

**THE CO-OPERATIVE UNIVERSITY OF KENYA (CUK)**

DIRECTORATE OF COMPUTING AND e-LEARNING (DCeL)

**PROJECT PROPOSAL**

**HATE SPEECH DETECTION ON TWITTER**

BY

**RIEKO BRIAN ONGERI**

**BCSC01/0047/2018**

**SUPERVISOR: *[DAMARIS KANGOGO]***

Project *Proposal* submitted in partial fulfillment of the requirements for the award of the

*BACHELOR OF SCIENCE IN COMPUTER SCIENCE*

**©2022**

# DECLARATION AND APPROVAL

**DECLARATION**

I hereby declare that this Project *[Proposal/ Report]* is my own work and has, to the best of my knowledge, not been submitted to any other institution of higher learning for any award.

**Student Name: RIEKO BRIAN Registration Number: BCSC01/0047/2018**

**Signature: .................................................... Date: ..............................................................**

**APPROVAL**

This project *Proposal* has been submitted with my approval as the University supervisor.

**Supervisor Name: EDWARD**

**Signature: ..................................................... Date: ...................................................**

# ABSTRACT

Twitter has grown to several hundreds of millions of users in recent years, and it could provide a rich data source for detecting and classifying hate speech instigators and hate targets on the platform. Hate speech is well-known to be disseminated through microblogging platforms. As a result, hateful wording refers to statements that unlawfully disparage any group or individual based on traits such as color, race, gender, ethnicity, sexual orientation, religion, or nationality. Such content has the potential to frighten, intimidate, or silence platform users, and some of it has the potential to motivate users to commit crimes. The continuing rise of social internet platforms, especially Twitter, has forced the need for more immediate analysis of hatreds and other related antagonistic responses to various trigger events. Twitter users usually air their views about various topics of their interest. It's difficult to distinguish hateful text messages from the vast amount of content published by twitter users. Traditional natural language processing algorithms encounter significant hurdles when it comes to extracting high-quality features from noisy, highly dimensional, codeswitched, and huge unstructured data.

The dataset used in proposed project is outsourced. The dataset is publically available. During the 2017 Kenyan general election, approximately 400k messages were crawled from twitter, using a combination of problematic hashtags, ethnic epithets, hate patterns, and messages from pro-hate user accounts. A team of 27 human annotators painstakingly assessed a random sample of 50k texts into three categories: hate speech, offensive, or neither. A grid search across all potential feature combinations was undertaken to examine and select the best model. The psychosocial feature set (PDC) was effective at identifying hate speech with 82.5 percent accuracy.

Therefore, this project aims to automate content moderation to identify hate speech using machine learning techniques in Kenya.

TABLE OF CONTENT

[DECLARATION AND APPROVAL i](#_Toc102472647)

[ABSTRACT iii](#_Toc102472648)

[LIST OF ABBREVIATIONS vi](#_Toc102472649)

[CHAPTER ONE: 1](#_Toc102472650)

[INTRODUCTION 1](#_Toc102472651)

[1.1 BACKGROUND TO THE STUDY 1](#_Toc102472652)

[1.2 STATEMENT OF THE PROBLEM 4](#_Toc102472653)

[1.3 OBJECTIVES 5](#_Toc102472654)

[1.3.1 General Objective 5](#_Toc102472655)

[1.3.2 Specific Objective 5](#_Toc102472656)

[1.4 SIGNIFICANCE OF THE STUDY 5](#_Toc102472657)

[1.5 SCOPE OF THE STUDY 6](#_Toc102472658)

[1.6 ASSUMPTIONS 6](#_Toc102472659)

[1.7 LIMITATIONS AND DELIMITATIONS 6](#_Toc102472660)

[1.8 Definition Of Terms 7](#_Toc102472661)

[CHAPTER TWO: LITERATURE REVIEW 8](#_Toc102472662)

[2.1 INTRODUCTION: 8](#_Toc102472663)

[2.2. HATE SPEECH DEFINATIONS; 8](#_Toc102472664)

[2.2 RELATED SYSTEMS 10](#_Toc102472665)

[2.3 LIMITATIONS (WEAKNESSES OF THESE SYSTEMS) 11](#_Toc102472666)

[2.4 PROPOSED SOLUTION WILL HANDLE THESE WEAKENESSES 11](#_Toc102472667)

