

CHAPTER 1

INTRODUCTION

Racism is at its core “an ideology of racial domination” [1] in which the presumed biological or cultural superiority of one or more racial groups is used to justify or prescribe the inferior treatment or social position(s) of other racial groups [2]. It describes the prejudice, discrimination, or antagonism by an individual, or racial group against a person or people on the basis of their membership to a particular racial or ethnic group.

Among scholars in social sciences, the term ‘race’ is generally understood to be a social construct, biologically meaningless when applied to humans. Physical differences such as skin color have no natural association with group differences in ability or behavior. Race nevertheless, holds tremendous significance in structuring social reality.

Indeed, historical variation in the definition and use of the term provides a different view. The term was first used to describe peoples and societies in the way we now understand ethnicity or national identity. Later, in the seventeenth and eighteenth centuries, as Europeans encountered non-European civilizations; Scientists and philosophers began to give race a biological meaning. They applied the term to plants, animals, and humans as a taxonomic subclassification within a specie.

Consequently, race became understood as a biological, or natural, categorization system of the human species. Johann Friedrich Blumenbach was first to categorize humans based on a comparative anatomy of the human skull; he listed the Caucasian, Mongolian, Malayan, Ethiopian, and American races as the five distinct categorizations of the human race. His thesis was widely used, though with several variations till the 19th century. Later in the 20th century, Carleton S.

Coon, an American anthropologist proposed the Negroid (Black), Australoid (Australian Aborigine and Papuan), Capoid (Bushmen/Hottentots), Mongoloid (Oriental/Amerindians), and Caucasoid (White) races as the five categorizations of humans, based on the same study. This thesis endures till date with again, several variations.

There was much prejudice based upon this way of looking at the world. Racism, a non-scientific theory or ideology, was that a particular race was superior or inferior. It argued that in the races that make up the human race, there are deep, biologically determined differences and stated that different races should live separately and not intermarry.

As Western colonialism and slavery expanded, the concept was used to justify and prescribe exploitation, domination, and violence against peoples racialized as nonwhite. These in turn supported the horrors of African slavery, Apartheid, the Jim Crow laws, Nazism, and Japanese imperialism.

1.1. Background of Study

With the advent of the internet, one of the greatest inventions of our time - social media, has become a vital communications tool through which people exercise their rights to freedom of speech and extensively exchange information and ideas, including racist ones. As a result, social media plays a salient role in the spread of racial driven hate speeches. Individuals who believe in the ideology of racism (Racists) can be seen on the social media cyberspace sharing their prejudices and bias through comments that are often publicly accessible.

Indeed, a growing movement of people around the world have been witnessed who are advocating for change, justice, equality, accountability of people in power and respect for human rights. In

such movements, social media has often played an important role by enabling people to connect with each other and exchange thoughts, emotions, and information, ergo creating a sense of solidarity.

However, as social media have come to dominate socio-political landscapes in almost every corner of the world, new and old racist practices increasingly take place on these platforms. Racist speech thrives on social media, including through covert tactics such as the weaponization of memes and use of fake identities to incite racist hatred. In a review and critique of research on race and racism in [4], Jessie et al identified social network sites as spaces where race and racism play out in interestingly and sometimes disturbing ways. Since then, social media research has become popular among academics, with its own journal, conference, and several edited collections [3].

In parallel, scholars like Araidna et al [5] have grown increasingly concerned with racism and hate speech online due to the rise of far-right leaders in countries like the United States, Brazil, India, and the United Kingdom and the weaponizing of digital platforms by white supremacists.

Reddit gives rise to toxic subcultures, YouTube to a network of reactionary right racist influencers and coordinated harassment is pervasive on Twitter. Users also produce and reproduce racism through seemingly benign practices, such as the use of emoji and GIFs. Social media contributes to reshaping “racist dynamics through their affordances, policies, algorithms and corporate decisions”.

Microaggressions as well as overt discrimination can be found in platform governance and designs. Snapchat and Instagram have come under fire for releasing filters that encourage white people to perform “digital blackface” and automatically lighten the skin of non-whites. Facebook, by

tracking user activity, enabled marketers to exclude users with what they called an African American or Hispanic “ethnic affinity”. And TikTok has faced criticism, when it suspended a viral video raising awareness of China’s persecution of Uighurs. This shows that digital technologies not only “render oppression digital” but also reshape structural oppression.

Social media platforms’ policies and processes around content moderation play a significant role in this regard. Companies like Facebook and Twitter have been criticized for providing vast anonymity for harassers and for being permissive with racist content disguised in humor because it triggers engagement.

Racist discourses and practices on social media represent a vital, yet challenging area of research. With race and racism increasingly being reshaped within proprietary platforms like Facebook, Twitter, Instagram, and YouTube. The proliferation of racial biases and discrimination spread on social media through comments propagate the ideology of racial domination, exposing the members of the discriminated racial or ethnic group to physiological and psychological distress.

David et al [6] reviewed some evidence that shows the adverse effect of racism on mental and physical health of African Americans living in the United States in three ways; First, racism in societal institutions can lead to truncated socioeconomic mobility, differential access to desirable resources, and poor living conditions that can adversely affect mental health. Second, experiences of discrimination can induce physiological and psychological reactions that can lead to adverse changes in mental health status. Third, in race-conscious societies, the acceptance of negative cultural stereotypes can lead to unfavorable self-evaluations that have deleterious effects on psychological well-being.

Another research by Joanne [7] suggests that racism, or discrimination based on race or ethnicity is a key contributing factor to the onset of diseases. Yin P. et al in [8] found that experiencing racism is associated with poor mental and physical health. The stress associated with experiencing racism can have long lasting physical effects. Stress can elevate blood pressure and weaken the immune system, which, in turn, raises the risk of developing long-term health conditions. Stress as a result of racism can also lead to behaviors that may cause further risk to physical health. Racial discrimination can be linked to higher rates of smoking, alcohol use, drug use, and unhealthy eating habits.

The authors in [9] associates' racism with higher stress levels, increasing a person of color's risk of developing high blood pressure. In fact, it reports that black people living in the United States are more likely to suffer hypertension than any other racial group.

Some social media giants like Facebook and Twitter have made efforts to secure their cyberspace by removing inappropriate contents. Contents such as graphic violence, nudity, fake accounts, hate speech, cyber bullying, fake news, and terrorist propaganda designed to instigate unnecessary fear among people often in order to control political outcomes or worse.

With the growing amount of the social media user population, it is increasingly unachievable for a social media network to manually vet all contents and comments across the platform to ensure a safe cyberspace for its users. Rather, machine learning approaches are applied to detect and remove contents that do not follow the community guidelines i.e., inappropriate contents. Sentiment analysis in natural language processing can be used to filter out fake news, hate speech, and prevent cyberbullying. In the same manner, video analysis in computer vision can be used to detect and remove graphic violence and some other explicit contents.

Although artificial intelligence may be the future, it is still far from human intelligence. Inaccurate classifications by a model applied to remove inappropriate contents on a social media platform may result in the neglect of actual offensive or inappropriate content and instead restrict the users' freedom of expression even when they are in line with the community guidelines.

This consideration drove the big tech company, Facebook to establish a content monitoring department in the company, hiring over 3000 employees in 2017 to assist with the removal of videos showing murder, suicide, and other forms of violence. US congresswoman Alexandria Ocasio-Cortez slammed its CEO, Mark Zuckerberg for this action, calling the job undesirable and detrimental to the mental health of these employees who work long hours and make swift decisions while sifting through traumatic contents.

Indeed, the remedy to the current insufficiency of AI models is not human effort, but rather to build even better models.

1.2.Statement of Problem

Racial conflicts have become even more frequent than before. Social media companies are continuously being slammed for their inadequate response to this problem. 2020 witnessed a worldwide movement calling for racial equality and justice. The movement began after an African American male was suffocated and murdered by an NYPD police officer. Since then, there has been significant publications on social media and the role it has played in amplifying racism.

