Plagiarism Checker X Originality Report

Similarity Found: 8%

Date: Monday, July 25, 2022
Statistics: 500 words Plagiarized / 6614 Total words
Remarks: Low Plagiarism Detected - Your Document needs Optional Improvement.

DESIGN AND IMPLEMENTATION OF A SYSTEM THAT DETECTS RACIST WORDS

BY

INYANG DIVINE IME
MATRIC NO: 17010301039

SUBMITTED TO

THE DEPARTMENT OF COMPUTER SCIENCE AND MATHEMATICS,
COLLEGE OF BASIC AND APPLIED SCIENCES,
MOUNTAIN TOP UNIVERSITY, IBAFO, NIGERIA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD DEGREE OF BACHELOR OF SCIENCE (B.SC.) IN COMPUTER SCIENCE

JULY 2021

DECLARATION

I hereby declare that this project was carried out by me and is a report of my own research work. It has not been presented in any previous application for a higher degree of this or any other University. All citations and sources of information are clearly acknowledged by means of reference.

_____ _____

INYANG DIVINE IME Date

CERTIFICATION

This is to certify that this project, Design and Implementation of a system that detects racist words was carried out by me, Inyang Divine Ime (Matriculation Number: 17010301039) and duly supervised by Dr C. P. Igiri.

_____ _____

Dr C. P. Igiri Date

(Supervisor)

_____ _____

Dr. M. O. Adewole Date

(Ag. Coordinator of Department)

DEDICATION

I dedicate this project to God Almighty for giving me life, good health and all I needed to make this work a success and secondly to my dear parents, Barrister & Mrs I.A. Inyang for their guidance, understanding and sacrifice. I also dedicate this work to my course-mates and friends for their support in the course of my four years study of Computer Science in Mountain Top University. May the Almighty God bless you all! Amen.

ACKNOWLEGDEMENT

The success and final outcome of this project goes to the Almighty God for wisdom and understanding. I appreciate my Supervisor Dr C. P. Igiri who took keen interest in my project work and guided me all along, and never relented to attend to me anytime I came to her for assistance.I would like to acknowledge the Acting Coordinator of the Department of Computer Science and Mathematics, Dr. M. O. Adewole, and owe him my deepest gratitude for the efforts, constant encouragement, guidance and support of all the academic and non-academic staff of the Department of Computer Science and Mathematics. For the teachings that have brought out positive values in me and making my stay a worthwhile one. I extend my gratitude to Mountain Top University for setting greater heights for me. I say God bless you richly.I heartily would like to thank my parents and my guardians, Barr & Mrs Asuquo Inyang and Evang. & Mrs Emah Otono, thank you all for your moral and financial support. I am grateful for all the investments into my education and future. I would not forget to remember all the students in the Department of Computer Science and Mathematics, for making my stay a worthwhile one, I say God bless you all richly.

TABLE OF CONTENT

LIST OF FIGURES

ABSTRACT

In recent years, there has been a constant need to automate simple everyday operations and activities for improvement, efficiency and advancement in all areas including communication and ethics. This project covers all fields related to hate speech detection as well as its forms and kinds, it also includes the use of essential case tools for description of the system as well as necessary software development tools and techniques. The main purpose is to develop an efficient system for detecting hate speech, particularly racist words. This project works on hate speech detection using web software development technologies and techniques. It involves the use of standard modern means of identification and detection of racist word or slangs. All the functionalities are displayed on the interface for user interaction which can be integrated on social media platforms. Limitations include repetitive software grammar training and testing as well as a centralized database integration database.Future recommendations include artificial intelligence integration and data mining and a reporting protocol.

CHAPTER ONE

INTRODUCTION

1.1 Background to the study

In the 21st Century the problem of racism and racial discrimination has become more serious and has been receiving worldwide condemnation. As the world of Technology develops, racists have taken advantage of this and have used it to continue form of racist acts. Racists can use their social media accounts to create posts with racist contents, reply and comment to people's post with racist remarks/comments. This racist acts have become one of the negative effects of ICT (Information Communication Technology).

Racial discrimination is a common stressor in the lives of adolescents of color in the U.S. Previous empirical research suggests that the majority of minority youth perceive themselves to be the victims of racial-ethnic discrimination (Benner & Kim, 2009; Harris-Britt, Valrie, Kurtz-Costes, & Rowley, 2007; Huynh &Fuligni, 2010; Martin et al., 2011; Medvedeva, 2010; Neblett et al., 2008; Pachter, Szalacha, Bernstein, &Coll, 2010; Seaton, Caldwell, Sellers & Jackson, 2008).

Early writings on the topic of race online argued that the internet could reduce or eliminate racial discrimination that people of color typically experience in offline settings (Glaser & Kahn, 2005; Kang, 2000).Social media platforms have been creating solutions to combat this menace. Platforms like Facebook, Twitter, Instagram, etc. have created a way of reporting social media accounts for comments, replies and posts that are considered offensive to the reporter. After this report is made, the administrators find out or ascertain if the reported comments, replies or posts is offensive/abusive and then proceed to block, restrict or delete the reported social media

accounts. This method has been in place for a long time now and has been able to put a check to the activities of cyber-racists or bullies but to an extent.

1.2 Statement of the Problem

The freedom people have in expressing their ideas/beliefs online through social media has had positive impact but it has also led to misuse. Racists and cyber bullies now use this freedom as a tool for making racist contents in the form of texts (posts, comments, replies, etc.) or speech(voice-notes, videos, etc.). The presence of this racist contents online have been a pain in the neck to the internet community. This project is aimed at devising a system that can put a stop to the presence of this racist contents online in form of texts.

1.3 Aims and Objectives of the study

The aim of this project is designing and implementing a system that can detect racist content in form of text. It is designed to perform the following objectives:-

i. To detect racist words.

ii. To label racist words.

iii. To flag a sentence/text with racist content/words.

1.4 Scope of the study

The scope of the project covers the development of a system for use in the ICT world but will focus on Mountain Top University as a case study. It will involve testing of the system by

different individuals making sentences as inputs and the system will determine if the inputs contain racist words/content and give the result as output. It also covers writing the background programming to ensure that the interface works with the database through the underlying codes to perform the required actions. It also involves the improvement and optimization of the system.

1.5 Significance of the study

This research work is set to be of benefit to the ICT society and the world at large. With the rise in racist activities online, this work will have an impact in curbing the presence of this racist contents. The application of this system will provide a secure and reputable environment for social media users. Different social media platforms can make use of this system to improve the use of their platforms by making the system to not only detect racist words/content but also put a stop to the use of certain prohibited words/texts and contents.

1.6 Definition of Terms

Racism: is the belief that groups of humans possess different behavioural traits corresponding to physical appearance and can be divided based on the superiority of one race over another. It may also mean prejudice, discrimination, or antagonism directed against other people because they are of a different race or ethnicity.

Racial discrimination: is any discrimination against any individual on the basis of their skin colour, or racial or ethnic origin. Individuals can discriminate by refusing to do business with, socialize with, or share resources with people of a certain group.

Social media: is a computer-based technology that facilitates the sharing of ideas, thoughts, and information through the building of virtual networks and communities.

Hate Speech:abusive or threatening speech or writing that expresses prejudice against a particular group, especially on the basis of race, religion or sexual orientation.

CHAPTER TWO

LITERATURE REVIEW

2.1 Hate speech detection

This is an abusive or threatening language that can cause harm to a specific culture or group or that may be offensive to some people. Religion, race, sexual orientation, and caste are all examples of pre-bias. The term "cyber-hate" refers to online discussions of hatred and violence. (2018) (Zhang & Luo). Hate speech is defined as any type of language, writing, or behavior communication that attacks or refers to an individual or group in a derogatory or discriminatory manner. (Zhang & Luo, 2018) Identifying hate speech on social media is a critical task. The unchecked spread of hatred can have serious consequences for our society, particularly for those or groups who are marginalized. Social media is a significant online platform for the dissemination of hate speech. Because social communication messages contain a lot of miswritten text and paralyzing signals, automatic detection becomes more difficult (e.g., emoticons and hashtags). Another challenge is the task's context, as well as the lack of agreement on what constitutes hate speech, which makes the task even more difficult for people.