[CHAPTER THREE: METHEDOLODY 12](#_Toc102472668)

[3.1 INTRODUCTION 12](#_Toc102472669)

[3.2 PROJECT DESIGN 12](#_Toc102472670)

[3.2.1 Data Collection 12](#_Toc102472671)

[3.3 DESIGN PROCEDURES 13](#_Toc102472672)

[3.4 SYSTEM REQUIREMENT 15](#_Toc102472673)

[3.5 DATA PREPARATION 15](#_Toc102472674)

[3.5.1 Data Cleaning 16](#_Toc102472675)

[3.5.2 Tokenization 16](#_Toc102472676)

[3.5.3 Removing Stop Words 17](#_Toc102472677)

[3.5.4 Filtering Out Punctuations 17](#_Toc102472678)

[3.5.5 Stemming 17](#_Toc102472679)

[3.5.6 Case Normalization 17](#_Toc102472680)

[3.5.7 Splitting Into Train And Testing Datasets. 17](#_Toc102472681)

[3.5.8 Exploration Data Analysis 18](#_Toc102472682)

[Deployment 18](#_Toc102472683)

[Budget 19](#_Toc102472684)

[Implementation Schedule 20](#_Toc102472685)

[REFERENCES 21](#_Toc102472686)

# LIST OF ABBREVIATIONS

**KNCH**R: Kenya National Commission on Human Rights

**NCIC**: National Cohesion and Integration Commission.

**IE:** Information Extraction.

**SMS:** Short Message Service.

**NLTK**: Natural Language Toolkit

# CHAPTER ONE:

## INTRODUCTION

This chapter provides an overview to social media, with a focus on Twitter. This section also covers the background study, clarifies the problem statement, and defines the study's major purpose and specific objectives. The chapter ends with an overview of the study's significance, justification, scope, and assumptions.

# 1.1 **BACKGROUND TO THE STUDY**

On January 6, 2021, the former US president was permanently banned from Twitter. On charges of inciting violence, the former president was permanently banned from social media sites such as Twitter and Facebook, as well as YouTube and Instagram. His supporters saw his tweets as hinting that the election was unfair and that they should protest (Twitter, 2021).

Social media platforms are firmly entrenched as a venue for Kenyans to process campaign, news and engage in various types of social activism. But not all Kenyans use these platforms in similar ways. A new study by the Mozilla Foundation looked at how Kenyans use Twitter to spread misinformation, from how frequently they tweet to the accounts they follow or mention in their own posts. (Odonga Madung).

The last published figure regarding monthly active Twitter users amounted to 330 million (statista, 2021) before the company discontinued reporting on the metric. This metric shows Twitter has such a robust audience. The number of tweets sent per day as of 2021 was 500 million tweets (internetlivestats, 2022)

The development and use of social networking sites (SNSs) has changed the way people share information and communicate with peers. People can express their ideas on a variety of subjects that influence their daily lives through postings (Facebook), tweets (Twitter), emoticons, and other means. Twitter is accessible to billions of people around the world. The platforms are being used by the media, groups, politicians, activists, corporations, and even governments to interact with their populations.

In recent years, the rise of hate speech on social media platforms has proven to be a challenging and intractable issue for certain African governments, who have resolved to use force to curb it, especially during electioneering periods. While the Internet and social media networks have been praised for giving people a new way to publicly express their opinions, thoughts, and feelings they have also increased the development of online hate speech content, making manual monitoring difficult.

Governments, non-governmental organizations, and civil society organizations have upped their pressure on social media corporations like Facebook and Twitter to improve their hate speech content moderation systems. The two social media platforms currently rely on users to report hostile content by clicking the report button next to the offending item.

Given the speed with which users report and flag information on social media, in comparison to the over 2000 human languages used on these platforms, and the small number of content moderators, reviewing and flagging every instance of reported hate speech is nearly impossible.

Nonetheless, these platforms have evolved into vehicles for the rapid and low-cost propagation of hate speech, such as racism, racial slurs, religious attacks, insults, and sexist sentiments. According to a research by a cloud-based web filtering and scanning service, 80% of blogs include unsuitable information. Furthermore, compared to other media such as spoken speech, written text has been shown to be more persistent in its shape (Waldron, 2014). Written content can quickly travel to a wider audience, resulting in offline social disorder, harm, and undiscovered challenges outside of public internet settings.