Moreso, the government of the United Kingdom have threatened to make social media companies legally accountable for the racist content on their platform after the witnessed increase of racist abuse on footballers in 2021. English football clubs have also threatened a boycott of social media

in a bid to eradicate online hate. In [10], Janice G. blames social media for amplifying white supremacy and suppressing anti-racism, pointing out that black social media users have experienced a great deal of censorship online. Often times, those who are outspoken about white supremacy and racism have found their content removed or taken down for violating community guidelines. This problem may be caused by a bias in the data used to train their models or by human error – Afterall, the data was labelled by humans. In any case, the scope of this problem is not known.

Racism is a pressing issue and social media undoubtedly plays a key role in the spread of racism both in sports and in societies in general. To solve this problem, I will be building a classification model using machine learning to detect racist comments on social media platforms.

1.3.Aim and Objective.

In this project, I propose a machine learning model for the automatic detection of racist comment across social media platforms.

To achieve this, I:

1. Track down past events and social media trends which are likely to have triggered racist reactions.
2. Retrieve annotated comments from public social media sites like Facebook, Instagram, Twitter, YouTube and TikTok.
3. Created an unbiased dataset of racist comments across social media platforms.
4. Prepared the data for supervised learning and cleaning the dataset using certain natural language processing techniques in order to achieve best results.

5. Trained the model to detect racist comments based on the dataset using the Naïve Bayesian Classifier, Logistic Regression, and Support Vector Machines.
6. Tested the model to determine its accuracy.
7. Compared the accuracy of the model when each of the classifiers were used.

1.4. Significance of Study

This project gains its significance from giving solidarity to the efforts already being used in creating a safe and racist free cyberspace for social media users, as well as protecting the social media communities from the physiological and psychological impacts of racism by curbing the spread of racist ideologies and hopefully racially inspired violence.

1.5. Scope of Project

Indeed, there is no one size fits all. Social media is available to people of diverse ethnic groups, ergo available in various languages. Racist expressions are a subgroup of generally offensive content, other subgroups include Sexism, misogyny, hate speech, amongst many others. These contents typically exist in the form of pictures, videos, text and in some cases, audio.

This project will however, only consider comments made in English language as well as focus on text based racist expressions often seen in the comment sections on Instagram, YouTube and TikTok, posts on Facebook, and tweets on Twitter.

Racist ideas can be expressed in a variety of ways, there is no clear definition of what exactly constitutes a racist utterance; what is racist to one person is highly likely to not be considered racist

universally [5]. Derald et al [11] describes racial microaggressions as the new face of racism. Microaggressions may sometimes be subtle and constructive making it difficult for even humans to detect. Although only racial utterances that constitute violence are illegal, any form of racism is unacceptable. Consequently, the data used in this project will contain racial microaggressions as well as racist contents disguised in humor.

1.6. Definition of terms

It is important to have an understanding of some of the terms and keywords that will be recurring in this report.

Race

/reɪs/

A term used to group categorize humans based on shared physical or social qualities into categories generally viewed as distinct by society.

Racism

/'reɪsɪz(ə)m/

Also called Racialism is the belief that humans may be divided into separate and exclusive biological entities called “races”; that there is a causal link between inherited physical traits and traits of personality, intellect, morality, and other cultural and behavioral features; and that some races are innately superior to others. The term is also applied to political, economic, or legal institutions and systems that engage in or perpetuate discrimination, prejudice, or antagonism by an individual, or racial group against a person or people on the basis of race or

otherwise reinforce racial inequalities in wealth and income, education, health care, civil rights, and other areas.

Microaggression

/mʌkrəʊəˈɡreɪ(ə)n/

A comment, action, or incident that subtly and often unconsciously or unintentionally expresses a prejudiced attitude toward a member of a marginalized group such as a racial or ethnic minority.

Social Media

A term used to generally describe forms of electronic communication such as websites for social networking and microblogging through which users create online communities to share information, ideas, personal messages, and other content.

Machine Learning

the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so.

Supervised Learning

The machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.

Natural Language Processing

NLP is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular; how to program computers to process and analyze large amounts of natural language data. The result is a computer capable of understanding the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

Naïve Bayesian Classifier

A family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features.

Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes

Support Vector Machines

Also support vector networks, are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.

1.7. Thesis Outline

The remaining part of this report is argued as follows:

Chapter 2 will contain an extensive review of the past works done by other researchers on this or a closely related subject matter such as hate speech, sexist, and general abusive language classification. It also maps the methodological paradigms that researchers have employed to solve this problem and discusses the successes and challenges that they have encountered.

Chapter 3 discusses the methodology used in the study, highlighting the text strategies deployed in gathering the corpus, and the criterium used in labelling the comments. This chapter will also discuss the data cleaning and natural language processing procedures used to prepare the dataset.

Chapter 4 will focus on the implementation of the methodologies highlighted in chapter 3. The theoretical and mathematical background of the classifiers to be used will be discussed in this chapter. Also contained in this chapter is an itemized list of software tools and libraries used throughout the project. Images and tables will be attached in this chapter to foster a practical understanding of the solution.

Finally, chapter 5 concludes this report and briefly summaries the entire work. The results of the implementation will be reported in this chapter and a brief comparison of the classifiers used in this study is made using tables and other demographic representations. To conclude this chapter,

the achievements and challenges of this study will be presented, and future recommendations will be discussed based on the reported results.

CHAPTER TWO

2. Literature Review

Fighting abusive online contents have become very important in a society where social media plays a fundamental role in shaping the ideologies of people. Racist comments, hate speech and offensive language are all subgroups of generally abusive online contents. There has been notably significant research on the racial discourse among scholars, Ariadna et al [5] mapped and discussed the recent developments in the study of racism and hate speech in the subfield of social media research. Systematically examining over a hundred articles, they addressed three research questions: Which geographical contexts, platforms, and methods do researchers engage with in studies of racism and hate speech on social media? To what extent does scholarship draw on critical race perspectives to interrogate how systemic racism is (re)produced on social media? And finally, what are the primary methodological and ethical challenges of the field? They found a lack of geographical and platform diversity, an absence of researchers' reflexive dialogue with their object of study, and little engagement with critical race perspectives to unpack racism on social media.

Although not so much literature exists on racist speech detection on social media, the detection of other forms of offensive speech is not so different in principle from detecting racist speech since they all deal with identifying speech that meet a certain predefined condition. Hence, in the review below, we include works that seek to detect offensive speeches with emphasis on how they perform on test data. We discuss the related works along two themes; those based on machine learning and those based on deep learning.

2.1. Machine Learning

Joni S. et al in [13] collected a total of 197,566 comments from four social media platforms: YouTube, Reddit, Wikipedia, and Twitter. 80% of the comments were labeled as non-hateful and the remaining 20% labeled as hateful. They experimented using several classification algorithms:

Logistic Regression, Naïve Bayes, Support Vector Machines, XGBoost, Neural Networks and feature representations such as Bag-of-Words, TF-IDF, Word2Vec, BERT, and their combination. In this study, it was observed that while all the models significantly outperform the keyword-based baseline classifier, XGBoost using all features performed the best at $F1 = 0.92$. Feature importance analysis indicates that BERT features were the most impactful for the predictions.

Another study by Marzieh M. et al in [14] introduced a transfer learning approach for hate speech detection based on the existing pretrained language model - BERT (Bidirectional Encoder Representations from Transformers) and evaluated the proposed model on two publicly available datasets that were annotated for racism, sexism, hate or offensive content on Twitter. Next, a bias alleviation mechanism was used to mitigate the effect of bias in training set during the fine-tuning of the pre-trained BERT-based model for hate speech detection. Toward that end, an existing regularization method was used to reweight input samples, thereby decreasing the effects of high correlated training set's n -grams with class labels, and then fine-tuned the pre-trained BERT based model with the new re-weighted samples. To evaluate the bias alleviation mechanism, they employed a cross-domain approach where they used the trained classifiers on the aforementioned datasets to predict the labels of two new datasets from Twitter, AAE-aligned and White-aligned groups, which indicated tweets written in African- American English (AAE) and Standard American English (SAE), respectively. The results from this study show the existence of systematic racial bias in trained classifiers, as they tend to assign tweets written in AAE from AAE-aligned groups to negative classes such as racism, sexism, hate, and offensive more often than tweets written in SAE from White-aligned groups. However, the racial bias in the classifiers reduced significantly after the bias alleviation mechanism was incorporated.

