SHAHZEB KHAN
01-134172-056
MUHAMMAD ABDULLAH
01-134172-030

Harassment Comments Filtering on Social Media

Bachelor of Science in Computer Science

Supervisor: Burhan Ud din Abbasi

Department of Computer Science
Bahria University, Islamabad

June 2021

Certificate

We accept the work contained in the report titled "HARASSMENT COMMENTS FILTRING ON SOCIAL MEDIA", written by Mr. SHAHZEB KHAN AND Mr. MUHAMMAD ABDULLAH as a confirmation to the required standard for the partial fulfillment of the degree of Bachelor of Science in Computer Science.

Approved by . . . :

Supervisor: Burhan Ud din Abbasi

_____

Internal Examiner: Name of the Internal Examiner (Title)

_____

External Examiner: Name of the External Examiner (Title)

_____

Project Coordinator: Name of the Project Coordinator (Title)

_____

Head of the Department: Name of the HOD (Title)

_____

June 23^{rd}, 2021

Abstract

Harassment is aggressive behaviour or intimidation. It is a form of discrimination that includes unwanted physical or verbal behaviour to humiliate or offend someone. Humans tend to compete with others in a foul way due to several factors. Jealousy may lead to actions that harm other people; harassment is one such action. Harassment can be physical or verbal; in verbal, there comes another form where the harasser uses the platform and that is social media. With the ongoing advanced world, social media has embedded in our daily life. It has become a frequent rostrum for harassers where they comment on different posts to hurt others. Then comes the effect of harassment comments on social media over its users, though there are many bad effects some of them are quitting social media, quitting career when the victim is some celebrity. In the case of women and underage groups, it can have a bad impact during their learning phase of life. The tool that we are going to develop will make social media users get rid of those harass comments it will provide them harassment-free environment. We have made two features in our tool; one is for those who are over 18 years of age and they will have an option where they can see those harass comments, and those who are below 18 their comments will be filtered out automatically and they would not be able to see them until our extension tool Is attached to their browser. We hope that this project will be able to fulfil the solution to the problem that has been mentioned.

Acknowledgments

In the name of Allah, the Most Gracious and the Most Merciful. Alhamdulillah, all commendations to Allah for the qualities and His approval in finishing this undertaking. We might want to offer our most profound thanks and are appreciative to our administrator Sir Burhan Ud Din Abbasi for allowing us to chip away at this rising innovation. We are humbly thankful to our supervisor Burhan Ud Din Abbasi who made his efforts with us in making this application throughout the final year project. He puts his additional knowledge and efforts for our help and was always there for our guidance.

SHAHZEB KHAN, MUHAMMAD ABDULLAH
Islamabad, Pakistan

June 2021

"We think someone else, someone smarter than us,
someone more capable, someone with more resources will solve that problem.
But there isn't anyone else."

Regina Dugan

Contents

1 Introduction **1**
- 1.1 Project Background . . . 1
- 1.2 Problem Description . . . 1
- 1.3 Problem Objective . . . 2
- 1.4 Methodology . . . 2
- 1.5 Problem Scope . . . 2
  - 1.5.1 Inclusions . . . 3
  - 1.5.2 Exclusions . . . 3
- 1.6 Application Area . . . 3

2 Literature Review **5**
- 2.1 Online Harassment . . . 5
- 2.2 Natural Language Processing . . . 5
- 2.3 Semantic Analysis . . . 6
  - 2.3.1 Lexicon Based . . . 6
  - 2.3.2 Machine Learning Based . . . 6
  - 2.3.3 Hybrid Based Approach . . . 7
- 2.4 Extension Development . . . 7
  - 2.4.1 Why extension Development is needed? . . . 7

3 Requirement Specifications **9**
- 3.1 Existing System . . . 9
  - 3.1.1 Comment Moderation Tool . . . 9
- 3.2 Proposed System . . . 10
- 3.3 Functional Requirements . . . 10
  - 3.3.1 Comments Scrapping . . . 10
  - 3.3.2 Semantic Analysis . . . 10
  - 3.3.3 Comments Filtration . . . 10
- 3.4 Non-Functional Requirements . . . 10
  - 3.4.1 Usability . . . 10
  - 3.4.2 Performance . . . 11
  - 3.4.3 Security . . . 11
  - 3.4.4 Extension Policy . . . 11
- 3.5 Use Case . . . 11

4 Design **13**
4.1 System Architecture . . . 13
4.2 Design Constraints . . . 14
4.3 Design Methodology . . . 14
4.4 High Level Design . . . 15
4.5 GUI Design . . . 16

5 System Implementation **19**
5.1 Languages Used . . . 19
5.1.1 Python . . . 19
5.1.2 JavaScript . . . 19
5.1.3 JSON . . . 20
5.1.4 DOM . . . 20
5.2 Development Environment . . . 20
5.2.1 PyCharm . . . 20
5.2.2 Visual Studio . . . 20
5.3 Technology and Algorithms . . . 21
5.3.1 Artificial Intelligence . . . 21
5.3.2 Natural Language Processing . . . 21
5.3.3 Logistic Regression Model . . . 21
5.4 Server . . . 22

6 System Testing and Evaluation **23**
6.1 Performance Testing . . . 23
6.2 Usability Testing . . . 23
6.3 Usablity Table . . . 28
6.4 Regression Model Testing . . . 28

7 Conclusions **29**

References **31**

List of Figures

2.1 Broad Classification of NLP 6

3.1 Use Case Diagram 11

4.1 Context Diagram 13
4.2 Data Flow Diagram Level 1 14
4.3 UML Class Diagram 15
4.4 Component Diagram 15
4.5 Sequence Diagram 16
4.6 Code Organization 16
4.7 Instagram Post 17
4.8 Message After Extension Turned On 17
4.9 Hidden Harassing Comments 18

5.1 Venn Diagram of Technology 22

6.1 Usability Testing Fig 1 24
6.2 Usability Testing Fig 2 24
6.3 Usability Testing Fig 3 25
6.4 Usability Testing Fig 4 25
6.5 Usability Testing Fig 5 26
6.6 Usability Testing Fig 6 26
6.7 Usability Testing Fig 7 27
6.8 Usability Testing Fig 8 27
6.9 Regression Model Testing 28

List of Tables

3.1 Difference between Existing and Proposed system. 9

5.1 JSON data types and values. 20

6.1 Performance Table 23
6.2 Usability Accuracy Check 28

Acronyms and Abbreviations

GUI: Graphical User Interface
DFD: Data Flow Diagram
NLP: Natural Language Processing

Chapter 1

Introduction

1.1 Project Background

In 1997, the first-ever social media site to be built was Six Degrees [1]. The purpose of this site was to promote communication and stay connected to friends, acquaintances, and family members online. Later, other social media apps were introduced, the number of users was increasing day by day. The reason for this popularity was the amazing benefits like a person can connect to others from one part of the world to another, and these were provided by the social media platform. Man's creation is never absolute. Though social media had many benefits, it can also cause many troubles. The misuse of social media increased, people got addicted and started wasting their time, similarly, lack of face-to-face communication also started. Along with these issues, online sledging became involved.

