog.hubspot.com/service/system-usability-scale-sus

Likert Scale: Examples and Definition. (n.d.). *Mentimeter.* https://www.mentimeter.com/blog/awesome-presentations/likert-scale-definition-and-how-to-use-it

Department of Health and Human Services. (n.d.). *System Usability Scale (SUS) | Usability.gov.* https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

Thomas, L. (2023, June 22). Simple Random Sampling | Definition, Steps & Examples. *Scribbr.* https://www.scribbr.com/methodology/simple-random-sampling/#:~:text=Simple%20random%20sampling%20is%20a,possible%20of%20this%20random%20subset

Quasi-Experimental Design: Types, Examples, Pros, and Cons. (2022, June 17). *MasterClass*. https://www.masterclass.com/articles/quasi-experimental

Windisch, S., Wiedlitzka, S., & Olaghere, A. (2021). PROTOCOL: Online interventions for reducing hate speech and cyberhate: A systematic review. *Campbell systematic* reviews, 17(1), e1133. https://doi.org/10.1002/cl2.1133

Bowker, J. & Ophoff, J. (2022) 'Reducing exposure to hateful speech online'. In: K. Arai (ed.) Intelligent Computing: Proceedings of the 2022 Computing Conference. vol. 3. Springer, Cham, pp. 630-645, Computing Conference 2022, United Kingdom, 14-15 July 2022.

Saya-ang, K. (2023, October 12). multilabel-tagalog-hate-speech. *Hugging Face*. https://huggingface.co/datasets/syke9p3/multilabel-tagalog-hate-speech/tree/main

Schmid, U. K., Kümpel, A. S., & Rieger, D. (2022). How social media users perceive different forms of online hate speech: A qualitative multi-method study. *New Media & Society*, 146144482210911. https://doi.org/10.1177/14614448221091185

APPENDICES

Appendix A. Gantt Chart

MONTH / ACTIVITY	JUNE	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER	JANUARY	FEBRUARY	MARCH
Planning and Initiation										
First Iteration Planing										
First Iteration Execution										
First Iteration Evaluation										
Second Iteration Planning										
Second Iteration Execution										
Second Iteration Evaluation										
Release and Deployment										
User Testing and Feedback										

Appendix B. Actual Thesis Expenses

THESIS EXPENSES

Quantity	Specifics	Approximate Cost	Actual Cost
1	Google Chrome Developer Account	5 USD	278.33 PHP

Prepared by:

Beatriz G. De Guia **Jordan Limwell C. Marcelo**

Eman Joseph T. De Leon **Lois Alysson R. Marquez**

Noted by:

Mr. Feliciano C. De Guia Jr. **Mrs. Ma. Lourdes C. Marcelo**

Mrs. Elsie T. De Leon **Mrs. Jiranee Rose R. Marquez**

Approved by:

Mr. John Irwin T. Vendivil **Engr. Regina R. Mape, MIT**

APPENDIX C. USER'S MANUAL

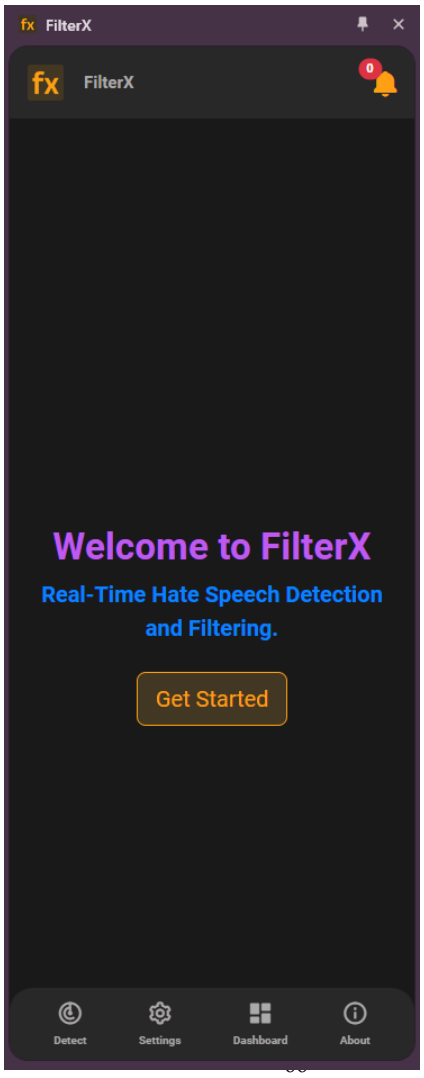
Appendix C. User's Manual

Introduction

FilterX is a web extension designed to moderate online hate speech in real-time. Using advanced natural language processing (NLP) and machine learning, it analyzes text content on web pages and filters out hateful content based on user-defined sensitivity settings.

