***Middle East University***

***Faculty of Information Technology***

***Computer Science Department/ Computer Information System Department***

*Graduation Project Title: *

Student Name: Mohammad Almasri StudentID: 201910294

Student Name: Ahmad Hammad Student ID: 201910338

Supervisor:

Dr. Bassam Al- Sharjabi

2022/2023

**Table Of** **Contents**

[** 1](#_Toc125576757)

[Abstract 5](#_Toc125576758)

[Chapter 1: Introduction 5](#_Toc125576759)

[1.1 – Background on Hate Speech 5](#_Toc125576760)

[1.2 – Project Introduction 5](#_Toc125576761)

[1.3 – Problem Statement 7](#_Toc125576762)

[1.4 – Project Scope 7](#_Toc125576763)

[1.5 – Project Questions 7](#_Toc125576764)

[1.6 – Project Objectives 7](#_Toc125576765)

[Chapter 2: Background and Related Work 9](#_Toc125576766)

[2.1 – Hate Speech 9](#_Toc125576767)

[2.2 – Machine Learning 10](#_Toc125576768)

[2.3 – Serious Gaming and Gamification 10](#_Toc125576769)

[2.4 – Related Work 12](#_Toc125576770)

[Serious Gaming and Gamification 12](#_Toc125576771)

[Hate Speech 13](#_Toc125576772)

[Chapter 3: Methodology and System Models 15](#_Toc125576773)

[3.1 – Project Methodology 15](#_Toc125576774)

[3.2 – Architecture 16](#_Toc125576775)

[3.3 – System Models 17](#_Toc125576776)

[3.4 – Machine Learning Workflow and Experiments 24](#_Toc125576777)

[Experiment 1 25](#_Toc125576778)

[Experiment 2](#_Toc125576779) 25

[Experiment 3](#_Toc125576780) 26

[Experiment 4](#_Toc125576781) 28

[Experiment 5](#_Toc125576782) 32

[Experiment 6](#_Toc125576783) 32

[Chapter 4: Conclusion 40](#_Toc125576784)

[References 41](#_Toc125576785)

**List of Tables and Figures**

Figure 2.1 – Hate Speech Severity Levels…………………………………………………………………………………………………...10

Figure 2.2 – Classification Game Website Layout……………………………………………………………………………………………..……13

Figure 3.1 – Project Methodology Flow Diagram...……………………………………………………………...………………………………..15

Figure 3.2 – DiscH Architecture Diagram………………………………………………………………...……………………………….16

Figure 3.3 – Data Collection Sequence Diagram.…………………………………………...……………………………………………………17

Figure 3.4 – User Registration Sequence Diagram.…………………………………………………………………………………………………18

Figure 3.5 – User Login Sequence Diagram.…………………………………………………………………..……………………………..18

Figure 3.6 – Post Question Sequence Diagram.……..………………………………………………………………………..…………………19

Figure 3.7 – Answering Question Sequence Diagram.…………………………………………………………………………………………………20

Figure 3.8 – Model Training.………………………………………………………………………………………………….20

Figure 3.9 – Use Case Diagram.………………………………………………………………………………………………….21

Figure 3.10 – Reward Process Activity Diagram.……………………...………………………………….……………………………………….22

Figure 3.11 – Hate Speech Labeling Activity ………………………………..………………………………………………………………………..…23

Figure 3.12 – Machine Learning Model Development Workflow.…………………………………………..…………………………………………………….24

Table 4.1 – NB and SVC model Results……………………………………………………………………………………….……….…...25

Table 4.2 – Annotation results…………………………………………………………………………...……………………..….26

Figure 4.1 – Class Distribution for 500 Responses………………………………....……………..……………………………………………..…27

Figure 4.2 – Class Distribution for 4.8k Responses……………………………….…………….…………………………………………….…27

Table 4.3 – Models Comparison between 500 Responses and 4.8k ……………………………………………………………………...…………………….……………28

Experiment 4 Figures…………………………………………………………….........................28 – 31

Experiment 5 BERT Vs. AraBERT ……………………………………………………………..........32

Experiment 6 Figures ……………………………………………………………………………..32 - 38

# Abstract

Discrimination of Hate (DiscH) is system used in social media, governments, which depends on Artificial Intelligence, Machine Learning, Natural Language Processing and Serious Gaming as it has the capability of doing so. The purpose of this project was to build new and better Arabic comment datasets with more correct and defined labels with the help of users by using Serious Gaming techniques. One of main reasons why this project is made is how fast Arabic language lexicons and bag of words change within time. By doing this we get more accurate datasets thus more useable and more accurate AI and Machine Learning models.

# Chapter 1: Introduction

## 1.1 – Background on Hate Speech

Hate is a human emotion caused or enhanced through exposure to certain information, usually involving enduring dislike, loss of empathy and even a desire for harm against particular targets.(Bahador, 2020) Hate speech, denied by many as a realized rational concept, spawned many incidents and related crime around the world, worsened by the upswing of political discourse against other individuals. this led to internal issues to escalate in society, perchance establishing red-alert attestation of human rights mass contravention and further grave misconducts.(Fino, 2020) Often, hate speech is treated as way to suppress and threaten particular arguments, to ridicule and scorn whole groups in society while stifling disagreement. This type of speech can be spread by media outlets or online, creating conditions propitious to aggression and other human rights infringements against groups deemed by others as being dishonorable and discreditable.(“Equality and Hate Speech,” n.d.) Sometimes, when comments are posted online, different perceptions of those comments are invoked by the public online demographic, among those perceptions, are hate assessments, and this coupled with the fact that people’s views don’t change easily denotes heated arguments, and may lead to more alarming quandaries, especially on issues where core values and beliefs are already set. thus, a clear outlining is needed of whether some comments indicate hate, and what type of hate it is. (Bahador, 2020) Forms of hate speech can include throwing insults between two online users, hurtful and offensive language with slurs and detrimental stereotypes that lead to toxic environment formulation. The matter can get especially threatening when entities seek to incite people to violence against marginalized groups.(*What Is Hate Speech?*, n.d.)

## 1.2 – Project Introduction

Data has accumulated at a huge rate since the last ten years, and has resulted in huge amounts of data unutilized, leaving much potential wasted and missed on. (“Artificial Intelligence: What It Is and Why It Matters,” n.d.) Artificial Intelligence and particularly machine learning takes proper advantage of this data with the help of sophisticated algorithms to make many revelations, that stretch from self-driving fully autonomous cars, connected home devices through internet of things, general intelligence upsurge with development of a brain-controlled robots, AI-enabled systems are transfiguring and advancing nearly all facets of human civilization. (“ARTIFICIAL INTELLIGENCE,” n.d.; “Artificial Intelligence: What It Is and Why It Matters,” n.d.) In short, We can think of AI as the simulation of human intellect all through programming computer machines to think in the same way as humans and mimic their actions, this allows for machines such as our personal computers to learn from experience, adjust to new inputs and perform human-like tasks. (“Artificial Intelligence: What It Is and Why It Matters,” n.d.; Frankenfield, n.d.)

With the advantage of AI being applicable in numerous sectors and industries, It can be leveraged to enable continuous learning and discovery of online hate speech statements on the internet. even though online hate speech can be more tolerable than hate speech expressed, namely, in real life, it is unfortunately more difficult to control and detect immediately. Computer machines are dependable in performing repetitive and tedious tasks, this can be used accordingly in situations of reporting and detecting hate speech, which would be a tedious task for a human being to carry out, especially since millions of online comments are being written down every minute.(“Artificial Intelligence: What It Is and Why It Matters,” n.d.) the detection job can be done competently and effectively on machines, and it’ll get better with time as technology advances and computing benchmarks are incessantly broken through.