2.2.Deep Learning

In recent years, researchers have shifted attention towards deep learning approaches. Many of the works have deployed convolutional neural networks, an example can be seen in [15], where Skreekanth M. et al proposed an ensemble-based system to classify an input post into one of three classes: Overtly Aggressive, Covertly Aggressive, and Non-aggressive. In this approach, three deep learning methods, namely, Convolutional Neural Networks (CNN) with five layers - Input, convolution, pooling, hidden, and output, Long Short Term Memory networks (LSTM), and Bi-directional Long Short Term Memory networks (Bi-LSTM) were used. A majority voting-based ensemble method was deployed to combine the three classifiers (CNN, LSTM, and Bi-LSTM). The model was trained on a Facebook comments dataset and tested on both Facebook comments (in-domain) and other social media posts (cross-domain). Their model achieved a weighted F1-score of 0.604 for Facebook posts and 0.508 for other social media posts.

Suvadip P. et al in [16] trained an end-to-end deep learning model to classify a given Tweet as racist, sexist or neither with a certain degree of tolerance. They explored a two-step approach of performing classification on abusive language and then classifying into the specific subgroups of abusive language and finally compared it with a one-step approach of doing one multi-class classification for detecting sexist and racist languages.

Three CNN-based models were implemented to classify sexist and racist abusive language: CharCNN, WordCNN, and HybridCNN. The major difference among these models were whether the input features were characters, words, or both. Each convolutional layer computed a one-dimensional convolution over the previous input with multiple filter sizes and large feature map

sizes. Maxpooling was then performed after the convolution to capture the feature that was most significant to the output.

Similarly, in [23] a public English Twitter dataset of 20,000 tweets in the type of sexism and racism, their approach showed a promising performance of 0.827 F-measure by using HybridCNN in one-step and 0.824 F-measure by using the character n-gram logistic regression and support vector machines in two-steps. The model surpassed current state of the art performance on the dataset (for end-to-end deep learning models), reaching a weighted macro F1 score of 0.85 and an AUROC of 0.91, primarily through improved model architecture and representation. Prior to this work, all neural models used thus far only achieved a 0.81 weighted macro F1 score at most on the dataset, rather than other possible traditional machine learning algorithms which have been shown to perform better when coupled with embeddings generated from neural networks.

The model showed improvements in performance and a significant advantage of using the proposed deep learning models. However, they observed that the dataset was very noisy, around 20% of clean tokens generated from the dataset did not have pre-trained word embeddings due to online and twitter-based idiosyncrasies, highlighting the nature of conversational language on Twitter and the social media cyberspace as a major challenge in racist and sexist classification.

2.3.Context Aware Models for Comment Classification

Beyond user comments, Lei G. in [24] examined context dependent comments. In this work, they presented an annotated corpus of hate speech with context information well kept, then they proposed two types of hate speech detection models that incorporate context information, a logistic regression model with context features and a neural network model with learning components for context. The data corpus consisted of 1528 annotated Fox News User comments, 435 labeled as hateful that were posted by 678 different users in 10 complete news discussion threads in the Fox News website. Their evaluation shows that both models outperform a strong baseline by around 3% to 4% in F1 score and combining these two models further improve the performance by another 7% in F1 score.

Deep Context-Aware Embedding has been used in [25] for the detection of hate speech and abusive language on twitter. To improve the classification performance, they enhanced the quality of the tweets by considering polysemy, syntax, semantic, OOV words as well as sentiment knowledge and concatenated to form input vectors. BiLSTM was used with attention modeling to identify tweets with hate speech.

2.4.Aggressive Comment Detection in Multilingual Corpus

Antagonistic contents propagated via social media networks have the potential harm and suffering on the individual and country levels and escalate to social tension and disorder beyond the cyberspace. Social media is available to people of diverse ethnic groups, ergo available in various languages. Consequently, aggressive comments are a multilingual problem. Research by Areej A.

et al in [17] presented some challenges and recommendations for the Arabic hate speech detection problem.

Raghad A. et al in [18] experimented with several neural network models based on convolutional neural network (CNN) and recurrent neural network (RNN) to detect hate speech in Arabic tweets. They also evaluated a language representation model - BERT (Bidirectional Encoder Representations from Transformers) on the task of Arabic hate speech detection. In this study, a new hate speech dataset that contained 9316 annotated tweets was compiled. Then, they conducted a set of experiments on two datasets to evaluate four models: CNN, gated recurrent units (GRU), CNN + GRU, and BERT. The result of this experiment on the dataset and an out-domain dataset showed that the CNN model gave the best performance, with an F1-score of 0.79 and area under the receiver operating characteristic curve (AUROC) of 0.89.

Another paper [19] addressed the issue by building a text analytics model with machine learning that can be used to filter racist comments in Sinhala language. A Two-Class Support Vector Machine was trained with a set of carefully chosen comments from Facebook that were labelled as racist and non-racist based on intent. The trained model was then able to classify racist comments with a 70.8% accuracy in their experimental results.

A database of comments was constructed by collecting random Sinhala language comments that appeared on public social media (Facebook) posts and were annotated by giving labels as ‘racist’ or ‘non-racist’ based on the intent of the comment.

In Belgium, Stephan T. et al in [20] published another research on racist comment detection in the dutch cyberspace. They presented a dictionary-based approach to racism detection in Dutch social media comments, which were retrieved from two public Belgian social media sites likely to attract racist reactions. These comments were labeled as racist or non-racist by multiple annotators. For this approach, three discourse dictionaries were created: first, they created a dictionary by retrieving possibly racist and more neutral terms from the training data, and then augmenting these with more general words to remove some bias. A second dictionary was created through automatic expansion using a word2vec model trained on a large corpus of general Dutch text. Finally, a third dictionary was created by manually filtering out incorrect expansions. Finally, they trained multiple Support Vector Machines, using the distribution of words over the different categories in the dictionaries as features.

The best-performing model used the manually cleaned dictionary and obtained an F-score of 0.46 for the racist class on a test set consisting of unseen Dutch comments, retrieved from the same sites used for the training set. The automated expansion of the dictionary only slightly boosted the model's performance, and this increase in performance was not statistically significant. They argued that the fact that the coverage of the expanded dictionaries did increase indicated that the words that were automatically added did occur in the corpus but were not able to meaningfully impact performance.

Later in [21], the same authors presented two experiments on the automated detection of racist discourse in Dutch social media. In both experiments, multiple classifiers are trained on the same training set. This training set consists of Dutch posts retrieved from two public Belgian social

media pages. The posts were labeled as racist or non-racist by multiple annotators, who reached an acceptable agreement score. The different classification models all used the Support Vector Machine algorithm, but used different sets of linguistic features, which can be lexical, stylistic or dictionary based, as in [20]. In the first experiment, the models are evaluated on a test set containing unseen comments retrieved from the same pages as the training set (and thus also skewed towards racism). In the second experiment, the same models from the first experiment were tested on an alternative test set, containing more neutral comments, retrieved from the social media page of a Belgian newspaper.

In both experiments, the best performing model relies on a dictionary containing different word categories specifically related to racist discourse. It reaches an F-score of 0.47 in the first experiment, and 0.40 in the second for the racist class and ROC Area Under Curve scores of 0.64 (experiment 1) and 0.73 (experiment 2).

Thomas D. et al in [22] suggested that lexical detection methods tend to have low precision because they classify all messages containing particular terms as hate speech and previous works using supervised learning have failed to distinguish between the two categories. They used a crowd-sourced hate speech lexicon to collect tweets containing hate speech keywords. Crowdsourcing was used to label a sample of these tweets into three categories: those containing hate speech, only offensive language, and those with neither. Then, they trained a multi-class classifier to distinguish between these different categories.