2.1.1 Automated approaches for detecting hate speech

Most social media platforms have user rules that prohibit hate speech, but in order for these rules to be enforced, extensive manual work is required to review all reports. Several platforms, including Facebook, have recently increased the number of content moderators. Automatic tools and approaches could speed up the review process or allocate human resources to positions that require close human examinations. This research provides an overview of automatic hate speech detection approaches based on text in this section. (S. MacAvaney, Yao; H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O;, 2019)

Methods based on key words

To identify hate speech, a keyword-based approach is used. An ontology or dictionary is used to identify text that contains potentially hateful keywords. Hate Base, for example, keeps a derogatory database of terms for various groups in 95 languages. Despite the fact that terminology evolves over time, such well-maintained resources are beneficial. As our study of hate speech definitions demonstrates, it is not always sufficient to simply use a hateful slur to constitute hate speech (MacAvaney et al., 2019)

Machine Learning Classification

Machine learning models generate a classifier to detect hate speech based on labels annotated by content reviewers with samples of labelled text. Various models have been proposed and have proven to be successful in the past. This study describes a variety of open-source systems that have been presented in recent research. (MacAvaney et al., 2019)

Source of Metadata

Additional social media data may help to better understand the characteristics of posts, potentially leading to a more effective approach to identification. Information such as the user's demo, location, timing, or even social participation on the platform can provide a more granular understanding of the position. (MacAvaney et al., 2019).

2.1.2 Hate speech kinds and forms

The definition of hate speech is not universally accepted, nor is it entirely agreed upon in its individual aspects. According to Ross and colleagues, a clear definition of hatred speech can aid in the study of hatred speech detection by making it easier to write down hate speech and thus increasing your reliability. The line between hatred and free expression, on the other hand, is blurry, making some hesitant to define hate speech precisely. The American Bar Association, for example, does not provide an official definition, instead claiming that speech that contributes to a criminal act may be punished as part of a hate crime. Similarly, this paper does not propose a specific definition but rather examines existing definitions to gain insights into what hate speech is and what technical challenges definitions can cause. This study synthesizes key definitions of hate speech from various sources, as well as some aspects of the definitions that make detecting hate speech difficult. Hate speech is defined as any direct attack on people because of their so-called "protected" race, ethnicity, national origin, religion, sex, gender identity, serious disease, or disability. Naturally, hate speech is a broader term that includes insults, discrimination, dehumanization, demonization, and incitement to violence.

There are three major types of hate speech.

Hate speech frequently emerges from the conceptual framework "us vs.," in which people distinguish between the "out-group," a group to which they believe they belong, and the "in-group." In this analysis, hate speech directed at out-groups is classified into three categories. The first is dehumanization and demonization of the group and its members, which is frequently associated with hate speech. (Bahador, 2017)

1. Dehumanization and diabetes

Dehumanization entails degrading populations and equating them to culturally despised entities such as pigs, rats, monks, or even dirt/silt. The Tutsi minority in Rwanda was the cockroaches leading up to and during the 1994 genocide, and they have recently become well-known for this. Dehumanization, if successfully transmitted, may have at least two political consequences. For starters, it has the power to unite members into an odious single entity, depriving them of their distinct individuality and humanity. .(Bahador, 2017)

4Demonization, on the other hand, includes portraying the group as superhuman or even fatal diseases, such as a monster, robot, or cancer, that pose a death threat to the group. The destruction of the adversary is presented in this way as not only acceptable, but also desirable and beneficial to the group's survival. (Bahador, 2017)

2. Intimidation and violence

While dehumanization and demonization are the most negative aspects of certain groups, they do not justify violence against them. However, another noteworthy type of hate speech involves the spread of violence. Violence against a specific group is a crime in many jurisdictions. (Bahador, 2017)

3. Prompt notification

For the purposes of this analysis, an early warning may be considered a third category of speech, which frequently borders on hate speech. Dehumanization or incitement are rarely the starting points of group-based hate speech, which is more subtle and measured. However, recognizing these early warning signs may be beneficial in preventing an escalation to more venomous words.

2.2 Communication systems

In order to transmit a signal in a communication system, the signal must first be processed, from signal reproduction to signal shaping to encoding and modeling. Following the preparation of the transmitted signal, the signal is transmitted to the channel transmission line. They face numerous issues as a result of signal transmission through these media, such as noise, attenuation, and distortion. (Vedantu, 2020). In order to transmit a signal in a communication system, the signal must first be processed, from signal reproduction to signal shaping to encoding and modelling. Following the preparation of the transmitted signal, the signal is transmitted to the channel transmission line. They face numerous issues as a result of signal transmission through these media, such as noise, attenuation, and distortion. . (Vedantu, 2020). Social media is a type of internet communication. Users of social networking sites can communicate, share information, and create digital content.

2.2.1 Types of current communication systems

Commercialization of video

In recent years, video has become increasingly popular on social media platforms such as YouTube, Facebook, Snapchat, and Instagram. It also makes recording videos with smartphones and cameras easier than ever before. (A. Bozkurt, A. Karadeniz, S. Kocdar;, 2017)

It is important that you transmit your message in a variety of formats and video is one of the most common ways to get involved.

Direct E-mail is similar to social networking direct messaging, but is generally more formal. It is the most popular way to communicate among companies with more than 200 billion emails sent each day. (Trickey & Stanley, 2018)

Direct Social Media Message (DM)

Social media need not necessarily be fully public. Nearly every social media channel offers a direct message option; some messaging services even have their own Facebook Messaging app. (Trickey & Stanley, 2018)

Blogging

A blog is a conversational website that allows people to post messages, news, knowledge, or any other type of information on the global Web. The majority of blogs have a comment section where you can interact with people who are interested in your blog post. That is why it is an excellent communication platform. (Trickey & Stanley, 2018)

Call for Participation

Voice is even more personalized than the other channels mentioned. One of the most widely used communication instruments is the telephone or mobile phone, which allows both parties to hear the other person's tones and emotions right away. (Trickey & Stanley, 2018).

Message Immediately (IM)

Although some Instant Messages formats are associated with social media, such as the Facebook Messaging system, there is a wide range of Instant Messaging platforms that are not associated with social networks, such as Hangouts and WhatsApp. (Trickey & Stanley, 2018).

2.2.3 Impact of hate speech on current communication systems

There have been reports of incidents on almost every continent. A large portion of the world now communicates with nearly a third of the world's people through Facebook alone. Experts say that people who are prone to racism, misogyny, or homophobia have increasingly found niches capable of strengthening their views and containing violence. Social media platforms also allow users to make their performances public.

Hate speech and its exposure can have profound psychological effects on a campus's reputation, climate, and morale, such as stress, anxiety, depression, and desensitization, in addition to being precursors to hate crimes and violence.

Haste speech is defined as any type of speech, writing, or conduct communication that attacks or uses derogatory or discriminatory language on the basis of a person or a group, i.e., on the basis of a person's religion, ethnicity, nationality, or other characteristics.

2.3 Hate speech detection systems

The increased use and sharing of social media has greatly benefited humanity. This has, however, resulted in a number of issues, including the spread and sharing of hate speech messages. As a result, recent studies have used a variety of functional techniques and machine learning algorithms to solve this emerging problem on social media sites to detect hate speech messages across multiple datasets. To the best of our knowledge, no study has been conducted that compares the variety of functional techniques and algorithms used to evaluate the features of engineering and the machine-learning algorithm on a standard dataset.

2.3.1 Types of hate speech detection systems

Approaches and baselines for detecting hate speech are included in logistic regression, vector support machine, and Nave Bayes. These models are frequently used in text classification. The likelihood of Naive Bayes label models is based on the assumption that features do not interact. Vector machines, which are supported by SVM and Logistic Regression, are a linear classifier that predicts classes based on function combinations.

2.3.2 Hate speech recognition systems

Recognizing Multilingual Online Hate

The exponential growth in the use of the Internet and social media over the last two decades has altered human interaction. This has resulted in many positive outcomes, but it has also resulted in risks and harms. Although the volume of harmful content online, such as hate speech, is unmanageable, the academic community's interest in automated methods of detecting hate speech has grown. Six publicly available datasets are analyzed in this study and classified into three classes - abusive, hateful, or neither - by combining them into a single homogeneous dataset. This research generates a basic model that improves the model's performance through the use of various optimisation techniques. After achieving a competitive performance score, this study develops a tool that identifies and scores a page with an effective metric in near-real time and uses the same feedback to retrain our models. This study demonstrates that our multilingual English and Hindi model is competitive and achieves comparable or superior performance when compared to most monolingual models. (Vashistha & Zubiaga, 2021).

Hate speech detection and multimodal learning

The goal of multimodal machine learning is to integrate and model various modes of communication, including verbal, audio, and visual communication. Deep education and representation are on the rise. The incorporation of multiple modalities into a unified learning framework is simple and efficient.

2.4 Related works

This project provides an overview of the approach and relevant discoveries in several of the group's most important articles. Technical or academic publications, lectures, books, texts, patents, technical reports, theses, or websites that provide comparable work on the development and application of a hate speech detection system are included.

In a research project carried out by (Erico, Salim, & Suhartono, 2020) Many countries have adopted legislation to ensure that companies have to deal within a time period with this type of content. It has been discussed. This systematic study examines hate speech detection and is used for the testing of hate speech and abusive language. The work also gives an overview of past investigations, including methods, algorithms and major characteristics. This systematic literature review will answer research questions. In this literature review, this study uses two research questions that form the basis of the next research. The correct classification of a piece of text as a hate speech requires a lot of properly labelled information.