Across all different parts of the world, we as human beings can communicate with each other through the use of internet. (“Importance Of Internet Technology For Easy Life,” n.d.) through internet interaction, people can learn about their difference in languages, cultures, beliefs, and diversity of other countries. There is an opportunity to make valuable usage of the internet users’ collective knowledge to optimize the task of hate speech detection. Enhancements will be enabling detection of certain types of hate speech, identifying new terms used in online hate speech, and more accurate classification of the internet comments. All of this creates an adaptable, evolving, human-aided hate speech detection system that is incorporates both AI capability and human knowledge. Collecting info from internet users requires building a website platform for serving contributing users that help deliver feedback and improvements to the hate speech detection system.

With building a website, a question comes to mind, How to attract website users? How do you convince them that this website is worth their time? This is where serious gaming comes to play. Serious gaming is about making gamified user-experience that has a purpose other than entertainment, player practice skills and achieve progress beyond simply enjoying a leisure activity. the main purpose is typically to promote learning and behavior change. (Pilote & Chiniara, 2019; “The Route to a Successful Serious Game,” n.d.) Games have an attractive fun factor that appeals to many audiences, several principles are integrated into games in order to be successful, among them (“The Route to a Successful Serious Game,” n.d.):

* Addictive gameplay
* Engaging for the player and has staying power with a progression system.
* Creative concept to hook new players.
* Fun context.

In this project, we built a hate speech detection system that incorporates an AI model to recognize and classify hate speech examples. These examples were verified and validated by website users. The website plays a role in classifying new hate speech words and terms, categorizing it more indicatively, and retrain the AI model on a bigger dataset to further optimize its performance and facilitate it to recognize wider sorts of hate speech. Serious gaming principles are mainly used to gravitate new website users and making the website user experience more enjoyable.

## 1.3 – Problem Statement

Many acts of cyberbullying, violence, and toxic behavior have been affecting individuals adversely, whether online in the world wide web or outdoors in various societies, mainly through the use of verbal words that express hate and dislike. An issue of underreporting arises when it comes to hate speech, due to the task being tedious and requiring huge efforts. A system must be produced for recognizing hate speech and streamlining the reporting procedure to avoid any further potentially damaging and injurious acts instigated from conducting hate speech, furthermore with human intervention and feedback, we can optimize the system to be more accurate and effective through the use of a gamified website.

## 1.4 – Project Scope

The project name is Discernment of Hate (DiscH), the main purpose for this project was to build an AI model that detects hate speech and labels them. The model is deployed as an API available for use, utilizing data generated through the website platform built, integrating serious games principles and ranking system that motivates users to interact with the platform, building the newly-labeled dataset. model is improved in time by continuous aid of the users, through updating new terms in Arabic language, and adding or removing classes depending on their evaluations.

We have used the Levantine Hate Speech and Abusive (L-HSAB) as starting point to build our model, we also got data spanning several topics from twitter API. These comments were used to evaluate and optimize the AI model with the help of users. it should be noted that data outside social media platforms is not used in this project, mainly focusing on colloquial Arabic comments. Our acceptance criteria are to make the AI model learn and make progress as well as attracting users through continuous progression and related-serious gaming principles.

## 1.5 – Project Questions

* Is the available Arabic dataset enough for our purpose?
* What AI model is used for hate speech classification?
* Is serious gaming going to be suitable for this dataset?
* How are we going to use serious gaming in this project?
* How do we define correct hate speech types (labels) based on users’ responses on the website?
* How are we going to evaluate the generated dataset and our AI model?

## 1.6 – Project Objectives

* Artificial intelligence model for classifying hate speech statements.
* Improving and updating the lexicons of Arabic Language which is frequently updated over time.
* Making hate speech classes more convenient and accurate.
* Optimization of the main AI model through validation of hate speech comments.
* Making an interactive website that asks users to evaluate (classify) comments and based on that we make our model learn from them.
* Ranking system for our users who help us with labeling.
* Reward system to reward completion of website activities.

# Chapter 2: Background and Related Work

In this iteration, we focused on designing the database schema for later implementation and imported hate speech data from labeled and unlabeled data sources. We then used the data collected to make sample questions (what type of hate speech is this?), after that we determined the main types of hate speech and why we used them, we also created the mechanism for how to handle recorded responses to the question to make the defining label for the hate speech example.

## 2.1 – Hate Speech

The main task of our project is to recognize different types of hate speech, so we decided to define several types of hate speech, in such a way that they are inclusive of all perceived types of online hate comments seen on the internet, the main types of hate speech we’ll be working with involve:

1. Racist: showing a certain race to be superior to another, believing particular personality traits are related completely to race and heritage, expressing animosity to human race, discrimination on account of race. (Smedley, 2021)
2. Violent: speech used to incite violent actions and movements, may induce breaking laws and committing felonies. (Bahador, 2020)
3. Religious affiliation: any speech that which nominates religious affiliation. (“Teaching about Controversial Issues and Use of Controversial Materials,” 1997)
4. Mockery: social act approach to joking about individuals and their distinction from the standards of the group. Usually done in a sarcastic manner (Dickinson et al., n.d.)
5. Sexual harassment: any form of unwanted, offensive, humiliating, or intimidating sexual conduct. (“What Is Sexual Harassment?,” n.d.)

We choose these types as they make for most internet comments seen online and have varying degrees of severity levels that can classified within the hate levels shown in Figure 2.1, with each type being distinct enough from the other, covering its own region of comments. These types also are descriptive enough of the online comment, whilst not being overly explicit, which would inaugurate data collection repercussions and impede on the AI model training process. the model will be able to distinguish between different types of hate speech, not merely classifying binary category of hate/not.

Figure 2.1 reveals how grave the consequences of hate speech can be, with some the comments may directly insinuate criminal action such as murder. Related description of each level and word examples are also provided.

A picture containing table

Description automatically generated

Figure 2.1: Hate Speech Severity Levels (Bahador, 2020)

## 2.2 – Machine Learning

Basically, machine learning is a branch of artificial intelligence, it is focused on two interrelated questions: How can one construct computer systems that automatically improve through experience? and What are the fundamental statistical computational-information-theoretic laws that govern all learning systems, including computers, humans, and organizations (Jordan & Mitchell, n.d.). Machine learning can be defined as applying statistical methods to input data to predict outputs. The four common types of machine learning are supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

The first type of machine learning is supervised learning where the algorithm generates a function that maps input to desired output, this type of learning using labeled data to train the model to validate it. The second type of machine learning is unsupervised learning where the algorithm tries to find the similarities between the unlabeled data. The third type of machine learning is semi-supervised learning which is a combination of supervised and unsupervised learning. The fourth type of machine learning is reinforcement learning where the algorithm learns a policy of how to act given an observation, it is based on rewarding the desired behavior and punching the undesired ones.(Ayodele, 2010).

## 2.3 – Serious Gaming and Gamification

interactive games that allow players to carry out activities that enable them to practice skills and achieve aspects beyond simply enjoying a leisure activity (Pilote & Chiniara, 2019) They have another purpose besides entertainment, which is to promote learning and behavior change (“The Route to a Successful Serious Game,” n.d.). Educational, Medical, and Marketing sectors benefit greatly from the use of serious, using it as an essential tool for strategic planning, training for crisis situations and education. The concept can be utilized to simulate realistic problems, giving a sense of realism that heightens the experience and helps deliver the main message across to the players. (Melger, https://www.deltares.nl/en/software-solutions/deltares-serious-game-portal/)

Serious gaming is an important concept because it is the main way we’ll attract the customer’s attention to the website, through an entertaining gamified interface experience, then we introduce the customer to our main task that we would like him to do, increasing his awareness, and enabling contribution to the task by making him interact with the website like it is a game he is trying to play. And this can work effectively as games are widely known as an accessible media to spend time and have fun, which translates well to an easy way for enabled interactivity with the website. (“The Route to a Successful Serious Game,” n.d.)

