Dissertation Submitted in Partial Fulfilment for The Degree of Master of Science in Big Data

MBACHU CHIZARAM
Department of Computing Science, Faculty of Natural Sciences, University of Stirling

Abstract

Problem: While we appreciate the numerous advantages and opportunities that using social media platforms provides, it is undeniable that many people are embarrassed, insulted, tormented, and harassed by anonymous users, strangers, or peers. These insults may be communicated openly or subtly, and while it may be easy to spot an insult outright in some cases, it may not be in others, this brings about the term microaggression. Microaggressions are subtle, small insults that are difficult to identify. To address this issue, the detection of microaggressions is coming to play, this detection has been a top concern in recent years, with the goal of reducing online toxicity and bullying, which are prevalent in today's environment.

Objectives: While it is the intended goal of this dissertation is to develop a model using different machine learning algorithms to accurately detect microaggressions, I will be concentrating on comments that contain hate speech rather than microaggressions for a variety of reasons that will be explained in greater detail in other chapters as we progress through the dissertation. Hate speech, like microaggression, is difficult to define but is considered as any sort of verbal expression calling for cultural, racial, or religious violence. The difference between hate speech and microaggression is basically in its usage, while hate speech is overtly expressed, microaggression is covertly expressed

Methodology: A pre-labelled dataset from https://www.kaggle.com/arkhoshghalb/twitter-sentiment-analysis-hatred-speech). was used and the terms used in the labelling were gotten from https://hatebase.org/ an organization that maintains a glossary of hate speech vocabulary. Each tweet is annotated for whether it contains hate speech or not, and the total number of tweets in this collection is a little over 31,000. A supervised machine learning method for detecting hate speech using machine learning algorithms such as Multinomial Nave Bayes, Random Forest Classifier, Extra Trees Classifier, Light Gradient Boosting (LGBM), Extreme Gradient Boosting (Xgboost), Adaptive Boosting (Adaboost), Logistic Regression, Logistic RegressionCV, Stochastic Gradient Descent (SGD) Classifier, and Support Vector Machine algorithms were all used.

Achievements: The findings show that our suggested methodology is a viable solution for detecting hate speech in online social media platforms. Finally, I compared the outcome of the chosen model to model used by other academics.

I. INTRODUCTION

Background and context to the problem

Microaggressions and hate speech are becoming more concerning occurrences that impact individuals of all ages and backgrounds, particularly on social media platforms. Social media has become a vital avenue of communication for practically everyone, especially young adults, and has become increasingly popular. As a result of the widespread use of social media among adults, the spread of microaggressions and hate speech has emerged as a big issue for those who use the platform to express themselves. Research on the detection of microaggressions has gained prominence in recent years because of the above occurrence, it has also become a hot subject, and firms such as Facebook, Google, and Yahoo are eager to contribute to research aimed at identifying successful and cost-efficient approaches in resolving the issue.

Before I go any further, I would provide some background information on what microaggressions and hate speech are. Microaggression can be defined as "brief and commonplace daily verbal, behavioural, or environmental indignities, whether intentional or unintentional, that communicate hostile, derogatory, or negative racial slights and insults …" [2].

According to the Federal Communications Commission, hate speech is defined as an abusive or threatening speech or writing that expresses prejudice toward a certain group, notably based on their race, religion, or sexual orientation.

Because of this wide-ranging problem of increased microaggression and hate speech among users, researchers have been working on developing methods and strategies that would identify and prevent hate speech from being used on a regular basis by those who use technology. The identification of microaggressions and hate speech has attracted a significant amount of attention, and machine learning approaches have been employed to identify the terms when they are spoken in public forums.

Machine learning (ML) is a type of data analysis that automates the process of developing analytical models. ML algorithms are frequently divided into two categories: supervised and unsupervised. To predict future occurrences, supervised machine learning algorithms apply what has been learned in the past to new data by using labelled examples from the past (supervised machine learning algorithms). For example, when using machine learning to identify hate speech or microaggression on social media, it extracts some features from the texts and then creates classifier algorithms to detect hate speech or microaggression based on the features retrieved. While unsupervised learning is often employed for cluster and unlabelled datasets, it will not be utilised in my dissertation since the properties of my datasets do not lend themselves to this kind of learning.

It is crucial to remember that while the primary goal of this dissertation is a determination of an effective model for the determination of microaggression using machine learning models. However, it must be noted that there is in a dearth of datasets for detection of microaggressions this may be due to the subjectiveness of microaggressions. Furthermore, the contextual nature of microaggression makes detection even more problematic [4].

Due to this unavailability, most study/research often utilize hate speech datasets to determine the potential of machine learning models to detect microaggression. This has been completed with varying degrees of success. This dissertation leveraging on existing literature on the detection of microaggression using machine learning attempts to achieve two cardinal purposes.

(a) An analysis of the existing literature on detection of microaggression using machine learning. As a side note, a. brief analysis of hate speech will be provided to establish a conceptual basis for their application as a viable alternative to actual microaggression dataset. This is achieved under section 2.
(b) Thereafter, section 3 provides an analysis of methodology adopted in this dissertation. Essentially, machine learning models will be applied to hate speech datasets. Unlike other previous studies, this dissertation applies stacking to ensure better recall and precision. The hypothesis here is that an application of stacking to datasets may provide more recall, precision and perhaps an accurate result when microaggression datasets are developed. Fundamentally, it is theorized that the same results made for the application of the machine learning models to hate speech datasets will, in theory, provide a similar result if applied to microaggression datasets.

1.2 Scope and objectives

The main objectives of this dissertation are.

- Define and understand microaggression and hate speech, the methods of detection as well as methods used in previous studies.
- Collect sufficient dataset to conduct a detailed analysis.
- Clean and extract the relevant features of the data for use in building a machine learning model

- Develop a model to determine if a comment is hate speech or not, and analyse the results.
- Utilize the most accurate model available to facilitate discussion of potential solutions for eliminating hate speech on social media platforms.

1.3 Limitation of study

One of the dissertation's drawbacks is the presence of errors in the labelling of the dataset that was used. If a dataset is not properly labelled, it is possible that the classifier will not perform well. Bias is also introduced into the dataset because of the labellers' actions, and these biases may be based on colour, gender, educational status, or religious convictions, among other factors.

Another limitation of this study is the fact that there are just a handful datasets large enough to train machine learning models that are freely available on the internet today. A big enough dataset (e.g., 20 million to 100 million tweets) is needed alongside human workforce to categorise the samples in getting acceptable results. When it comes to hate speech, trained annotators may be quite good in labelling a dataset.

Also, a major constraint of this dissertation will be the difficulty in categorizing words as microaggressive, the development of specific talents in categorization will be required, particularly when they must be separated from abusive language, or 'hate speech'.

1.4 Achievements

Although this study cannot conclusively state that microaggression detection is impossible, we do state that the model used to construct this model is capable of being employed in the future, provided that the problem with dataset availability is addressed.

1.5 Overview of Dissertation:

Immediately following the introduction, Chapter 2 addresses pertinent work by other researchers in the detection of microaggressions and hate speech. Chapter 3 provides a historical context for the dataset, as well as the source and explanation of the data that was used. Detailed descriptions of the study methodology and model used to detect microaggressions may be found in Chapter 4. In Chapter 5, we will review the conclusions of our inquiry. Finally, in Chapter 6 of this dissertation, the study makes some findings about the relevance, limits, and implications of the proposed model, as well as future suggestions.

II. STATE-OF-THE-ART

2.1 Introduction

Conscious and deliberate distinctions based on race, gender, religion, sexual orientation, or socio-economic class has been a resilient feature of society since time immemorial. The existence of these distinctions is expressed through prejudices, stereotypes, discrimination to outright social exclusion [2]. In some cases, societal distinctions have led to disastrous consequences on societal development and indeed economic stability. This was seen during the South African Apartheid from 1948 to 1994 charactered by societal segregation between the black natives and white colonialists, and the German holocaust between 1941 and 1945 which saw the systematic murder of six million Jews.

But these prejudices and discrimination have historically been outright and unequivocal without any need for analyses. However, more recently, with the wave of global disapproval for distinctions like racism, discrimination etc took a new form. This new form is referred to as microaggressions. The rationale behind the use of microaggression to re-emphasize or portray orthodox societal or social economic distinctions is mainly because the consequences, usually imposed by law and other societal morals and values, which accompanies outright and unequivocal expression of distinctions act as a strong deterrent hence the use of more subtle methods for expression.

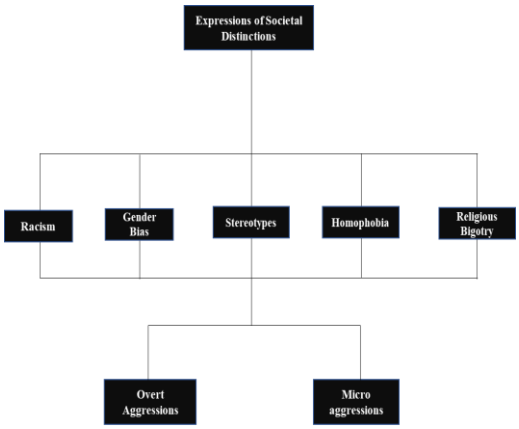


Figure 2. 1 Subdivisions of expressions of societal distinctions

These microaggressions, as the existing literature on the subject refer to them are employed due to their subtle nature and their difficulty in detection. Usually, victims of micro-aggression are oblivious of the aggression until after the fact and even then, such aggression due to its interpretative ambiguity are open to discourse, if detected at all, and usually ends with a shrug from the victim using the words 'They probably meant something else' to avoid making a mountain out of ant hill.

Thus, the hallmark of microaggressions is that they are subtle, ambiguous, difficult to detect and nearly impossible to measure. In other words, micro aggressors can express traditional societal distinctions without the consequences which would ordinarily accompany more overt aggression, and even where a potential case of microaggression is detected, they are subject to varying interpretations which may further water down the act.

It is against this backdrop that this chapter seeks to provide an analysis of the existing literature on microaggression focusing on its meaning, categorization of types of microaggressions, theories on various methods applied for the detection and measurement of microaggressions. To ensure clarity, it is important to note the following:

a. This chapter shall use the term 'societal distinctions' to refer to a wide array of prejudices, stereotypes, discriminations usually directed to persons of marginalized status (this conclusion is in full understanding of the possibility of the expression of societal distinctions to be channelled towards individuals in a non-marginalized individuals).

b. It must be noted that societal distinctions are theoretical constructs, thus they only take form when expressed. Therefore, this dissertation will be focusing on the expression of microaggressions with a precise attention to how they are expressed. It is the opinion of this author that identifying the methods by which microaggressions are expressed constitutes the first step towards detection and indeed measurement of microaggression within numerous social situations i.e., workplace, school etc.

c. It is presumed that various detection methods may be more effective to identify certain sub-classes of microaggression than others. However, methods for the measurement of micro-aggression after detection may be applicable across sub-types of microaggression.

2.2 Conceptualisation of Microaggression

As stated in paragraph 2.1, societal disapproval and legal prohibition of discrimination or bias practices across facets of the society led to the disappearance of overt expressions of societal distinctions. But in the wake of its disappearance, came microaggressions which was rather covert and subtle in nature. But the important question is what are microaggressions?

The provision of an answer to this question is important because a definition helps excludes certain expressions of societal distinctions which may not be completely overt but also not absolutely covert due to continues use i.e., certain expressions which may have been considered as micro-aggression is now considered as an overt expression of bias due to its continuous use; the popular ones are; assuming an Asian person knows maths or a black person is athletic and making statements based on these assumptions to the individual. This in the past was considered microaggression but due to continuous

use, is considered overt expression of a stereotype. In essence, it is my opinion that microaggressions stop being microaggressions once they are no longer subtle or lose their ambiguity, in that case it becomes an outright hate speech. Notwithstanding, the literature on the subject seem to agree on one fundamental conclusion that such activity is substantially unbeneficial to society [4].