In a research by (Hettiarachchi, Weerasinghe, & Pushpanda, 2020) to detect hateful content in Romanized Sinhala Language social media comments and documents automatically Most researchers studied hate speech recognition in or in English, but they identified Sinhala words written in English as Romanised Sinhala language here. Hate speech and other hateful texts are

becoming increasingly problematic, and machine learning and computer science are being used to combat this. In this study, the various extraction methods and four learning algorithms, N-gram differences and unigram, bigram and trigram, and the Min-Df value, are compared. The study investigates and compares various features of the various classifiers when classifying hate speech comments on Facebook.

In a research paper presented by (Al-Hassan & Al-Dossari, 2019) The history of hate speech and the methods used to detect it. The most recent contributions to hate speech and anti-social behavior will also be examined. Finally, challenges and recommendations have been presented for Arabic hate speech detection issues.

In a research paper by (Joni Salminen, Maximilian Hopf, & Shammur A. Chowdh, 2020) There is a scarcity of online hate detection models that use multi-platform data. To address the research gap, this study gathered a total of 197566 comments from four platforms: YouTube, Reddit, Wikipedia, and Twitter, with 80 percent of them being non-hateful and 20 percent being hateful. The research then tests various classification algorithms (Logistic Return, Nave Bayes, Support Vector Machines, XGBoost, and Networks) (Bag-of-Words, TF-IDF, Word2Vec, BERT, and their combination). Despite the fact that all models significantly outperform the keyword baseline classification, XGBoost performs best when all features are used (F1=0.92). According to functional analysis, the BERT functionality is the most effective for forecasting. Conclusions support the generality of the best model, as platform-specific results from Twitter and Wikipedia are comparable to their source documents.

In another research article by (Lazaros Vrysis, et al., 2021)The project aims to monitor and shape anti-refugee and anti-migrant speech in Greece, Italy, and Spain. This is how a web interface for the creation and query of a multi-source database containing hate content is implemented and tested. The sources chosen include Twitter comments and posts, YouTube videos, and Facebook posts, as well as a list of websites with commentary and articles. Users can use the interface to scan social media for keywords, search an existing database, annotate records using a dedicated platform, and provide new content. Furthermore, new methods and machine learning models are used for hate speech detection and sentimental text analysis. You can use an internet-compatible graphical user interface to access the online interface. A multifactor questionnaire was created to collect user feedback on the web interface and its various functionalities in order to evaluate the interface.

In a thesis paper by (Themeli, 2018) To effectively handle the abusive task of language discrimination online, use multiple text representation and grading techniques. BagofWords (BoW), n-gram bag of word and character, feelings, syntax and grammar analysis, word embedding, and n-gram charts are among the techniques investigated. We also put various classification algorithms to the test, including Naive Bayes, Logistic Regression, Random Forests, K-Nearest Neighbours, and ARN. Our goal is to evaluate representation and classification algorithms in terms of their performance contribution to the Hate Speech detection task. Furthermore, the utility of n-grams (NGGs) is highlighted in an efficient, low-dimensional text representation that generates vectors of similarity that appear to be deep characteristics with significant input into classification results. Aside from binary classification experiments, we also test our method in multi-class classification experiments with respect to abusive language discrimination tasks.

In a research project by (Vashistha & Zubiaga, 2021) Six available data sets are analyzed in this study by combining them into a single uniform data set. After classifying them as abusive, hateful, or none of the above, we build a basic model and use various optimization techniques to improve model performance scores. We create a tool to identify and evaluate a page with an effective metric in near-real time after achieving a competitive efficiency score, and we retrain our model using the same feedback.

In a research by (Perifanos & Goutsos, 2021)By combining computer vision and natural language methods for abusive context detection, this study presents a new multimodal method for detecting hate speech. Our research focuses on Twitter messages, specifically hateful, xenophobic, and racist Greek language directed at migrants and refugees. We combine transferring learning with a better definition of transformer bidirectional encoder representations (BERT) in our approach (Resnet). Our contribution includes the creation of a new hate speech classification dataset comprised of tweet IDs and a visual appearance code that would be displayed on our web browser.

According to a research done by (Abro, Shaikh, Ali, Khan, & Mujtaba, 2020) This paper compares the performance of three feature engineering techniques and eight machine learning algorithms on three distinct classroom datasets in order to assess their performance. The test results show that the bigram functions perform best with 79 percent overall precision when used in conjunction with the vector machine algorithm. Our research has practical implications for detecting automated hate speech messages and can be used as a foundational study.

CHAPTER THREE

3.0 METHODOLGY

3.1 METHOD OF IDENTIFICATION OF SYSTEM USER REQUIREMENTS.

This section gives an overview of the system and user requirements of the system, and how these requirements were gathered. The identified system and user requirements were sourced through the study of related works, social media platforms and informal discussions with users of the system. Observations were carried out on systems that shared similar functions. From these observations, features that needed to be implemented in the system were identified. After developing a demo version of the system, the software was made available for users to experiment with. This activity helped to refine the requirements of the system and add new requirements that had not previously been identified.

3.1.2 USER REQUIREMENTS

This section gives an overview of the requirements of the system, for the users of the system. These requirements are functionalities the system is supposed to make available to the users of the system. They are highlighted below;

I. It shall provide a text area or text field for the user to input his/her text.

II. It shall provide a submit button for the user to send his/her text after input.

III. It shall provide a response message to be displayed to the user when his/her text has been submitted and found not to have racist content.

IV. It shall provide an error message to be displayed to the user when his/her text has been submitted and found to have racist content.

3.1.3 SYSTEM REQUIREMENTS

This section provides a detailed description of the requirements for developers of the proposed system. The system requirement of the system is described in terms of functional and non-functional requirements. The functional and non-functional requirements of the system are listed below:-

3.1.3.1 FUNCTIONAL REQUIREMENTS

They are the functionalities provided by the proposed software. They include;

i. The system shall provide a screen for a user's input to be seen.
ii. The system shall provide an avenue for the developer to be monitoring the system.
iii. The system shall provide an avenue for the developer to receive usage report of the system.
iv. The system shall provide an avenue for the developer to perform regular system check to ensure effectiveness and efficiency.

3.1.3.2 NON-FUNCTIONAL REQUIREMENTS

The identified non-functional requirements are listed below:-

I. The system shall be able to detect racist content in a text or sentence even if the words are misspelled.
II. The system shall be able to detect the use of an emoji in a sentence or text.
III. The system shall provide an interface that is easy to use and easy to understand to the users.

IV. The system shall render all of its services in a timely fashion.

3.1.3.3 SOFTWARE AND HARDWARE REQUIREMENTS

The system software and hardware requirements were put together from an informal study of the users and the already existing systems that this system can be used with. The users of this system are internet users or social media users. With all this in mind, the software and hardware requirements of this system have been put together. They are the standard requirements of a PC or Laptop that will use the system. Not enough RAM (Random Access Memory) can affect the speed and efficiency of the system. The hard disk also has to be sufficient to store the file. The requirements include:-

a. Processor: Core i3 (Minimum).
b. Processor speed: 2.5GHz (Minimum).
c. RAM: 4GB (Minimum).
d. Hard disk: 500GB (Minimum).
e. Monitor Display: LED.
f. Mouse: Touchpad with multi-touch gesture support, USB or PS/2 .
g. Internet connectivity 3G (Minimum).

3.2 SYSTEM DESIGN MODELS.

Several UML models were used to specify the design of the system and structure its overall architecture. These models provided graphical representations of the interactions between several users of the system and the system, the hierarchy of activities in the system and the flow of

control between these activities, and the overall architecture of the system. These graphical representations are illustrated in the following parts of this section

3.2.1 USE CASE DIAGRAMS

This section gives an illustration of the various users of the designed software and their interactions with the system using use case diagrams.

a) Use case diagram: the role of the user is to send messages or text through the system and receive responses from the system if the message has racial content. The figure below illustrates the user use case scenarios.