Close analysis of the predictions and the errors in their experiment, shows when the model could reliably separate hate speech from other offensive language and when this differentiation was more difficult. They found that racist and homophobic tweets are more likely to be classified as hate speech but that sexist tweets are generally classified as offensive. However, tweets without explicit hate keywords are more difficult to classify.

CHAPTER THREE

Methodology

3.0. Introduction

This chapter discusses the methodology used in the study, highlighting the strategies deployed in gathering the corpus, and the criterium used in labelling the comments. This chapter will also discuss the data cleaning and natural language processing procedures used to prepare the dataset.

3.1. Analysis of Existing Systems

Based on the literature review, many advancements and successes have been recorded using both machine learning and deep learning approaches in similar classification and detection problems based on a given particular dataset. However, even though the datasets used in these researches have been divided into training and testing sets to ensure that the model has not overfitted on the training data, there still remains a high possibility that the model was overfitting on that particular dataset, making it unreliable when given a totally different dataset.

One of the reasons for this is the intrinsic bias of the datasets used in training these models. Due to the lack of a definitive or deterministic structure of racist language, many data mining strategies used to gather data for this task are likely to fail in collecting the much-needed versatile data. For example, comments without explicit racist keywords will be ignored if a keyword search data mining algorithm is used to gather the data. In modern racist discuss, racist comments on social media can be more constructive and difficult to detect, some even buried under the disguise of dark humor. Consequently, the systems that have been implemented for the detection of such

comments by social media companies are notorious for ignoring racist comments disguised as humor.

3.2. Justification for Proposed System

In order to address the problem discussed, the dataset used in this project covers different types of racist comments including racist microaggressions as they have the same psychological and physiological effects as racist comments. The data was collected from multiple social media platforms such as Facebook, Instagram, Twitter, YouTube, and TikTok. To remove the biases and lack of versatility introduced by text mining methods, I have employed a more tedious method of manually gathering a dataset of 2,000 comments, 1,000 of which are annotated as racist comments and the other 1,000 non-racist comments.

3.3. The Proposed System

The proposed system introduced three machine learning baseline models for the detection of racist comments across social media platforms trained using data collected from popular social media sites and built using three machine learning classifiers infamously used in text classification problems. First the Naïve Bayesian classifier, then Logistics Regression, and finally Support Vector Machines.

3.4. Gathering the Dataset

The strategy used to gather the data was to handpick racist comments from social media platforms by following real-life events that were likely to attract racist reactions. One of such events occurred

in 2010 when Rima Fakhri Slaiby, a Lebanese American model was crowned Miss America. White supremacists who were not so excited by the idea of an Arab descent being Miss America, took to social media to express their disappointment leaving a long trail of racist comments.

2008 and 2012 presented another trail of racist comments to be followed after Barack Hussein Obama of Ethiopian descent defeated fellow aspirant John Sidney McCain in 2008 at the United States presidential election to become the first African American president of the United States and was reelected in 2012 for a second term. Many racist comments were found on social media against him during both campaigns and while he was in office.

Another such event was the “Black Lives Matter” protest in 2020 which occurred in the middle of a worldwide COVID-19 pandemic after the tragic murder of African American George Perry Floyd Jr which many believed was racist incited. Angry protesters who were not pleased by the actions of the police and the response of the government on the matter were seen on the streets all over the United States and several other nations calling for an end to racism and police brutality. Some police departments responded by using force to displace these protesters leaving many injured and some arrested for breaking COVID-19 safety guidelines, this however resulted in further escalation. Another group of individuals disguised as protesters took the opportunity to vandalize both government and private properties and looted several stores, large and struggling businesses included. As expected, social media was the place to get news and share thoughts of the ongoing situation. Store owners whose stores had been vandalized and looted were less discreet about their racial prejudices. Naturally, this event left a long trail of racist comments and microaggressions.

The comment section of footballers after a bad performance at a game was another vital premise to gather the racist comment dataset from. I noticed that footballers who are racialized as non-white are likely to become victims of online racist abuse after a bad game. Some examples are African European Manchester United player Anthony Martial who was racially abused on social media platforms after his side's 1-1 draw against West Brom, Marcus Rashford after missing a penalty against Crystal Palace, and several other footballers such as Paul Pogba, Wilfred Zaha, Yan Dhanda, Axel Tuanzebe, Raheem Sterling, Antonio Rudiger, Romelu Lukaku and many others, all victims of racial abuse simply for belonging to a different racial group.

Social media influencers who have started a trend on “What’s the most racist thing you’ve ever been told” presented a unique opportunity to gather racist comments, most of which were disguised in humor. Racist comments were also found scouring through the post comment section of celebrities who have been victims of racial abuse. Finally, the dataset was completed by collecting racist comments from anonymous social media profiles who generally refer to themselves as racist police. They scout for racist comments on the social media platforms and post a screenshot of such comments with the username of the user who posted it in display as some form of social justice, though misguided as this tends to result in more people sending abusive comments and messages to these users.

Comments	Label
if you voteobama youre just another nigger enabler how bout jiggaboos learn somethin bout hard work instaed of welfare and food stamps gop	1
yes obama needs to stay in the jungle and biden need to stay in jail	1
I really want to meet obama someday so I can call him a nigger	1
How does an ugly Arab bitch win miss America. Pathetic!!!	1
This what years of allowing foreign invaders has caused. We even have a nigger president	1
whites helpline facilitate the invasion of their own countries traitors have no soul they should be burned at the stake	1
why does a gorilla always have a frown, because in ten million years it will be a nigger	1

Fig 3.1.1. Some examples from the racist comment dataset

3.5.Labeling the Dataset

A supervised learning approach was used. Each comment on the dataset was labeled 1 or 0, 1 signifying that it is a racist comment and 0 signifying that it is a non-racist comment. Since the data gathering was manual, it was fairly easy to label the racist comments. However, for non-racist comments, regular non-offensive comments, as well as offensive, aggressive, sexist, and misogynistic comments which are not racist were used to build up the non-racist dataset. The particular reason for this action is to ensure that the model is not basing its decisions on the wrong keywords rather than the actual racist ones. For example, from the dataset, it can be observed that most racist comments occur with offensive keywords even though offensive comments are not necessarily racist.

3.6.Processing the Dataset

To get the best results from the model, some natural language processing techniques were applied to the data to improve its quality.

3.6.1. Removing Punctuations

The punctuations present in the data do not add any value or meaning to the data. Instead, the punctuation when attached to any word, will create a problem in differentiating it from other words.

3.6.2. Case Normalization

This process converts all the characters in the dataset to lowercase as most programming languages are case sensitive and will treat “Hello” and “hello” differently even though they mean the same thing.

3.6.3. Tokenization

Datasets are often a table of strings that the computer cannot understand. Tokenization is a way of separating strings into smaller units called Tokens. Here, these tokens are words. The result of this process is a list of words i.e. Each sentence is a list of words.

3.6.4. Removing Stop Words

Stop words include “and, I, he, she, but, was, were, etc.” which do not add meaning to the data as they will not help in identifying whether a comment is racist or not.

3.6.5. Lemmatization and Stemming

Lemmatizing is the process of reducing a word to its root form. The main purpose is to reduce variations of the same word, thereby reducing the number of words that will be included in the model. The reason I have used this approach instead of stemming even though it is faster than lemmatizing is that stemming removes the end of a word without considering the context of the word. Whereas lemmatizing considers the context of the word and shortens it into its root form based on its dictionary definition.