2.2.1 What are Microaggressions?

Microaggressions are "brief and commonplace daily verbal, behavioural, or environmental indignities, whether intentional or unintentional, that communicate hostile, derogatory, or negative racial slights and insults …" [2]. Building on this definition provided by [2], [5], defined microaggression as a "commonplace, interpersonally communicated "othering" messages related to a person's perceived marginalized status [5]. Further conceptualizing what microaggressions are, Nordmarken emphasized that microaggressions are often manifestations of derogatory stereotypes invisible to many deliverers and recipients but often leads to oppression on an interpersonal level [5].

Distinguishing microaggressions from direct aggressions, Periyakoil et al, stated that microaggressions on the surface may appear innocuous and insignificant, especially to individuals who are not from the targe marginalised group. However, these microaggressions can have terrible effects on the psychological wellbeing on persons from the target marginalized group [6].

In essence, microaggressions by their very nature are 'micro', in other words, they consist of subtle messages manifesting themselves as seemly unconscious and unintentional messages as snubs, insults or other indignities, frequently embedded within a stream of communication [7]. This characteristic of microaggression appearing unintentional is a fundamental feature which feeds into the ambiguity of microaggressions. This in most cases leaves the victim confused as to the intentions of the aggressor

Type of Microaggression	Examples	Meaning
Racial	Where a White person tells a person of colour "You are a credit to your race"	This connotes that people of colour are generally not as intelligent as whites [8]
Homophobia	Saying to a gay man that "Are you gay? Because you look quite masculine"	This connotes gay people cannot be masculine.
Gender Bias	Saying to a woman "You are quite smart for a lady"	This connotes women are generally not smart
Religious Bigotry	Saying to an Indian "it is interesting that you can eat beef"	This connotes all Indians are Hindus and do not eat beef.
Stereotypes	Saying to a black man "You must be good at sports"	This connotes all black guys must be athletic.

Table 2. 1 Examples of microaggression

Therefore, the major characteristics of microaggressions can be summarized as follows:

a. Microaggression are subtle discriminatory expressions, usually in the form of negative verbal utterances, snubs, or slights against another individual or a group of individuals usually from a marginalized group [2].
b. Microaggression reveal the prejudice or bias against another group either based on gender, race, sexuality etc. therefore whether such microaggression was intentional or unintentional is immaterial [5].

c. Microaggression are usually ambiguous and shrouded in compliment or flattery making its detection by the victim difficult. As it is in the case of unintentional microaggression, the aggressor might not know that he has been aggressive.

However, notwithstanding the form or nature of microaggression, its peculiarity, as opposed to outright aggression, creates difficulty because they are difficult to detect and therefore challenging to address. Also, it must be noted that the effects of microaggression on victims are like the effect of overt aggression some of these effects includes depression, fear, hypervigilance, fatigue, feeling of alienation and may lead to chronic health problems. This underlines a fundamental fact, methods for early detection and reprimand for microaggressions are essential, hence the importance of this dissertation.

2.2.2 Categories of Microaggressions

As emphasized in paragraph 2.1 and Table 2.1, microaggression may occur under various categories or themes, these include racial microaggressions, homophobia, gender bias based microaggressions, religious bigotry based microaggression and general stereotyped informed microaggression. It must be noted that there may be more categories of microaggression, but it is my opinion that other categories of microaggression could be subsumed under these broad types and where they do not fall under them, they will usually fall under general stereotypes.

A categorisation of Microaggressions is essential to this chapter and this dissertation in general because it provides the foundation for an analysis of the question whether methods for the detection of microaggression will change for different types of microaggression. Essentially, this means that would the methods for detecting racial microaggression be the same as the methods for detecting gender bias based microaggressions? This will be answered under chapter 5 of this dissertation. But it is important to comment that at the core various types of micro-aggression adopt text or lexical forms as its major ascertainable form of expression, thus it can be concluded that methods adopted for each form of microaggression should have similar results.

Also, another question which begs for an analysis of various categories of microaggression is the question whether the methods for detecting microaggressions will change due to the context within which they occur. To provide more clarity, would the methods for detecting racial aggression in the workplace be effective for detecting racial aggression in various levels of educational institutions i.e., primary schools, high schools, and tertiary institutions? This would also be answered in chapter 5 of this dissertation and the conclusion appears to be the same as the prior paragraph at least if the primary input form is textual or lexical.

2.2.3 Racial Microaggressions

The literature on racial microaggression can be traced to psychiatrist Chester Pierce in the 1970s, where it was debunked as a new form of racism. In essence, Pierce concluded that racial microaggression underline a more subtle variant of racism where the racist adopts brief everyday exchanges like subtle snubs, dismissive looks, gestures, and tones to send derogatory messages to members of racial minority groups. These actions are often covert making detection difficult [9].

So, what are racial microaggressions? Racial microaggression can be defined as racially related acts or attitudes that are perceived as hostile, or as subtle verbal, non-verbal or visual racial insults that are experienced by minorities [10].

Furthermore, DeCuir-Gunby et al. categorized forms racial microaggression into micro-assaults, micro-insults, and micro-invalidations [9]. In essence, microinsults are explicit verbal or non-verbal attacks intended to hurt its recipient. This often takes the form of racial slurs or displaying racially charged symbols. In most cases, microinsults are considered overtly racist and are categorized as hate speech so they will not be considered as microaggressions for the purpose of this dissertation. Sue et al seems to be in support of this conclusion. Essentially, the intention of the racist is clear and ascertainable by the victim without room for conjecture unlike micro-insults or micro-invalidations which may, in some situations, be missed by an oblivious recipient. On one hand the racist is trying to be explicit on the other, the racist attempts to be discreet [2].

Also, Derthick disregarding micro-assaults as a form of racial microaggression stated that only micro-insults and micro-invalidation capture the true definition of racial microaggressions are more covert, seeking to communicate racist belief but often disguised as compliments [11]. For example, where a white person says "For an African you speak good English" or a more subtle racial microinsults like when a white person waits to ride the next elevator when a person of colour is on it [8].

Furthermore, it must be noted that there is a dearth in the literature on the impact of racial microaggression unlike its overt counterpart which has been thoroughly documented. But existing sparse literature on the subject suggest that racial microaggression towards minorities leads to poorer physical and mental health [10]. Also Gee et al., posited that perceived stigmatization expressed to microaggression towards marginalized groups is associated with depression and anxiety, decreased psychological wellbeing and other physical issues [12]. Finally, racial microaggression usually fall under 9 themes which are as follows.

a. Ascription of intelligence – This occurs where intelligence is assigned to a person of colour based on their race e.g. Where a white person says to a person of colour, "You speak good English" [8].
b. Myth of meritocracy – This seeks to infer that a person race doesn't play a role in life success i.e., where a white person tells a person of colour that "Everyone can succeed in this society, if they work hard enough" [8].
c. Alien in own land – This occurs where people from a marginalized group are presumed to be born in other countries due to the colour of their skins e.g., when a white person asks a black British citizen "where are you from" [8].
d. Colour blindness – This includes statements and expressions that suggest that a white person doesn't want to acknowledge race i.e. "When I look at you, I do not see colour" [8].
e. Criminality/assumption of criminal status – This involves statements where a person of colour is assumed to be a criminal simply because of their race or colour. For example, when a white lady crosses over to the other side of the road on seeing a black man approaching her on the same side of the road.
f. Denial of individual racism – This generally include a wide array of verbal or non-verbal expressions by mostly white individuals to deny their racial biases [8].
g. Pathologizing cultural values/communication styles – This form of microaggression presupposes that the values and communication styles of the dominant group/ whites are ideal. A good example of this is seen when a white person tells a black person such individual is loud and should speak more calmly [8].
h. Second-class citizen - This generally occurs where a person is accorded preferential treatment due to the person's race. For example, when a person of colour is mistaken for a service worker [8].
i. Environmental microaggressions – This includes expressions on a more macro-level i.e., systematic, or environmental. A good example is where a university with a huge international community has all its buildings named after white heterosexual upper class males [8].

2.2.4 Gender and Sexual Microaggression

a. Gender microaggression

In a study of the common types of gender-based microaggression in medicine, [6] suggested that gender based microaggression are verbal or non-verbal everyday behaviours arising from unconscious bias, covert prejudice, or hostility due to the gender of an individual.

Akin to racial microaggression, gendered microaggression have been categorized into gender micro-assaults, gender micro-insults and micro-invalidations [13]. To ensure clarity, we have set out examples of gender based microaggressions against its form below:

Note that there is significant debate around micro-assaults as a form of microaggression due to its more overt nature. To this end, this shall be excluded.

Gender aggressions possess various themes, some of which includes sexual objectification, second-class citizenship, use of sexist language, assumption of

inferiority (e.g., when a man says "You are strong for a woman"), restrictive gender roles, denial of reality of sexism, denial of individual sexism, invisibility, sexist humour/jokes and environmental invalidations or macro-aggressions which take place on a systemic level (For example unequal pay; glass ceiling etc) [14].

b. Sexual Microaggression

Sexual Microaggression has the same definition as racial and gender microaggression with the underlying subject being the sexuality of the subject. Like the other forms of microaggressions, they can be divided into micro-insults and micro-invalidations. As with racial and gender based microaggressions, micro assaults were excluded because they are not considered to be true forms of microaggression. The table below provides some examples:

Forms of Sexual Microaggression	Examples
Microinsults	"You don't sound/look gay" "It can't be real sex unless it's with the 'opposite' sex" "Being bisexual is greedy, you just want the best of both worlds" "How did you turn gay?"
Microinvalidations	"You're being oversensitive" "You're just confused", "It's just a phase" "You're not a real man" "Why are you making things hard for yourself?" "Why do you have to flaunt it?" "I'm not homophobic, but..."

Table 2. 2 Examples of sexual microaggression (The University of Edinburgh, 2021)

Other expressions of Microaggression can be divided into a broad array of stereotypes however each conforms with similar characteristics of subtilty, ambiguity as racial, gender and sexual microaggressions.

2.2.5 Detecting Microaggressions

As emphasized above, the covert nature of microaggression creates the problem of detection and the hardship of detection creates another problem of measurement. However, over the years various approaches have been utilized to detect microaggression from more traditional methods utilizing questionnaires to more automated methods which utilized machine learning. While this dissertation will focus on automated methods for detecting micro-aggression, an overview of traditional approaches shall be provided.

Nonetheless, before providing an analysis of methods, it is important to note the following:

a. Methods adopted for the detection of microaggression are quite restrictive in scope of application and the target subjects. What this means is that each method was developed to detect specific kinds of microaggressions within specific contexts. For example, the Racial Microaggression Scaling (RMAS) method which involved the use of questionnaires, as was common with all orthodox methods, focused on racial microaggressions using specific words or answers to specific questions covering various themes of racial aggression [15].
b. This restrictive attribute is also seen in machine learning approaches. Fundamentally, machine learning approaches require the use of classifiers to sieve relevant data containing potential cases of micro-aggression from a wide array of data sources like twitter [16]. The result here is that the resulting data provides situations of microaggression without providing precise information on the context of such aggression. But if the detection of microaggression without a consideration of the type, which is often apparent, or the context is not the aim of the study, then the use of machine learning perhaps provides the best methods for collation of relevant data and an

analysis of such data to spot instances of microaggression.

The aim of this dissertation is to develop a model using different machine learning algorithms to accurately detect microaggression (in this case hate speech is the focus). Therefore, this dissertation shall focus on building a model to detecting "hate speech" using machine learning algorithms. Another focus of this dissertation is to ascertain if an alternative approach can be used to increase the accuracy of some parameters that will be presented in the next chapters.

2.2.6 Methods for detecting Microaggression

2.2.6.1 The Racial Microaggression Scaling (RMAS)
The RMAS is a more traditional method adopted to assess the themes and categories of racial microaggressions. With regards to detecting microaggression. Yearwood suggests that there are three tools which could be utilized to assess exposure to microaggression. The first tool is the ethnic microaggression scale often used with adolescents by adopting 12 items that measure denial of racial identity, emphasis on differences and negative treatment [4]. The second tool posited by Yearwood called the inventory of microaggressions against black individuals was used to measure racial microaggressions in 385 undergraduates who self-identified as Black or African American. This tool adopts 14 items which measures the exposure of test subjects to insults and invalidations by others [4].

