Creative Research Publishers
https://www.crossref.org/06members/50go-live.html

Journal of Behavioral Informatics
Digital Humanities & Development Research
Vol 7, No 1, 2021 Series

A Classification Model Based on Machine Learning for Detecting Racist Comments on Social Media Platforms

Allenotor, D. & Oyemade, D. A.
Department of Computer Science
Federal University of Petroleum Resources Effurun (FUPRE)
Effurun, Delta State, Nigeria
E-mails: allenotor.david@fupre.edu.ng; oyemade.david@fupre.edu.ng
Phones: +234-8100534069, +234-8039209152

ABSTRACT

Racial conflicts have become even more prevalent than before. As a result, social media companies are continuously being slammed for their inadequate response to the problem caused by racial discrimination. For example, the year 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 research efforts focusing on social media and the role it has played in amplifying racism. Similarly, 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. To solve this problem, we will track down past events and social media trends which are likely to have triggered racist reactions and retrieve annotated comments from public social media sites like Facebook, Instagram, Twitter, YouTube and TikTok. We will create an unbiased dataset of racist comments across social media platforms. We will be building a classification model using machine learning to detect racist comments on social media platforms. We propose a machine learning model for the automatic detection of racist comment across social media platforms. The results we obtained from our research shows that the support vector machine-trained model performs the best with an accuracy of 88.19%. The models proposed in this research outperformed most of the pre-existing models for the same task.

Keywords: Race, Racism, Cyberbully, Hate Speech, Support Vector Machine, Confusion Matrix

Journal Reference Format:
Allenotor, D. & Oyemade, D. A. (2021): A Classification Model Based on Machine Learning for Detecting Racist Comments on Social Media Platforms Journal of Behavioural Informatics, Digital Humanities and Development Research. Vol. 7.No. 1, Pp 121-136. ICT University USA Endowed Research Series Publication in collaboration with SMART-Africa. Available online at https://www.isteams.net.behavioraljournal.
Article DOI No - dx.doi.org/10.22624/AIMS/BHI/V7N1P9

1. INTRODUCTION/BACKGROUND

Several research have made efforts to describe racism. Wilson in [1] and Matthew in [2] described racism is an ideology of racial domination 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. They further observed that prejudice, predisposition, 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 are among the propelling forces behind every form of racism. Research efforts and scholars in social sciences use the term race to generally describe a social construct, which is biologically meaningless when applied to humans. Physical differences such as skin color have no natural association with group differences in ability or behavior. Race, yet, holds tremendous significance in structuring social reality.

Indeed, historical variation in the definition and use of the term race, 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. There is 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. However, as Western colonialism and slavery expanded, the concept was used to justify and prescribe exploitation, domination, and violence against peoples racialized as nonwhite [1]. These in turn supported the horrors of African slavery, Apartheid, the Jim Crow laws, Nazism, and Japanese imperialism.

With the advent of the Internet—social media, 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 [3]. Undeniably, a growing movement of people around the world have been witnessed who are advocating for change either in justice, a change in equality or a change in 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 Jessie [4] identified social network sites as spaces where race and racism play out in interestingly and sometimes disturbing ways. Since then, social media research efforts have increased among academics. In parallel, research efforts such as [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 [10]. 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.

Microaggressions ([11] and [12]) 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”. 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 such as 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 and Ruth [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 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] and [10] 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 where it is estimated that about 4.62 billion people around the world now use social media [xxx], 424 million new users have come online within the last 12 months. The average daily time spent using social media is 2 hours 27minutes, 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 (ML) approaches are applied to detect and remove inappropriate contents that do not follow the community guidelines.

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. This consideration drove the big tech company, Facebook to establish a content monitoring department in the company, hiring over 3,000 employees in 2017 to assist with the removal of videos showing murder, suicide, and other forms of violence. The 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. Certainly, the remedy to the current insufficiency of AI models is not human effort, but rather to build better models.