People are expected to use the platform to express their ideas as we approach the election period of 2022. Different people hold various political viewpoints. As a result of these viewpoints, disagreements arise among various organizations. The outcome of the debates cannot be either positive or negative. Negative outcomes can be dangerous, and they must be contained.

When it comes to the role of social media in politics, Kenya is a particularly intriguing case study. A variety of constraints afflict Kenya's major media. These include a political corps of journalists who are firmly tied to ethnic sympathies, reliance on government advertising, and official intimidation. The recent elections reveal that it is losing its position as the country's preeminent agenda setter. Social media is progressively taking over that function.

Hate speech on social media is especially common in Kenya during national election campaigns, especially during presidential elections. There have been an increasing number of campaign-related incidents across the country that have provoked online public reactions bordering on hate speech during this time period. Politicians' invocations of negative ethnic feelings (Ajulu, 2002), which frequently provoke intense public reactions and counter-reactions from Kenyans on social media platforms, are among the most obnoxious of these.

Social media gives users an easy-to-use way to communicate and network, thus generating huge data useful in various fields. Online social networking sites are becoming more popular each day. Among all these sites, Twitter is the fastest growing site than any other social networking site. Kenya is a multicultural country with over forty-two ethnic tribes, each with its unique way of communicating. Almost all ethnic communities in Kenya have some stereotypes about them; these stereotypes may be positive or negative (National Cohesion and Integration Commission, 2013). Over the last decades, people are getting more engaged with widespread social networks. Microblogging applications opened up the chance for people to express and share their thoughts extensively and in a real-time manner. Such expressions afford researchers the ability to investigate the online social emotions in different events. People now can speak freely; this allowed them to exchange all sorts of thoughts, emotions, and knowledge. However, cyberspace is not always safe; it can be a reason for disseminating aggressive and harmful content. Hate speech is a common online form for expressing prejudice and aggression.

There is no international legal definition of hate speech, and the characterization of what is ‘hateful’ is controversial and disputed. In the context of this document, the term hate speech is understood as any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor. This is often rooted in, and generates intolerance and hatred and, in certain contexts, can be demeaning and divisive. In Kenya, hate speech has been defined as any form of speech that degrades others, promotes hatred, and encourages violence against a group based on criteria, including religion, race, color, or ethnicity.

Hate speech is the subject of an expanding amount of study, which include offensive language detection *(SocialCom 2012)* cyberbullying, radicalization, and terrorism.

Standard approaches and applications in natural language processing, computational linguistics, and machine learning are all being challenged by user-generated content on social media. It's noisy, irregular, full of duplicate and missing values, huge, diversified in data types, generated in real time, and subject to all of the other problems that big data brings. Codeswitching provides a hurdle when parsing sentences and performing contextual analysis on words and phrases using traditional monolingual methodologies.

This project is based on a case study of hate speech on Kenyan social media during the campaign period leading up to the general elections in August and the rerun elections in November 2017. Because of the representativeness, scale, and accessibility of public access to tweets in the dataset to be used in this project, Twitter was chosen as the social media network for this study.

To build automate content moderation to identify hate speech using machine learning techniques using a dataset tweets from Kenya's 2017 elections.

## 1.2 STATEMENT OF THE PROBLEM

The number of registered users is steadily growing, as is the volume of online user-generated content. Manual flagging to remove hostile information from online media is difficult. As a result, accurate, automated systems for detecting hate speech in internet media are required. To improve the effectiveness of machine classification of big data analysis, it is vital to study a methodology that more correctly captures essential traits inherent in complex forms of hate speech. This type of data, particularly which derived from user-generated content on social media, must be treated methodically in order to transform it from a highly dimensional and low-quality state to one that is low-dimensional and high-quality enough to train a machine classifier. As a result, it is hoped that the machine classifier's deployment will improve data-driven decision-making by key national security stakeholders by serving as early-warning systems that automatically monitor hate speech on social media, particularly during perennial trigger events like presidential elections, referendums, and occurrences like terrorist attacks and gender-based violence.

Therefore, this project aims to automate content moderation to identify hate speech using machine learning techniques using the available public dataset.

## 1.3 OBJECTIVES

This project’s main objective was to perform sentiment analysis in detecting, analyzing, and classifying hate tweets in social media, particularly Twitter, by providing a novel framework for sentiment analysis.