Finally, and perhaps the more relevant tool is to this dissertation, is the Racial Microaggression Scale (RMAS) developed by Torres-Harding, Andrade, and Diaz in 2012. Utilizing 32 items, the RMAS assesses the experience of invisibility, criminality, undesirable culture, sexualization, experience of environmental insults etc to assess exposure of environmental insults [4].

Adopting the RMAS method for detecting Racial Microaggression, Torres-Harding et al. provided light on the procedure by referring to the RMAS method as an implied consent procedure where the data were collected in person using paper and pencil questionnaires, and through an online survey [10].

2.2.6.2 Machine Learning Methods (MLMs)
Machine learning methods can be divided into supervised learning methods, unsupervised learning, and reinforcement learning. Essentially, supervised learning involves a learning method where the machine learns under guidance. The aim of supervised learning is to predict the target variable using specific predictor variables. This form can be divided into classification task and regression task. On one hand, the classification task utilizes a target variable consisting of different categories, while a regression task occurs where the target variable is limitless.

Conversely, we have unsupervised learning which predicts data without labels. Here the machine is trained on unlabelled data without any guidance to either solve association or clustering problems. For context, an association problem basically involves discovering patters in data, finding co-occurrences etc. A good example is the relationship between fuel and car. On the other hand, a clustering problem involves grouping data that are similar together, in this case data is split into different groups based on similarity and traits.

Finally, reinforcement learning is where an agent interacts with its environment by producing actions and discovers either an error or a reward. This type of learning is based on the principle of learning by positive and/or negative feedback. It aims to train a model to and the best chain of decisions (policy), solving a specific problem. Decisions which are beneficial are rewarded whereas disadvantageous decisions are punished.

But more relevant to this essay is supervised learning methods. It is proposed that there is a potential for a previously supervised learning method to become unsupervised after substantial training. But the question remains, how does this apply to this dissertation? With the rise of social media and other big data banks, the research on microaggression has become more popular, this is because social media i.e., Facebook, Twitter and Instagram provide researchers with a wide array of data. The datasets provided by social media is not just relevant for the size of data, but also the fact that it is publicly

available, and the size is sufficient to train ML models to achieve suitable results.

2.2.7 Machine learning algorithms for detecting microaggressions

Over the years there have been quite several machine algorithms developed for various reasons ranging from the use of feature extraction sentiment analysis to collate and interpret drug reviews using RNN-BiLSTM Models [17], the use of machine learning to classify google play store application review. Sentiment analysis is a natural language processing (NLP) approach for determining whether a piece of data is positive, negative, or neutral. Sentiment analysis is vital in this dissertation because it will help detect emotions behind a tweet and decide whether the message tweet is microaggressive in nature with the use of machine learning. Also, Dey et al., utilized machine learning methods to analyse the sentiments of customer reviews on Amazon products. This study incorporated multiple product review factors including product quality, content, time of review related to product durability and historically older positive customer reviews [18]. In this study, the sentiment of the consumer was analysed by the Naïve Bayes classifier (NB) and the support vector machine (SVM) classified the sentiments of users in binary categories. In essence, the system adopted by Dey et al., is as emphasised in the diagram below.

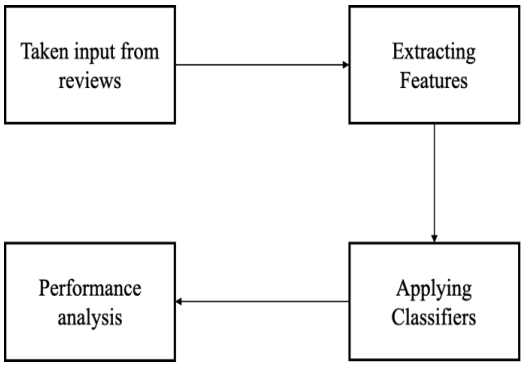


Figure 2. 2 Basic block diagram of Dey et al NBC-SVM system (adopted from Dey et al. 2020)

But more specifically is a study conducted in 2018 by Gaydhani et al. using an N-gram and term frequency-inverse document frequency (TFIDF) detect hate speech and offensive language on twitter using machine learning [16]. TF-IDF (Term Frequency-Inverse Document Frequency) is another feature extraction method commonly used when doing text analysis, here, it converts text into a format that is usable by machine learning algorithms and assigns a weight to each term in the text not just by its count but by its importance in the whole dataset thereby reducing the effect of less weighted terms [16]. Term frequency is how many times a word appears in a text while Inverse Document Frequency gives less importance to words that appear a lot in a text. As emphasised by Gaydhani et al., most machine learning approaches used to detect toxic language or by analogy, microaggressions often involves the extraction from the text. Here lexical features such as dictionaries, bag of words etc are utilized [16].

Note that the dissertation's emphasis on microaggressions will move at this point to predominantly focusing on hate speech, which is critical to understanding the topic. Several limitations were encountered during the study, including a lack of datasets that were solely dedicated to microaggressions and a lack of comparisons to current literature on the detection of microaggressions using machine learning. As a result, I am comparing the findings of my research to the current literature on hate speech recognition using machine learning techniques.

2.3 Hate Speech

Hate speech has been defined by Zhang as ‘any communication that disparages a person or a group based on some characteristics (to be referred to as types of hate or hate classes) such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics [19]. As opposed to microaggressions, hate speech are overt, express, and easily identifiable. This has contributed to the availability of labelled datasets simply by using data analysis to quantify or qualify linguistic characteristics of hate speech contents on social media.

2.4 Related Work

Despite the frequent occurrence of microaggression, relatively little research has been conducted on the issue. I will look at some study done by other academics on hate speech, read their conclusions, and briefly discuss their findings.

[16] recognized hate speech on social media using the supervised machine learning approach. The researchers' major goal was to get an understanding of how different combinations algorithms work on the dataset to aid in the selection on which approaches to utilize for programmatic hate speech identification. He determined that a combination of TF-IDF, SMOTE, and Logistic Regression was sufficient for delivering the best results using the classifier developed.

The feature engineering techniques are some of the most common methods in the detection of microaggression. [20] and [21] employed n-grams, which are words or character combinations in the feature engineering method. While [20] concludes that while the problem of hate speech detection is far from being solved, the use of a character n-gram based approach provides a solid foundation in hate speech detection. [21] best result was provided by a character 4-gram model, which achieved 78% accuracy, and the researchers concluded that differentiating profanity from hate speech is a difficult assignment.

[22] carried out an experiment utilizing two (2) datasets, and the results showed that there were 21 binary and ternary models for each dataset. The accuracy of the binary models created using Random Forest (RF) and word2vec was found to be higher than that of the Support vector machine(SVM) and naïve Bayes (NB) models.

A 73% F1-score was achieved by the SVM model combined with word2vec, which demonstrated superior performance than the NB and RF models.

The ternary models outperformed the binary models when it came to managing misclassification between hate and non-hate tweets and comments, while the binary models outperformed the ternary models overall. Furthermore, the ternary SVM model with word2vec produced an F1-score of 53%, which was much higher than the scores obtained by the models with NB and RF.

As a result, the models based on SVM employing word2vec performed marginally better than the NB and RF models, which was the main conclusion.

Davidson et al [23] employed TF-IDF, bigram, unigram, and trigram to identify hate speech on social media. They also used part-of-speech and emotion lexicon features. As a result of their research, logistic regression and Linear SVM were shown to be the most accurate models, with an overall accuracy of 0.91, recall of 0.90, and F1 score of 0.90. They concluded that, even though the model performed well about 40% of hate speech tweets were misclassified, resulting in a bias in the classification of hate speech tweets.

2.5 Conclusion

From an analysis of the literature on microaggressions, hate speech and the approaches utilized by various studies, the following conclusions can be drawn:

a. For microaggression to be classified as such, it must be subtle and reveal prejudice or bias against another group either based on gender, race, sexuality etc. These negative expressions which may be verbal or non-verbal are also usually shrouded with ambiguity.
b. Whether a microaggression was intentional or not is immaterial and does not change its nature.
c. There are various forms for microaggression, but all forms retain the core features of microaggressions but with various themes i.e., race, sexual orientation, gender etc.
d. Due to the negative impact of microaggression on the lives of the marginalized individual and the society at large, various approach has been proposed to accurately detect microaggression with varying success.
e. Machine learning methods appears to show the greatest promise in detecting microaggression and this dissertation specifically in chapter 3 shows how machine learning algorithms are utilized to detect microaggression and more importantly, new ways are suggested to improve detection.
f. There is little literature on the detection of microaggression, and although the detection of hate speech and the detection of microaggression seem to be comparable, it has not yet been shown that the model for the detection of hate speech can also accurately detect microaggression.

III. DATASET

This chapter sheds light on the dataset that is used. First, basic features of the set are discussed then an explanation to why this specific dataset has been chosen for this work is explained giving us an understanding on the qualitative and quantitative features of the dataset. The chapter also explained the different classifiers used.

We will be using 2 different methods to test the accuracy of our model, the first is the method of building a single model while the second is using the stacking method. I will go further into the methods as we go further.

Python was used for data processing, manipulation, visualisation, and the development of a machine learning model. Python is a powerful and adaptable data visualisation, exploration, and manipulation tool. Python offers numerous internet support networks, some of which I used and are mentioned in the attestation early in this dissertation. Microsoft Excel was also exploited for some data visualization.

3.1 Data Collection

This work was done using the dataset from (https://www.kaggle.com/arkhoshghalb/twitter-sentiment-analysis-hatred-speech).

It is a collection of 31,962 tweets that have been labelled as either hate speech or non-hate speech.
The terms used in the labelling was gotten from https://hatebase.org/ an organization that has a hate speech terminology database. If a tweet contains words classified as hate speech, the tweet will be labelled as a hate speech and labelled otherwise if it does not contain a hate speech. The labels of the tweets are not perfect and there are some mislabelled tweets, such as the tweet in row 18 (“retweet if you agree”) is labelled as a hate speech. Looking at row 13 (“i get to see my daddy today!! #80days #getti...”), this is well labelled as non-hate speech tweet.

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is s...
1	2	0	@user @user thanks for #lyft credit i can't us...
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation
...	...	...	...
31957	31958	0	ate @user isz that youuu?ð□□□ð□□□ð□□□ð□□□ð□□□ð...
31958	31959	0	to see nina turner on the airwaves trying to...
31959	31960	0	listening to sad songs on a monday morning otw...
31960	31961	1	@user #sikh #temple vandalised in in #calgary,...
31961	31962	0	thank you @user for you follow

31962 rows × 3 columns

Figure 3. 1 Dataset Overview

The dataset comprises three columns, as seen in Figure 3.1. The id column displays the tweets' sequential numbering. The label column distinguishes between the two types of tweets, with 0 denoting non-hate speech tweets and 1 denoting hate speech.

3.2 Data Exploration

When compared to hate speech tweets, Figure 3.2 reveals that a high percentage of tweets are labelled as non-hate speech which has caused the dataset to be imbalanced. An imbalanced dataset occurs when a class of labelled dataset significantly out numbers the other class, this problem of data imbalance is common in classification task. This is crucial information since machine learning algorithms may place a lower value on labels with a small number of them. This big disparity could be due to the difficulties of taking personal bias into account when labelling a tweet or because of the type of dataset being used. This imbalance can cause machine learning algorithms to perform poorly.

It is possible to overcome data imbalance by selecting the most relevant performance metric (e.g. accuracy, recall, F1 and precision) for the task at hand. Alternatively, we can employ a sampling strategy such as undersampling or oversampling, which involves oversampling the minority class of the labelled data set or eliminating "undersampling" the majority class of the dataset to bring each of them up to parity.

In the instance of my dataset, an undersampling strategy was used for the non-hate speech class to

bring the dataset to the level of the hate speech class, which was incredibly low in the first place.