We can’t consider a website platform as a game without properly gamifying it. Gamification can be done through adding game elements and applying game-design principles in a non-game context (in our case, the website) in order to create similar experiences to those experienced when playing games in order to motivate and engage users. This we’ll be implemented in the general user interface of the website by stylizing the design and making it simpler while also adding the fun factor in its general navigation and usage. Implementation wise, we’ll use a specialized frontend framework called vue.js to achieve our design goal.

The game elements we’ll use to achieve serious gaming are:

* **Points**: rewards for the player for successful accomplishment of the activity
  + Activities include choosing an answer(question response) that is representative of the majority of responses, good feedback on the answer, general everyday logging on the website
  + used as form of the progression for the online user, while also acting as currency to purchase certain personalized features for the user.
* **Badges**: a summarized visual representation of the online user’s status of accomplishments
  + gold, silver, bronze badges
    - gold: best recognized response to the question, with good feedback
    - silver: best recognized response to the question, with okay feedback
    - bronze: best recognized response to the question, with no feedback provided
  + these badges are shown publicly to any user on the website
* **Leaderboards**: list the best current players on the website, ranking them according to their accomplishment of activities
  + Ranking based on the highest number of gold badges claimed each month.
* **Performance graphs**: shows metrics such as number of questions answered and badges claimed each month, and other metrics which allow the player to evaluate his own performance over time
  + Metrics: monthly question answered / monthly badge claims / overall good feedbacks / overall bad feedbacks / overall okay feedbacks
  + Percentage difference is used to show the metrics’ relative change compared to last month, to let the user focus on improvements.
* **Meaningful stories**: incorporated for the purpose of enriching unstimulating contexts, to keep the user’s attention and motivation, and giving gravity to the main task via a running narrative.

## 2.4 – Related Work

### Serious Gaming and Gamification

Because games have familiarity with a large number of people, we can utilize the game environment to make people learn in an effective context that make the player involved and engaged, making abstract concepts seem like relevant occurrences of everyday life. (Bellotti et al., n.d.)

A paper by F. Bellotti described effective ways to design serious game suggested discussing with stakeholders to consider different fields and technologies relevant for the design of the game, developing the product while exploiting the strengths of the technologies, focusing primarily on the benefit for the user, and incorporating sound educational principles. Though the paper acknowledges that the current serious games lack advancement in certain aspects, it does encourage comprehensively analyzing the requirements and design principles of building serious games, which will result in a much stronger product, strongly suggesting looking into evaluation tools, methodologies, computing and communication architectures, and various technologies. (Bellotti et al., n.d.)

In another paper by Giorgio Maria, it was shown that affordable annotated dataset creation is possible through mixed approaches that depend on crowd-sourcing and active user learning, making it an alternative to the traditional approach of annotation-by-experts that is known to be time-consuming and costly. This was achieved through a gamified classification-task-based website, which was used to build accurate classifiers that are also efficient, not depending on large amounts of data. The general layout of the game is a two-dimensional plane that has many points, separated by the user specifying a straight-line separating data into classes, the line being a probabilistic model that has hyperparameters determined by the user. the data points are accordingly labeled based on the line separation specified by the user. players can adapt the two parameters m and q in order to optimize the separation of points and find a better model. Each level in the game is presented with a main goal, which is finding the best classifier with the least number of resources. The game is implemented in a interactive web application, its general layout is shown in figure 2.2. The resultant model (separation result) is measured based on F1-score and compared to SVM. General results show that good-performing classifiers can be trained/validated with just 25% of the original dataset, although the user-built model showed worse results than a fully-tuned and well-trained SVM model. (Maria, n.d.)

Chart

Description automatically generated

Figure 2.2: Classification Game Website Layout

Serious gaming was also deployed for Major Depressive Disorder (MDD) recognition, the paper presented a serious game called “The Delivery” that collects data from players through in-game conversations, environment interactions, and task completions. After building a dataset, machine learning models were trained to recognize patterns of players with MDD vs players without MDD. The paper reported data collected from 26 gameplay sessions (24% players with MDD, 76% players without MDD), and application of 4 classification models: decision trees, random forests, k-Nearest Neighbors, and support vector machines. Evaluation results for the models showed promising potential in the recognition task, even with the dataset being imbalanced and relatively small, with SVM classifier having the best overall outcomes (80.8% accuracy, 71.2% F1-score). (Tsionas, n.d.)

### Hate Speech

Online communities are growing so fast, users uploading and share a ton of data in various forms text, images, or even videos. It’s hard to make sure that everyone is treated in a decent way, so a new problem of online hate speech has been produced (Pitsilis, 2018).

The article by Mullah, reviewed advances made so far in hate speech detection in social media. They analyzed different approaches classical machine learning, Ensemble and deep learning. This study found that there is more research work done in hate speech using machine learning than ensemble and deep learning, so more research and exploration can be done to for improvements. The research also discussed in the weaknesses and strengths which can help in guiding the researchers’ choice. This article also identified Cultural variations, pandemic or natural disaster, data sparsity, imbalance dataset challenge and dataset availability concern (Mullah, 2021).

The targeted audience for this research review is mostly newcomers in the domain of hate speech (text) classification in the social media. This review provides all the required steps needed to follow in conducting text classification tasks using machine learning and some open challenges in the domain (Mullah, 2021).

# Chapter 3: Methodology and System Models

We worked on outlining and designing the general project methodology for the system, while updating the system mechanisms already established in iteration 1. modeling and creation were also accomplished on general processes for some the ideas elaborated on iteration 1, such as serious gaming and gamification, certain frameworks like Django, and hate speech detection. Another deliverable is the workflow schedule and pipeline design for the system regarding automation of certain processes as an improvement with regards to efficiency. Finally, we talked about general project constraints and challenges, while also providing broad information on the implementation of the project with relation to frameworks and particular software.

## 3.1 – Project Methodology

For the project, we initially collected data and worked on determining the hate speech labels, after manual labeling was done for a certain set of comment, and an AI model was trained and tested to evaluate the results of the new dataset built from using the gamified website platform made for answering hate speech labeling questions. The following shows more detail about the general process:

**Data Collection:** was collected from multiple sources labeled and unlabeled, and the machine learning model was trained and tested labeled dataset.

**New hate Speech Label Creation:** we defined new labels of hate speech more descriptive for the dataset we are going to generate, and they are: Racist, Violent, Religious affiliation, Mockery, Sexual harassment, and Normal (non-hate).

**Manual Labelling of dataset:** we manually labeled the comments using new labels on the website platform to build the new dataset. After a certain amount of time, new labels are formed based on response data and best-label-criteria functions to make the new labeled dataset.

**Train/ Test AI model and keeping correctly classified part of dataset**: training the model on the formed dataset which gets preprocessed (tokenization, removing stop-words, feature extraction with TF-IDF, bag of words) and appended to previously formed dataset (main model dataset with new labels) for model training and evaluation. Based on the results, correctly classified data is appended to main model dataset for the next training process.

**Evaluation of new Dataset (SG):** For misclassified comments we re-label it in the website platform for further feedback from website users, while keeping a set of correctly classified comments in the website for label validation.

The dataset goes on iterative process of labeling, training and testing the model, which is going to improve the dataset, and subsequently, the AI model.

Figure 3.1 shows the general methodology modeled in flowchart design.