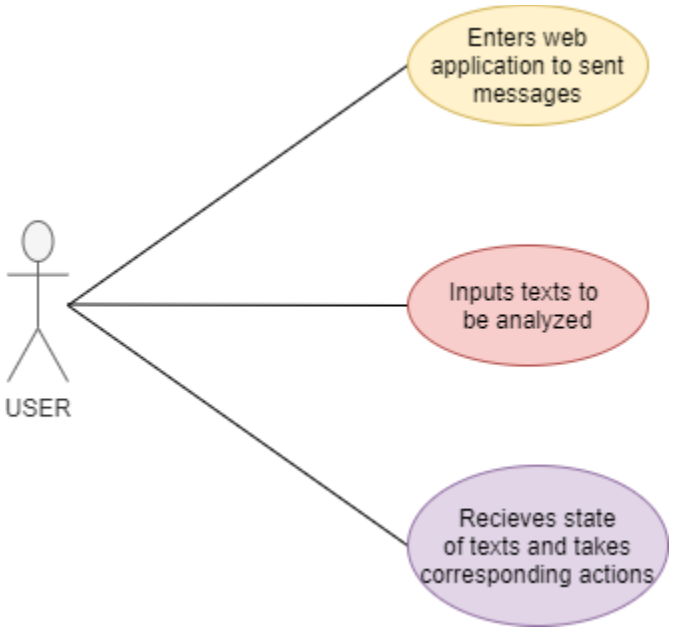

Figure 3.1: Use Case Diagram

b) Admin use case diagram: the role of the admin is to manage the interactions between the system and the user. They are responsible for monitoring the system and receiving usage report of the system. The figure below illustrates the administrator use case scenarios.

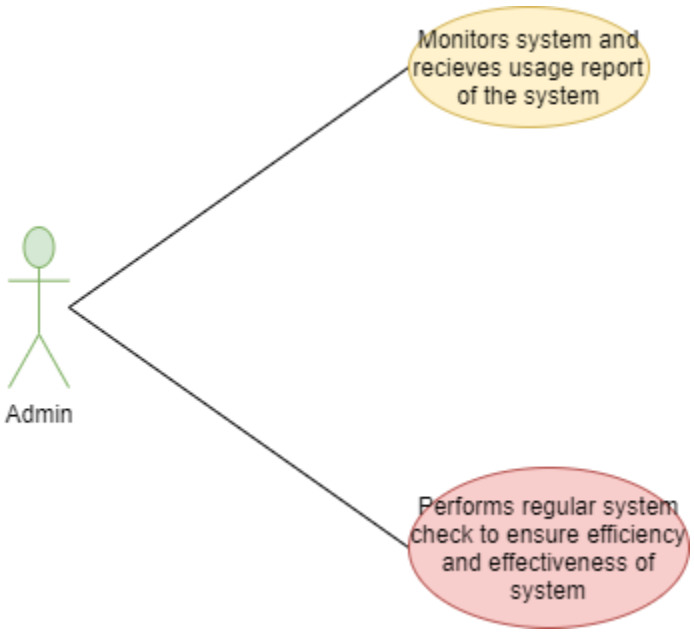

Figure 3.2: Admin Use Case Diagram

3.2.2 SEQUENCE DIAGRAMS

The sequence diagram illustrates the sequence of activities that happens in the system. The sequence of activities in the proposed system is illustrated in the figure below. The diagram graphically represents several sequences of activities that occur within the proposed system i.e., user's entry into the web app, input of text/message and response gotten. It also shows the sequence of activities that occur when the text is inputed.

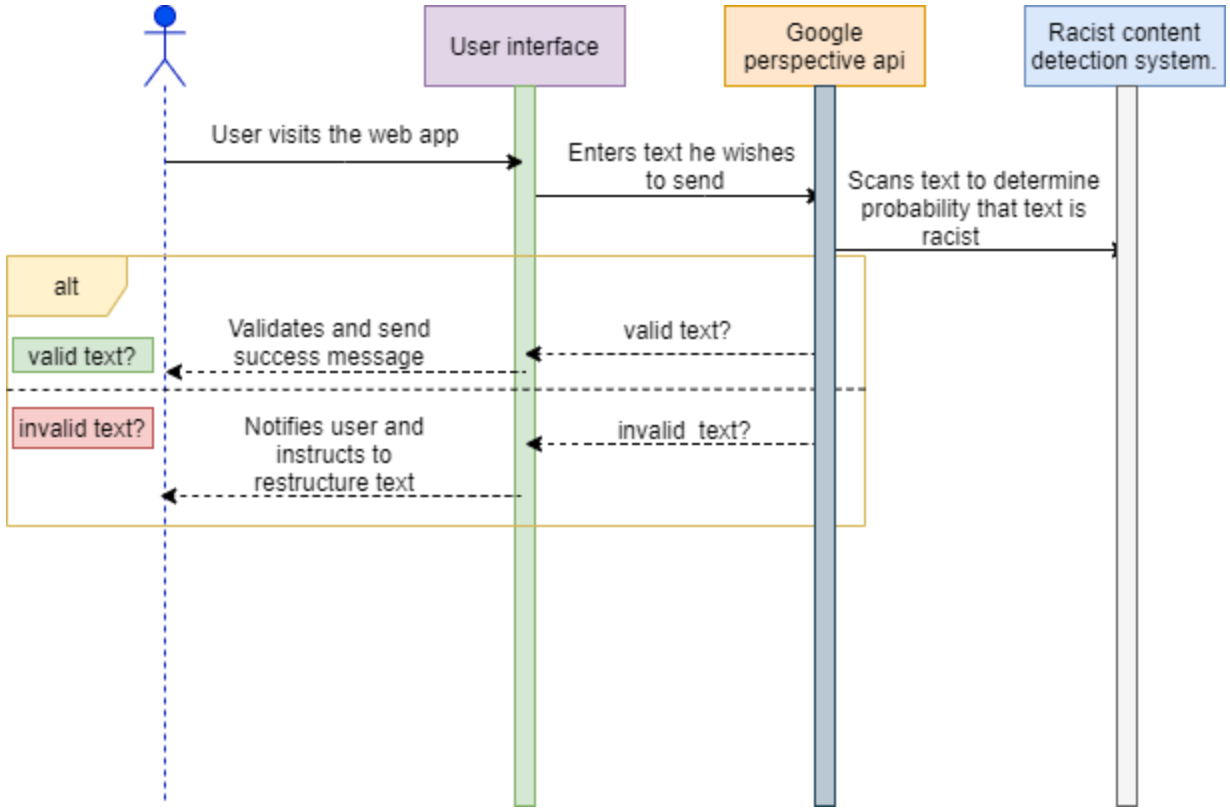

Figure 3.3: Sequence Diagram

3.4 ACTIVITY DIAGRAMS

The activity diagram illustrates all the activities that occur within the system and the flow of control between these activities. It shows the standard order of activities in the software system.

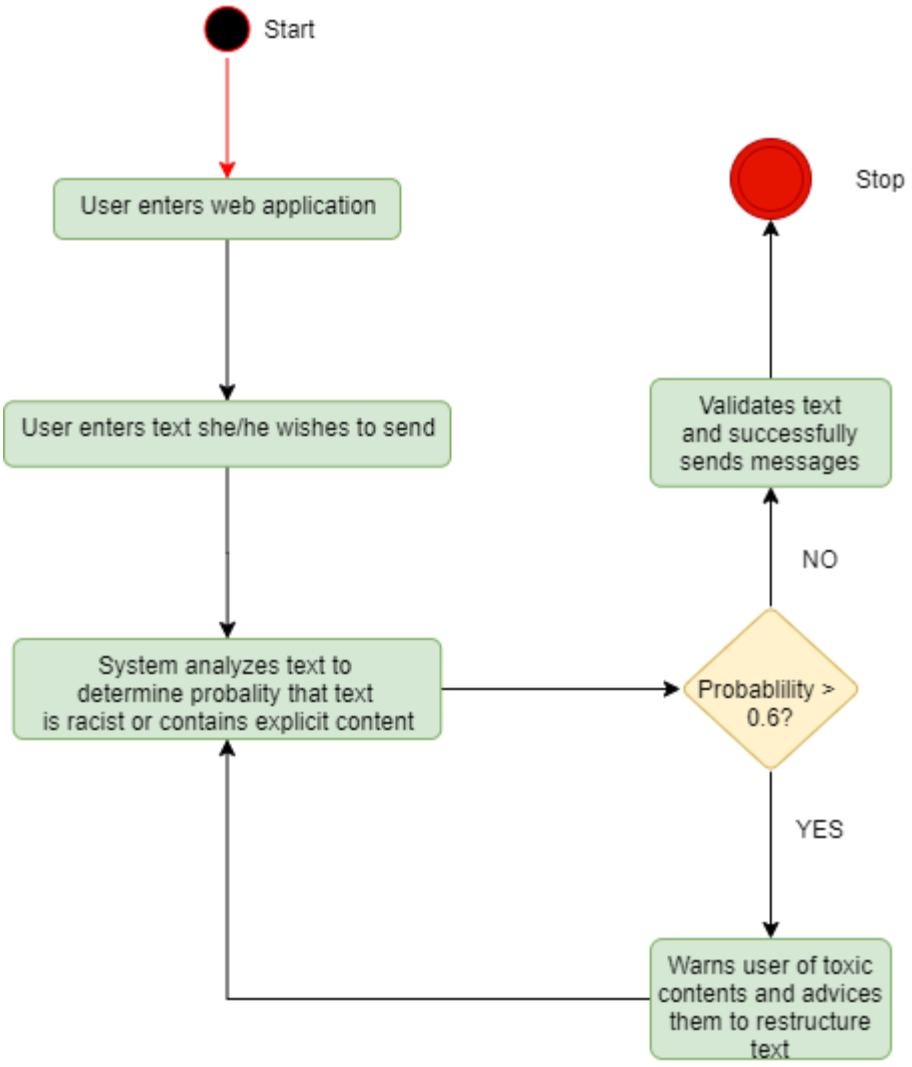

Figure 3.4: Activity Diagram

3.5 ARCHITECTURAL DIAGRAMS

The architectural diagram illustrates the organization and architectural structure of all the components in the system. It also helps to understand how the components of the system communicate with each other. The proposed software consists majorly of six components. The users of the system, the user interface they communicate with, react-native UI components, application logic implemented in javascript, firebase backend, and an API. The architectural diagram of the proposed system is graphically illustrated in the figure below:-