Installation

1. Open your web browser.
2. Visit the web store (e.g., Chrome Web Store).
3. Search for "FilterX".
4. Click "Add to Chrome" or the equivalent button for your browser.
5. Confirm any prompts to complete the installation.
6. Once installed, you will see the FilterX icon in your browser's toolbar. Clicking this icon opens the FilterX user interface (UI), which provides access to all features of the extension.



After installation:

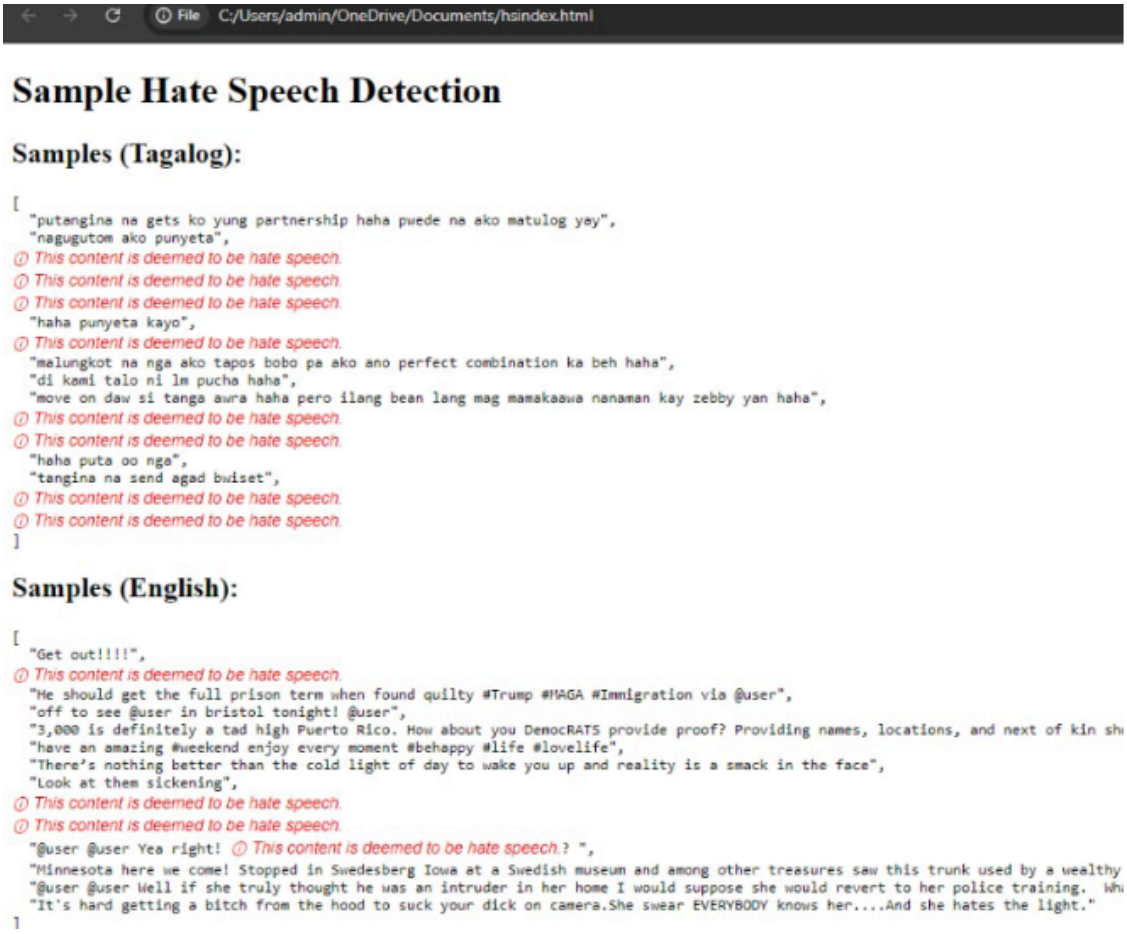
7. Click the FilterX icon to open the UI.
8. Follow the on-screen instructions to complete the initial setup, including setting your preferred language(s) for hate speech detection.
9. Navigate to any web page.
10. Click the FilterX icon to open the UI.
11. Press the "Scan" button to initiate a scan of the current web page.

Real-time Scanning

12. FilterX can automatically scan and filter content in real-time as you browse:

13. Ensure the "Real-time Scanning" option is enabled in the settings menu.

14. FilterX will continuously monitor the web pages you visit and filter out hate speech according to your sensitivity settings.



Display Modes

FilterX offers two display modes for filtered content:

1. Highlighting: Detected hate speech will be highlighted on the web page.
2. Censoring: Detected hate speech will be replaced with asterisks or removed from the web page.