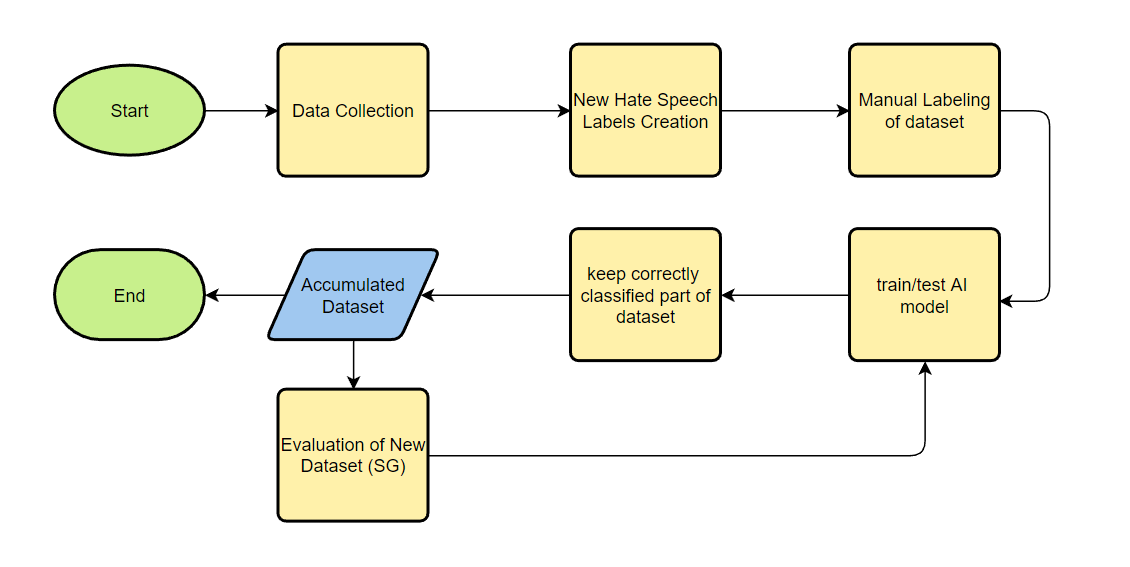


Figure 3.1: Project Methodology Flow Chart

## 3.2 – Architecture

A pipeline is built for an ETL process by moving the data from different sources, transforming it to an appropriate format, and loading it to destination stores (PostgreSQL database). Sources of data include twitter API, YouTube online comments on selected videos, and labeled datasets from websites such as Kaggle or scientific paper links. the pipeline is scheduled to work in fixed intervals (E.g. ), so the general procedure is automated. Another schedule is set for model training, so that it is trained based on new dataset composed from online responses from the website stored in the database.

We used django as a web framework for the frontend and backend components, which allows for web development using python programming language. Html, CSS, JavaScript, and jQuery is used for building the website pages of the frontend. we tried to make the design attractive for a better user experience, implementing serious gaming principles and a progression system using points, medals, and user dashboard. This keeps the user engaged and motivated when using the website.

We used django as it is a complete package for developing web application, having many ready-to-use utilities and packages for versatile and secure design that follows traditional principles while also being maintainable, the various of utilities of the framework suits rapid web development that is also enhanced by its dependance of python language, one of the easiest and most accessible programming languages on the current market that most people working in the AI industry have worked with one or another.

Figure 3.2 below shows the architecture diagram for the project.

Graphical user interface, application

Description automatically generated

Figure 3.2: DiscH Architecture Diagram

## 3.3 – System Models

Here we modeled our system using sequence and use case UML diagrams, these diagrams represent the overall general system operations that are organized in these phases: data collection, user registration, user login, post question, answering question.

Data collection, shown in figure 3.3, is about gathering the data from labeled and unlabeled data sources, processing it accordingly, and importing it into the database for it be displayed as a question in the website. For the dataset collection model, the term “subset data” was changed to “unimported”, this better explains the fact we are trying to specify data not imported into database before. Additionally, a new reply message was specified after the import into database message, to indicate that the imported data should be moved to an imported data directory.

User registration and login, shown in figures 3.4 and 3.5, represent the overall login and registration mechanism for the website.

The general process behind question uploading (figure 3.6) is first initiated by a scripted database export to the database backend, which then gets formulated as a labeling question (for each question). Finally, for each response recorded on the website, it is processed and imported or updated back on the database.

In posting question, we changed some messages to better clarify the process such as “export hate speech comment” to “export question data” (data in database is already in question format). And “formulate labeling question” to “add hate speech types” (more descriptive, as question data is added on hate speech types as choices for the question response). We also decided to remove return feedback reply message as response already implies if it has feedback (on the condition It was written with the choice). Figure 3.6 shows the updated model for post question.

A picture containing text

Description automatically generated

Figure 3.3: Data Collection (Sequence Diagram)

Graphical user interface

Description automatically generated with medium confidence

Figure 3.4: User Registration (Sequence Diagram)

Diagram

Description automatically generated

Figure 3.5: User Login (Sequence Diagram)

A picture containing diagram

Description automatically generated

Figure 3.6: Post Question (Sequence Diagram)

Every response recorded is considered as some user’s answer to a hate speech labeling question, the answer being one of hate speech types of the user specified to the hate speech comment (racist, violent, religious affiliation, mockery, sexual harassment, normal), figure 3.7 shows the overall process. note that there is a scheduled update for the responses-to-label process that is responsible for building the new dataset that consists of the new labels constituent of the hate speech types we specified for each hate speech question. Serious gaming elements is utilized in our system, incorporating user achievement and performance data.

After amassing a large enough dataset, we then use the dataset to build a new model that is capable of predicting new hate speech labels, going from only being able to predict hate/non-hate to racist/violent/religious affiliation/mockery/sexual harassment/normal which are the same as the answer choices for each labeling question, thus enabling the website as a user-based platform for building bigger and more expansive datasets. Figure 3.8 shows the general process behind model training.