Every day thousands of people face negative, abusive, and threatening comments posted on social media which results in quitting their social media accounts. Few victims have tried to kill themselves as well [2]. On the other hand, if we talk about underage children, these comments can put a negative impact on their lives. So, to protect a social media user from this kind of activity, we are going to develop a system that will filter out these harassing comments and provide a user-friendly social media platform.

1.2 Problem Description

Due to the misuse of social media platform, the issue of cyberbullying is increasing day by day. Social media harassment has affected users. Its effects vary from the type of reactions that are tended by the victims,

- One quits his or her social media account and isolate themselves.
- According to the report there are one million suicide attempts [2] made each year.

- The most targeted victims are females and celebrities
- Under-age group of people can have a bad impact on their learning phase of life by these comments.

1.3 Problem Objective

To develop a browser extension that will filter out the harassment comments. All those comments which are typed in English language on social media to avoid foul comments using Natural Language Processing and JavaScript.

Here are listed other main objectives of our project:

- To provide celebrities with harass free social media environment.
- To provide women an open, online sexual harassment-free platform.
- To provide social media a third-party approach and idea that may help them in their future releases.

1.4 Methodology

The users will install our extension on their browser. After installation, if they press the filter button of extension, foul/harassment comments will be filtered. The methodology we are using is that first, we will scrape the comments from the website URL then we apply NLP to filter out harass/bully comments that are in English and save it to a file. After that, JavaScript will read that file and remove that element of the website that is showing that comment. NOTE, we are only changing elements of the website that means it is only client-side. We are not hacking Facebook and deleting harassment comments from their databases.

For the data set, we will analyze commonly used harassed words. For English, we may get a dataset or built-in vocabulary.

1.5 Problem Scope

After the analysis and research on the victims, our target will be for the following group of users:

- Celebrities.
- Under age children.
- Females.

1.5.1 Inclusions

All those comments which are typed in the English language will be filtered out.

1.5.2 Exclusions

We preferred to make an extension rather than a desktop or web application because an app can tend a user to login forcefully through that environment. There are a lot of users that log in from their mobile phones and tablets but this project is related to those users who manage their pages through a desktop and it is obvious that master accounts and pages are handled through the desktop.

1.6 Application Area

Our targeted areas are “Instagram”, “YouTube” and "Facebook".All of them are commonly used social media platforms hence later on this system can be embedded into their sites and applications as a new tool.

Chapter 2

Literature Review

2.1 Online Harassment

Maeve Duggan a research director[3] stated in his research "Online Harassment 2017" that roughly 4 out of 10 Americans experienced online harassment. Offensive name-calling, physical threats, and sexual harassment are the most common ways. He also stated that 27 percent of people were harassed using offensive name-calling, intentional embarrassment victims were 22 percent, physical threats were given to 10 percent and 6 percent people were harassed sexually.

All these factors imposed an impact on the victims. 23 percent of them stopped posting on social media 13 percent stopped using social media whereas 30 percent stood up and took part in multiple organizations to take action on harassment. [3]

2.2 Natural Language Processing

NLP is a subfield of Artificial Intelligence in which we implement multiple solutions to the issues related to natural language, text, or speech. The ultimate goal of NLP is to understand Natural Language and to process the following tasks:

- Interpret an input text
- Translate from one language to another
- To provide answers to the given questions in the text
- Make inferences from the text

The field of NLP was originally referred to as Computational Linguistics; Computer Science - is concerned with developing internal representations of data and efficient

processing of these structures. It has several applications that include machine translation, NLP text processing, user interface, artificial intelligence, and many more. [4]

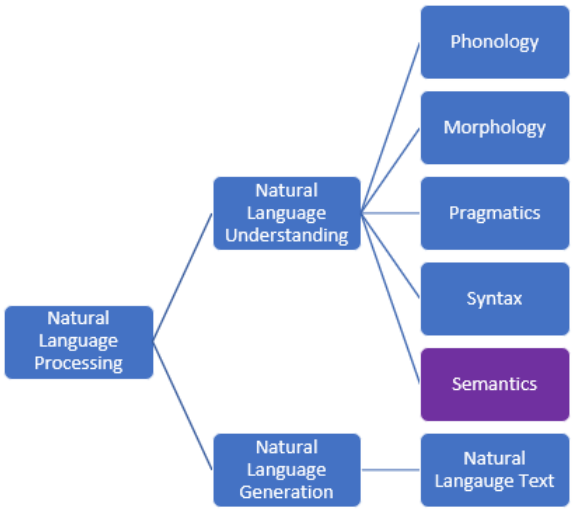


Figure 2.1: Broad Classification of NLP

The NLP technique that we are going to use in our project is Semantic analysis.

2.3 Semantic Analysis

Semantic analysis is a procedure that detects whether the data provided is positive or negative. Data provided is in the form of text only. Approaches for sentiment analysis:

2.3.1 Lexicon Based

- Dictionary-based: Dictionary is made and words are matched in the dictionary but words are collected manually.
- Corpus-based: Here a large data set is required and words are context-specific

2.3.2 Machine Learning Based

- Supervised: Relies on labeled training models. Classification and Regression lies in this category.
- Unsupervised: This technique does not use previously listed data to train the classifier. Clustering lies in this category.

2.3.3 Hybrid Based Approach

- It is a combination of both Lexicon-based and Machine Learning-based approaches its advantage is that it can achieve the best of both. [5]

Our project will also extract the textual data from the comments and semantic analysis will be applied to that data. Similarly, we will extract data by using JavaScript which will take data from the complete element.

2.4 Extension Development

The extension is a small program that customizes the browsing experience. It is used to add additional functions and features to web browsers.

2.4.1 Why extension Development is needed?

- To modify the user interface.
- To block adds on web pages.
- To translate text from one language to another.
- To extract data for legal processing.

JavaScript is the well-known first-class scripting language for web pages. It is one of the most important technologies of the World Wide Web. It provides interactive elements and complex features to web pages that engage a user. It acts as the main building block of extension development. We will use JavaScript to develop an extension first which will contain NLP code that will do further functions. JavaScript will also be used to extract data from the elements [6].