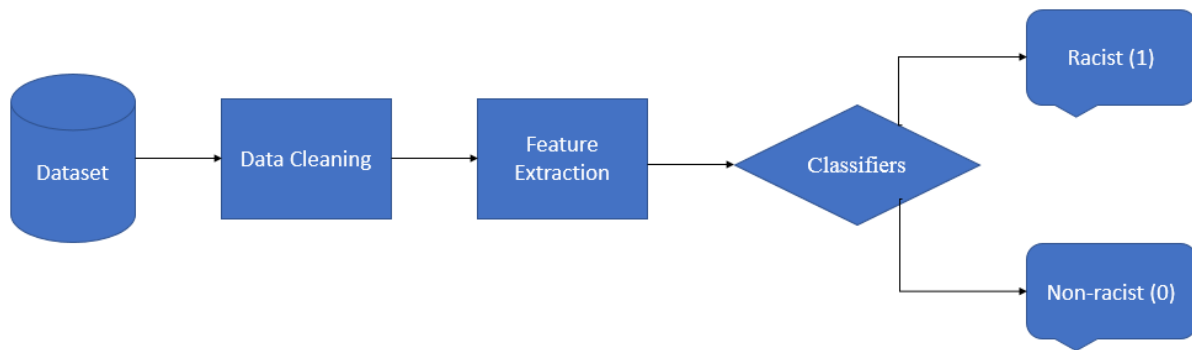


Fig. 3.1.2. Flowchart of Procedures for Comment Classification

CHAPTER FOUR

Implementation

4.0. Introduction

This chapter focuses on the experimental setup, development, training, and testing of the models for racist comment classification based on the data corpus discussed in chapter 3. It provides an overview of the choice of platform, programming language, tools, libraries, feature extraction technique, software and hardware requirements, and the performance of the classifiers used.

4.1. Implementation Tools

The models were built using the python programming language in Google Collaboratory Notebook, a product of google research that allows users to write and execute python codes in a cell-oriented paradigm through the browser. It is a free Jupyter Collaboratory Notebook environment that runs on google cloud servers and supports popular and powerful python libraries without requiring tedious setup and configuration. Some of the libraries used in this project include Pandas, NumPy, Scikit-Learn, and NLTK.

4.1.1. Pandas

Pandas is a fast and efficient data frame object for data manipulation and analysis with integrated indexing, data structures, and operations for manipulating numerical tables, Pandas allows importing data from various file formats such as comma-separated values CSV, JSON, SQL, and Microsoft Excel.

4.1.2. NumPy

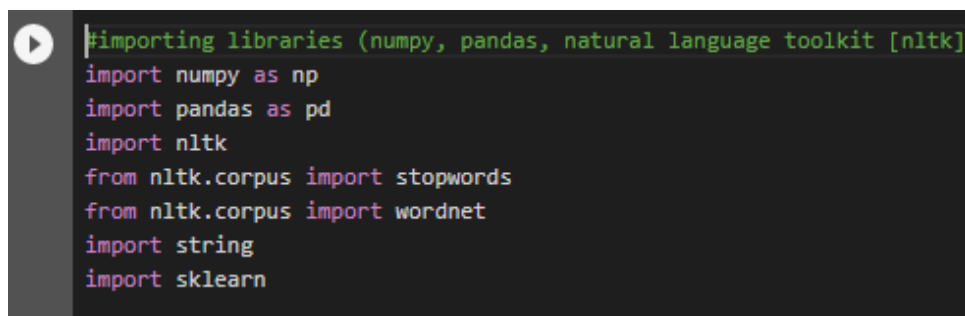
NumPy is a python library that adds support for large, multi-dimensional arrays and matrices along with a large collection of high-level mathematical functions to operate on these arrays.

4.1.3. Scikit-Learn

Also “*sklearn*”, is a machine learning library that features various classification, regression, clustering, and performance evaluation algorithms that were of utmost importance in this project.

4.1.4. Natural Language Tool Kit (NLTK)

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language. Libraries in the NLTK suite that were used in this project include “*stopwords*” for removing stop words and “*wordnet*” for lemmatization.



```
#importing libraries (numpy, pandas, natural language toolkit [nltk])
import numpy as np
import pandas as pd
import nltk
from nltk.corpus import stopwords
from nltk.corpus import wordnet
import string
import sklearn
```

Fig 4.1. Importing Libraries

4.2. Components of Implemented Models

The data corpus used in this project is a dataset of 2000 comments, a combination of 1000 racist comments and 1000 non-racist comments manually gathered from social media networks and labeled 1 for racist and 0 for non-racist comments. However, after removing all the missing rows (NaN rows, which are very common in real-life data analysis) the dataset was reduced to 1988 comments in total.

```
<bound method DataFrame.dropna of                                     Comment  Label
601  FB loves niggers, kikes, faggots, dykes and al...          1
925  There are many reasons to believe that refugee...          1
1317                illness is fucking up this country          0
1028                i need a trippy bitch who fuck on Hennessy    0
1034  im done with bitter bitches its a wrap for tha...          0
...                ...                ...
870  if you see a 539 check your pocket, they are a...          1
1542  if you think about it, trans people are the re...          0
1776                REAL HOT GIRL SHIT                          0
1975                cotton candy duo                            0
230  damn this comment section is whiter than snow ...          1

[1988 rows x 2 columns]>
```

Fig 4.2. The dataset

4.2.1. Text Processing

In order to prepare the dataset for training, text processing techniques available in the NLTK natural language processing package were used to apply the processing techniques explained in chapter 3.

```
601  [FB, love, nigger, kike, faggot, dyke, othe, s...
925  [many, reason, believe, refugee, burden, u]
1317  [illness, fucking, country]
1028  [need, trippy, bitch, fuck, Hennessy]
1034  [im, done, bitter, bitch, wrap, angry, bird, t...
Name: Comment, dtype: object
```

Fig 4.2.1. Processed data

4.2.2. Feature Extraction

In textual form, the computer would be unable to process the data corpus. After processing the data, each sentence has been reduced to a list of words. These words have to be encoded as integers or floating-point values for them to be just as input in a machine learning classifier. For the purpose of this project, the CountVectorizer package was used to convert the textual tokens to a vector of token counts i.e., reducing each token to a vector representing its frequency of appearance.

```
[ ] #Convert text to matrix
from sklearn.feature_extraction.text import CountVectorizer
bow = CountVectorizer(analyzer=processCorpus).fit_transform(df['Comment'].values.astype('U'))
```

Fig 4.2.2. Applying CountVectorizer on the Tokenized data

4.2.3. Splitting the Dataset

It is common and good practice to split the dataset into training and validation sets. One primary importance of this is to ensure that the model does not overfit the training set. It is for this reason only 80% of the data corpus was used in training the models while the other 20% was used to validate the performance of the model.

```
[ ] #Split data into training and testing dataset (test 20%, train 80%)
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(bow, df['Label'], test_size = 0.20, random_state = 0)
```

Fig 4.2.3. Splitting the dataset

4.2.4. Training the Models

Three machine learning classification algorithms available in the scikit-learn package, naïve Bayesian classifier, logistic regression, and support vector machines were used to train the models.

4.2.4.1. Naïve Bayesian Classifier

The multinomial naïve bayes algorithm was used to fit the training data. Multinomial Naive Bayes algorithm is a probabilistic machine learning method, commonly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.

Bayes theorem, proposed by Thomas Bayes is calculated using the mathematical formula:

$$\mathbf{P(A|B) = P(A) * P(B|A) / P(B)}$$

Where the probability of class A is being calculated when predictor B is already provided and:

P(A) = Prior probability of class A (The class).

P(B) = Prior probability of B (The word).

P(A|B) = Occurrence of predictor B given class A probability.

The feature extraction process implemented creates a “bag-of-words” containing the frequency of each occurring word in each comment. This is then used by the algorithm to obtain the prior probabilities of each word given its appearance in either class.

Fig 4.2.4.1 below shows that the model reports a 0.95 F1 score on both racist and non-racist comment classification with a precision of 0.96 and 0.94 on nonracist and racist classification respectively. The model reports an accuracy score of ≈ 0.95 (95%) on training data.

	precision	recall	f1-score	support
0	0.96	0.93	0.95	782
1	0.94	0.96	0.95	808
accuracy			0.95	1590
macro avg	0.95	0.95	0.95	1590
weighted avg	0.95	0.95	0.95	1590
Confusion Matrix:				
[[731 51]				
[32 776]]				
Accuracy score: 0.9477987421383648				

Fig 4.2.4.1. Naïve Bayesian Classifier Training Performance Report

4.2.4.2. Logistic Regression

In the second model, a binary logistic regression classifier was used to fit the trained data. Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, which means there would be only two possible classes (1 or 0, or in this case, racist or non-racist). Mathematically, a logistic regression model predicts $P(Y=1)$ as a function of X . The mathematics beyond this point becomes quite complicated.