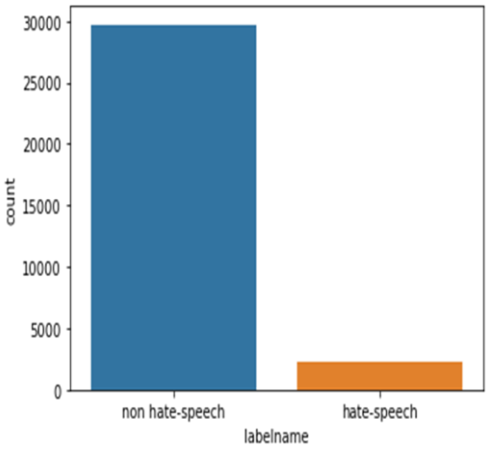


Figure 3. 2 A count plot of class distribution

Figure 3. 2 A word cloud of hate speech tweets.

Figure 3. 4 A word cloud of non-hate speech tweets.

Figures 3.3 and 3.4 show a visual depiction of the hate speech and non-hate speech terms that were most often used in the sample.

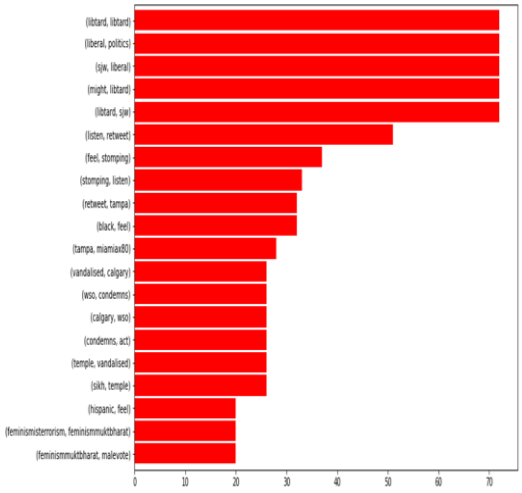


Figure 3. 5 20 most used hate speech bigrams in the dataset.

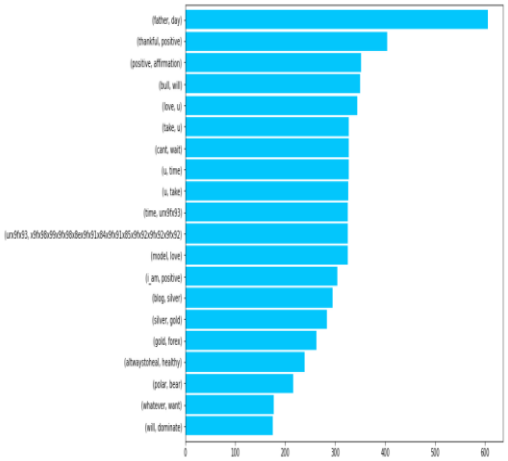


Figure 3. 6 20 most used non-hate speech bigrams in the dataset.

An n-gram is a word sequence consisting of n words of different lengths. A unigram is a single word, but a bigram is two words, and so on and so forth.

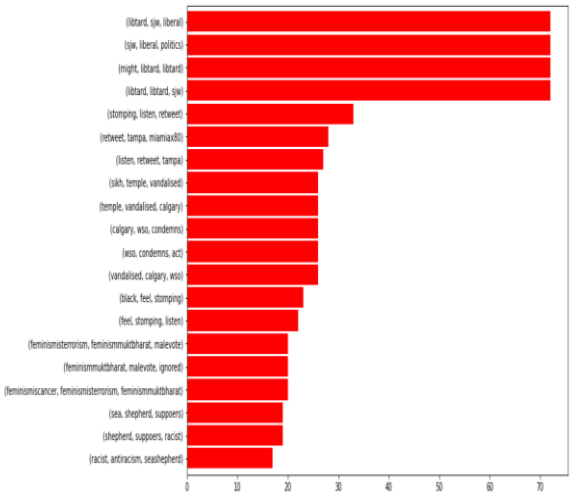


Figure 3. 7 20 most used hate speech trigrams in the dataset.

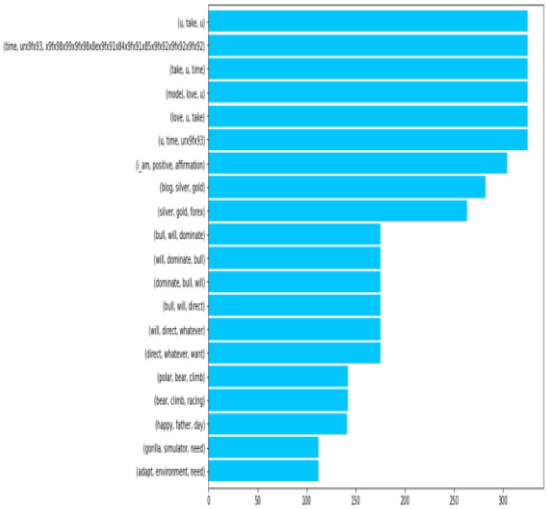


Figure 3. 8 20 most used non-hate speech trigrams in the dataset.

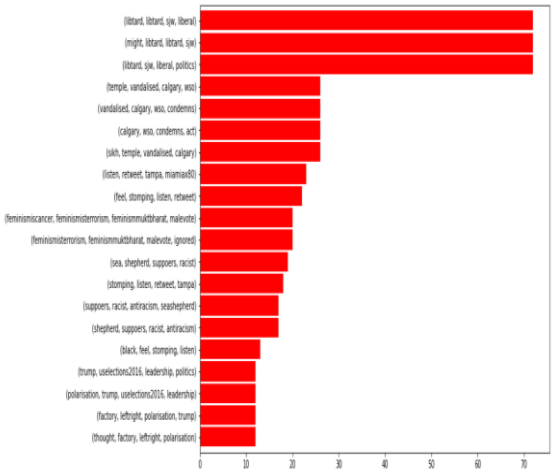


Figure 3. 9 20 most used hate speech quadgrams in the dataset.

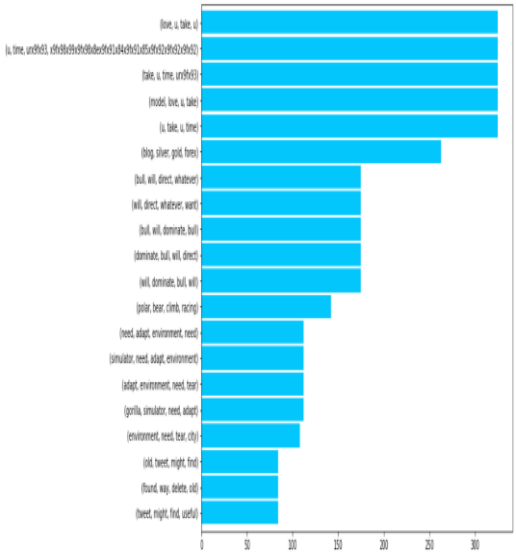


Figure 3. 10 20 most used non-hate speech quadgrams in the dataset.

The 20 most common n-grams that make up hate speech and non-hate speech tweets are depicted in figures above.

Understanding the dataset and how it was labelled is made easier by looking at figures above. The use of n-grams in tweet labelling is intended to give additional information; the greater the degree of n-grams, the more context is provided. In addition to improving predictions, utilising a 3-gram or trigram training model will be more effective in detecting hate speech as compared to when using a single n-gram in building a model.

Higher n-grams are given more weights when coupled with TF-IDF since they appear less frequently when the two are combined. Words with higher weights are assigned greater significance; for example, during classification, a tweet might be classed as either hate speech or non-hate speech depending on which words have the highest weights in the tweet.

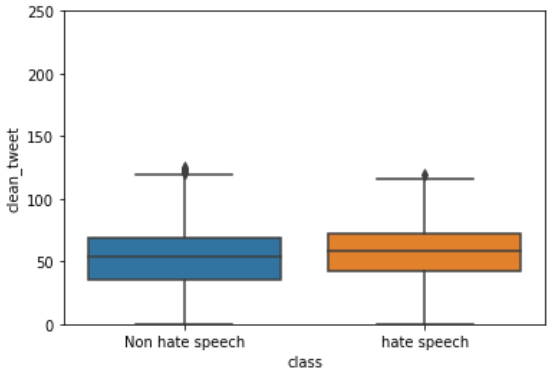


Figure 3. 11 Tweet length by class

The figure above shows that shorter tweets are hate speech tweets in this dataset. This is not due to imbalance of the dataset, but the difference in length could be a factor when TF-IDF is used, this is because when considering TF-IDF, the more a term is used the less important it becomes.

3.3 Data Cleaning

The dataset utilised in this study is made up of tweets, which include special characters such as "@," "#," and punctuation that are not needed because we are interested in the text of each tweet. Regular expressions are character combinations that are used to search for patterns in text [24]. The regular expression library in the python programming language is used to filter out special characters from the dataset, leaving only the text and taking out stopwords, stopwords are words that are usually thought to be ineffective. Most search engines disregard these terms because they are so prevalent that incorporating them would significantly expand the size of the index while providing no improvement in precision or recall, all stop words have been removed and the remaining text has been converted to lowercase and non-alphanumeric character have been removed as well.

The word "user," which was intended to represent account ids for privacy reasons, was omitted from the complete dataset. This was done since the word "user," as described above, was meant to represent account ID but can be misunderstood to signify a hate speech in most circumstances. To avoid this misunderstanding and from misinterpreting the improper term as hate speech, the word "user" has been eliminated.

	id	label	tweet	clean_tweet
0	1	0	@user when a father is dysfunctional and is s...	father dysfunctional selfish drags kids dysfun...
1	2	0	@user @user thanks for #lyft credit i can't us...	thanks lyft credit use cause offer wheelchair ...
2	3	0	bihday your majesty	bihday majesty
3	4	0	#model i love u take with u all the time in ...	model love u take u time ur
4	5	0	factsguide: society now #motivation	factsguide society motivation
...	...	...	...	...
31957	31958	0	ate @user isz that youuu?ð ð ð ð ð ð...	ate isz youuu
31958	31959	0	to see nina turner on the airwaves trying to...	see nina turner airwaves trying wrap mantle ge...
31959	31960	0	listening to sad songs on a monday morning otw...	listening sad songs monday morning otw work sad
31960	31961	1	@user #sikh #temple vandalised in in #calgary,...	sikh temple vandalised calgary wso condemns act
31961	31962	0	thank you @user for you follow	thank follow

31962 rows × 4 columns

Figure 3. 12 comparison of the dataset after cleaning

3.4 Feature extraction

Feature extraction is a technique for extracting essential features from a dataset so that the machine learning algorithm can perform better [39]. After the data was cleaned, features were extracted, the suggested feature extractions are designed to extract the most significant characteristics of a dataset from it. This algorithm works in conjunction with the input of pre-processed and tokenized dataset terms to execute data extractions. All the retrieved characteristics are employed in the training of the models, as well as in the prediction of whether posts and comments are hateful or not. The feature extraction techniques used in this study, such as word n-grams, the Vader Emotion Intensity Analyser, and the TF-IDF, are well-known in the field of text mining methodologies. In each technique, the feature vectors that will be utilised to train the machine learning classifier are provided.

Word N-grams.
An N-gram is a word sequence of n lengths. N-grams come in a variety of lengths, and in this dissertation, I used a combination of one to four N-gram. Because machine learning algorithms cannot use text, these texts are turned into a bag of words including the number of times (count) a word appears, but instead of a single word, a combination of one to four N-grams is employed in this case [56].