Chapter 3

Requirement Specifications

3.1 Existing System

There is no such official system so far developed regarding this project. Instagram has proposed a system that is restricted to business accounts only. It has limitations yet it is favorable and discussed in the below subsection.

3.1.1 Comment Moderation Tool

Few of the social media apps have started working on the filtration of the harassed and reported comments. Instagram is a popular social media platform and they have introduced an option to their business accounts users, which hides unusual comments. Similarly, all those comments and emojis which are manually set by the user as inappropriate, are also filtered out. [7]

Though this update holds value in itself, yet many other bugs need to be solved to convert this static tool into a proper dynamic way.

Existing System	**Proposed System**
The existing system was stuck to business and major accounts only	On the other hand, our target users will be of any category whether they are business accounts, under age group or females.
The existing tool doesn't work on the current comments, it takes pre-defined words and emojis and just deletes them	Our system will use the NLP algorithm, where the semantic analysis will decide whether to remove the comment or not.

Table 3.1: Difference between Existing and Proposed system.

3.2 Proposed System

The purpose of our system is ultimately the same as Instagram's comment moderation tool but our system will be an enhanced version of it. Few points will distinguish our work from existing. They are stated below.

3.3 Functional Requirements

3.3.1 Comments Scrapping

First of all, our system will extract all the comments from the user's profile. Data collection is important because all other processing will be done on this data. Comments will be extracted from the div element.

3.3.2 Semantic Analysis

After that in the back-end, our NLP module will apply semantic analysis on the comments where harass comments will be detected and those comments will be passed to the front system.

3.3.3 Comments Filtration

This last part will remove those div elements which contain harassment. As we have two main users, hence we also have added functionality that will be different for different users. It is explained below:

- Business Accounts (Celebrities): They will have an option that will allow them to see who commented on their profile.
- Under Age Users: They will not have any kind of option to see harass comments.

3.4 Non-Functional Requirements

Following are the non-functional requirements of our application

3.4.1 Usability

The interface of our system is very simple. There is an icon attached to the extension. Whenever the user clicks that icon an Alert pops out. After that system starts working and harassed comments are filtered out.

3.4.2 Performance

Performance is a measure based on the hardware being used. Performance is one of the main risks that we will be facing because we have to scrap all the comments, then those comments will be saved in a file that will match the harassed words with the dictionary that we have provided, and then those comments will be filtered. It may take 40- 50 seconds to filter them out but we are still researching to move that time to 10 seconds at least.

After the final testing, there was a great turnaround related to the performance of our software. Initial loading time was reduced to 10 seconds. After that overall time taken was only 1-2 seconds to filter out all the comments.

3.4.3 Security

Privacy of the users will be kept safe all the work will be done on the client-side and the user's initials will not be disturbed. Each user will have his/her unique id and password. This will make the system secure.

3.4.4 Extension Policy

Browser extension policies will be kept in mind and extensions will be developed under the complete enterprise policy.

3.5 Use Case

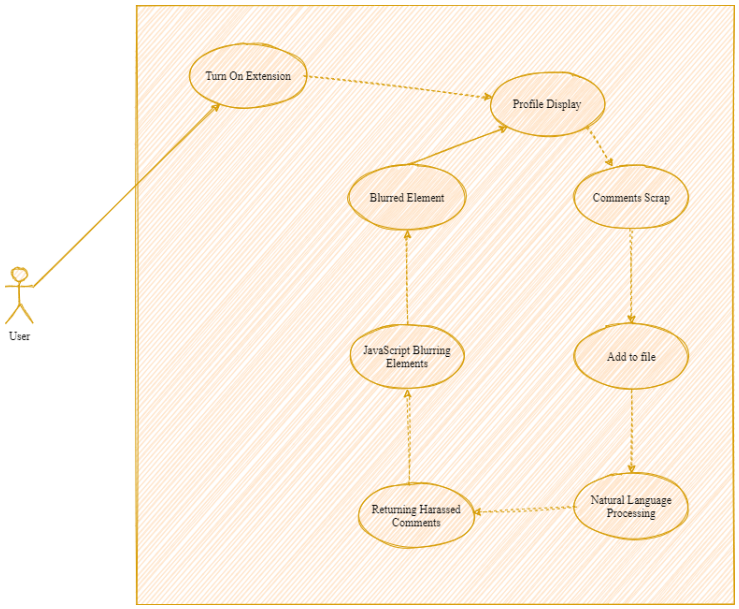


Figure 3.1: Use Case Diagram

Chapter 4

Design

Systems design is the process of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements.

4.1 System Architecture

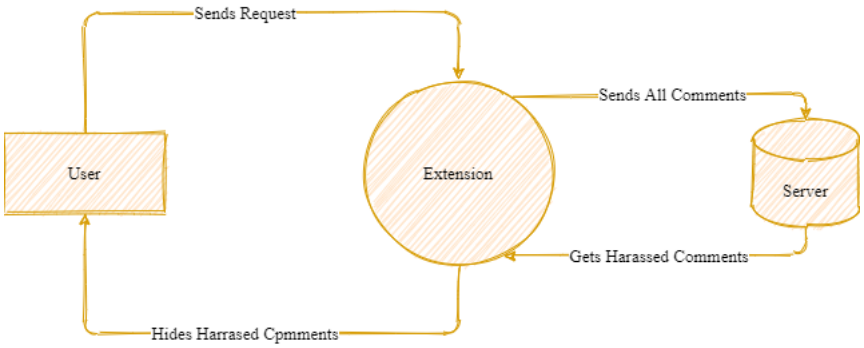


Figure 4.1: Context Diagram

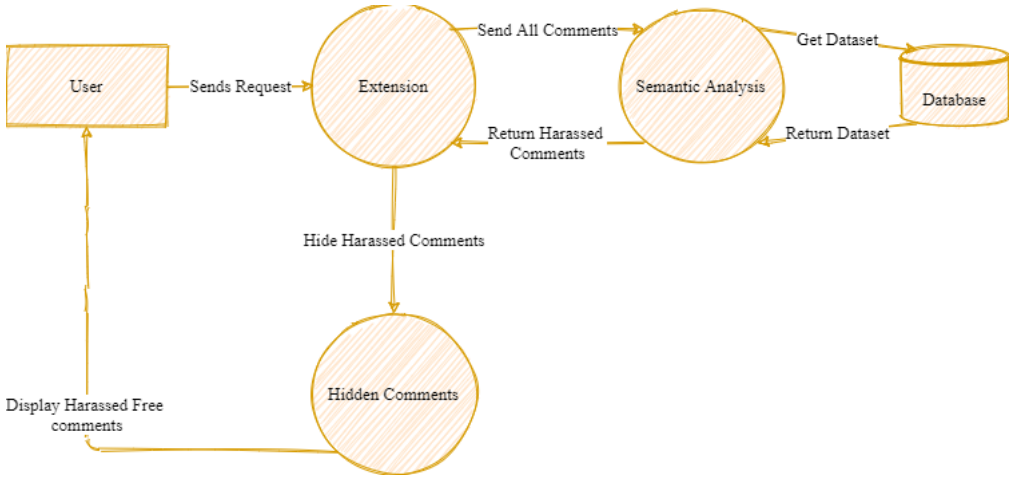


Figure 4.2: Data Flow Diagram Level 1

4.2 Design Constraints

In our design, we have avoided complex GUI. We have provided the user with a simple icon which on click starts working. This is because of the following reasons:

- Most of the functions don't have any kind of GUI. This is because extensions are made to assist web applications and web applications already have GUI. So having GUI over GUI is not recommended.
- It can put an unattractive impression on a user because most of the users want a simple and interactive GUI.