Fig 4.2.4.2 shows that the model reports a 0.97 F1 score and precision, respectively on both racist and non-racist comment classification. The model reports an accuracy score of ≈ 0.97 (97%) on training data, an improvement from the multinomial naïve bayes classifier.

	precision	recall	f1-score	support
0	0.97	0.97	0.97	782
1	0.97	0.97	0.97	808
accuracy			0.97	1590
macro avg	0.97	0.97	0.97	1590
weighted avg	0.97	0.97	0.97	1590
Confusion Matrix:				
[[758 24]				
[26 782]]				
Accuracy score: 0.9685534591194969				

Fig 4.2.4.2. Logistic Regression Training Performance Report

4.2.4.3. Support Vector Machine

The last model was trained using the linear Support Vector Machine (SVM) classifier. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate an n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called “support vectors”. The linear kernel was selected in this case as the dataset is a 2-dimensional space and linearly separable. All other values were left at default.

Fig 4.2.4.2 shows that the model reports a 0.99 F1 score and precision, respectively on both racist and non-racist comment classification. The model reports an accuracy score of ≈ 0.99 (99%) on training data, an improvement from both the multinomial naïve bayes and Logistics regression classifiers.

	precision	recall	f1-score	support
0	0.99	0.99	0.99	782
1	0.99	0.99	0.99	808
accuracy			0.99	1590
macro avg	0.99	0.99	0.99	1590
weighted avg	0.99	0.99	0.99	1590
Confusion Matrix:				
[[774 8]				
[8 800]]				
Accuracy score: 0.989937106918239				

Fig 4.2.4.3. Support Vector Machines Training Performance Report

4.3. Hardware and Software Specification

The hardware and software used in this project are listed below:

Hardware Specification

- i. 8GB RAM (Random Access Memory)
- ii. Intel Core i7-3517U
- iii. Keyboard and Mouse
- iv. 256GB HDD (Hard Disk Drive)

Software Specification

- i. Windows 8.1
- ii. Google Chrome Browser
- iii. Google Collaboratory Workbook
- iv. Programming Language: Python
- v. MS Excel

4.4. Model Validation

To ensure that the models have not overfitted the training data, all the models were validated using the test data to see just how well they do with new data.

The confusion matrix is useful to understand how accurate the prediction of the model is for the testing data. A Confusion matrix is an $N \times N$ matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.

For a binary classification such as this, the confusion matrix would be a 2×2 matrix with four values: True Positive, False Positive, True Negative, and False Negative.

Table 4.4.1 Confusion Matrix Explained

NON- RACIST (0)	TP	FN
RACIST (1)	FP	TN

Where:

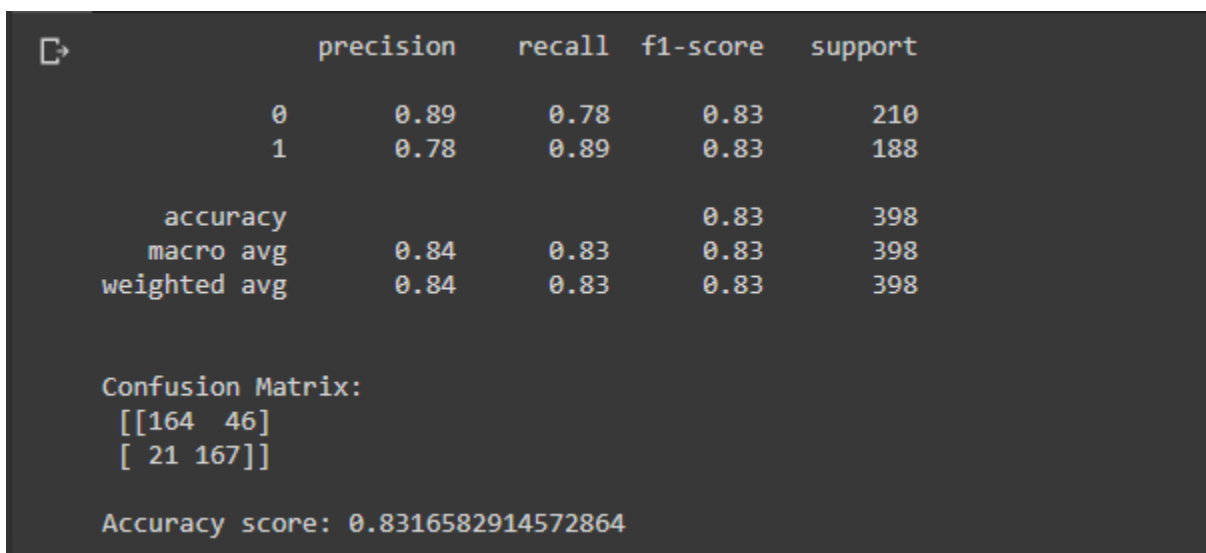
True Positive (TP): The predicted value matches the actual value i.e., the actual value was positive, and the model predicted a positive value.

False Positive (FP) – Type 1 error: The predicted value was falsely predicted i.e., the actual value was negative, but the model predicted a positive value. This is also known as the Type 1 error.

True Negative (TN): The predicted value matches the actual value i.e., the actual value was negative, and the model predicted a negative value.

False Negative (FN) – Type 2 error: The predicted value was falsely predicted i.e., the actual value was positive, but the model predicted a negative value. This is also known as the Type 2 error.

Fig 4.4.1 shows that the multinomial naïve bayes trained model reports a 0.83 F1 score on both racist and non-racist comment classification with a precision of 0.89 and 0.78 on nonracist and racist classification respectively.

A terminal window with a dark background and light-colored text. It displays a table of model performance metrics, a confusion matrix, and an accuracy score. The table has columns for precision, recall, f1-score, and support. The confusion matrix is a 2x2 array. The accuracy score is a decimal value.

	precision	recall	f1-score	support
0	0.89	0.78	0.83	210
1	0.78	0.89	0.83	188
accuracy			0.83	398
macro avg	0.84	0.83	0.83	398
weighted avg	0.84	0.83	0.83	398

Confusion Matrix:
[[164 46]
[21 167]]

Accuracy score: 0.8316582914572864

Fig 4.4.1. Multinomial Naïve Bayes Trained Model Validation Performance Report

The confusion matrix shows that the model has classified 164 non-racist comments corrected and 46 incorrectly, it has also classified 167 correctly and 21 incorrectly. The model reports an accuracy score of ≈ 0.83 (83%) on validation.

Table 4.4.2. Confusion Matrix of Multinomial Naïve Bayes Trained Model Validation

NON- RACIST (0)	164	46
RACIST (1)	21	167

Fig 4.4.2 below shows that the logistics regression trained model reports a 0.88 and 0.86 F1 score with a precision of 0.86 and 0.88 on nonracist and racist classification respectively.

	precision	recall	f1-score	support
0	0.86	0.90	0.88	210
1	0.88	0.84	0.86	188
accuracy			0.87	398
macro avg	0.87	0.87	0.87	398
weighted avg	0.87	0.87	0.87	398
Confusion Matrix:				
[[189 21]				
[31 157]]				
Accuracy score: 0.8693467336683417				

Fig 4.4.2. Logistic Regression Trained Model Validation Performance Report

The confusion matrix shows that the model has classified 189 non-racist comments corrected and 21 incorrectly. it has also classified 157 correctly and 31 incorrectly.

Table 4.4.3. Confusion Matrix of Logistic Regression Trained Model Validation

NON- RACIST (0)	189	21
RACIST (1)	31	157

It is clear from the confusion matrix that the logistic regression trained model made more incorrect racist comment classifications than the naïve bayes trained model even with a higher accuracy score of ≈ 0.87 (87%) on validation.