Vader Sentiment Intensity Analyzer.
The NLTK python library's Vader (Valence Aware Dictionary and Sentiment Reasoner) sentiment intensity analyser is a sentiment analysis tool that is used to evaluate text sentiments and provide scores based on whether the sentiment is good, negative, or neutral. It also calculates a compound score by adding the positive, negative, and neutral scores together. It is standardised between -1 and +1, indicating the degree to which a sentence is negative or positive [55].

	label	clean_tweet	Negative	Positive	Neutral	Compound
0	0	father dysfunctional selfish drags kids dysfun...	0.655	0.000	0.345	-0.7650
1	0	thanks lyft credit use cause offer wheelchair ...	0.000	0.379	0.621	0.6705
2	0	bihday majesty	0.000	0.000	1.000	0.0000
3	0	model love u take u time ur	0.000	0.512	0.488	0.6369
4	0	factsguide society motivation	0.000	0.545	0.455	0.3400
...	...	...	...	...	...	...
31957	0	ate isz youuu	0.000	0.000	1.000	0.0000
31958	0	see nina turner airwaves trying wrap mantle ge...	0.153	0.302	0.545	0.4588
31959	0	listening sad songs monday morning otw work sad	0.508	0.000	0.492	-0.7351
31960	1	sikh temple vandalised calgary wso condemns act	0.355	0.000	0.645	-0.5106
31961	0	thank follow	0.000	0.714	0.286	0.3612

31962 rows × 6 columns

Figure 3. 13 result of the Vader sentiment intensity analyzer

TF-IDF Vectorizer.
Another popular feature extraction method for text analysis is TF-IDF (Term Frequency- Inverse Document Frequency). This also changes text to a format that machine learning algorithms can understand. It gives each phrase in the text a weight based on its value in the entire dataset, decreasing the impact of less weighted terms [18]. Term frequency refers to how many times a word appears in a text, whereas Inverse Document Frequency devalues words that appear frequently in a document.

3.5 Classifiers
To compare the performance of 10 machine learning models, training tweets were obtained from a database, classified, and processed by ten different machine learning classifiers. The test tweets, which had been pre-processed and their characteristics collected, they were put into the classifier after it had been trained to identify polarity, then the accuracy of the classifiers was compared. The number of classifiers to be employed is at the discretion of the researcher. These ten classifiers were chosen because of their high performance in tweet classification and because they have been used by different authors on various occasions.

The writer of the article [49] identified the five (5) most frequent classifiers used in a classification assignment. [49] cited Logistics Regression, K-Nearest neighbors, Decision Tree, Support Vector Machine, and Naive Bayes. In this dissertation, four (4) of the classifiers specified by the author will be applied.

[50] investigated the application of several classifiers with the goal of obtaining the best results in the resolution of health-related concerns. [50] employed four (4) different classifiers on four (4) different datasets to arrive at their conclusions. The researcher investigated the usage of K-nearest neighbour, Decision Tree, Random Forest, and support vector machine as well as other machine learning techniques. To detect online toxicity, [51] used regularised logistic regression, random forest, Naive Bayes, and support vector machines (SVM), while to detect hate speech on social media, [52] used logistic regression, Naive Bayes, and support vector machines (SVM).

In conclusion, it is plausible to conclude that the use of classifiers in classification tasks is growing more widespread as the machine learning technique becomes more prevalent.

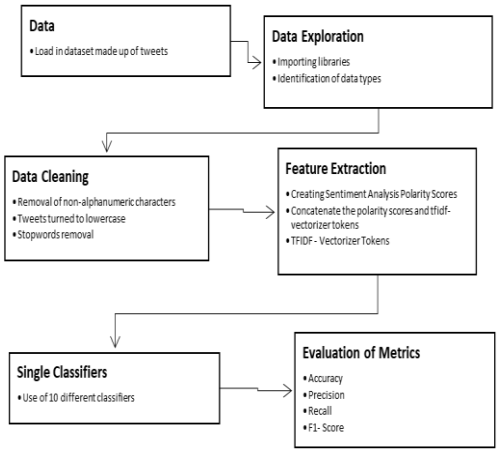


Figure 3. 14 Process Flow using a single classifier

3.6 Stacking Classifier

Stacking is a machine learning approach for combining different classification algorithms. It is also referred to as stacked generalization in most cases. It is used to combine the predictions of several classification algorithms and it was first introduced by [25].

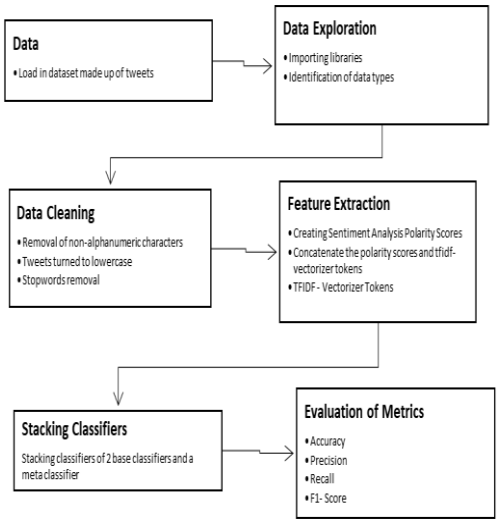


Figure 3. 15 Process Flow using stacking classifier

3.7 Evaluation

To evaluate the performance of our classifiers, this dissertation used a confusion matrix. A confusion matrix is a table that is used to validate the performance of a classifier based on test data.

After doing that, I proceeded in calculating the accuracy of the classifier, precision, recall and F1 score. For our dataset, if a tweet is micro aggressive and the classifier was able to detect and classify it as one, then a True Positive (TP) is declared. But if wrongly detected and classified as non-micro aggressive then a false negative (FN) is declared. In the same light, if the tweet is non-micro aggressive and the classifier was able to detect and classify appropriately as one, then a True Negative (TN) is declared, otherwise, a False Positive (FP) is declared.

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 3. 16 Confusion matrix

Accuracy

This is how correct the classification process is. It is the degree to which the outcome of a predicted value is true and basically refers to the closeness of a prediction to its true value.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

This is the detection exactness, which shows how good a model is at whatever it predicts. It measures the number of predicted values that are the same with the true values, for example, a model predicted 10 items and turned out to be false on one item, the accuracy is 9/10 (90%).

$$Precision = \frac{TP}{TP + FP}$$

Recall

This is the ratio of the correct predictions to the total number of correct items in the dataset. For example, if there are 10 items and the model detects 5 items and they are correct, the recall is 50%. If it detects all 10, the detector can be said to have 100% recall.

$$Recall = \frac{TP}{TP + FN}$$

F1 Score

This shows the balance between precision and recall, because there is usually a trade-off between precision and recall (Ying Chen et al, 2020).

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision\ +\ Recall}$$

3.8 Machine Learning

For detecting hate speech, I used a variety of machine learning approaches to determine which one was the most efficient and accurate. Choosing the appropriate classifier is one of the most crucial phases in machine learning, since it is impossible to choose the most successful model without first understanding how each algorithm performs and what the desired result is.

My goal is to evaluate 10 machine learning algorithms to see which one is the most effective and accurate in detecting hate speech. My initial step will be to examine each classifier and compare its properties to those of my dataset to see whether they are both a good match.

3.8.1 Linear Support Vector Machine

Support Vector Machine commonly known as (SVM) is a classifier that finds an optimal hyperplane that maximizes the margin between two classes. This classifier is widely utilized in text classification. In SVM, the goal is to design a hyperplane that classifies all training vectors in two different classes and select the class that leaves the maximum margin as the best choice.

This classifier was chosen because it is good for text classification, handles data with a lot of features, and performs well with datasets with outliers, my dataset satisfies most of the criteria.

3.8.2 Random Forest

A random forest classifier is a machine learning algorithm made up of several decision trees classifiers. It merges a collection of independent decision trees to get a more accurate and stable prediction. Random forest is a popular classifier in text classification and has been used by authors such as [26], [27], were [27] used in the detection of ransomware attacks on computer systems.

In addition to being a versatile classifier that is relatively simple to comprehend, random forest is also capable of properly managing large datasets, which is a property of my dataset as well. Random forest algorithm is more accurate in forecasting outcomes when compared to the decision tree approach. It also performs well with datasets that are imbalanced and include many outliers, which is a characteristic of mine as well.

3.8.3 Extreme Gradient Boosting (Xgboost)

XGBoost is a machine learning algorithm created to improve performance and speed. It is an implementation of gradient tree boosting. Authors like [28],[29] have used it in their works. It is an ensemble technique that aggregates several weak models to create one strong model. The weak model learns from its mistakes with each iteration.

This classifier was chosen because it is a very popular solution for Kaggle competitions, it is very good for classification problems, it is very flexible and performs better than random forest, also there are claims that it is 10 (ten) times faster than already existing solutions [30].

3.8.4 Logistic Regression CV

Logistic Regression CV: This is an implementation of the logistic regression using either liblinear, newton-cg, sag or LBFGS optimizer. The LBFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) optimizer has been used in this work because it helps in managing memory and faster with large datasets. The L2 penalty was used as it does not discard any datapoint no matter how close to zero it is [31],[32]. have used this algorithm for text classification purposes and has shown great performance. I selected logistic regression over other methods for text classification because of its

simplicity and widespread use [33]. LR is well-known for achieving strong results in online hate detection while requiring just a little amount of model complexity [34]. Therefore, it seems logical to use the logistics regression when comparing different models. According to Gunasekara and Nejadgholi [35], "traditional machine learning classifiers such as linear regressions models have also been successfully utilised to identify abusive online language." In addition to [36], LR has been utilised for online hate detection at least by Xiang et al. [37], and it has been used for online hate detection by researchers.

3.8.5 Multinomial Naïve Bayes
The Bayes theorem is employed in this method to predict the likelihood of an event occurring based on data evidence. The Nave Bayes predictors are self-contained, i.e., they assume that all features are independent of one another.
I chose this classifier because it is good for categorical classification, it is very effective as well.

3.8.6 LightGBM
Using the decision tree approach, this is a high-performance gradient boosting framework that is fast, distributed, and scalable. This model divides the tree into leaves, and while it is prone to overfitting, it may be easily adjusted by determining the depth at which the tree is divided. This classifier was chosen because it is both fast and extremely efficient; it is also highly prevalent in classification tasks, which is why it was picked.

3.8.7 AdaBoost
To improve the performance of a weak classifier, Adaboost employs a combination of many low-performing classifiers to obtain a higher level of performance. This classifier does not overfit and has a small number of parameters, which will allow the model to run more quickly. This classifier uses a collection of decision trees to train duplicates of the original dataset; it focuses particularly on datapoints that are difficult to classify [38]. [39] also pointed out the fact that it performs better than other algorithms in classification problems.

Adaboost is usually employed for classification problems, and it has a long history of being successful as a classification technique. Therefore, it was selected.

3.8.8 Extra Tree Classifier
Extra tree classifier, like random forest, is an ensemble machine learning model; here, the classifier integrates predictions produced by several decision trees; the choice on which trees to utilise is made at random. This classifier's execution is quicker and, in most cases, surpasses the random forest technique. One of the reasons this classifier was selected is its speed. [40] pointed out its speed computationally comparing to the random forest method.

3..8.9 Logistics Regression
It is a forecasting analysis. A sigmoid function is used to describe the connection between one independent variable and one or more independent variables [41].

This is a widely common classification method that is most often employed when the dataset has a binary output, meaning when the output is classified as either 0 or 1. My datasets have a lot of characteristics that this attribute fits neatly within. As previously stated by other researchers, logistic regression is simple to use, analyse, and comprehend, and it is also very efficient to train.

3.8.10. SGD Classifier
Stochastic Gradient Descent (SGD) is a very simple and effective method in text classification, it prides in being very east to implement and very efficient in its results. It is a simple yet effective optimization technique. In situations where the datapoints to be trained are arrays of floating-point integers for each feature, SGD performs well. Because TFIDF-vectorizer turns text to floating points, it is an excellent classifier to employ in text classification tasks used by [42] and [43] because it converts text to floating points.

This classifier outperforms all other members of the gradient family in terms of performance. This system works very well with enormous datasets and can be updated on a regular basis. Given that we are dealing with big datasets, the qualities listed above are very applicable to our dataset, which has the potential to rise dramatically from one year to the next.

3.9 Experimental setup of single classifiers used
For the detection of hate speech, a collection of tweets was used as dataset where 70% of data was considered for training and 30% for testing.