Similarly, a user must have opened any post or picture containing comments. Without that our system will not work.

4.3 Design Methodology

Complete object-oriented methodology for our project has been used. Initially, the extension gets 3 values in the form of a JSON file.

- id as a number.
- text (comment) as a string.
- prediction as a string.

After that those values are sent to a python file as an array where it converts this array into a python dictionary.

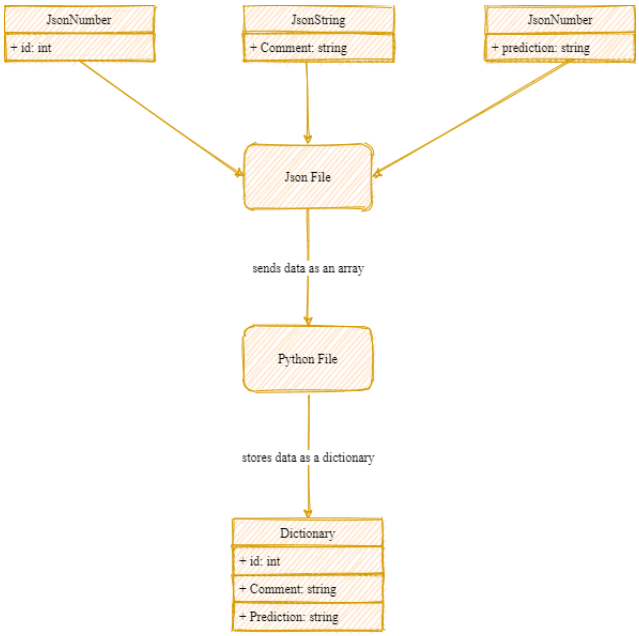


Figure 4.3: UML Class Diagram

4.4 High Level Design

This section describes in further detail elements discussed in the Architecture. High-level designs are most effective if they attempt to model groups of system elements from several different views. Typical viewpoints are:

1. Conceptual or Logical:

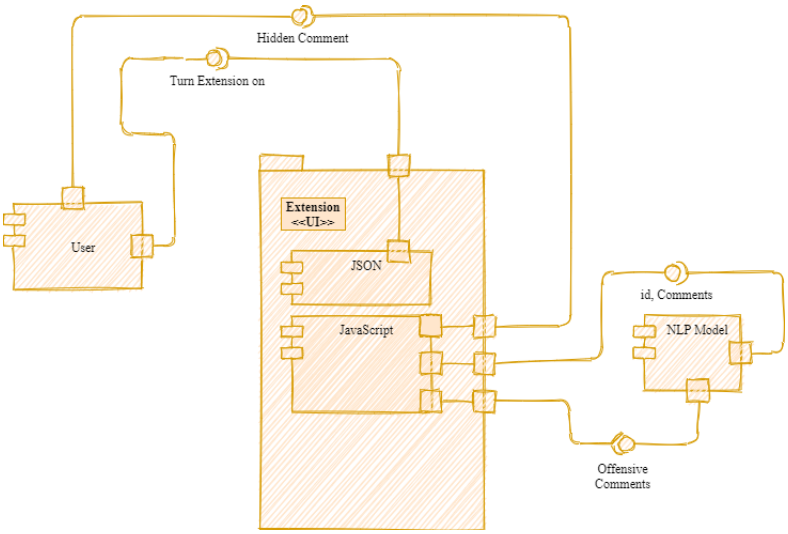


Figure 4.4: Component Diagram

2. Process:

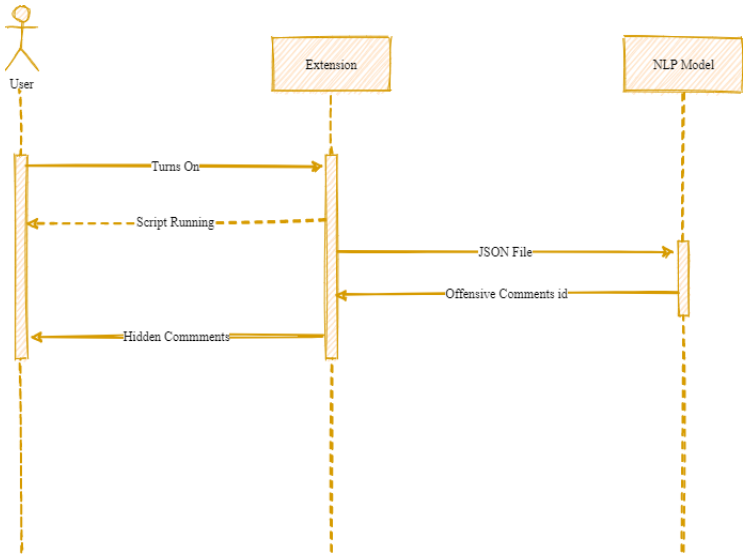


Figure 4.5: Sequence Diagram

3. Module:

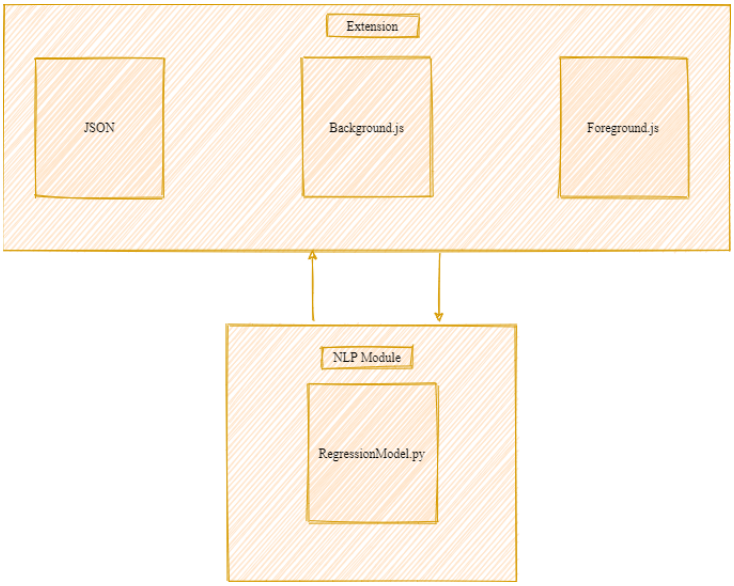


Figure 4.6: Code Organization

4.5 GUI Design

Figure 4.7: Instagram Post

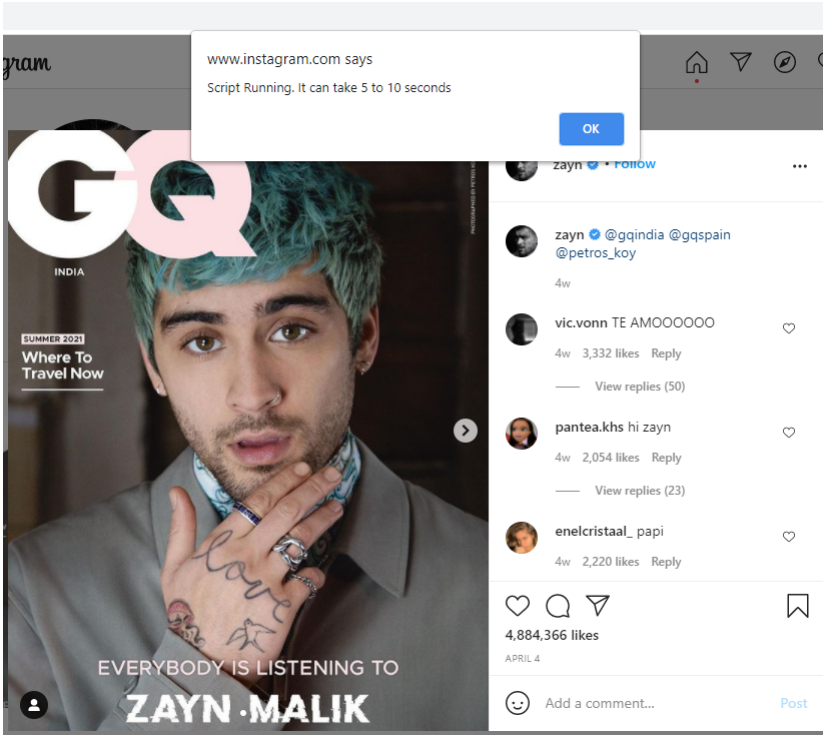


Figure 4.8: Message After Extension Turned On

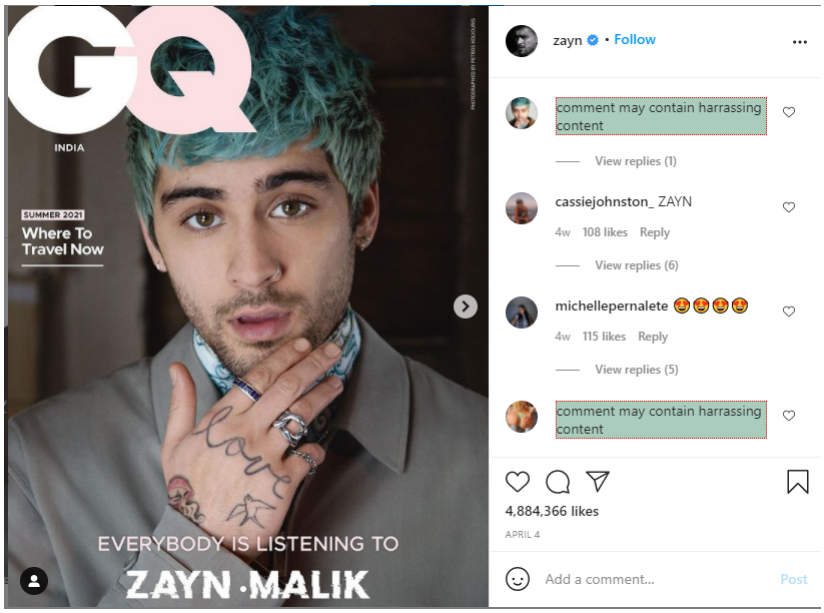


Figure 4.9: Hidden Harassing Comments

Chapter 5

System Implementation

In this chapter, we will discuss the implementation of our proposed system. From tools and technologies to, development environment to algorithms and languages that are used will be discussed here.

5.1 Languages Used

There are two languages that we have used in our system: Python and JavaScript.

5.1.1 Python

It is a multi-purpose programming language used for almost every kind of development. In recent times Python has got its name over the development of AI and Machine learning. One can feasibly deploy their algorithms and models using Python.

As our project is more related to Artificial Intelligence hence we have used Python as our main language to train and test our model. The text (comments in our case) that is extracted is used by our model is matched with the dataset and sent back to our front system is all through Python language.

5.1.2 JavaScript

For our second part of the project that is extension development, we have used JavaScipt. JavaScript is more related to web development, Hence for the extracting of the comments from the webpage JaveScript has played its role.

Many other languages can be used to develop an extension but one of the key reasons to use JS is it can easily asses website data.

5.1.3 JSON

JSON stands for JavaScript Object Notation that is used for storing and transporting the data. Data can be in the form of strings, numbers, objects, arrays, Boolean and null. In our case, we are getting an id as a number and comment plus prediction as a string. So basically, it looks like accordingly given in the table below.

Name	Data Type	Example Data
Id	Number	0,1,2,3..
Text (Comment)	String	"You are a looser."
Prediction	String	Initially Null. After that "Offensive or Unoffensive"

Table 5.1: JSON data types and values.

5.1.4 DOM

DOM stands for Document Object Model.It acts as an application programming interface (API) for HTML and XML. DOM is the very initial component of our project. A webpage that we are using for comments filtration is converted into a DOM file. After that comments are sent into extension in the form of a JSON file.

5.2 Development Environment

There are plenty of IDE's present online and free. The platform is very important for any kind of development and one should always choose the Development Environment which has top ratings. This is because good rated IDE has a ratio of executing code more accurately.

5.2.1 PyCharm

It is the best IDE that is used to develop python applications. We have also used this platform as its community version is free and reliable.

5.2.2 Visual Studio

Visual Studio is used as a development environment for JavaScript. Hence our extension and all of its complete development are done over the visual studio. Once again the reason for choosing Microsoft Visual Studio is due to its overall ratings, as it is the best hybrid platform used to edit and code many languages.