However, Fig 4.4.3 below shows that the linear kernel support vector machine trained model reports a 0.89 and 0.88 F1 score with a precision of 0.89 and 0.87 on nonracist and racist classification respectively.

	precision	recall	f1-score	support
0	0.89	0.88	0.89	210
1	0.87	0.88	0.88	188
accuracy			0.88	398
macro avg	0.88	0.88	0.88	398
weighted avg	0.88	0.88	0.88	398
Confusion Matrix:				
[[185 25]				
[22 166]]				
Accuracy score: 0.8819095477386935				

Fig 4.4.3. Support Vector Machine Trained Model Validation Performance Report

The confusion matrix shows that the model has classified 185 non-racist comments corrected and 25 incorrectly. it has also classified 166 correctly and 22 incorrectly.

Table 4.4.4. Confusion Matrix of Support Vector Machine Trained Model Validation

NON- RACIST (0)	185	25
RACIST (1)	22	166

The matrix shows that the SVM trained model made more incorrect non-racist comment classifications than the logistic regression trained model but outperforms the naïve bayes trained model. More so, it makes more racist comment misclassifications than the naïve bayes trained

model while outperforming the logistic regression trained model. However, the overall accuracy score of this model surpasses the other two with an accuracy score of ≈ 0.88 (88%) on validation.

Table 4.4.5. Summary of Results

Classifier	Precision	Recall	F1 score	Accuracy (%)
Naïve Bayes (Multinomial)	0.78	0.89	0.83	83.17
Logistic Regression	0.88	0.84	0.86	86.93
SVM (Linear kernel)	0.87	0.88	0.88	88.19

Table 4.4.5. show the summary of results on all the classifiers on solely racist comment detection on the validation data.

CHAPTER FIVE

Conclusion

This chapter briefly summaries the entire work. The results of the implementation are discussed in this chapter and a brief comparison of the classifiers used in this study are made. To conclude this chapter, the achievements and challenges of this study will be presented, and future recommendations will be discussed based on the reported results.

In this work, I have proposed 3 models to identify racist social media comments using machine learning algorithms. Racist discourses and practices on social media represent a vital, yet challenging area of research. With race and racism increasingly being reshaped within proprietary platforms like Facebook, Twitter, Instagram, and YouTube. The proliferation of racial biases and discrimination spread on social media through comments propagate the ideology of racial domination, exposing the members of the discriminated racial or ethnic group to physiological and psychological distress.

Many advancements and successes have been recorded using both machine learning and deep learning approaches in similar classification and detection problems. However, even though the datasets used in these research have been divided into training and testing sets to ensure that the model has not overfitted on the training data, there remains a high possibility that the model was overfitting on that particular dataset, making it unreliable when given a totally different dataset.

One of the reasons for this is the intrinsic bias of the datasets used in training these models. Due to the lack of a definitive or deterministic structure of racist language, many data mining strategies used to gather data for this task are likely fail in collecting the much-needed versatile data. For example, comments without explicit racist keywords will be ignored if a keyword search data

mining algorithm is used to gather the data. In modern racist discuss, racist comments on social media can be more constructive and difficult to detect even for humans, some even buried under the disguise of dark humor. Consequently, the systems that have been implemented for the detection of such comments by social media companies are notorious for ignoring racist comments disguised as humor.

To address this problem, the dataset used in this project covers different types of racist comment including racist microaggressions as they have the same psychological and physiological effects as racist comments. The data was collected from multiple social media platforms such as Facebook, Instagram, Twitter, YouTube, and TikTok. To remove the biases and lack of versatility introduced by text mining methods, a more tedious method was used to manually gathering a dataset of 2,000 comments, 1,000 of which are annotated as racist comments and the other 1,000 non-racist comments. Three different models are then trained and tested with 80% and 20% respectively of the data corpus using three machine learning classifiers infamously used in text classification problems. First the Naïve Bayesian classifier, then Logistic Regression, and finally Support Vector Machines.

The results obtained from this project in Table 4.4.5 shows that the support vector machine trained model performs the best with an accuracy of 88.19%. The Confusion matrix in Table 4.4.4 shows that the model made more incorrect non-racist comment classifications than the logistic regression trained model but outperforms the naïve bayes trained model. More so, it makes more racist comment misclassifications than the naïve bayes trained model while outperforming the logistic regression trained model. The logistic regression model is the next best with an accuracy of 86.93%. the multinomial naïve bayes model performed the least with an 83.17 accuracy score.

	Features	Classifier	Precision	Recall	F1-score
Stephan et al.	Dictionary-based, n-gram	SVM	0.24	0.02	0.04
Dulan et al.	TFID, n-gram	SVM	1.00	0.36	0.53
Stephan et al.	BOW	SVM	0.49	0.43	0.46
Leo et al.	Char, word, LIWC, NRC	LR	0.57	0.48	0.52
		LSTM	0.52	0.40	0.45
		Bi-LSTM	0.59	0.44	0.50
Suvadip et al.	TFIDF	LR	-	-	0.79
		SVM	-	-	0.81
Ji et al.	Maxpooling	LR	0.81	0.60	0.69
		SVM	0.82	0.53	0.64
		FastText	0.76	0.63	0.69
		CharCNN	0.69	0.75	0.72
		WordCNN	0.70	0.76	0.73
		HybridCNN	0.71	0.77	0.74
This Project	BOW	NB	0.78	0.89	0.83
		LR	0.88	0.84	0.86
		SVM	0.87	0.88	0.88

Table 5.0. Comparison of Existing Models and Proposed Models

Table 5.0 shows that the models proposed in this project outperformed most of the preexisting models for the same task. The primary reason for this success is the quality and vastness of the training data. Although it can be argued that the non-neural algorithms used for classification have performed their best due to the small size of the dataset.

5.1. Limitations

It is emphatically difficult and time-consuming to manually gather racist comments and likewise labeling them. Automatically generating the data has been shown in this project to affect the performance of the model on new data.

Several researchers have proposed several models and with each a new data collection method. The benchmark datasets commonly contain oddities that result in a high preference for the classification of some samples to a specific class. These greatly affect the versatility of the data that is collected and hence, the reliability of the model. For instance, if researcher A has collected a dataset of racist comments based on the explicit mention of the word “*Nigger*” or “*Nigga*”, even if this data is split and then trained and tested. The model will report a high performance on classification as it is dealing with the same dataset i.e., a similar kind of racist comment.

Furthermore, there is a fine line between protecting social media cyberspace from abusive users and restricting freedom of speech. This fact is the primary drawback in the deployment of these kinds of models by social media companies as misclassifications may be misconstrued as a civil right violation.

5.2. Recommendations

Some ways to tackle the limitations discussed in 5.1 above include cross-domain validation, the use of neural networks, and building context-aware models.

- i. **Cross-Domain Approach:** Evaluating models not solely based on the test data portion of the dataset used to train them but also on totally different datasets that have been gathered

using a different method from the dataset that was used in training will give an unbiased report on the true performance of the model.

- ii. **Neural Network:** Combining classifiers in a neural network has been shown to achieve better results in classification problems especially with larger datasets.
- iii. **Context-awareness:** One major reason for misclassification is the possibility for one comment to be of both classes based on its context. The context accompanying a comment can be useful in identifying whether it is racist or not.