### 1.3.1 General Objective

This project’s main objective is to perform in depth sentiment analysis on social media posts, specifically Twitter, to help identify hate speech.

### 1.3.2 Specific Objective

1. To flag out various stereotypical /hate words used to propagate and disseminate hate speech on Twitter.

3. To develop web based of the hate tweets classification model.

4. To test and deploy the model.

## 1.4 SIGNIFICANCE OF THE STUDY

Hate speech should not be looked down upon. Building good relationship with people from different races, tribes, and countries is something that everyone should be capable to do regardless of their origin, tribe, and opinions in the internet regardless of the platform they are using.

If the platforms are not well used they can lead to chaos not in the platforms but also these conflicts can be brought to real life. People have committed suicide through being body shamed on these platforms .The internet will be there forever and the number of people using it are increasing spontaneously .People should be educated about hate speech and how to respond to it in case its used against them.

If the topic on hate speech is taught properly everyone will be at peace and the internet will be used in a proper way and not in a way of promoting violence.

It is highly recommended to Twitter and other social media websites to implement strict policies and mechanisms to control hate speeches to control them to create a peaceful social media environment.

The model build will be able to classify majority of the tweets into categories that can be analyzed into various classes that worked upon.

This project will be beneficial to general public cooperate organizations and even the government

## 1.5 SCOPE OF THE STUDY

The model is guaranteed to perform very well in a specific context or domain, such as the entangled political and ethnic hate speech prevalent in Kenya. The content of hate speech that will be analyzed in this project is mostly text; it excludes multimedia elements such as photos, graphics, movies, meme and audio.

## 1.6 ASSUMPTIONS

Twitter understands many languages in particular English and Swahili.

Many of the Kenyan users on the platform use both English and Swahili.

There are many Tweets, Retweets and comments from various users.

Users on the platform use both English, Swahili in their tweets.

## 1.7 LIMITATIONS AND DELIMITATIONS

There are always many tweets per minute. It becomes very hard to get through these many tweets .Analyzing these tweets is a very hard work among different users , These different come from different ethnicity backgrounds with different languages.

There are different people tweeting on different subjects/topics per minute. These projects is mainly focused on hate different hate speech detection through a web app. It will be focused on English and Swahili language among many other languages. Hate speech detection will be done through analyzing of tweets from different people.

The use of abbreviation can be difficulty to encounter among different languages in tweeter but the use of Google to identify these languages is beneficiary to many different people.

# 1.8 Definition Of Terms

These are to be identified as the project proceed ……

# CHAPTER TWO: LITERATURE REVIEW

## 2.1 INTRODUCTION:

This chapter provides a comprehensive summary of existing research on detecting hate speech in text documents, with a focus on short text messages commonly found on social media platforms.

The chapter begins by defining hate speech before going over some of the most common examples. In this chapter too, the various definitions of hate speech are addressed, and the content analysis technique is utilized to find some common terminology that may help define the study's primary themes. It then goes into some of the current efforts to address automatic hate speech detection and closes by giving the recommended solution.

## 2.2. HATE SPEECH DEFINATIONS;

Sentiment analysis of tweets has become an important research subject due to the development in use and popularity of informal language, as well as the use of social media platforms, particularly Twitter (S. Vosoughi, 2015). Twitter is a well-known platform for sharing information and opinions. This platform is primarily used prior to, during, and following live events like elections and religious festivals among others.

However, most users have used the platform to disseminate hate messages among the users and the general public

We're interested in detecting, recognizing, analyzing, and preventing the propagation of hate speech emotions on social media platforms in Kenya, particularly Twitter. The wording used in the majority of these hate speech texts or letters is stereotyped. Several studies have looked into the detection of flaming and virulent messages on social media, as well as the spread of nasty messages on dark web forums. Hate speech employs derogatory and threatening rhetoric directed at specific groups of people based on religion, race, nationality, color, or gender. The source of the hate message is usually a member of a supposedly competing group, a fellow Twitter user, or someone from another ethnic group in Kenya. The classification of opinions and emotions in a text has been widely utilized using subjective language analysis (Wiebe, 10 2005).

Hate speech is defined in a variety of ways in the literature, and there is no universal definition. Other organizations, such as international and domestic legislation, have attempted to define hate speech and even identified specific targets based on what are legally known as protected traits, such as race, ethnic origin, religion, or gender. This section is not intended to define hate speech or provide a global definition; rather, it is intended to develop a workable definition in order to establish a shared understanding in this study. Hate speech definitions can be found in dictionaries, government and nongovernment organizations' existing legislation and policies, as well as published studies.