From this dataset, a tweet is classified as either showing hate speech or not and this was achieved by using selected machine learning algorithms listed above. The setup of the machine learning technique was implemented on a jupyter notebook environment and is made up of several steps.

The data is first loaded with the pandas library, after which the python natural language toolkit NLTK is used to clean the data and construct a corpus from the dataset. NLTK is a package that has numerous features for dealing with human language. The cleaning of the data involves the use of regular expression to remove non-alphanumeric characters, since only words and not numbers or special characters are of interest or importance, the data is also converted to lowercase, if words are the same but are in different cases these are seen as two different words which can leave words that are not useful in the dataset. The NLTK library has a functionality that helps in removing words that are not useful or that have no importance in the data, in this data one of such word is "user" as it is used to hide the identity of authors in the data for privacy.

In the next step, data exploration is done to extract meaning and information form the data. Another step is the feature extraction step where sentiment analysis scores were created for each tweet in the data, the data in text form cannot be used by the machine learning algorithm so a TFIDF-vectorizer token of one-to-four word n-grams is created from the text and the output which is usually a number for each word in the tweet is concatenated with the polarity score of each tweet and this is split into training and test data which is supplied to the machine learning classifier. Polarity scores are an important feature which can be added to the training data to help in better classification, a table showing a sample of polarity score for some tweets is shown in chapter 3.3. After the training is completed, the predictions of the classifier are compared with the test data set aside and the metrics used in these comparisons are accuracy, precision, recall and F1-score.

3.10 Stacking
In Stacking, the predictions of combined classifiers are used as input for the prediction of the final classifier. This process is carried out in two steps:

The first step involves the use of several classifiers known as the base classifiers, these base classifiers are trained using the training data from the datasets and their output (classification) is used as input for the second step.

The second step uses a final classifier called a meta-classifier which is trained on the output of the base classifiers. The base classifiers are also called Level-0 classifiers and meta-classifiers as Level-1 classifiers. The Level-1 classifiers job is to determine how to effectively combine the predictions of the Level-0 classifiers [44].

There are several advantages to using stacking classifiers.

- It has shown to improve prediction accuracy and detection speeds.
- It has the potential to help balance bias-variance trade-off better than individual classifiers. As Zhang rightfully said and I quote "Accuracy (F-score) was improved in mobile malware detection through the use of a stacked classifier" [45].

Using a classifier in conjunction with others in a stacking classifier outperformed the individual classifier [46].

However, there are the down sides to it as well, the disadvantages of using stacking classifier are that it introduces computational complexity and can take a longer time to train and require more memory. And as Giraud said, stacking classifiers may perform better or just as good as individual classifiers but not always [47].

3.11 Experimental setup for stacking classifiers
The stacking classifier used in this dissertation is made up of the two (2) best performing classifiers and a meta-classifier. The meta-classifier used is the Logistic Regression. The choice for Logistic regression was influenced by its high performance in NLP problems and it has shown better performance in several studies such as [39], [42] and [48].

I used as a first level (base), the predictions of the two (2) best performing classifiers and then used Logistics Regression at the second level to predict the output from the earlier first-level predictions and after that I used the next 2 best performing classifiers and the process goes on again till we test all 10 classifiers. Aside the fact that we are using the meta classifier to predict models, the experimental setup for the stacking classifier is same as described above when using a single classifier.

IV. RESULTS

The results for the various classification algorithms used in this dissertation is presented in the various diagrams below. From the results gotten with the dataset, it is seen that there are performance differences when the data is balanced and when the data is not balanced. The performance differences may lead to the answer of which metric is more important.

Results for Imbalanced Dataset

The accuracy ratings of the different classifiers, as shown in Figure 4.1, are quite high, with the lowest being 94% for the least accurate classifier. Apart from the high accuracy, the classifiers were able to correctly identify non-hate speech tweets, but they had difficulty correctly identifying hate speech tweets since the recall rate for hate speech tweets is quite low.

Linear SVC performed the best in classifying the hate speech tweets. The precision came out to be 80% and recall was 61%. This means that most of its predicted sentiment was accurate when compared to its training set by a small amount. Hence 80% of the hate speech tweets were relevant and 61% of the hate speech tweets were retrieved. This means very few false positives were found for the hate speech.

The classification of tweets from test data that had not been seen by the classifiers before, trained on imbalanced data. Out of 17196 tweets from the test dataset, it was only able to classify 945 as hate speech, which is 5.5% compared to 7.01% classification of hate speech tweets from the original dataset. From the result it is seen that high accuracy with poor recall does not give a very good result.

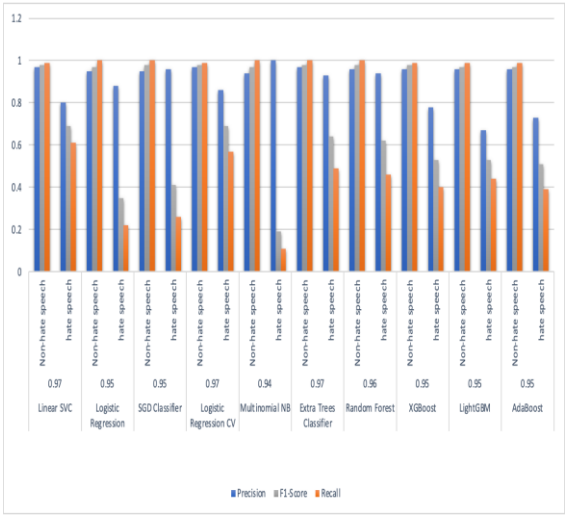


Figure 4. 1 Evaluation results for imbalanced dataset

Results for Balanced Dataset

From Figure 4.3 the accuracy scores of the various classifiers are not as high as the accuracy from the imbalanced dataset. Machine learning classifiers are built on the premise of balanced class distribution, but when the dataset is not balanced it tends to classify mainly on the majority class and this gives it very high accuracy. The highest accuracy gotten is 88% but the recall is higher than that of the imbalanced dataset. In 6 out of the 10 classifiers used there is better recall for hate speech tweets.

The classification of tweets from test data that had not been seen before by the classifiers trained on balanced data. Out of 17196 tweets it was only able to classify 3332 as hate speech tweets, which is 19.38% compared to 7.01% classification of hate speech tweets from the original dataset.

When the performance of the classifiers is compared it seen that for better classification of tweets, the recall metrics is better than the accuracy metrics. To get better recall, balancing the labels (class) of the dataset is advised.

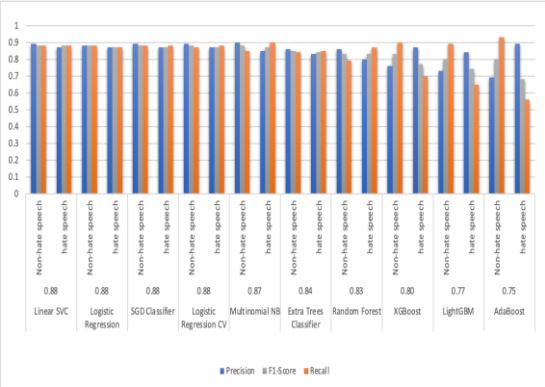


Figure 4. 3 Evaluation results for balanced dataset Results of Stacking Classifier

The outcome of the stacking classifier has shown that combining high performing classifiers does not always result in superior performance than individual high performing classifiers in terms of classification accuracy.

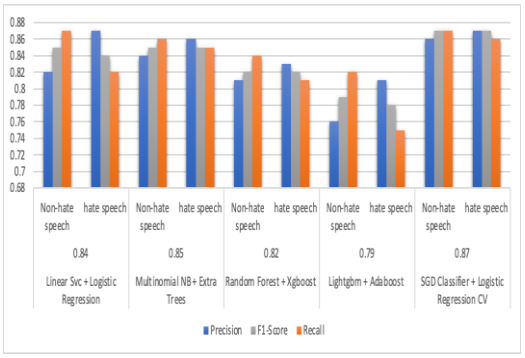


Figure 4. 5 Evaluation results from stacking classifiers

4.4 Comparison of results to other authors

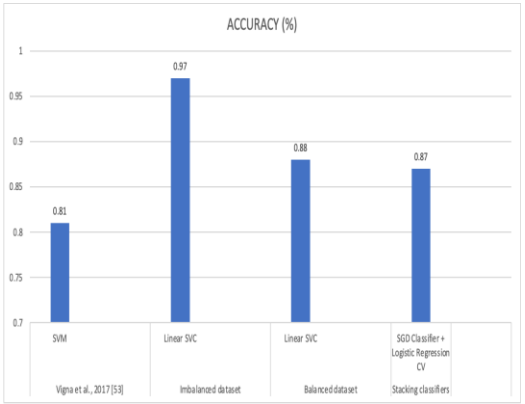


Figure 4. 6 Comparison of results w.r.t. accuracy

Accuracy measures how many tweets are correctly identified in each of the class. First, comparing my results with those presented by Vigna et al [53], [53] achieved an overall accuracy of 81% and used the support vector machine classifier to achieve that. While in my result, the linear SVC gave me the highest accuracy for the imbalanced dataset as 97% and the 88% for the balanced dataset, using the stacking classifier, the combination of SGD classifier and logistics regression CV gave me the highest accuracy of 87%. It is safe to conclude that the linear SVC is the top performer in this category. Although as usual a concrete conclusion cannot be given, however it is safe to recommend the linear SVC in building a model when working with a dataset with the characteristics that mine possess and looking at the accuracy of a model.

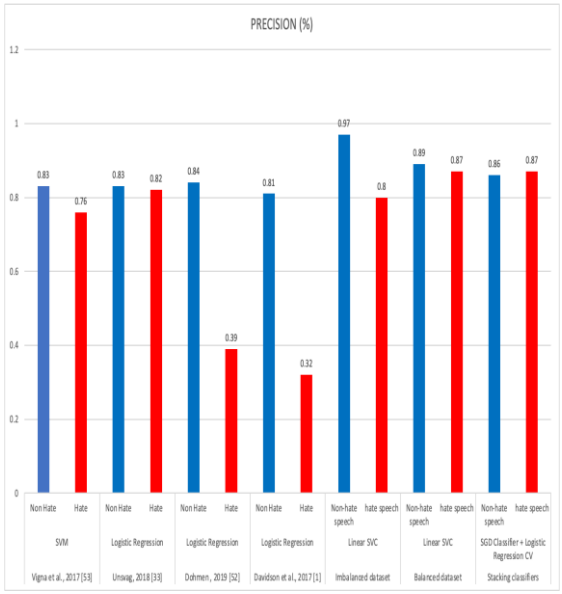


Figure 4. 7 Comparison of results w.r.t. precision

Precision measures how many tweets, among those classified as expressing hate speech. Fig. 4.7 above compares the work of different researchers, from the result above, it is seen that linear SVC outperforms all other classifiers and is ranked as the best results in terms of precision, achieving a precision of 87%. I am not considering the result of the imbalanced dataset because of the misclassification it encountered.

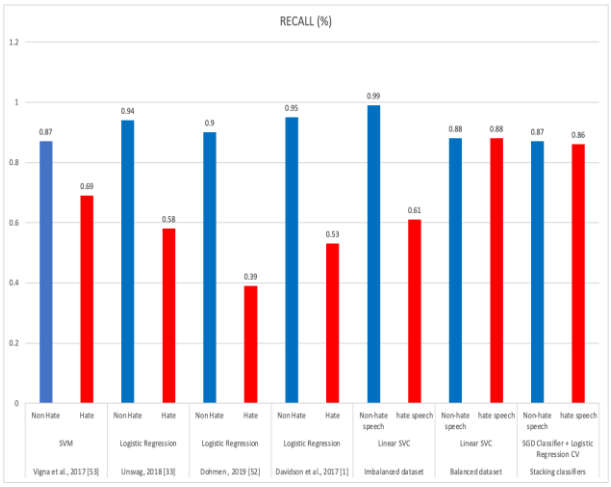


Figure 4. 8 Comparison of results w.r.t. recall

The number of successfully identified tweets in the entire set is represented by the recall metrics. A low recall rate indicates that many relevant tweets are left unidentified. Just like in the case of precision, the linear SVC proved to be the highest when comparing it with other researchers, I also noted that for 3 consecutive times logistics regression has been the highest performer, this is not strange because this classifier is one of the most commonly used , however if we look at the works of [53],[33],[52] and [1], it is evident that the linear SVC classifier was not selected as a model to test. This brings me to the conclusion that perhaps the linear SVC compared to the logistics regression is a better performer and more researchers should investigate this classifier when building a model with a similar dataset.