5.3 Technology and Algorithms

This part outlines the base of our project. Technology stands for skills and methods we have used to achieve the goals. In our case, the core technology is Artificial Intelligence.

5.3.1 Artificial Intelligence

AI is one of the emerging branches of Computer Science. Its main concept is to make the computer act like human beings. Furthermore AI has other branches depending upon the type of experiments and algorithms that are being processed. The one that we are using is Natural Language Processing.

5.3.2 Natural Language Processing

NLP is a subfield of Artificial Intelligence in which we implement multiple solutions to the issues related to natural language, text, or speech. The ultimate goal of NLP is to understand Natural Language and to process the following tasks:

- Interpret an input text
- Translate from one language to another
- To provide answers to the given questions in the text
- Make inferences from the text

The field of NLP was originally referred to as Computational Linguistics; Computer Science - is concerned with developing internal representations of data and efficient processing of these structures. It has several applications that include machine translation, NLP text processing, user interface, artificial intelligence, and many more. [4]

5.3.3 Logistic Regression Model

It is the statistical model used to predict the probability of a certain class or event. In our system, this model predicts the probability of harassment where it checks whether the comment is “offensive” or “unoffensive”. The main library over which we have developed our model is **Scikit learn**

The overall venn diagram of our project is also shown below. It shows from a higher level to a lower level the techniques we have used.

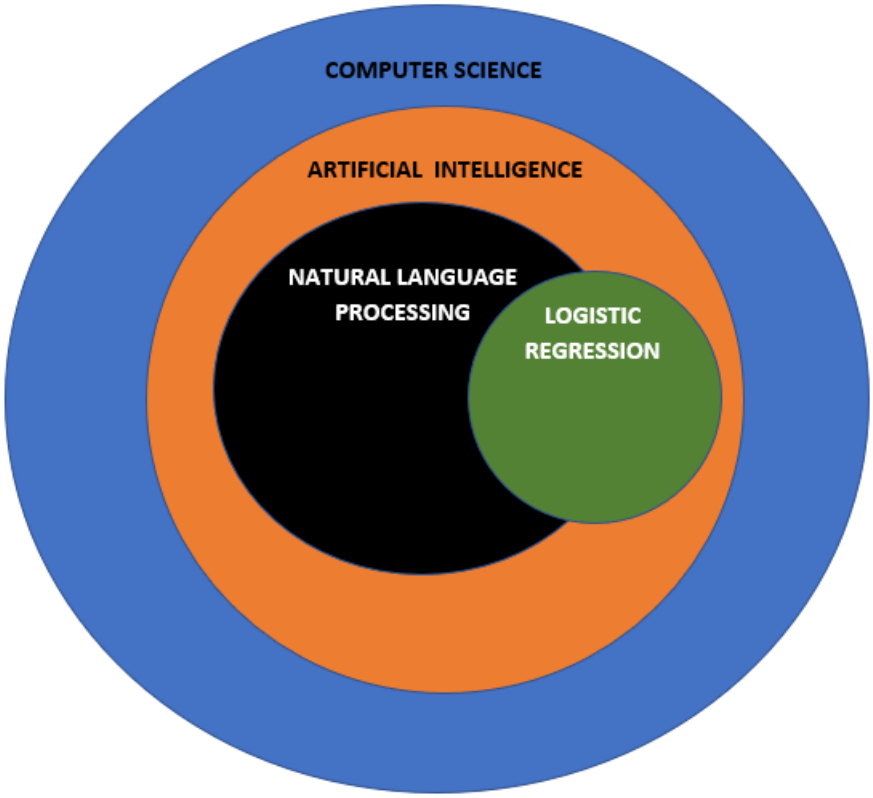


Figure 5.1: Venn Diagram of Technology

5.4 Server

Our server is present on Flask. Flask is a python class datatype and web framework. For the deployment of the site, it provides servers.

Flask has a route system. Its basic concept is actually to execute the desired function whenever a user visits that web page embedded in our code. For Example in our case whenever a user clicks on the extension icon and if he is on Instagram then all the comments will be extracted from that Instagram page and the server will be joined using the routing system.

We have deployed our server on heruko.com. It is an open-source platform that helps developers to build and run their applications.

Chapter 6

System Testing and Evaluation

In this chapter, we are going to include multiple testings we have conducted and their results and outcomes.

6.1 Performance Testing

At First when the user clicks the extension icon for the very first time. It takes around 11 seconds to filter out harassed comments.

After that system works itself (there is no need to click the extension button again). It stores previous values and consider them as “offensive comments”. Performance table is given below.

Iteration	**Event Fire**	**Time Taken**
Initial	On pressing Extension Icon	10.73 Seconds
Final	On Scrolling the Comments	1.95 Seconds

Table 6.1: Performance Table

6.2 Usability Testing

For usability testing, we created a post and commented few harassed comments on it. Total 36 comments were posted out of which 20 were containing harassing content. All of those comments are shown in [Figure 6.1, 6.2, 6.3, 6.4].

After that, our system detected 17 out of those 20 comments as “harassing content”. They can be shown in [Figure 6.5, 6.6, 6.7, 6.8].

Similarly [Table 6.2] shows the overall accuracy percentage of usability testing.