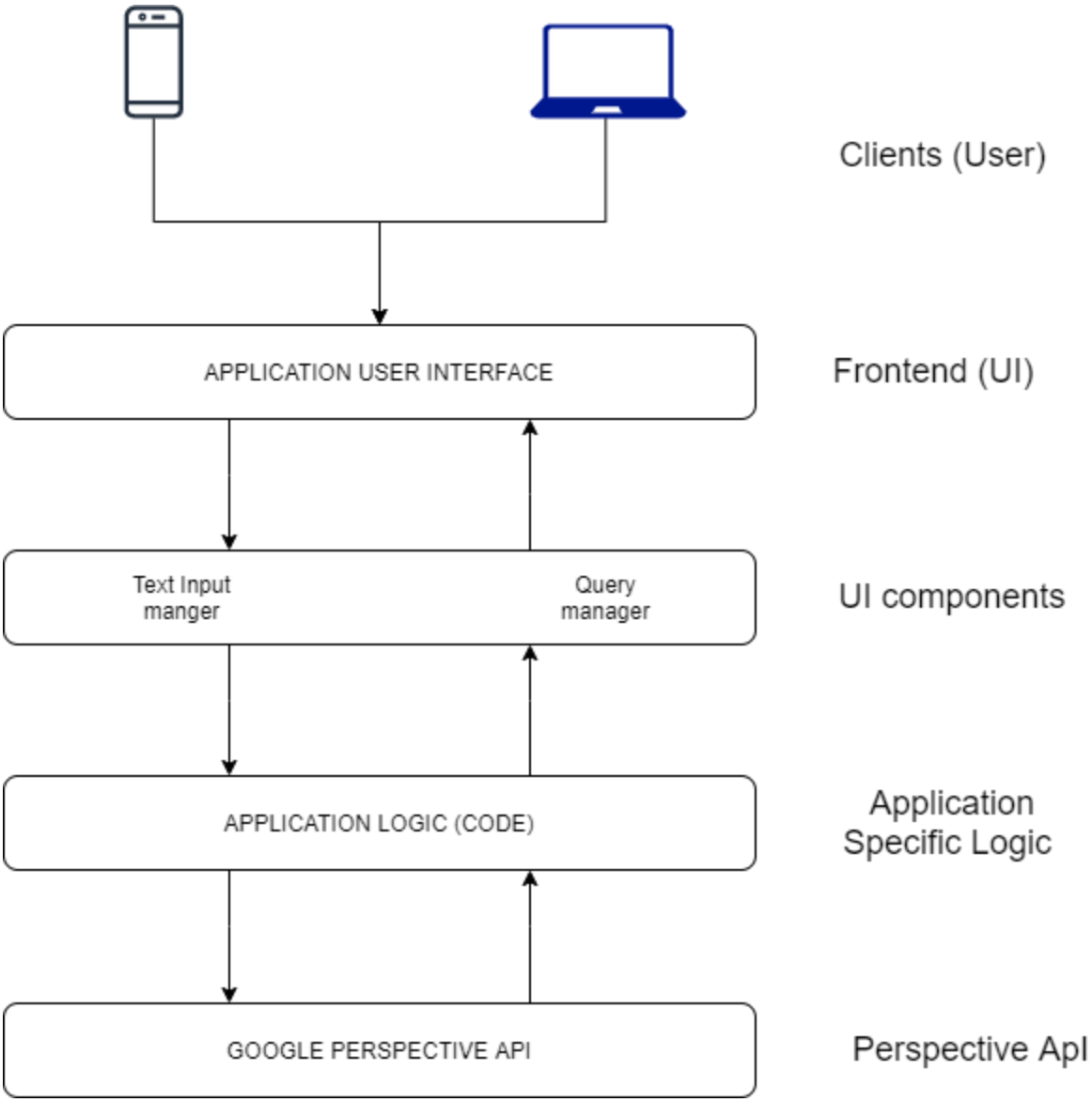

Figure 3.5: Architecture Diagram

3.3 PERSPECTIVE API BY GOOGLE

Perspective is a free API that uses machine learning to identify "toxic" comments, making it easier to host better conversations online.

3.4 SYSTEM TESTING APPROACH

The designed system will be tested for correctness using two methods of testing namely

i. Unit testing: The goal of having these unit tests set up is to ensure that whatever new code or new functionality that gets added to the code base, don't end up breaking the system. Unit tests were written for all the major classes of the system and tested before every build of the application was run.

ii. Acceptance testing: this will be done on completion of the project to ensure that the designed system meets the specified needs of the identified users. Whatever omissions to the software or bugs that are discovered at this stage of testing will be duly noted and fixed.

3.5 DEVELOPMENT TOOLS

The system was developed using React Native, Javascript and Perspective API by Google.

CHAPTER FOUR

4.1 IMPLEMENTATION AND RESULT OF DESIGNED SYSTEM INTRODUCTION

This chapter shows the implementation of the system that detects racist words. The tools used in system design and development of the system's primary idea and functionality to accomplish its defined mission.

4.2 SYSTEM MODULES

The system comprises of the following modules:

1. The home page: This is the first page that appears when the system is launched from the web browser. It displays the text area for the user to type his/her sentence and press the submit button.
2. The valid message page: This is the page that displays when a valid message is sent. It displays the text 'Sending message to receiver'. The text input by the user is successfully sent.
3. The invalid message page: This is the page that displays when an invalid message is sent. It displays the text 'You have used inappropriate words. Please re-phrase.' . The text input by the user is not sent. The user has to re-phrase the text or sentence before the message can be sent.
4. The emoji page: This is the page that appears when an emoji is used in a text/sentence. If the emoji can be perceived by users as having racial meanings, the system stops the text/sentence with the message from being sent and displays 'You have used inappropriate words. Please re-phrase.' ,if it doesn't have any racial meanings it displays 'Sending message to receiver……' .

5. Development page: This project was built and tested on an emulator, where I checked for errors in my codes ill display what it looks like below also the database I made use of React Native.