The following are some of the definitions of hate speech from various sources;

“Abusive or threatening speech or writing that expresses prejudice against a particular group, especially on the basis of race, religion, or sexual orientation.”[The Oxford dictionary]

“A form of other-directed speech which rejects the core human rights principles of human dignity and equality and seeks to degrade the standing of individuals and groups in the estimation of society.”[UN’s International Committee on the Elimination of Racial Discrimination]

“All forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility towards minorities, migrants and people of immigrant origin.”[The European Court of Human Rights]

“Words of incitement and hatred against individuals based on certain group characteristics they share. It includes speech that advocates or encourages violent acts against a specific group, and creates a climate of hate or prejudice, which may, in turn, foster the commission of hate crimes.” Some of the definitions of hate speech from social media companies include: [Kenya National Cohesion and Integration Commission (NCIC) Act, 2008] “Content that promotes violence against or has the primary purpose of inciting hatred against individuals or groups based on certain attributes, such as race or ethnic origin, religion, disability, gender, age, veteran status, sexual orientation/gender identity.”

“Content that attacks people based on their actual or perceived race, ethnicity, national origin, religion, sex, gender or gender identity, sexual orientation, disability, or disease is not allowed.

We do, however, allow clear attempts at humor or satire that might otherwise be considered a possible threat or attack. This includes content that many people may find to be in bad taste (example: jokes, stand-up comedy, popular song lyrics, etc.).” [Facebook hate speech policy]

“You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.”[Twitter hateful conduct policy]

From the definitions of various organizations and legislations we can conclude that;

Hate speech has three distinct aspects, according to various definitions: To begin, it is an expression that threatens, incites, discriminates, degrades, attacks, intimidates, insults, offends, or stigmatizes, whether nonverbal through body language or verbal through writing, pictures, or graphics.

Second, hate speech can be divided into two categories: those directed directly at an individual and those directed at a group of individuals based on their membership in a protected social characteristic such as gender, ethnic origin, religion, or race (ElSherief et al., 2018). Third, the word is meant to elicit hatred, aggressiveness, prejudice, intolerance, and negative feelings and attitudes against the topic.

## 2.2 RELATED SYSTEMS

According to a survey of the literature, a rising number of studies are developing and manually annotating corpora for foul language, sentiment analysis, and hate speech detection. Most of these research use corpora in English (Orehek & Human, 2016) as well as European languages such as Portuguese (Fortuna et al., 2019), German (Ross et al., 2016), Spanish (Fersini et al., 2018), and Italian (YouTube, n.d.), New datasets in (Kumar et al., 2018), Arabic (Burnap & Williams, 2015), and Amharic (Kumilachew Tegegnie et al., 2017) have been created in a few past projects. However, no study has created and annotated corpora for codeswitched data, which is common in multilingual social media networks. Furthermore, some studies have used binary categories to annotate their corpora, with a few studies (Burnap & Williams, 2015) adopting multiple categories. Using a methodology similar to (Waseem & Hovy, 2016) the study used a three-category manual annotation of the corpus.

\*\*\*\*Elaboration\*\*

## 2.3 LIMITATIONS (WEAKNESSES OF THESE SYSTEMS)

Due to the diversity of languages spoken in various places across the globe. These research employed different languages and datasets. Each research concentrated on a single language, which was either German, Arabic, Hindu, or Portuguese.

Following a review of the literature on hate speech classification, it became obvious that previous studies used a variety of criteria. However, they are usually jumbled, making comprehension much more challenging.

In contrast to these previous studies, the annotation framework we developed may be used to detect hate speech and takes into account codeswitched messages in both English and Swahili, Kenya's official and national languages. Although a previous study [(Gupta et al., 2018)] proposed a framework for detecting abusive tweets, it was limited to Google's offensive word list and lacked the theoretical foundation needed to advance research.

## 2.4 PROPOSED SOLUTION WILL HANDLE THESE WEAKENESSES

The goal of the proposed project is to present a simple and effective method for qualitatively identifying and analyzing hate speech in short text documents using human-readable high-level psychosocial features, such as PDC-based features, which can then be mapped to machine-readable lower-level features like Term Frequency-Inverse Document Frequency (TF-IDF) and one-hot encoding vectors for training a machine classifier.