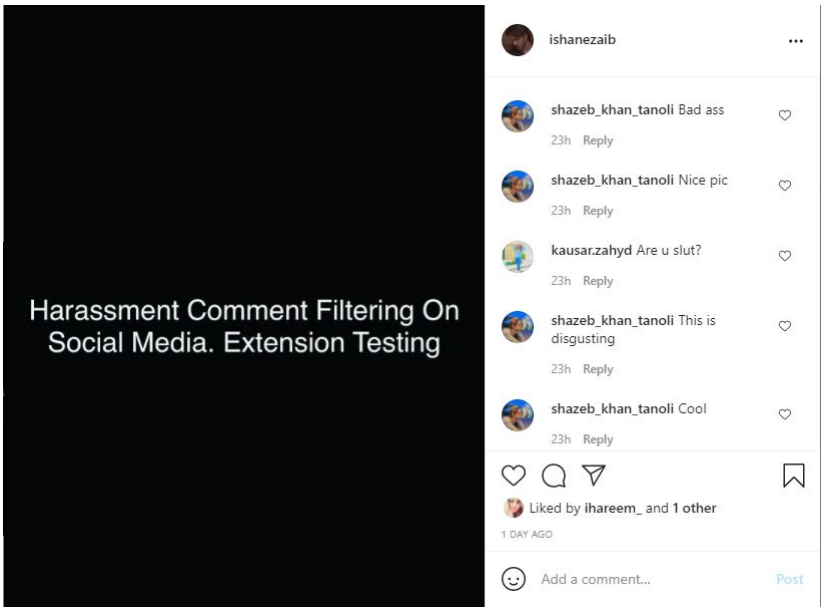


Figure 6.1: Usability Testing Fig 1

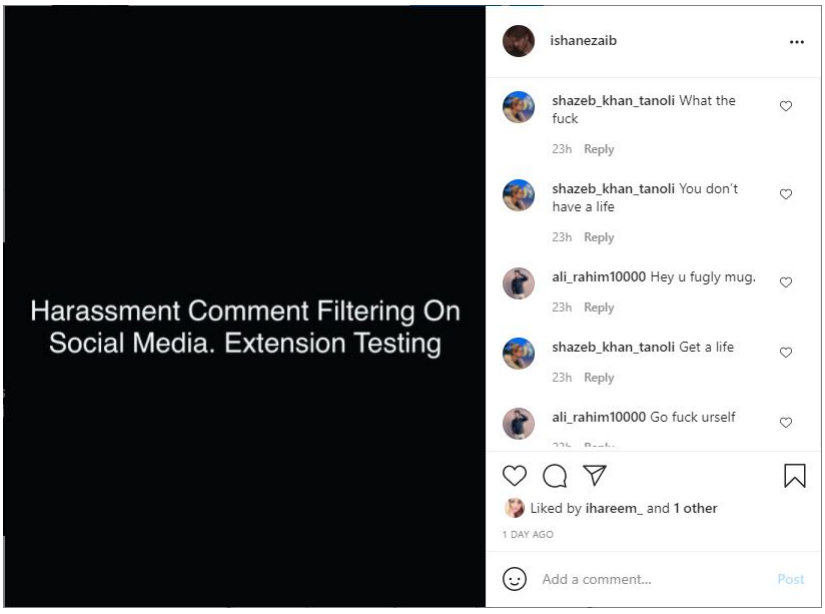


Figure 6.2: Usability Testing Fig 2

Figure 6.3: Usability Testing Fig 3

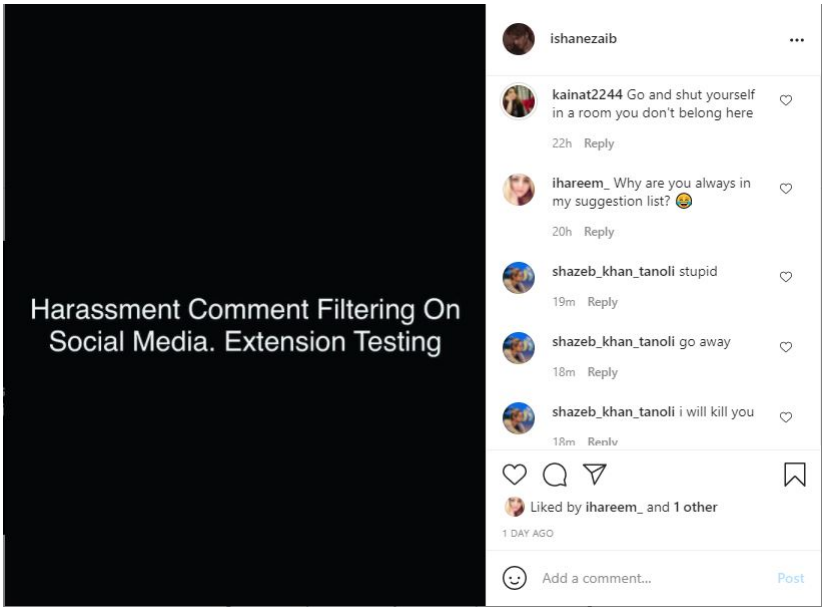


Figure 6.4: Usability Testing Fig 4

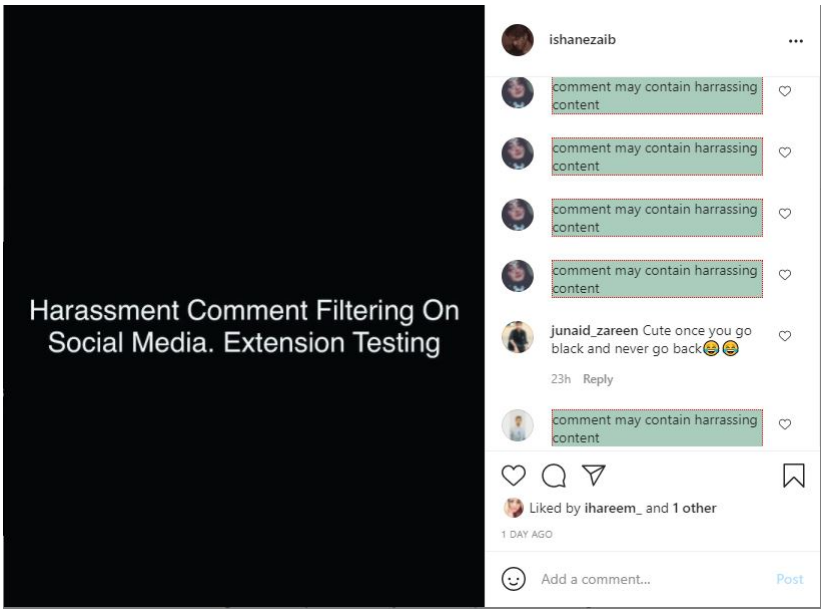


Figure 6.5: Usability Testing Fig 5

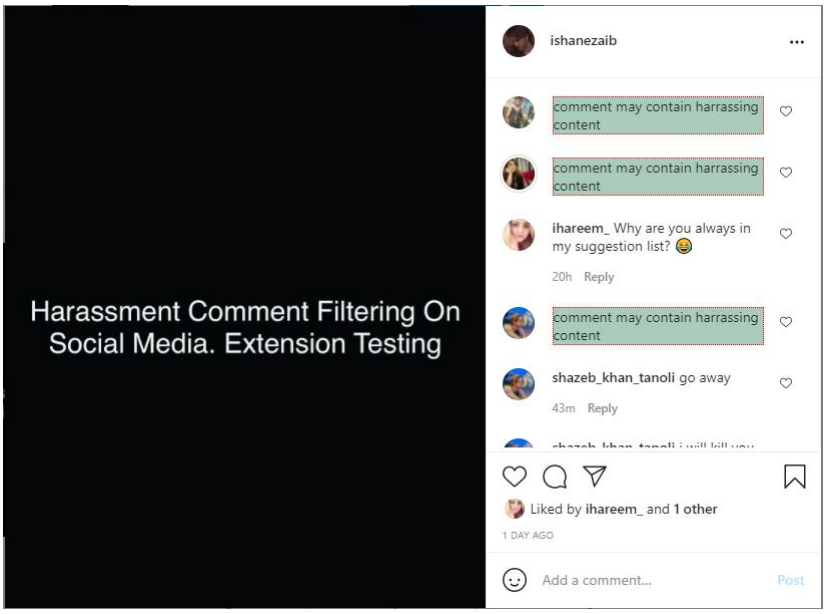


Figure 6.6: Usability Testing Fig 6

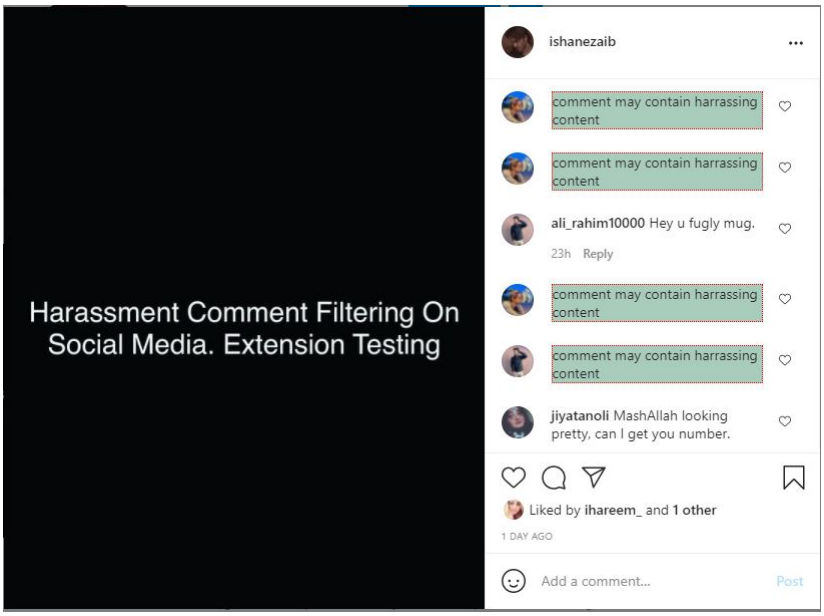


Figure 6.7: Usability Testing Fig 7

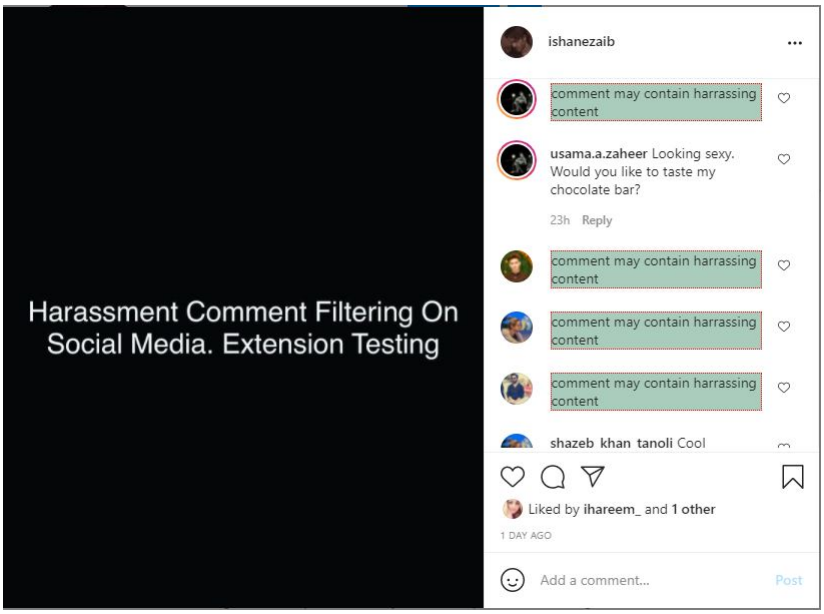


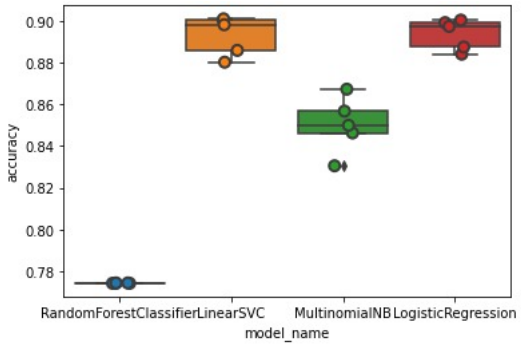
Figure 6.8: Usability Testing Fig 8

6.3 Usablity Table

Total Comments	Harassed Comments	Comments Detected	Accuracy Percentage
36	20	17	85

Table 6.2: Usability Accuracy Check

6.4 Regression Model Testing




```
[ ]  cv_df.groupby('model_name').accuracy.mean()
```

```
model_name
LinearSVC                 0.893072
LogisticRegression        0.893637
MultinomialNB             0.850140
RandomForestClassifier    0.774321
Name: accuracy, dtype: float64
```

Figure 6.9: Regression Model Testing

Chapter 7

Conclusions

Technology has evolved from time to time. Different ways of technology were adapted. Every time a new process was proposed. The aim was kept the same to enhance the work and to reduce the time. Machine learning is one of the main types of evolving technologies. Image recognition, speech recognition, medical diagnosis, classification, prediction all are real-life examples of machine learning.