The remaining parts of this article are organized as follows. Section 2 contains an extensive review of the past works done by other researchers on the research area or a closely related subject matter such as hate speech, sexist comments, and general abusive language classification. The idea is to create a roadmap which is intended to interact between the methodological paradigms that researchers have employed to solve this problem and the relative successes and challenges that have resulted as such. In Section 3, we will provide the methodology that we have adopted in our study. The objective is to highlight the textual strategies deployed in gathering the corpus, and the criterium which we use to identify the comments. We shall also discuss data cleaning and natural language processing procedures we used for the preparation of the dataset. Section 4 of this article will focus on the implementation environment of the methodologies we used. We also itemized the list of software tools and libraries used throughout the research in this section. Section 5 reports the results and presents discussions. Section 6 concludes the research.

2. RELATED WORKS

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 [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, the author addressed three research questions:

i. Which geographical contexts, platforms, and methods do researchers engage with in studies of racism and hate speech on social media?
ii. To what extent does scholarship draw on critical race perspectives to interrogate how systemic racism is (re)produced on social media? And finally,
iii. what are the primary methodological and ethical challenges of the field?

Ariadna [5] found that there is 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 there are only a handful of existing 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 related work that we present, 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 Related Literature

Joni et al. in [13] collected a total of 197,566 comments from four social media platforms: The YouTube, Reddit, Wikipedia, and Twitter. They labeled 80% of the comments they gathered as non-hateful and the remaining 20% they labeled as hateful. They experimented using several classification algorithms. These include Logistic Regression, Naïve Bayes, Support Vector Machines, XGBoost, Neural Networks and feature representations such as Bag-of-Words, TF-IDF, Word2Vec, Bidirectional Encoder Representations from Transformers (BERT), and their combination. In their study, they 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.

In another study, Marzieh, Farahbakhsh, and Crespi in [14] introduced a transfer learning approach for hate speech detection based on the existing pretrained language model - BERT 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.

Deep Learning Related Literature

In recent years, researchers have shifted attention towards deep learning approaches. Many of the works have deployed convolutional neural networks, an example is given in [15] where Skreekanth 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 and Jayadev 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 Ji and Pascale [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.

2.2 Context Aware Models for Comment Classification

Beyond user comments, Lei 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. The 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.

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

In another research effort, Raghad and Hend 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 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 article [19] addressed the issue by building a text analytics model with ML that can be used to filter racist comments in Sinhala language. A Two-Class Support Vector Machine (SVM) 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.

Tulkens et al. in [20] published another research on racist comment detection in the Dutch cyberspace. In their report, they presented a dictionary-based approach to racism detection in Dutch social media comments. They retrieved from two public Belgian social media sites comments likely to attract racist reactions. These comments were labeled as racist or non-racist by multiple annotators. Using 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. They created a second dictionary using automatic expansion—a word2vec model trained on a large body of general Dutch text. Thirdly, they created a third dictionary which manually filtered out incorrect expansions. Finally, they trained multiple Support Vector Machines (SVM), using the distribution of words over the different categories in the dictionaries as the main 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, which they 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 [20] 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 research efforts 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 [20], 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). 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 [22].

3. METHODOLOGY

The methodology of our approach is presented as follows.

(a) Dataset Gathering

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 Fakih 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 Barrack 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 protocols, 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. We 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, Anthonio Rudiger, Romelu Lukaku and many others) are 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 embark on data gathering on 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.

(b) Dataset Labeling

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. After the labelling stage of the collected dataset, to get the best results from the model, some natural language processing techniques were applied in order to improve the data quality. Then after, punctuations were removed before case normalization was done on the dataset. The collected datasets also passed through the process of tokenization were separation of the strings into smaller units called Tokens was done. Stop words such as include "and, I, he, she, but, was, were" were removed because they do not add any meaning to the collected dataset. The final stage of dataset preparation is the lemmatization and stemming. In this process, we were able to reduce variations of the same word, thereby reducing the number of words that will be included in the model. A simple flowchart of the procedure is hereby given in Figure 1 for clarity.