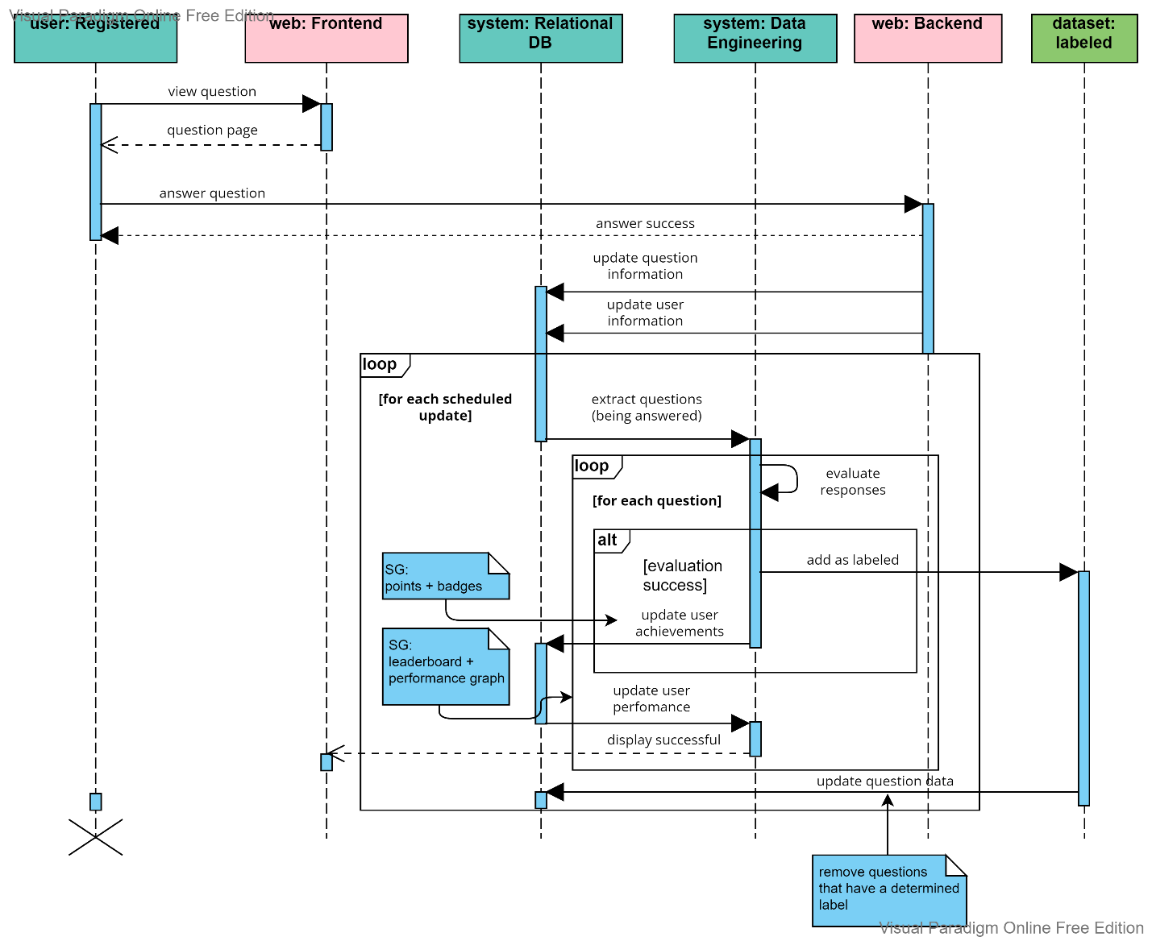


Figure 3.7: Answering Question (Sequence Diagram)

Graphical user interface, diagram

Description automatically generated

Figure 3.8: Model Training

In the use case diagram, shown in figure 3.9, the user can perform actions: logging in, registration, viewing ranking systems, and labeling comments. Our web developers can access these actions additionally to the new labeled database which will be imported to our database. The role of data scientist here is to evaluate the new labels, update, and optimize our machine learning model.

Diagram

Description automatically generated

Figure 3.9: Use Case Diagram

Figure 3.10 shows the basic process behind the reward system for the website. It shows that the website user gets rewarded points and badges for answering a labeling question, the amount of points depend on the accuracy of the label chosen, and good feedback is chosen based on the number of upvotes users gave the justification for the labeling. It should be noted that the reward is determined by a number of user responses accumulated to establish and distinguish good user responses from others, as to provide them with better rewards. This means that the reward won’t be provided immediately upon the achievement of the task.

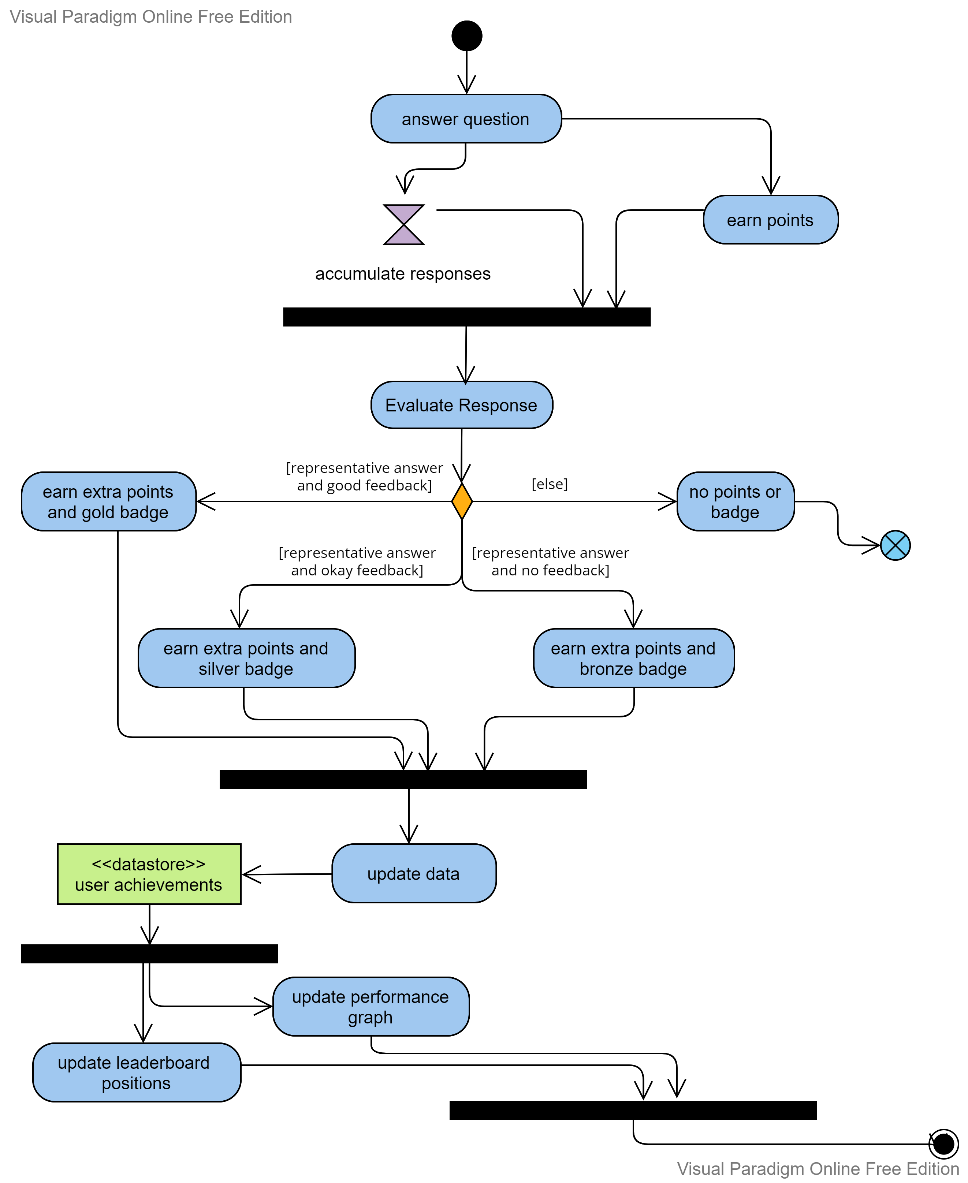


Figure 3.10: Reward Process Activity Diagram

We’ll entail more details about the reward process in the context of the system shown in a sequence in later iterations.

The main process of label a hate speech comment through the website is shown in figure 3.11. we can see that initially the user must visit the website and log in or register with an account, then answer one or more hate speech questions that ask, “what type of hate speech is this?”. Finally, after accumulating a certain number of responses for a certain question, the hate speech comment related to the question is labeled based on a decisive criterion, and the question related is removed from the website.