Natural Language Processing which was the main core of our project is mostly related to linguistics. Its real-life examples are email filtration, predictive text, language translation, digital phone calls, etc. In our system, we used this technique to provide a harass-free environment. Though it was a new thing for us, we did our utmost and got successful. We learned how to play will real-time text using NLP modules and algorithms and how a meaningful full sentence can be accessed and analyzed.

Our second main part was extension development. This tool helps web browsers by making them work more effectively. Its development helped us to gain knowledge about the processing of one programming language with another.

In the future, this project will surely help and assist those who want to work and research over NLP and text prediction.

References

[1] DREW HENDRIKS. Complete history of social media: Then and now. 2013. Cited on p. 1.

[2] Kalhan Rosenblatt. Cyberbullying tragedy: New jersey family to sue after 12-year-old daughter's suicide. 2017. Cited on p. 1.

[3] MAEVE DUGGAN. Online harassment 2017. 2017. Cited on p. 5.

[4] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: State of the art, current trends and challenges. 08 2017. Cited on pp. 6 and 21.

[5] Anvar Shathik and Krishna Prasad Karani. A literature review on application of sentiment analysis using machine learning techniques. pages 2581–7000, 08 2020. Cited on p. 7.

[6] Thomas Peham. How to develop a chrome extension in 2018. 2018. Cited on p. 7.

[7] Andrew Huchinson. 5 new features across twitter, facebook and instagram that you need to know about. 2016. Cited on p. 9.