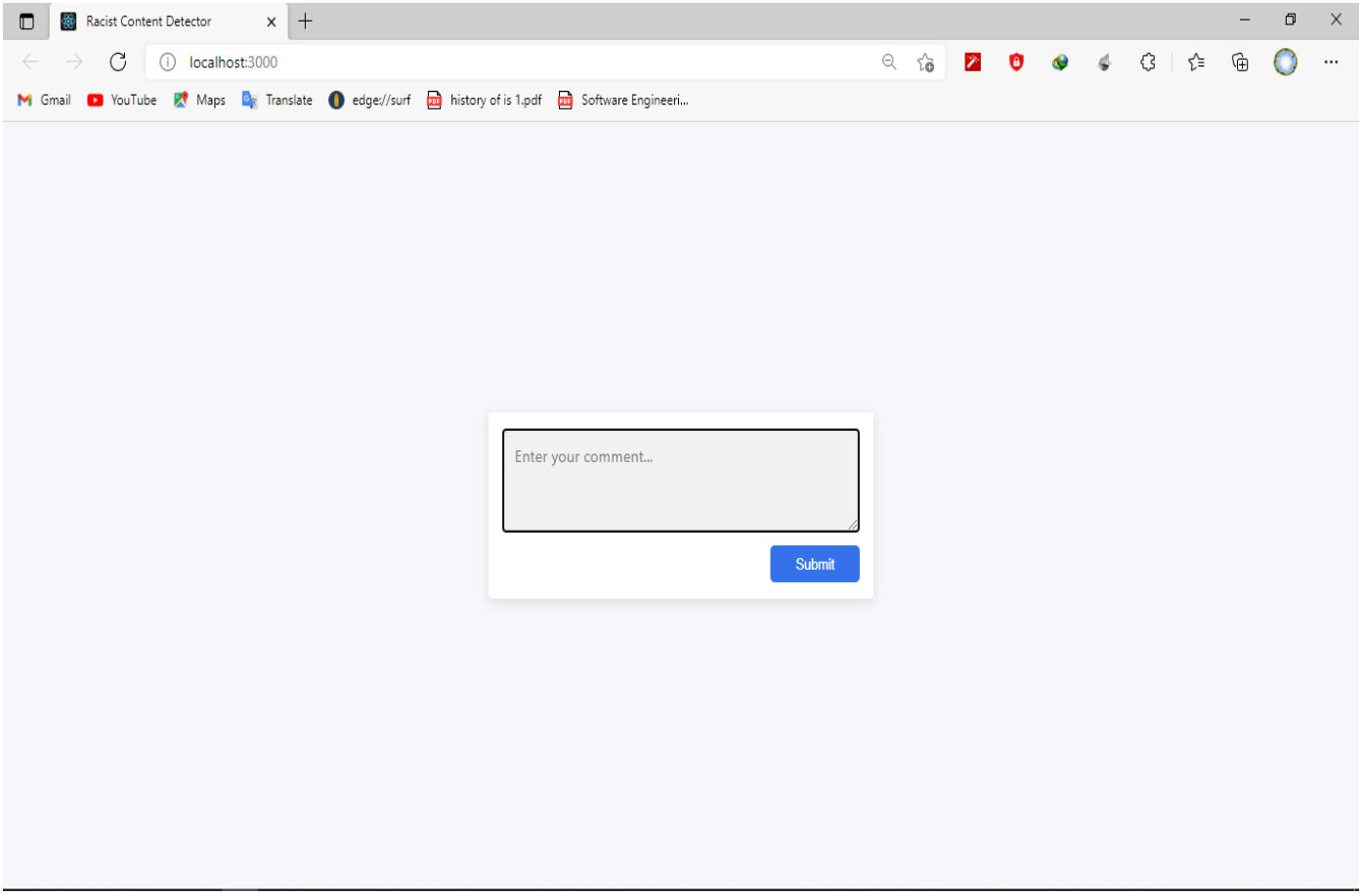

Fig 4.1: The Home Page

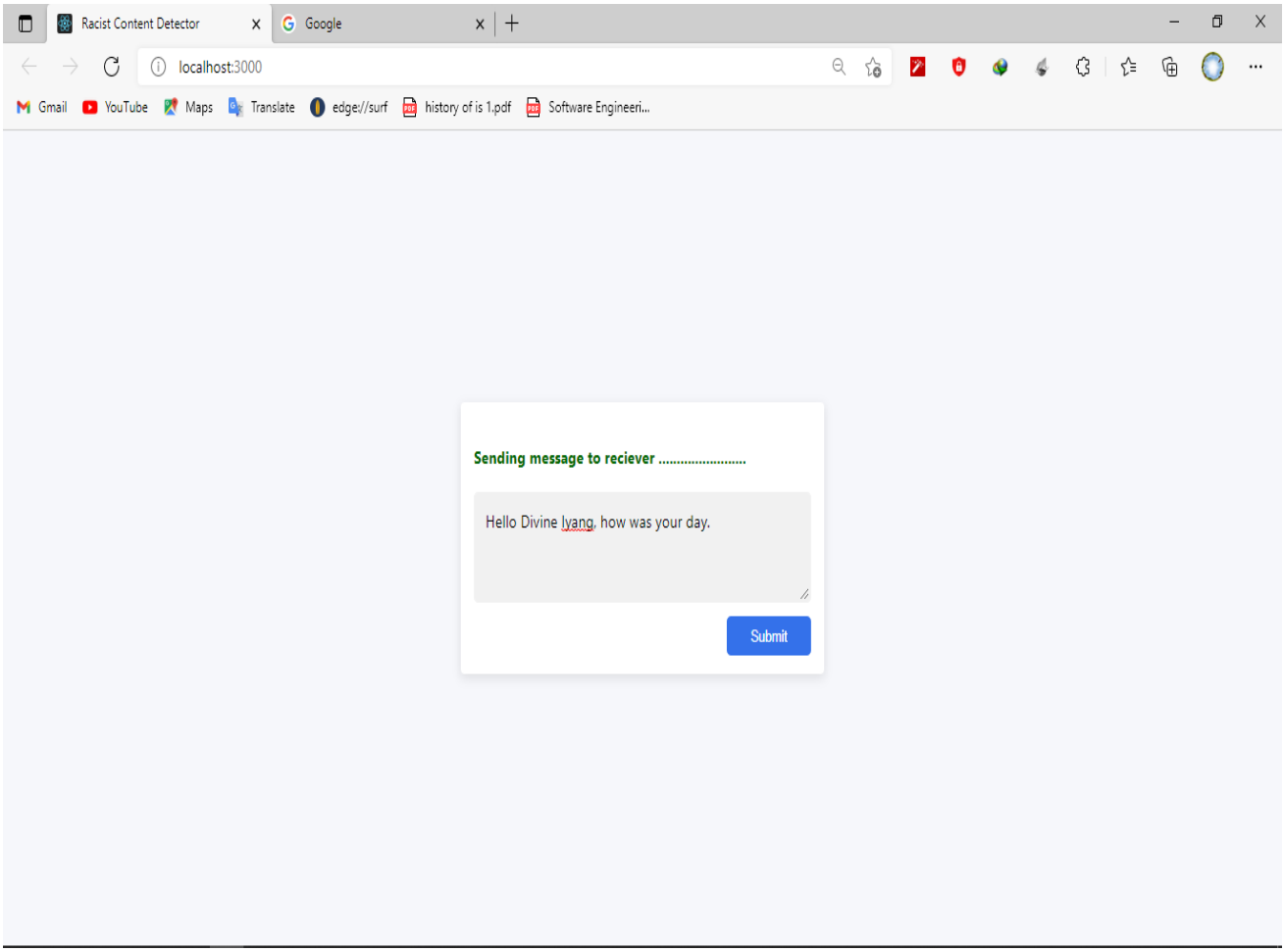

Fig 4.2: The Valid Message Page

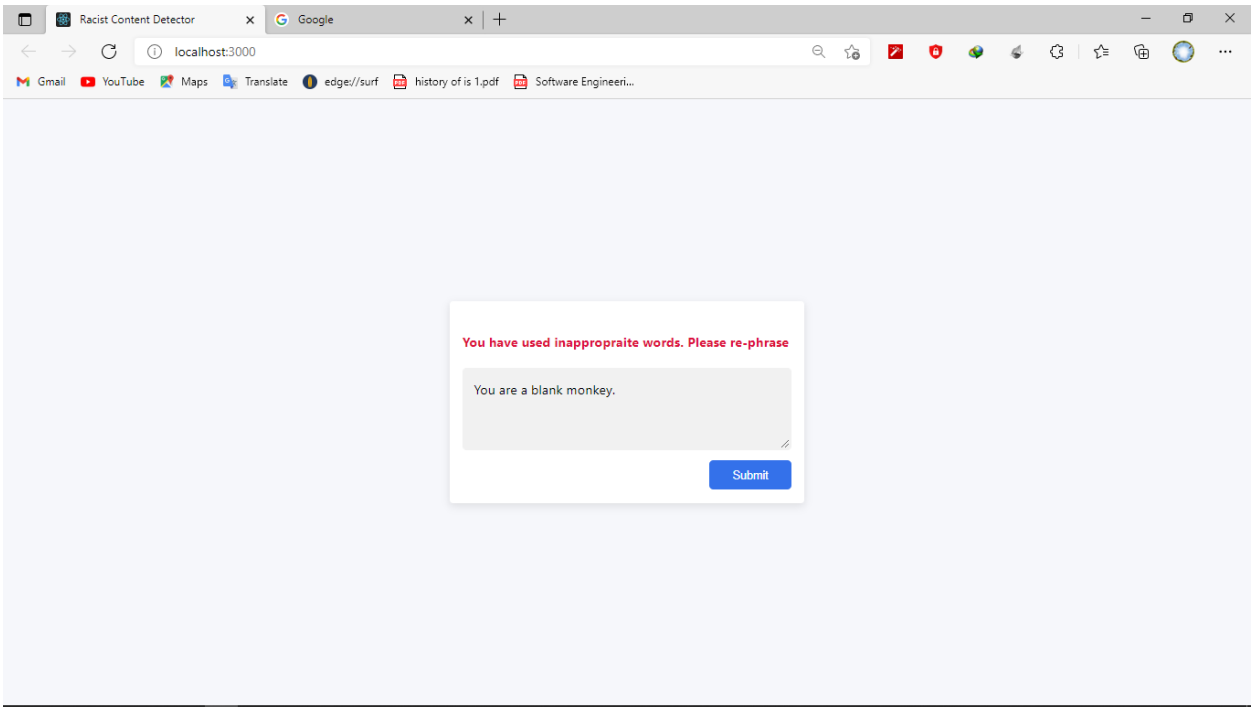

Fig 4.3 The Invalid Message Page

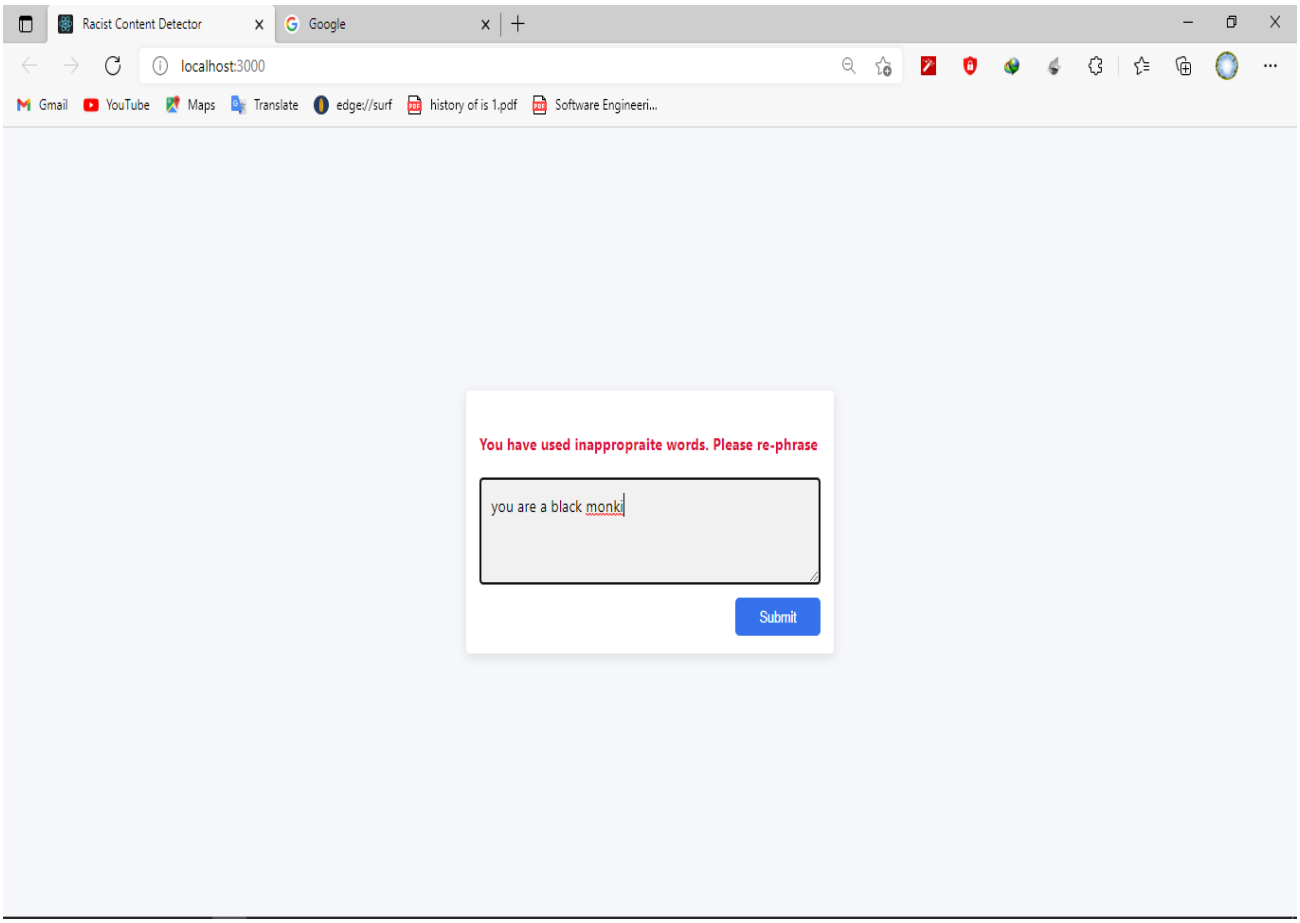