- To switch between modes, go to the settings menu in the FilterX UI and select your preferred mode under "Display Options

Appendix D. Curriculum Vitae of Researchers

Curriculum Vitae of

Beatriz G. De Guia

San Juan St. Tuktukan Guiguinto, Bulacan
Beyadeguia462@gmail.com
+63 – 9626 – 705 – 298

EDUCATIONAL BACKGROUND

Level	Inclusive Dates	Name of school/ Institution
Tertiary	June 2025	STI College Balagtas
TechVoc	N/A	
High School	March 2019	Parada Nation High School
Elementary	March 2015	Hen. Tiburio De Leon Elementary School

PROFESSIONAL OR VOLUNTEER EXPERIENCE

Inclusive Dates	Nature of Experience/ Job Title	Name and Address of Company or Organization
N/A		

AFFILIATIONS

Inclusive Dates	Name of Organization	Position
N/A		

SKILLS

SKILLS	Level of Competency	Date Acquired
Programming Java	Intermediate	August 2022
Web Development (HTML, CSS, PHP)	Intermediate	June 2022

TRAININGS, SEMINARS OR WORKSHOP ATTENDED

Inclusive Dates	Title of Training, Seminar or Workshop
October 2024	Cloud Computing and Virtualization
October 2024	Computer Science Careers
October 2024	Mastering the Basics of TypeScript: A Beginner Guide
November 2024	Introduction to Digital Citizenship

Curriculum Vitae of

Eman Joseph T. De Leon

0562 Tiaong Guiguinto, Bulacan
Emanjoseph04deleon@gmail.com
+63 – 9496 – 380 – 523

EDUCATIONAL BACKGROUND

Level	Inclusive Dates	Name of school/ Institution
Tertiary	June 2025	STI College Balagtas
TechVoc	month year	
High School	month year	
Elementary	month year	

PROFESSIONAL OR VOLUNTEER EXPERIENCE

Inclusive Dates	Nature of Experience/ Job Title	Name and Address of Company or Organization
month year		
month year		
month year		
month year		

Listed in reverse chronological order (most recent first).

AFFILIATIONS

Inclusive Dates	Name of Organization	Position
month year		
month year		
month year		
month year		

Listed in reverse chronological order (most recent first).

SKILLS

SKILLS	Level of Competency	Date Acquired
		month year
		month year
		month year

TRAININGS, SEMINARS OR WORKSHOP ATTENDED

Inclusive Dates	Title of Training, Seminar or Workshop
month year	
month year	
month year	
month year	

Listed in reverse chronological order (most recent first).

Curriculum Vitae of

Jordan Limwell C. Marcelo

Blk 10 Lot 9 Wind Street Amihana Homes San Nicolas Bulakan, Bulacan
Offial.jordanmarcelo@gmail.com
+63 – 9310 – 360 – 212

EDUCATIONAL BACKGROUND

Level	Inclusive Dates	Name of school/ Institution
Tertiary	June 2025	STI College Balagtas
TechVoc	N/A	
High School	2013 - 2019	Assumpta Academy
Elementary	2006 - 2013	Marcelo H. Del Pilar Memorial School

PROFESSIONAL OR VOLUNTEER EXPERIENCE

Inclusive Dates	Nature of Experience/ Job Title	Name and Address of Company or Organization
N/A		

AFFILIATIONS

Inclusive Dates	Name of Organization	Position
N/A		

SKILLS

SKILLS	Level of Competency	Date Acquired
Python	Intermediate	November 2021
Java	Novice	November 2021
Javascript	Novice	November 2023

TRAININGS, SEMINARS OR WORKSHOP ATTENDED

Inclusive Dates	Title of Training, Seminar or Workshop
N/A	

Curriculum Vitae of

Lois Alysson R. Marquez

Libis St. Duhat, Bocaue, Bulacan
loisalyssonmarquez@gmail.com
+63 – 9478 – 243 – 597

EDUCATIONAL BACKGROUND

Level	Inclusive Dates	Name of school/ Institution
Tertiary	June 2025	STI College Balagtas
TechVoc	N/A	
High School	2014 - 2019	Sto. Niño Academy
Elementary	2008 - 2013	Duhat Elementary School

PROFESSIONAL OR VOLUNTEER EXPERIENCE

Inclusive Dates	Nature of Experience/ Job Title	Name and Address of Company or Organization
N/A		

AFFILIATIONS

Inclusive Dates	Name of Organization	Position
N/A		

SKILLS

SKILLS	Level of Competency	Date Acquired
Java	Intermediate	2021
Python	Novice	2022
Web Programming	Intermediate	2022

TRAININGS, SEMINARS OR WORKSHOP ATTENDED

Inclusive Dates	Title of Training, Seminar or Workshop
N/A	