REFERENCES

1. Wilson W.J, *The Bridge Over the Racial Divide: Rising Inequality and Coalition Politics*, University of California Press Berkeley CA, 1999.
2. Matthew Clair, Jeffrey S. Denis, *Racism, Sociology of*, Harvard University Cambridge MA USA, 2015.
3. Neshapriyan M, *Social Media, and Freedom of Speech and Expression*, Legal Service India, 2015.
4. Daniels Jessie, *Race and Racism in Internet Studies; A Review and Critique*, New Media and Society, 2013, pp 695 – 719.
5. Ariadna M, Johan F, *Racism, Hate Speech, and Social Media: A Systematic Review and Critique*, Television and New Media, 2021, pp 205 – 224.
6. David R, Ruth W, *Racism and Mental Health: The African American Experience*, Ethnicity and Health, 2010, pp 243 – 268.
7. Joanne Lewsley, *What are the Effects of Racism on Health and Mental Health*, 2020.
8. Yin Paradies, Jehonathan Ben, Nida Denson, Amanuel Elias, Naomi Priest, Alex Pieterse, Arpana Gupta, Margaret K, Gilbert Gee, Robert K. Hills, *Racism as a Determinant of Health: A Systematic Review and Meta-Analysis*, Plos One, 2015.
9. National Centre for Health Statistics, *Health, United States 2015: With Special Feature on Racial and Ethnic Health Disparities*, Hyattsville MD, 2016.
10. Janice Gassam Asare, *Social Media Continues to Amplify White Supremacy and Suppress Anti-Racism*, Forbes, 2021.
11. Derald Sue, Jenifer B, Annie L, Kevin L. Nadal, *Racial Microaggression and the Asian American Experience*, Culture, Diversity and Ethnic Minority Psychology, 2007, pp 72 – 81.

12. Derald W. Sue, Christina M. Capodilupo, Gina C. Torino, Jennifer M. Bucceri, Aisha M. Holder, Kevin L. Nadal, Marta Esquilin, *Racial Microaggressions in Everyday Life, Implications for Clinical Practice*, 2007.
13. Joni Salminen, Maximilian H, Shammur A, Soon-gyo J, Hind Almerekhi, Bernard J. Jansen, *Developing An Online Hate Classifier for Multiple Social Media Platforms, Human-centric Computing and Information Sciences*, 2020.
14. Mozafari M, Farahbakhsh R, Crespi N, *Hate Speech Detection and Racial Bias Mitigation in Social Media Based on BERT Model, PLOS One*, 2020.
15. Sreekanth Madisetty, Maunendra Sankar Desarkar, *Aggression Detection in Social Media using Deep Neural Networks, Proceedings of the First Workshop on Trolling, Aggression, and Cyberbullying*, 2018, pp 120 – 127.
16. Suvadip Paul, Jayadev Bhaskaran, *Exposing Racism and Sexism Using Deep Learning, Stanford University Press*, 2017.
17. Areej Al-Hassan, Hmood Al-Dossari, *Detection of Hate Speech in Social Networks: A Survey on Multilingual Corpus, COSIT, AIAPP, DMA, SEC*, 2019, pp. 83–100.
18. Raghad Alshalan, Hend Al-Khalifa, *A Deep Learning Approach for Automatic Hate Speech Detection in the Saudi Twittersphere, Applied Science*, 2020.
19. Dulan S. Dias, Madhushi D. Welikala, Naomal G.J. Dias, *Identifying Racist Social Media Comments in Sinhala Language Using Text Analytics Models with Machine Learning, International Conference on Advances in ICT for Emerging Regions*, 2018.
20. Stephan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, Walter Daelemans, *A Dictionary-based Approach to Racism Detection in Dutch Social Media, CLIPS Research Center, University of Antwerp*, 2016.

21. Stephan Tulkens, Lisa Hilde, Elise Lodewyckx, Ben Verhoeven, Walter Daelemans, The Automated Detection of Racist Discourse in Dutch Social Media, *Computational Linguistics in the Netherlands Journal* 6, 2016. Pp 3 – 20.
22. Thomas Davidson, Dana Warmusley, Michael Macy, Ingmar Weber, Automated Hate Speech Detection and the Problem of Offensive Language, *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, 2017.
23. Ji Ho Park, Pascale Fung, One-step and Two-step Classification for Abusive Language Detection on Twitter, *Human Language Technology Center Department of Electronics and Computer Engineering Hong Kong University of Science and Technology*, 2017.
24. Lei Gao, Ruihong Huang, Detecting Online Hate Speech Using Context-Aware Models, *Texas A&M University Print*, 2018.
25. Usman Naseem, Imran Razzak, Ibrahim A. Hameed, Deep Context-Aware Embeddings for Abusive and Hate Speech Detection on Twitter, *Australian Journal of Intelligent Information Processing Systems*, 2019, pp 69 – 76.

APPENDIX A

Codes:

```
#importing libraries (numpy, pandas, natural language toolkit [nltk] and string)
import numpy as np
import pandas as pd
import nltk
from nltk.corpus import stopwords
from nltk.corpus import wordnet
import string
import sklearn

#Load dataset
from google.colab import files
uploaded = files.upload()

#Read the CSV file
df = pd.read_csv('combine.csv')

#Remove duplicates and null rows
df.drop_duplicates(inplace = True)
df.isnull().sum()
df.dropna

#Download stopwords package
nltk.download('stopwords')
nltk.download('wordnet')

#Cleaning the corpus
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()

def processCorpus(text):
    #remove punctuations and stopwords
    nopunc = [char for char in text if char not in string.punctuation]
    nopunc = ''.join(nopunc)

    clean_words = [word for word in nopunc.split() if word.lower() not in stopwords.words('english')]
    lemmatized = [lemmatizer.lemmatize(word) for word in clean_words]
    return lemmatized
```

```

df['Comment'].head().apply(processCorpus)

#Convert text to matrix
from sklearn.feature_extraction.text import CountVectorizer
bow = CountVectorizer(analyzer=processCorpus).fit_transform(df['Comment'].values.astype('U'))

#Split data into training and testing dataset (test 20%, train 80%)
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(bow, df['Label'],
test_size = 0.20, random_state = 0)

#Create and train Naive Bayes Classifier
from sklearn.naive_bayes import MultinomialNB
model_NB = MultinomialNB().fit(x_train, y_train)

#Export/save Naive Bayesian Model
import pickle

filename = 'model_NB.pkl'
pickle.dump(model_NB, open(filename, 'wb'))

#Evaluate Model
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
pred = model_NB.predict(x_train)
print(classification_report(y_train, pred))

print()
print('Confusion Matrix: \n', confusion_matrix(y_train, pred))

print()
print('Accuracy score:', accuracy_score(y_train, pred))

#Validate Model
predic = model_NB.predict(x_test)
print(classification_report(y_test, predic))

print()
print('Confusion Matrix: \n', confusion_matrix(y_test, predic))

print()
print('Accuracy score:', accuracy_score(y_test, predic))

from sklearn.linear_model import LogisticRegression

```

```

model = LogisticRegression()
model.fit(x_train, y_train)

Pkl_Filename = "model_LR.pkl"

with open(Pkl_Filename, 'wb') as file:
    pickle.dump(model, file)

#Evaluate Model
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
desc = model.predict(x_train)
print(classification_report(y_train, desc))

print()
print('Confusion Matrix: \n', confusion_matrix(y_train, desc))

print()
print('Accuracy score:', accuracy_score(y_train, desc))

#Validate Model
desc_test = model.predict(x_test)
print(classification_report(y_test, desc_test))

print()
print('Confusion Matrix: \n', confusion_matrix(y_test, desc_test))

print()
print('Accuracy score:', accuracy_score(y_test, desc_test))

import matplotlib.pyplot as plt
from sklearn import svm
from sklearn.svm import SVC

model3 = svm.SVC(kernel='linear')
model3.fit(x_train, y_train)
pred3 = model3.predict(x_train)

model3 = svm.SVC(kernel='linear')
model3.fit(x_train, y_train)
pred3 = model3.predict(x_train)

#Evaluate Model
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

print(classification_report(y_train, pred3))

print()

```

```
print('Confusion Matrix: \n', confusion_matrix(y_train, pred3))

print()
print('Accuracy score:', accuracy_score(y_train, pred3))

#Validate Model
pred_test3 = model3.predict(x_test)
print(classification_report(y_test, pred_test3))

print()
print('Confusion Matrix: \n', confusion_matrix(y_test, pred_test3))

print()
print('Accuracy score:', accuracy_score(y_test, pred_test3))
```