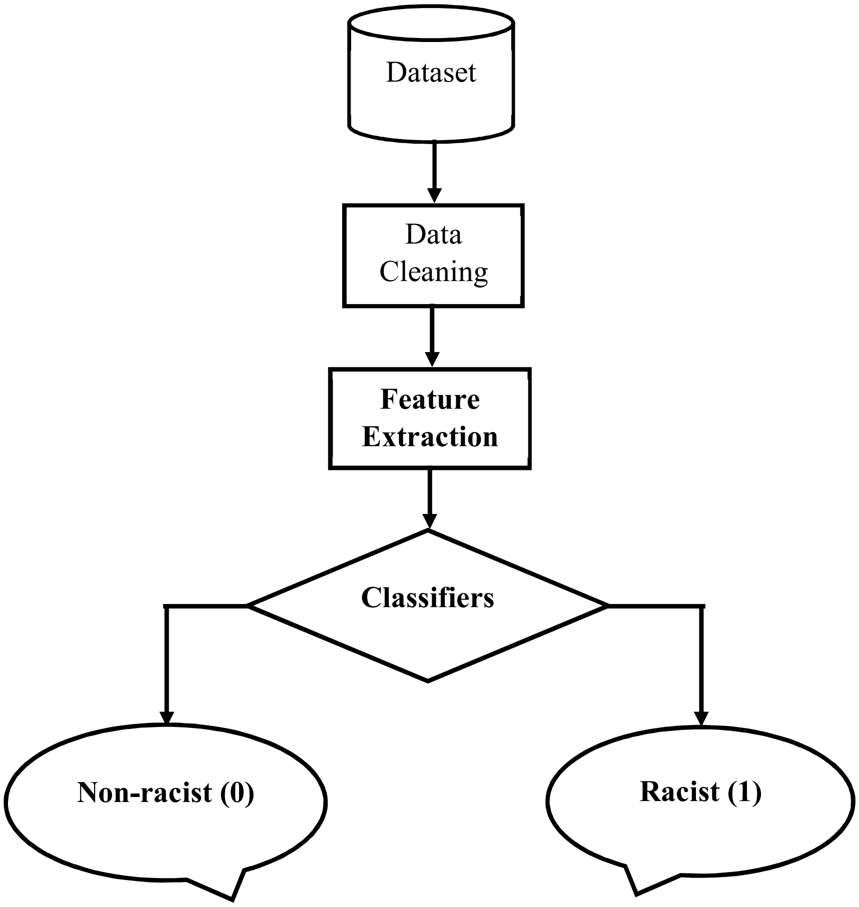


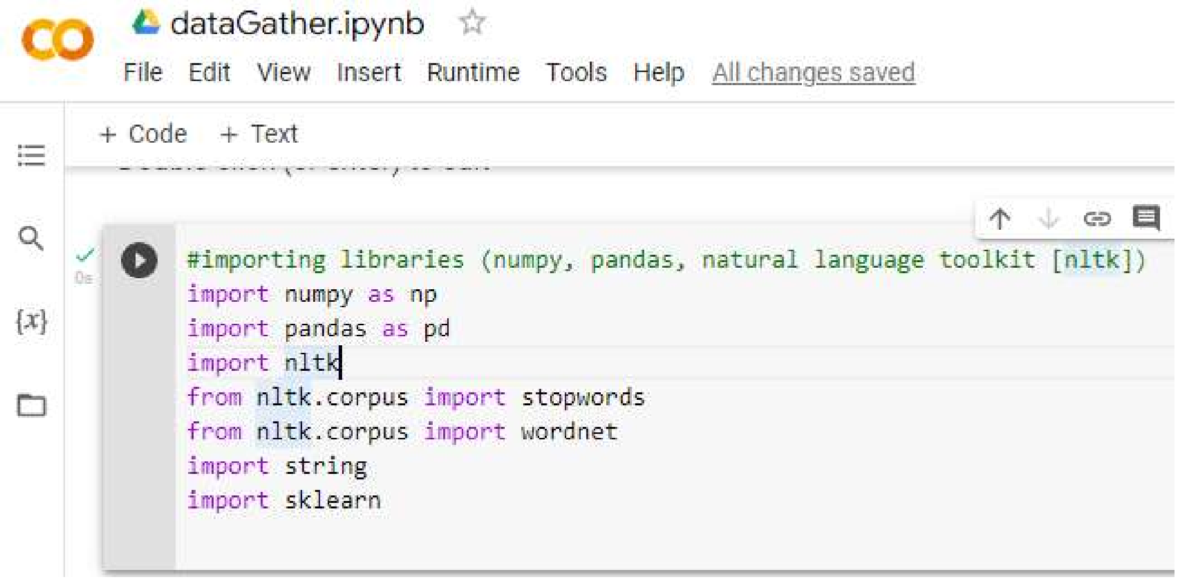
Figure 1. Flowchart of Procedures for Comment Classification

4. IMPLEMENTATION ENVIRONMENT

The following hardware and software were used in the course of this research. They include an 8GB RAM (Random Access Memory) that runs on Intel Core i7-3517U processor, a standard mouse and keyboard, and a 256GB of Hard Disk Drive. The software tools include Windows 10 Pro, Google Chrome Browser, Google Collaboratory Workbook, Programming Language: Python, and MS Excel.