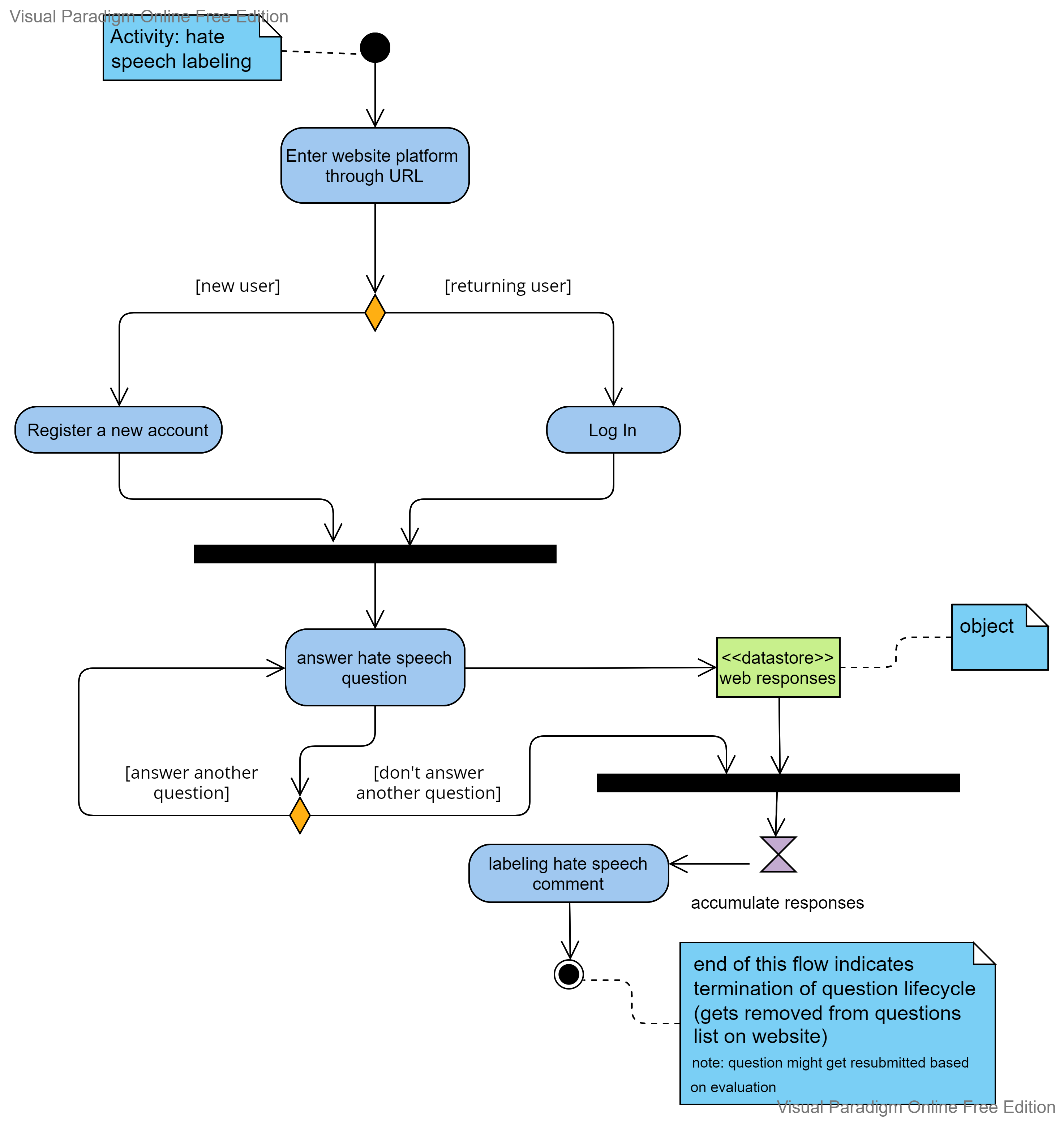


Figure 3.11: Hate Speech Labeling Activity Diagram

## 3.4 – Machine Learning Workflow

Figure 3.12 shows the general workflow conducted for machine learning life cycle, after getting responses we re-train, tune, test and validation the model, then we deployed the model into the production in the website. This process is iterative.

Diagram

Description automatically generated

Figure 3.12: Machine Learning Model Development Workflow

# Chapter 4: Experiments & Results

## 4.1 – Experiment 1

We will evaluate our labeled data sources to derive initial and leading inferences and interpretations about the data and the classification problem, before inputting it into the website platform. the project primarily focuses on the Naive Bayes (NB) and Support Vector Classifier machine learning methods (SVC). Following the preprocessing of the tweets, the models were trained, tweaked, evaluated, and decided on the better performance model based on the accuracy, precision, f1-score, and recall of these models. The Naïve Bayes classifier is based on probabilities, while SVC tries to draw hyperplanes between the data.

The dataset we are using for this iteration is called “Levantine Twitter Dataset for Hate Speech” (shortly for L-HSAB) and is an Arabian political dataset taken from politicians, social activists, and TV anchors.

After removing stop words and performed TF-IDF, the data was split into train and test, with test size being 20%. two experiments have been run with our initial dataset L-HSAB; they were:

1. Binary Classification: tweets are classified into hate or normal.
2. Multi-class Classification: tweets are classified into hate, abusive or normal.

A spreadsheet was made summarizing the outcome of both experiments, shown in table 4.1 below. We notice across both cases that SVC had better results, but since this isn’t the data from the DISCH platform responses, both models will be used for subsequent experiments.



Table 4.1: NB and SVC model results

Other Noted Results:

* Competent performance in binary classification results, with metrics ranging around 75% - 85%.
* Binary Classification Outcomes are generally better than multi-class, with notable declines in Precision, Recall, and F1-score.

## 4.2 – Experiment 2

This experiment was conducted for analyzing results and discovering helpful insights for improvements regarding our workflow, which would be implemented in later iterations.

**Data Annotation experiment**:

300 hundred tweets were brought over from L-HSAB dataset, then 5 chosen annotators were tasked to label comments either Racist, Violet, Religious affiliation, Mockery, Sexual harassment or Normal. Labeling was made via a simple choice on an excel sheet, given to each annotator, then results were collected thereafter. The maximum vote principle was used to determine the label. The profound effect was that the initial dataset is being labeled with multiple new classes, table 4.2 shows the output of the annotation experiment, we faced three cases:

1. Total agreement: the five annotators annotated a tweet with the same label.
2. Major agreement: three out of five annotators agreed to the same label.
3. Conflict: if there is group of annotators labeled something different than other group of annotators, and it resulted in two or more of the choices having the same number of votes, therefore there is no one maximum label vote.



Table 4.2: Annotation results

After generating the new dataset from the aforementioned experiment, we started removing stop words, performing TF-IDF, and splitting the data into train and test. We ran multiple experiments using NB and SVC, the accuracy with NB classifier was generally higher than SVC classifier, but accuracy alone is not the only measure here.

**Experiment conclusions**:

Our “Sexual Harassment” label was not found in the generated dataset, this is due there is few sexual harassment related comments. This probably means that there is not enough data for certain classes, indicating issues of imbalanced data. Thus, more annotated data is needed to work with, Models will be trained after further data collection in later iterations.

## 4.3 – Experiment 3

the labeling process was conducted through the deployed web application. The first 500 responses were labeled by us. And the rest was by other users in the university, resulting in an overall 4,800 responses. All scores are based on macro averages for the evaluation metrics. Figures 4.1 and 4.2 show the general distribution for the 500 case and the 4,800 case.

Chart, bar chart

Description automatically generated

Figure 4.1: Class Distribution for 500 Responses

Chart, bar chart

Description automatically generated

Figure 4.2: Class Distribution for 4.8k Responses

We have run two experiments one with 500 responses and the one was with 4.8k responses, by looking at table 4.3 we can see that:

* SVC performed better in 500 responses.
* Naïve Bayes was better in the 4.8k responses, the BERT model was almost doing the same job.
* The performance in both cases was almost identical having close scores, which doesn’t make any sense since the second case has almost 10 times the amount of data in the first case.



Table 4.3: Models Comparison between 500 Responses and 4.8k Responses

We have run two experiments one with 500 responses and the one was with 4.8k responses, by looking at table 4.3 we can see that:

* SVC performed better in 500 responses.
* Naïve Bayes was better in the 4.8k responses, the BERT model was almost doing the same job.
* The performance in both cases was almost identical having close scores, which doesn’t make any sense since the second case has almost 10 times the amount of data in the first case.

## 4.4 – Experiment 4

In this experiment, we tried turning the problem into binary classification for each of the hate speech classes outlined in the project, while also trying the ‘hate’ label to evaluate hate recognition instead of classification. The reason for this experiment is to try and observe whether there is a notable improvement or not when simplifying the given problem, and possibly give an insight as to the general inadequacy of the classification for the multi-class case.