CONCLUSION

5.1 Summary

This dissertation was originally intended to be about the detection of microaggression on social media platforms using machine learning; however, during the process of analysing and obtaining my dataset, due to the numerous challenges encountered, the most significant of which was the unavailability of datasets specifically on microaggression, the dissertation took a detour and became more focused on the detection of hate speech. The dataset that was utilized were the tweets made by various people on twitter.

These tweets were classified as either displaying hate speech or not by utilising phrases from hatebase.org that have already been classified. The features utilised to train the models were a mix of n-grams, polarity scores, and TF-IDF features, which were combined to form the final model. Ten (10) different machine learning classifiers were trained on these characteristics to see how well they would perform, and the top performers were merged using a stacking classifier to see if there would be any additional improvement in the findings.

When the dataset is imbalanced, the linear support vector classifier (SVC) was the top performer, achieving an accuracy of 97%, while precision was 80%, recall was 61%, and the F1 score was 69% for the hate speech class. When dealing with the balanced datasets, the top performer was the linear support vector classifier (SVC), which achieved an accuracy of 88% in testing, F1 score of 88%, recall of 87%, and precision scores of 87% for this category of hate speech. The combination of the SGD classifier and the Logistic Regression CV classifiers are the top performers for the stacking classifier, with an accuracy of 87%, precision of 87%, recall of 86%, and an F1 score of 87% achieved in the hate speech class, respectively.

The performance of the balanced dataset was not improved by the stacking classifier, but it did yield a result that was quite near to it. Even though the accuracy of the imbalanced dataset is greater than that of the balanced dataset, the balanced dataset performed better on the other assessment criteria. In addition, the precision, recall, and F1 score of the models trained on the balanced dataset demonstrate that they are more capable of distinguishing between non-hate speech and hate speech tweets than the models trained on the imbalanced dataset, which are only capable of classifying tweets that contain non-hate speech. These findings demonstrate that the combination of characteristics and model developed in this dissertation is efficient for the identification of hate speech in social media platforms.

5.2 Recommendations

It is important to pay attention to how a dataset for the identification of microaggression is labelled, considering cultural, ethnic, and religious sensitivity.

In addition, there should be an emphasis on developing datasets and models that are specific to a particular group; for example, members from a single community could label their own dataset and evaluate whether this is a better strategy than a one size fits all labelling system.

5.3 Future works

Because creating dataset is human effort and can be very stressful, the creation of dataset has been lagging. Researchers should concentrate on the creation of datasets for microaggression rather than on the development of models. To tackle this problem, we will make sure that we have exhausted all the publicly available dataset, after that we can explore the sharing agreement method where an agreement is made between the researcher and the focus area e.g a hospital, a school, a club etc. and data is collected from the recipients of the focus area. Also, the use of crowdsourcing platforms is another viable means of gathering data, additionally, the use of the traditional data collection methods such as through interviews, questionnaires and surveys is also an option that can never go out of style.

Also, for future study, researchers should move towards investigating alternative machine learning approaches, such as Neural Networks and deep learning and using bigger datasets of tweets. This is because deep learning approaches have been stated by [54] to outperform other machine learning approaches in performance, especially when it comes to dataset that are huge in size.

Micro-aggression

REFERENCES

[1] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber,"Automated Hate Speech Detection and the Problem of Offensive Language," https://github.com/t-davidson/hate-speech-and-offensive-language, 2017.

[2] D. W. Sue, J. Bucceri, A. I. Lin, K. L. Nadal, & G. C. Torino, "Racial microaggressions and the Asian American experience,". Cultural Diversity and Ethnic Minority Psychology, 13(1), 72–81, 200. Assessed on: May 05, 2021. [Online]. Available: https://doi.apa.org/doiLanding?doi=10.1037%2F1099-9809.13.1.72.

[3] T. Monnica Williams, "Microaggressions: Clarification, Evidence, and Impact," Perspectives on Psychological Science, vol. 15(1) 3-26, 2019. Accessed: Jan. 10, 2022 [Online]. Available: https://journals.sagepub.com/doi/pdf/10.1177/1745691619827499

[4] L. Edilma Yearwood, "Journal of Child and Adolescent Psychiatric Nursing," vol. 26 Issue 1, 2013, p.98 99. Assessed on: May 01, 2021 [Online]. Available: doi: 10.1111/jcap.12021

[5] Nordmarken Sonny, "Microaggressions," TSQ: Transgender Studies Quarterly 1(1-2): 129-134. Assessed on: May 10, 2021 [Online]. Available: doi: 10.1215/23289252-2399812

[6] S. Periyakoil Vyjeyanthi MD, Chaudron Linda MD, MS, V. Hill Emorcia PhD, Pellegrini Vincent MD, Neri Eric MS, C. Kraemer Helena PhD, "Common Types of Gender-Based Microaggressions in Medicine, Academic Medicine," Journal of the association of American Medical Colleges, Vol. 95 - Issue 3 - p 450-45, March 2020. Accessed on: Sept. 23,2021[Online].Available:https://journals.lww.com/academicmedicine/fulltext/2020/03000/common_types_of_gender_based_microaggressions_in.37.aspx

[7] Derald Wing Sue, Microaggressions and Marginality: Manifestation, Dynamics, and Impact. NJ: John Wiley & Sons, 2010.

[8] Wing, Capodilupo, Torino Bucceri Holder Nadal, Equilin, "Racial Microaggressions in Everyday Life: Implications for clinical practice," American Psychologist vol. 62, 4, 27-286, 2007. Accessed on: Sept. 21, 2021 [Online]. Available: https://sph.umn.edu/site/docs/hewg/microaggressions.pdf

[9] T. Jessica DeCuir-Gunby and W. Norris Gundy Jr, "Racial Microaggressions in the Workplace: A Critical Race Analysis of the Experiences of African American Educators," Urban Education vol. 51(4) 390-414. Accessed on: Sept. 21, 2021 [Online].

Available:https://doi.org/10.1177/00420859166 28610

[10] R. Susan Torres-Harding, L. Alejandro Andrade Jr., and E. Crist Romero Diaz,"The Racial Microaggressions Scale (RMAS): A New Scale to measure experiences of racial microaggressions in people of colour", Cultural Diversity and Ethic Minority Psychology, vol. 18(2) 153-164, Apr. 2012. Accessed on: Sept. 28, 2021 [Online]. Available: https://psycnet.apa.org/fulltext/2012-09819-005.pdf

[11] G. Wong, A.O. Derthick, E.J.R. David, Anne Saw, and Sumie Okazaki, "The What, the Why, and the How: A Review of Racial Microaggressions Research in Psychology", Race Soc Probl, vol. 6(2) 181–200, June 2014. Available: doi: 10.1007/s12552-013-9107-9

[12] C. Gilbert Gee, S. Michael Spencer, Juan Chen, David Takeuchi, "A Nationwide Study of Discrimination and Chronic Health Conditions Among Asian Americans", American Journal of Public Health 97, no. 7 pp. 1275-1282, July 1, 2007.Available:https://doi.org/10.2105/AJPH.2006.091827

[13] D.W. Sue, K.L. Nadal, C.M. Capodilupo, (2008), "Racial, gender and sexual orientation microaggressions: Implications for counselling and psychotherapy". In D. W. Sue & D. Sue (Eds.), Counselling the culturally diverse: Theory and practice (5th ed., pp. 105-130). Hoboken, NJ: John Wiley & Sons.

[14] Peter Hasson, UNH: gender roles and sexist jokes cause migraines, heart disease among women, Campus Reform Oct. 2015. Accessed on: Sept. 2, 2021.[Online].Available:https://campusreform.org/?ID=6897

[15] R. Susan Torres-Harding, L. Alejandro Andrade, Jr., and E. Crist Romero Diaz, "The Racial Microaggressions Scale (RMAS): A New Scale to measure experiences of racial microaggressions in people of colour", American Psychological Association, vol 18(2), 153 – 164, 2012. Accessed on: May 9, 2021. [Online]. Available: https://psychology.umbc.edu/files/2016/10/The Racial-Microaggressions-Scale_Torres-Harding-2012.pdf

[16] Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat, "Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach", Department of Computer Engineering, Maharashtra Institute of Technology, Pune, India, vol 18(2), 153 – 164, Sep. 2018. Accessed on: May 9, 2021. [Online]. Available: doi: arXiv:1809.08651v1Belima 2016

[17] Sanjay Dey, Sarhan Wasif, Dhiman Sikder Tonmoy, Subrina Sultana, Jayjeet Sarkar, and Monisha Dey, "A Comparative Study of Support Vector Machine and Naive Bayes Classifier for Sentiment Analysis on Amazon Product Reviews", International Conference on Contemporary Computing and Applications 5-7, Feb. 2020. Accessed on: May 9, 2021. [Online]. Available: doi: 10.1109/IC3A48958.2020

[18] Ziqi Zhang and Lei Luo, "Hate Speech Detection: A Solved Problem? The Challenging case of Long Tail on Twitter". Accessed on: Jan. 22, 2022. [Online]. Available: doi: arXiv:1803.03662v2

[19] Zeerak Waseem and Dirk Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter,". In Proceedings of the NAACL student research workshop, pages 88–93, 2016. Accessed on: Jan. 22, 2022. [Online]. Available: doi: arXiv:1803.03662v2

[20] Shervin Malmasi and Marcos Zampieri, "Detecting hate speech in social media," In Proceedings of the International Conference Recent Advances in Natural Language Processing, pages 467–472, 2017. Accessed on: Jan. 22, 2022. [Online]. Available: doi: 10.26615/978-954-452-049-6_062

[21] Sudhir Kumar Mohapatra, Srinivas Prasad, Dwiti Krishna Bebarta, Tapan Kumar Das, Kathiravan Srinivasan and Yuh-Chung Hu, "Automatic Hate Speech Detection in English-Odia Code Mixed Social Media Data Using Machine Learning Techniques", Appl. Sci. 2021, 11, 8575. Accessed on: Jan. 22, 2022.

Available: https://doi.org/10.3390/app11188575

[22] T. Davidson, D. Warmsley, M. Macy, I. Weber, "Automated hate speech detection and the problem of offensive language". In Proceedings of the Eleventh International AAAI Conference on Web and Social Media, Montreal, QC, Canada, vol. 11, 15–18 May 2017.Available:https://ojs.aaai.org/index.php/ICWSM/article/view/14955

[23] Karim, A. Azhari, M. Alruily, H. Aldabbas, and S. B. Belhaouri, "Classification of Google Play Store Application Reviews Using Machine Learning". Preprints version 1, 2020. Accessed on:Nov.22,2021.Available:doi:10.20944/preprints202007.0646.v

[24] H.David Wolpert, "Stacked Generalization". Neural Networks, Vol. 5, Issue 2, 1992, Pages 241-259. Nov. 22, 2021. Available: https://doi.org/10.1016/S0893-6080(05)80023-1.