(a) Implementation Tools

The python programming language in Google Collaboratory Notebook [26] was used for the implementation of ideas in this research. The python programming language in Google Collaboratory Notebook is 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. These were simply added by the import statement as shown in Figure 2.

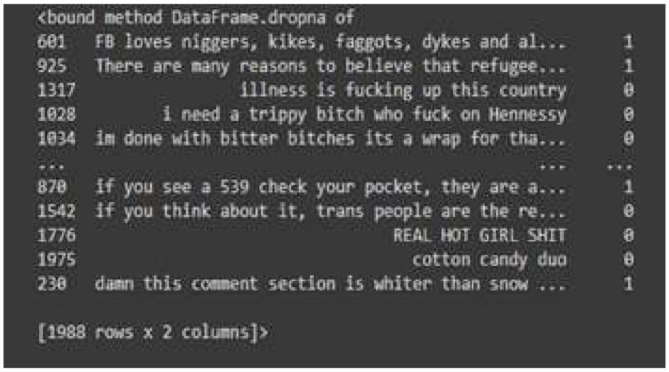


```
#importing libraries (numpy, pandas, natural language toolkit [nltk])
import numpy as np
import pandas as pd
import nltk
from nltk.corpus import stopwords
from nltk.corpus import wordnet
import string
import sklearn
```

Figure 2. Importing Libraries

(b) Components of Implemented Models

The data corpus used in this project is a dataset of 2,000 comments, a combination of 1,000 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. Figure 3 shows the corresponding reduced dataset.

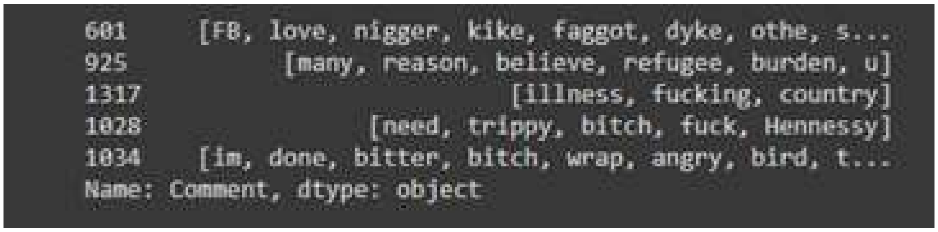


```
<bound method DataFrame.dropna of
601   FB loves niggers, kikes, faggots, dykes and al...     1
925   There are many reasons to believe that refugee...     1
1317             illness is fucking up this country         0
1028         i need a trippy bitch who fuck on Hennessy     0
1034  im done with 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]>
```

Figure 3. The Dataset

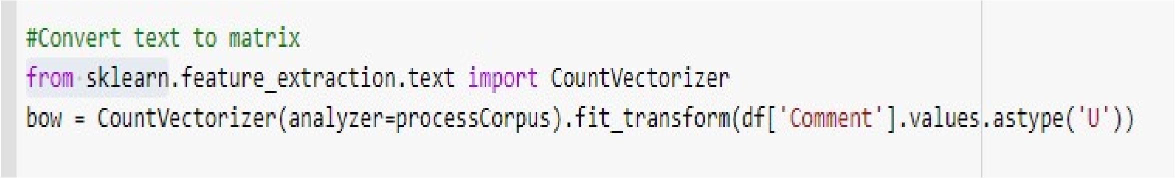
To prepare the dataset for training, text processing techniques available in the natural language processing NLTK package was used to apply the processing techniques. Figure 4 shows the resulting processed data.



```
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
```