We trained the models on 500 response data and 4800 response data. The results are shown below (table 4.4, table 4.5) for the measure of accuracy.



Table 4.4: Binary Classification 500 Responses – Accuracy



Table 4.5: Binary Classification 4.8k Responses - Accuracy

Stated results entail the following:

* Generally good accuracy across most labels (E.g., Racist/not Racist with +80% accuracy)
* Logistic Regression and Random Forest showed the best results by average.
  + Except for sexual harassment case for logistic regression
* A huge improvement in accuracy compared to multi-class results.

But then if we calculate the precision of the models to check positive case accuracy, the results degrade to very poor scales, as shown in tables 4.6 and 4.7 below.



Table 4.6: Binary Classification 500 Responses - Precision

  
Table 4.7: Binary Classification 4.8k Responses - Precision

Taking both accuracy and precision into account, we can conclude the following:

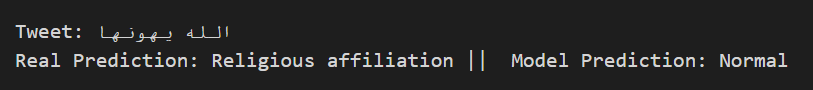
* Since we have good accuracy but bad precision on positive class (Hate, Violent, Mockery, etc.), the model can differentiate on comments for the negative class but is unable to do so for the positive cases, having very few true positives captured in its predictions.

the model can’t recognize different types of hate, but it can recognize hate from non-hate. This is evident by precision measure between hate/not hate and other cases.

### Analyze Incorrect classes & Reasoning

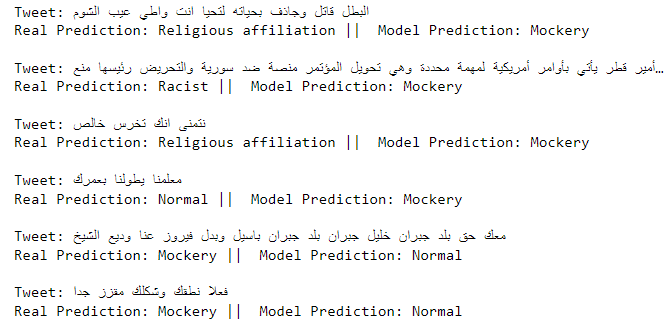
Annotation results are shown below in table 4.8, some notes are:

1. Inaccurate labeling by the website users, either by not understanding or differentiating between the labels or making random submissions.



Text

Description automatically generated with medium confidence 



* For example, in the cases shown above, we can see that the real prediction was off, and the model was predicting it correctly.

1. Conflicts between users’ answers and user bias on certain comments.

On 4.8k responses, we have 3092 unique question.

* Total Agreements: all responses agreed to the same label.
* Conflict: the response for the comment contains two classes.
* Huge conflict: the response for the comment has more than two answers



Table 4.8: Annotation results

### Balancing the Classes

Another experiment was run where classes were balanced, to analyze whether misclassification is attributed to the misbalance issue, we used under-sampling for the 500 responses, thus just 37 comments for each class. We also did the same for the 4.8k responses, which corresponds to 295 comments for each class. Results are shown in table 4.9 below.



Table 4.9: Models Experiments

Result: We can see that there no improvement even when balancing the classes, thus the issue can’t be ascribed by the misbalance, clarified by the results displayed above.

## 4.5 – Experiment 5

For the final experiment, we’ve tried using pretrained models employed from the hugging face platform, particularly text-classification models based on BERT (and ARABERT for Arabic language), through which we are going to fine-tune them with our data and evaluate the results, to see any significant improvement from the previous results.

The models were trained for 10 epochs on a batch size of 4 (due to computation constraints), using train and evaluation splits of the data, with fraction of 15% for the evaluation part. various evaluation metrics were used, such as precision, recall and accuracy. The final evaluation for the models is revealed in 4.10 below.



Table 4.10: BERT Vs. araBERT

The following can be concluded:

1. An underwhelming score across all measures, and especially the precision metric.
2. A significantly lower measure for precision than accuracy means that the problem of identifying categories of hate persists from previous experiments.
3. The model is generally learning misleading patterns found within our data, even after cleaning.

## 4.6 – Experiment 6

Due to users’ labeling bias and errors, and the overall meager classification results, another method was tried for making the dataset, by showing only 1000 questions to the users, we can make just these 1k show in the website to get a good number of responses on these questions, by doing that we can almost make sure that the questions and their bag of words were labeled correctly. We’re also limiting the number of people using the website to a couple of users to resolve the conflict matter. Table 4.11 shows the annotation results for the procedure.



Table 4.11: Annotation results

### Class Distribution

General class distribution after labeling is displayed below (Figure 4.3). It is noted that the normal class is most prominent among classes, and very few examples are labeled for both religious affiliation and sexual harassment, which is true for previous cases of class distribution figure, with the difference being that the skew level being higher and more imbalanced than previous annotation experiments.

Chart, bar chart

Description automatically generated

Figure 4.3: Class Distribution (1000 Responses)

### Word cloud

After finishing with our labeling, we extracted unique bag of words and made word cloud for each class, shown in figures 4.4-4.9

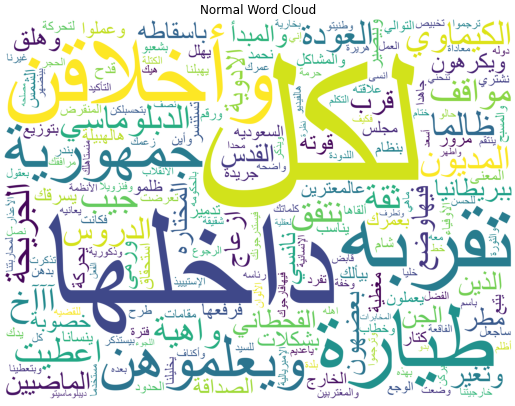


Figure 4.4: Word Cloud for Normal class

Text

Description automatically generated

Figure 4.5: Word Cloud for Mockery class

Text, whiteboard

Description automatically generated

Figure 4.6: Word Cloud for Violent class

Text, whiteboard

Description automatically generated

Figure 4.7: Word Cloud for Sexual Harassment class

Text

Description automatically generated

Figure 4.8: Word Cloud for Religious Affiliation class

Text

Description automatically generated

Figure 4.8: Word Cloud for Racist Class

### Misclassifications

Figures 4.9 and 4.10 indicate the reduction of misclassification occurring when training models on data resultant from 1000 responses. This shows the gain of applying patterns to model evaluation, to achieve better accuracy.

Chart, bar chart

Description automatically generated

Figure 4.9: Misclassification (before pattern)

Chart, bar chart

Description automatically generated

Figure 4.10: Misclassification (after pattern)

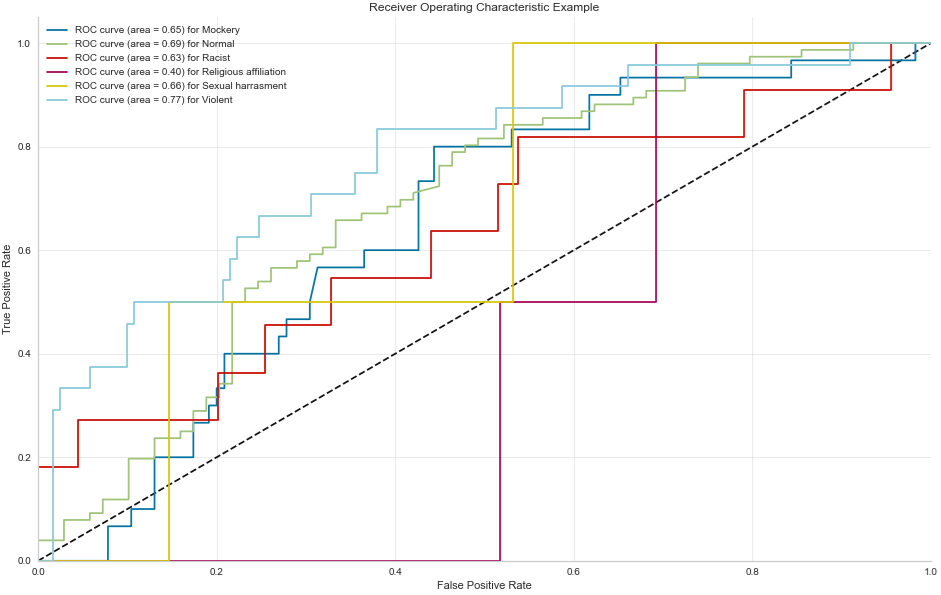


Figure 4.11: ROC Figure

Figure 4.11 illustrates receiver operating characteristic that can show us ratio the true positive and false positive, we can see that ROC score (0.40) for religious affiliation class which is bad, while the highest roc score (0.77) was for violent class.