[25] H.S.M. Abbas, , Z.H. Qaisar, , X. Xu, and C. Sun, , "Nexus of E-government, cybersecurity and corruption on public service (PSS) sustainability in Asian economies using fixed-effect and random forest algorithm", Online Information Review, Sep. 2021. Accessed on: Nov. 23, 2021. Available: doi: https://doi-org.ezproxy-s2.stir.ac.uk/10.1108/OIR-02-2021-0069

[26] A.M. Maigida, S.M. Abdulhamid, Olalere, M. et al., "Systematic literature review and metadata analysis of ransomware attacks and detection mechanisms". J Reliable Intell Environ 5, 67–89, May 2019. Available: doi: https://doi.org/10.1007/s40860-019-00080-3

[27] Bohdan, M. pavlyshenko. (2019). Machine-Learning Models for Sales Time. 1–11. https://doi.org/10.3390/data4010015

[28] Suryachandra, P. (2020). CLASSIFICATION OF THE FEATURE-LEVEL RATING SENTIMENTS FOR TELUGU LANGUAGE REVIEWSUSINGWEIGHTED.11(12),373383. https://doi.org/10.34218/IJARET.11.12.2020.040

[29] Chen, T., & Guestrin, C. (2016). XGBoost : A Scalable Tree Boosting System. 785–794.

[30] Nemati, A. (2018). Gender and Age Prediction Multilingual Author Profiles Based on Comments. 1–8.

[31] Zheng, Y. (2019). An Exploration on Text Classification with Classical Machine Learning Algorithm. 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 81– 85. https://doi.org/10.1109/MLBDBI48998.2019.00023

[32] Unsvag EF, Gambäck B. The efects of user features on twitter hate speech detection. In: Proceedings of the 2nd workshop on abusive language online (ALW2) (2018), 75–85

[33] Gröndahl T, et al. All you need is" Love": evading hate-speech detection. arXiv preprint arXiv:1808.09115. 2018

[34] Gunasekara I, Nejadgholi I. A review of standard text classifcation practices for multi-label toxicity identifcation of online content. In: Proceedings of the 2nd workshop on abusive language online (ALW2); 2018. P. 21–5

[35] Fortuna P, Nunes S (2018) A survey on automatic detection of hate speech in text. ACM Comput Surv 51:4. https:// doi.org/10.1145/3232676

[36] Xiang G, et al. Detecting ofensive tweets via topical feature discovery over a large scale twitter corpus. In: Pro ceedings of the 21st ACM international conference on Information and knowledge management; 2012. P. 1980–4

[37] Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences, 55(1), 119–139. https://doi.org/https://doi.org/10.1006/jcss.1997.1504

[38] Rustam, F., Mehmood, A., Ahmad, M., Ullah, S., Khan, D. M., & Choi, G. Y. U. S. (2020). Classification of Shopify App User Reviews Using Novel Multi Text Features. 8.

[39] Al-sariera, Y. A., Adeyemo, V. E., Balogun, A. O., & Alazzawi, A. K. (2020). AI Meta-Learners and Extra- Trees Algorithm for the Detection of Phishing

Websiteshttps://doi.org/10.1109/ACCESS.2020.3013699

[40] Wenando, F.A., T.B. Adji, and I. Ardiyanto, Text classification to detect student level of understanding in prior knowledge activation process. Advanced Science Letters, 2017. 23(3): p. 2285-2287.

[41] Abro, S., Shaikh, S., & Ali, Z. (2020). Automatic Hate Speech Detection Using Machine Learning: A Comparative Study. www.ijacsa.thesai.org. https://thesai.org/Downloads/Volume11No8/Paper_61-Automatic_Hate_Speech_Detection.pdf.

[42] Karim, A., Azhari, A., Alruily, M., Aldabbas, H., & Belhaouri, S. B. (2020). Classification of Google Play Store Application Reviews Using MachineLearning.July,122https://doi.org/10.20944/preprints202007.0646.v1

[43] Kaushalya. (2021). Comparison of Machine Learning Classifiers for Sentiment Analysis in Hotel Reviews. 2021.

[44] Ghalejoogh et al., 2019

[45] Zhang et al., 2015).

[46] Chand et al., 2016).

[47] (Giraud- carrier et al., 2004).

[48] Katona, E., Buda, J., & Bolonyai, F. (2021). Using N-grams and Statistical Features to Identify Hate Speech Spreaders on Twitter.

[49] Yulia Gavrilova, Classification Algorithms: A Tomato-Inspired Overview, Serokell, Aug. 5 2020. Accessed on: Jan. 20, 2022. [Online]. Available: https://serokell.io/blog/classification-algorithms

[50] H.S.M. Abbas, , Z.H. Qaisar, , X. Xu, and C. Sun, , " Performance Analysis of Algorithms on Different Types of Health Related Datasets", Journal of Physics: Conference Series, 1577,2020. Accessed on: Jan. 20, 2022. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1577/1/012051/pdf

[51] Helen Graham, "Detecting toxicity in online discussion," M.S. Thesis, Division of computing Science and Mathematics faculty of Natural science, University of Stirling 2018. Accessed on: Jan. 20, 2022.

[52] Carolin Dohmen, "Detecting hate Speech in Social Media – A Machine Learning Approach", B.S. Thesis, Faculty of Engineering and Computer Science, Hamburg University of Applied Sciences, 2019.

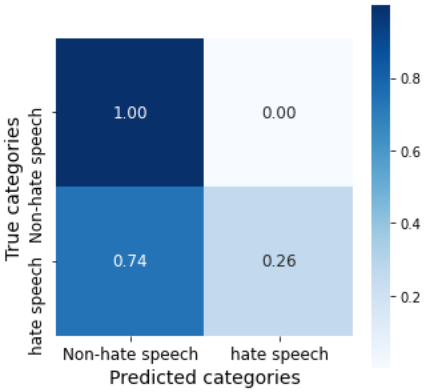
[53] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchil, and Maurizio Tesconi, "Hate me, hate me not: Hate speech detection on Facebook". In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, 2017. Accessed on: Jan. 1, 2022. [Online].Available:file:///Users/zara/Downloads/ITA-SEC17.pdf

[54] John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Ame, "Social Media Cyberbullying Detection using Machine Learning". International Journal of Advanced Computer Science and Applications, vol.10, No. 5, 2019. Accessed on: Jan. 20, 2022. [Online]. Available:https://thesai.org/Downloads/Volume10No5/Paper_87-Social_Media_Cyberbullying_Detection.pdf

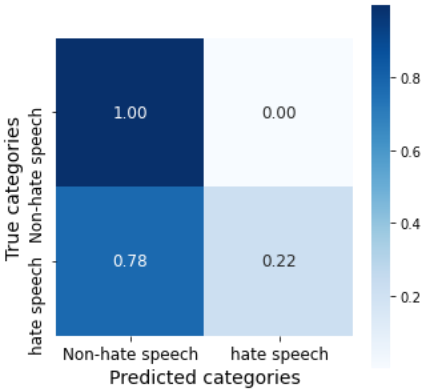
Result for imbalanced dataset

Classifier	Accuracy	Label	Precision	Recall	F1-Score
Linear SVC	0.97	Non-Hate Speech	0.97	0.99	0.98
		Hate Speech	0.80	0.61	0.69
Logistic Regression	0.95	Non-Hate Speech	0.95	1.00	0.97
		Hate Speech	0.88	0.22	0.35
SGD Classifier	0.95	Non-Hate Speech	0.95	1.00	0.98
		Hate Speech	0.96	0.26	0.41
Logistic Regression CV	0.97	Non-Hate Speech	0.97	0.99	0.98
		Hate Speech	0.86	0.57	0.69
Multinomial NB	0.94	Non-Hate Speech	0.94	1.00	0.97

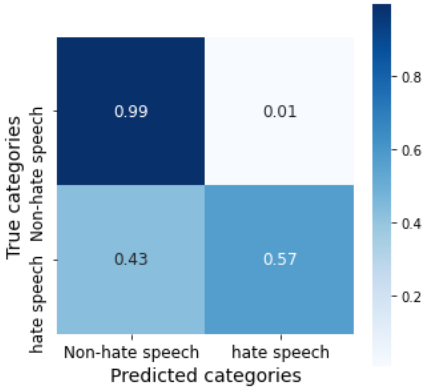
		Hate Speech	1.00	0.11	0.19
Extra Trees Classifier	0.97	Non-Hate Speech	0.97	1.00	0.98
		Hate Speech	0.93	0.49	0.64
Random Forest	0.96	Non-Hate Speech	0.96	1.00	0.98
		Hate Speech	0.94	0.46	0.62
XGBoost	0.95	Non-Hate Speech	0.96	0.99	0.98
		Hate Speech	0.78	0.40	0.53
LightGBM	0.95	Non-Hate Speech	0.96	0.99	0.97



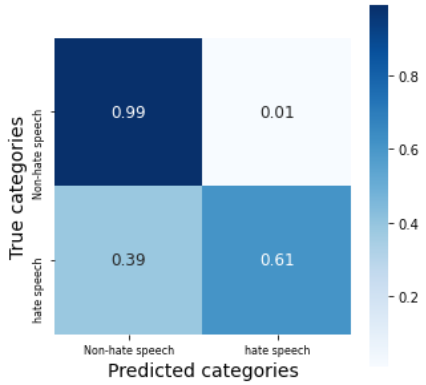
a) SGD Classifier



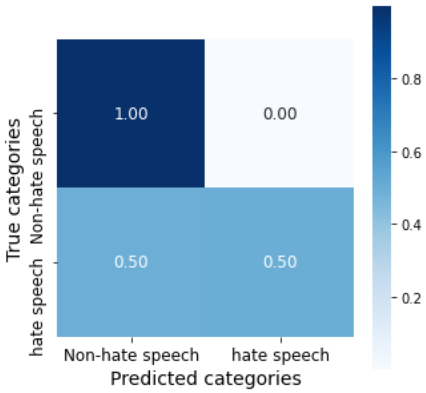
b) Logistics Regression



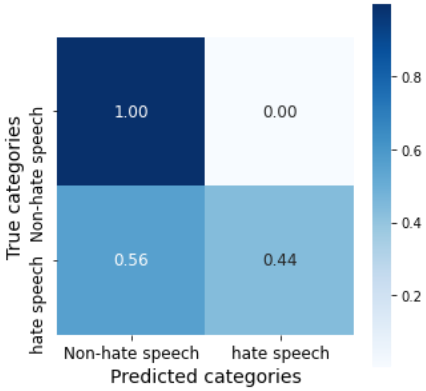
c) Logistics Regression CV



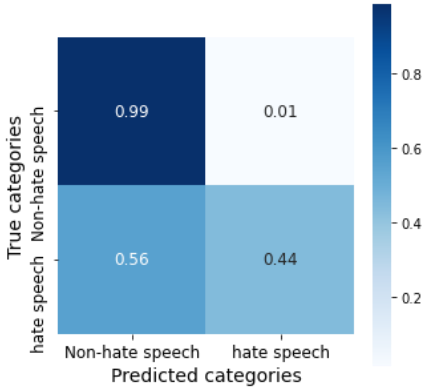
d) Linear SVC



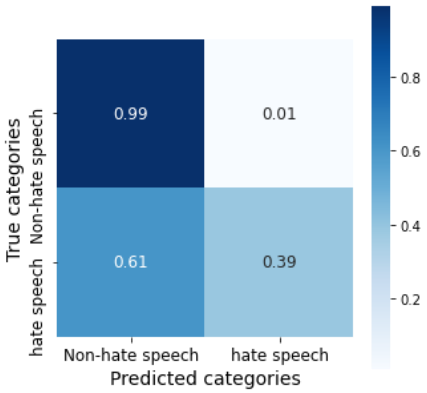
f) Extra Tree classifier



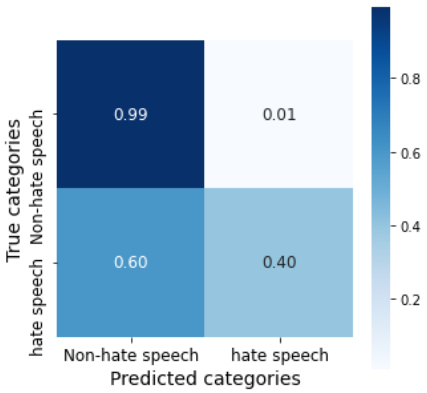
g) Random Forest



j) Lightgbm



h) AdaBoost



i) XgBoost