Fig 4.4 The Invalid Message with Wrong Spelling Page

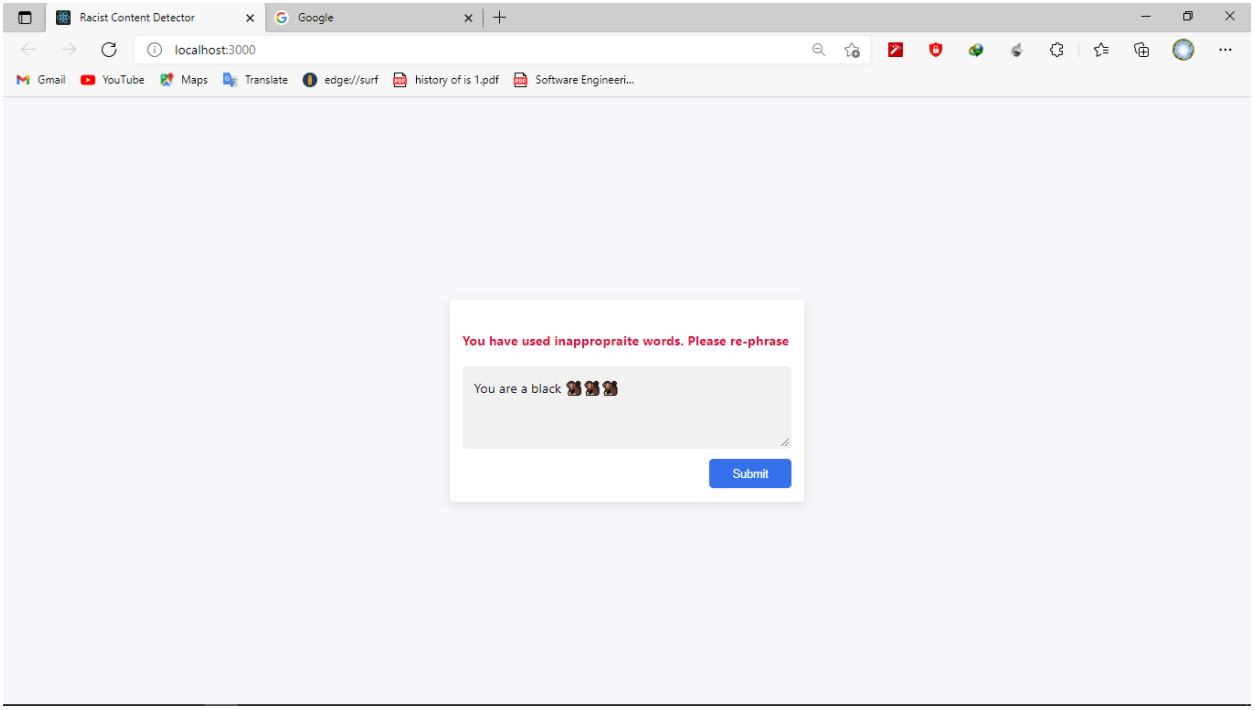

Fig 4.5 The Emoji Page

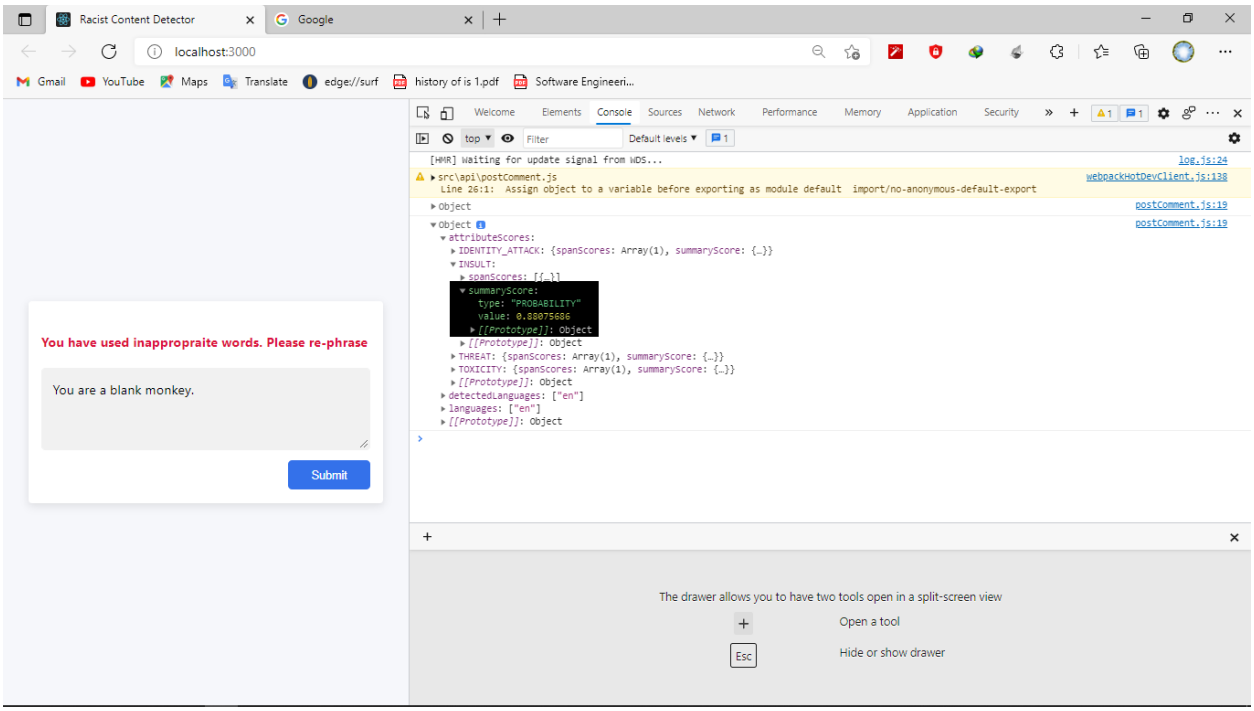

Fig 4.6 React Native software used, Perspective API and the Racist content detection system