Figure 4. Processed data

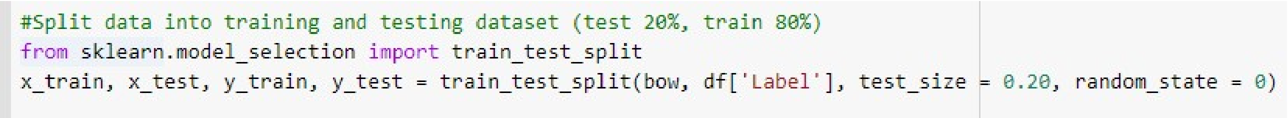
In the feature extraction mode, we encoded the data as integers or floating-point values so that they can be in their machine-readable form. For the purpose of this research, the CountVectorizer package was used to convert the textual tokens to a vector of token counts i.e., reducing each token to a vector representing its frequency of appearance. Figure 5 shows the snippets for count vectorizer.



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

Figure 5. Applying CountVectorizer on the Tokenized data

We went ahead to split the data after the tokenization stage. The data were splitted 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. Figure 6 shows the splitting snippet.



```
#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)
```

Figure 6. Splitting the dataset

5. RESULTS AND DISCUSSIONS OF EXPERIMENTS

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. Table 1 shows the naïve Bayesian classifier training performance report where we report 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.

Table 1. Naïve Bayesian Classifier Training Performance Report

	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				

Table 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.

Table 2. Logistic Regression Training Performance Report

	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				

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. Table 2 shows that the model reports a 0.99 F1 score and precision, respectively on both racist and non-racist comment classification. Similarly, Table 3 shows a model report of an accuracy score of ≈ 0.99 (99%) on training data, an improvement from both the multinomial naïve bayes and Logistics regression classifiers.

Table 3. Support Vector Machines Training Performance Report

	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				

Table 4. Multinomial Naïve Bayes Trained Model Validation Performance Report

	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				

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 on Table 4. Table 5 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.

Table 5. Logistic Regression Trained Model Validation Performance Report

	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				

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. 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, Table 6 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.

Table 6. Support Vector Machine Trained Model Validation Performance Report

	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				

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. 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 7. 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 7 shows the summary of results on all the classifiers on solely racist comment detection on the validation data. Table 8 shows that the models proposed in this research 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.

Table 8. Comparison of Existing Models and our Proposed Model

Research Efforts	Classifier	Precision	Recall	F1-score
Suvadip and Jayadev [16]	LR	-	-	0.79
	SVM	-	-	0.81
Dulan et al. [19]	SVM	1.00	0.36	0.53
Tulkens et al. [20]	SVM	0.24	0.02	0.04
Tulkens et al. [21]	SVM	0.49	0.43	0.46
Ji and Pascale [23]	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
Our Research	NB	0.78	0.89	0.83
	LR	0.88	0.84	0.86
	SVM	0.87	0.88	0.88

6. CONCLUSION

We proposed 3 models in this article to help identify and extract 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.

Unfortunately, many systems that have been implemented for the detection of such comments by social media companies are notorious for ignoring racist comments disguised as humor. It is to address this problem, that the dataset used in this research was spread to cover 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.

REFERENCE

[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. and Johan F., Racism, Hate Speech, and Social Media: A Systematic Review and Critique, *Television and New Media*, 2021, pp 205 – 224.

[6] David R. and 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, and 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, and 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, and Marta Esquilin, Racial Microaggressions in Everyday Life, *Implications for Clinical Practice*, 2007.

[13] Joni Salminen, Maximilian H, Shammur A, Soon-gyo J, Hind Almerekhi, and 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., and Crespi N., Hate Speech Detection and Racial Bias Mitigation in Social Media Based on BERT Model, *PLOS One*, 2020.

[15] Sreekanth Madisetty and 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 and Jayadev Bhaskaran, Exposing Racism and Sexism Using Deep Learning, *Stanford University Press*, 2017.

[17] Areej Al-Hassan and 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 and 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, and 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, and Walter Daelemans, A Dictionary-based Approach to Racism Detection in Dutch Social Media, *CLIPS Research Center, University of Antwerp*, 2016.

[21] Stephan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and 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 Warmsley, Michael Macy, and 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 and 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 and Ruihong Huang, Detecting Online Hate Speech Using Context-Aware Models, *Texas A&M University Print*, 2018.

[25] Usman Naseem, Imran Razzak, and 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.

[26] Google Collaboratory Notebook https://colab.research.google.com/