### Evaluation

The following evaluation for the best model trained (logistic regression) is shown in table 4.11 below, we see a substantial improvement after applying patterns, going from 58.62% accuracy to 98.62%, with the biggest improvement being in precision, recall, and f1 measure for the classes in general.

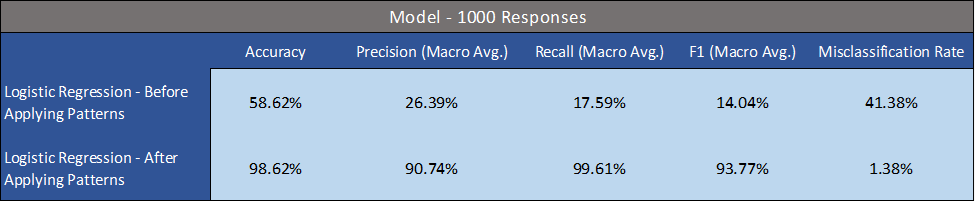


Table 4.11: Evaluation Results (before and after patterns applied)

Following Conclusions:

* A general improvement in performance for this labeled set (before patterns) compared to results of previous experiments. This entails that a high conflict rate in labeling results in a dataset with conflicting patterns that can’t be used effectively for training.
* Applying generated patterns and BOW after generating model inference provide superior results, especially for identifying target classes (positive case).

# Chapter 5: Conclusion

Experiments 1 and 2 showcases an important performance degradation when transforming the problem from a binary classification issue to multi-class. Shown by the decrease in general accuracy for the evaluation conducted.

Experiment 3 shows us that accumulating more responses does not necessarily improve data quality, as it depends on the conflict between users in the labeling process, and how many users have labeled the comments. This is evident by the fact that 500 response data has similar evaluation scores to that of 4.8k response.

Another important remark, revealed by the outcomes of experiment 4, is the model’s incapability, with the data provided, to distinguish between different types of hate speech, with poor precision scores contrasted by the high accuracy achieved by evaluated models, even when simplifying the problem to binary cases.

With the current data amassed through the website, the pretrained models used (BERT and ARABERT) did not provide a higher performance advantage over simpler ML models, as shown in experiment 5. Though a substantial improvement can be made when applying frequency patterns and BOW collected through the website, evidenced by conducting experiment 6, drastically improving the precision for the model through reclassifying mispredicted examples.

The data contains misleading patterns that may be linked back to the annotation process, or the general nature of colloquial Arabic. Some of the comments might be mislabeled based on the analysis for some of the comments. The model can competently recognize hate from non-hate, but it can’t differentiate between its classes. No improvement throughout various model results (ML or DL models), due to the low quantity and quality of the data, this could be solved in time as more labeling is done through the website, giving more representative online user answers.

# References

1. ARTIFICIAL INTELLIGENCE. (n.d.). *NIST*. https://www.nist.gov/artificial-intelligence
2. Artificial Intelligence: What it is and why it matters. (n.d.). *SAS*. https://www.sas.com/en\_us/insights/analytics/what-is-artificial-intelligence.html
3. Ayodele, T. O. (2010). Types of Machine Learning Algorithms. In Y. Zhang (Ed.), *New Advances in Machine Learning*. IntechOpen. https://doi.org/10.5772/9385
4. Bahador, B. (2020). Classifying and Identifying the Intensity of Hate Speech. *Items Insightrs from the Social Sciences*. https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/classifying-and-identifying-the-intensity-of-hate-speech/#:~:text=Hate%20speech%2C%20of%20course%2C%20is,demonization%2C%20and%20incitement%20to%20violence
5. Bellotti, F., Berta, R., & De Gloria, A. (n.d.). Designing Effective Serious Games: Opportunities and Challenges for Research. *Research Gate*. https://www.researchgate.net/publication/49588460\_Designing\_Effective\_Serious\_Games\_Opportunities\_and\_Challenges\_for\_Research
6. Dickinson, P., Francesco, B., Graham, W., & Agee, M. (n.d.). *The Phenomenon of “Mocking.”* https://www.nzac.org.nz/assets/Uploads/Journals/1-v2.-The-Phenomenon-of-Mocking.pdf
7. Equality and hate speech. (n.d.). *Article19*. https://www.article19.org/issue/equality-and-hate-speech/
8. Fino, A. (2020). Defining Hate Speech: A Seemingly Elusive Task. *Journal of International Criminal Justice*, *18*(1), 31–57. https://doi.org/10.1093/jicj/mqaa023
9. Frankenfield, jake. (n.d.). Artificial Intelligence (AI). *Investopedia*. https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp
10. Importance Of Internet Technology For Easy Life. (n.d.). *Asianet FIBER*. https://asianetbroadband.in/importance-of-internet-technology-for-easy-life/
11. Jordan, M., & Mitchell, T. (n.d.). *Machine learning: Trends, perspectives, and prospects*. https://www.cs.cmu.edu/~tom/pubs/Science-ML-2015.pdf
12. Maria, G. (n.d.). *Gamification for Machine Learning: The Classification Game*. http://ceur-ws.org/Vol-1642/paper7.pdf
13. Melger, E. (https://www.deltares.nl/en/software-solutions/deltares-serious-game-portal/). Serious Gaming. *Deltares*.
14. Mullah, N. (2021). Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review. *IEEE Access*. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9455353
15. Pilote, B., & Chiniara, G. (2019). Chapter 2—The Many Faces of Simulation. In G. Chiniara (Ed.), *Clinical Simulation (Second Edition)* (Second Edition, pp. 17–32). Academic Press. https://doi.org/10.1016/B978-0-12-815657-5.00002-4
16. Pitsilis, G. (2018). *Effective hate-speech detection in Twitter data using recurrent neural networks*. https://folk.idi.ntnu.no/heri/papers/pitsilis\_APIN2018.pdf
17. Smedley, A. (2021). Racism. *Encyclopedia Britannica*. https://www.britannica.com/topic/racism
18. Teaching about Controversial Issues and Use of Controversial Materials. (1997). *IMB*. https://www.flemingschools.org/cms/lib/CO02201757/Centricity/Domain/11/IMB%20Teaching%20about%20Controversial%20Issues%20and%20Use%20of%20Contraversial%20Materials.pdf
19. The route to a successful serious game. (n.d.). *GRENDEL GAMES*. https://grendelgames.com/method/
20. Tsionas, A. (n.d.). *Serious Game Development for the Diagnosis of Major Depressive Disorder Cases Using Machine Learning Methods*.
21. *What is hate speech?* (n.d.). https://challengehate.com/what-is-hate-speech/
22. What is sexual harassment? (n.d.). *REACH OUT*. https://au.reachout.com/articles/what-is-sexual-harassment