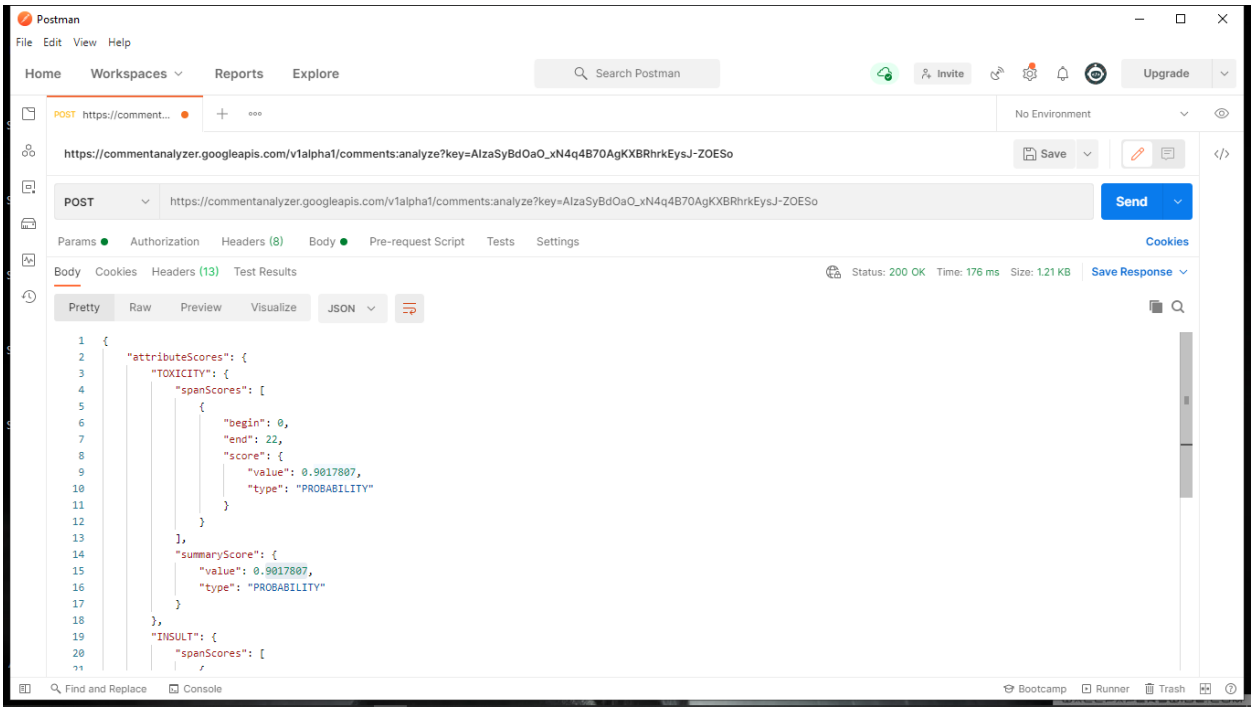

Fig 4.7 Testing the API with Postman

Postman is an interactive and automatic tool for verifying the APIs of your project.

CHAPTER FIVE
SUMMARY AND CONCLUSION

5.1 SUMMARY

Taking into consideration, the presence of racial contents online, it has become a matter of utmost necessity that a system is developed to put an end or at least massively reduce the presence of this contents. The system developed will alert users of the presence of racial words and emoji in their text/sentence to be sent or delivered and inform them to change the words or sentence to an acceptable text/sentence. This will massively reduce the presence of racial words seen on the internet, thereby making the online community more user-friendly, serene and ethically better.

5.2 LIMITATIONS

- Limited time to acquire the necessary knowledge/skills needed to develop the system.
- Lack of software for more efficient testing.
- Testing needed internet connection, which wasn't in adequate and efficient supply.
- The system only operates on PC, doesn't operate on mobile devices.

5.3 CONTRIBUTION TO KNOWLEDGE

There have been different works done in respect to putting an end to racial contents online. Most of the previous or already existing works were based on detecting racial contents posted online. This work deals with stopping racial words and emoji from being posted online or being seen online. It detects the racial content before it gets visible to the online community and stops it from being sent.

References

A. Bozkurt, A. Karadeniz, S. Kocdar;. (2017). Social Networking Sites as Communication, Interaction, and Learning Environments. *Perceptions and Preferences of Distance Education Students*, 348-365.

Abro, S., Shaikh, S., Ali, Z., Khan, S., & Mujtaba, G. (2020). Automatic Hate Speech Detection using Machine Learning: A Comparative Study. *(IJACSA) International Journal of Advanced Computer Science and Applications*.

Al-Hassan, A., & Al-Dossari, H. (2019). DETECTION OF HATE SPEECH IN SOCIAL NETWORKS: A SURVEY ON MULTILINGUAL CORPUS. *6th International Conference on Computer Science and Information Technology*.

Bahador, B. (2017, november 17). *Classifying and Identifying the Intensity of Hate Speech*. Retrieved from items.ssrc.org: https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/classifying-and-identifying-the-intensity-of-hate-speech/

Benner, A. D., & Kim, S. Y. (2009). Intergenerational experiences of discrimination in Chinese American families: Influences of socialization and stress . *Journal of Marriage and Family*, 71(4), 862-877. doi: 10.1111/j.1741-3737.2009.00640.x.

Erico, C., Salim, R., & Suhartono, D. (2020). A Systematic Literature Review of Different Machine Learning. *INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION*, 2549-9610.

Glaser, J., & Kahn, K. (2005). *Online prejudice and discrimination: From dating to hating.* Oxford: Oxford Press.

Harris-Britt , A., Valrie, C. R., Kurtz-Costes, B., & Rowley, S. J. (2007). Perceived racial discrimination and self-esteem in African American youth: Racial socialization as a protective factor. *Journal of Research on Adolescence*, 17(4), 669-682. doi: 10.1111/j.1532-7795.2007.00540.x.

Hettiarachchi, N., Weerasinghe, R., & Pushpanda, R. (2020). Detecting Hate Speech in Social Media Articles in Romanized Sinhala. *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer).* Colombo, Sri Lanka.

Joni Salminen, Maximilian Hopf, & Shammur A. Chowdh. (2020). Developing an online hate classifer for multiple social media platforms. *Human centric and Information sciences*.

Lazaros Vrysis, N. V., Kotsakis, R., Saridou, T., Matsiola, M., Veglis, A., Arcila-Calderón, C., & Dimoulas, C. (2021). A Web Interface for Analyzing Hate Speech. *Future internet*, 13-80.

MacAvaney et al. (2019). Hate speech detection: Challenges and solutions.

Perifanos, K., & Goutsos, D. (2021). Multimodal Hate Speech Detection in Greek Social Media. *Multimodal Technol. Interact.*

S. MacAvaney, Yao; H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O;. (2019). *Hate speech detection: Challenges and solutions*. Retrieved from https://journals.plos.org/: https://doi.org/10.1371/journal.pone.0221152

Themeli, C. K. (2018). *Hate Speech Detection using different text representations in online user comments.* NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS.

Trickey, N., & Stanley, T. (2018, december 18). *Types Of Modern Communication*. Retrieved from resourcetechniques.com: https://www.resourcetechniques.co.uk/news/web-design/types-of-modern-communication-100244

Vashistha, N., & Zubiaga, A. (2021). Online Multilingual Hate Speech Detection: Experimenting. *Information 2021*.

Vashistha, R., & Zubiaga, A. (2021). Online Multilingual Hate Speech Detection: Experimenting with Hindi and English Social Media. *Information*, 12(5).

Vedantu. (2020, june 27). *Communication Systems*. Retrieved from VEDANTU: https://www.vedantu.com/physics/communication-systems

Zhang, Z., & Luo, L. (2018, october 25). *Hate Speech Detection: A Solved Problem?* Retrieved from arxiv.org: https://arxiv.org/pdf/1803.03662.pdf