Lexical and other NLP-based features have been used in previous hate speech identification studies. These skills were not able to successfully capture hate speech in codeswitched communications on their own. As a result, classifier models that explicitly use these traditional qualities will underperform, resulting in a significant amount of false negatives, which is the opposite of how hate is represented in social media posts.

The proposed project will cover the codeswitching using the already provided dataset. By the end of the end we could have an automated Twitter hate speech detection system with classification modeling and an interactive web version that will be in a graphical representation.

# CHAPTER THREE: METHEDOLODY

## 3.1 INTRODUCTION

In this chapter, the methodology is outlined. Data collection process, Text preprocessing,

Sentiment detection, sentiment analysis, and model development are done to achieve the proposed project’s objectives of identifying and classifying hate speech.

Comprehending what hate speech was from different sources was the right move that could help the project get started.

## 3.2 PROJECT DESIGN

### 3.2.1 Data Collection

Approximately 400k unprocessed messages were gathered and saved in a comma-delimited file (CSV) format. These primarily comprised of Twitter text messages, often known as tweets, from Kenya's general elections in August 2017, as well as a follow-up election held 60 days later in October 2017. To build a large raw corpus, additional tweets were crawled from January to December 2017 as well as the March 2013 general elections.

The dataset for this project was sourced from a study called A Model for Classifying Hate Speech Text from Social Media Leveraging on Psycho-social Features and Machine Learning and a team UMATI project in 2017 after the elections and the election re-run. The data is available on the Keggle site.

The reason for using this dataset is that the data was sourced at the right time when majority of the people using had really different opinions.

The dataset is a .csv file with 50,176 text posts from Twitter where 6% of the tweets were labeled as hate speech

The labels on this dataset were voted on by crowdsource and determined by majority-rules.

## 3.3 DESIGN PROCEDURES

Design principles can be used to develop a road map that positions and guides us toward our goals. Design techniques assist us keep on track with the project while achieving our final goal.

The design procedure guides us on how to go about problem description, data collection, feature identification, model creation, and model evaluation in this project.

The first problem will be data gathering and text classification, which will produce mostly qualitative results. It will be crucial to automatically detect the underlying function in text corpora in order to develop a model (the dataset).

The suggested study will adhere to the defined objectives by doing research in the connected topic of hate speech on Twitter.

The first step will be to determine whether hate speech has a common definition.

The project can move forward after a working definition for hate speech is established.

As a result, a mixed-method approach will be used to identify prior related research and projects that are relevant to the proposed project, which will be guided by the suggested objectives.

Content analysis was utilized to find crucial terms from numerous definitions of hate speech and current hate theories, which will subsequently influence the construction of the study's conceptual framework. Using the architecture from these past projects, the annotation technique for leading the project to appropriately label the messages with either of the three preset types of hate speech, offensive, or neither was built.

Statistical machine learning methods were used to extract quantitative inferences from the annotated data and construct a classifier that accurately predicts new messages as belonging to the three predefined classes.

Statistical machine learning models were utilized in previous projects to generate quantitative conclusions from annotated data and construct a classifier that intelligibly predicts new messages as belonging to the three predefined classes, and this will also lead the proposed project.

Both methodologies were utilized to create a full picture of the hate speech phenomenon and to impact the experimental design of the study in terms of identifying salient aspects and training a computer system to detect hate speech in text messages.

This sparked the notion of producing a web-based interactive version that could be used.

Apart from that, an experimental study was conducted in which data was collected for a year, including Kenya's 2017 general elections. Hate speech has been known to spike on social media during trigger events like presidential election campaigns, which can last months before and after the official election results are published. In Kenya, during the presidential campaign season in August 2017, which included a rerun election later that year, short text messages potentially containing hate speech from social media were gathered. Given the empirical objective of the data, which entails identifying hate speech from social media content, a thorough understanding of the hate speech phenomenon, as guided by many hate theories from sociology and psychology, will be essential. The model will be trained on this.

From the dataset used in this project, human annotators first used a deductive technique to establish the classification category of each communication and label it, as instructed by the annotation framework (Ombui et al., 2019) based on the three aspects of the triangular theory of hate (Sternberg, 2003).Despite this, the study's assumptions were developed using an inductive technique based on the findings of a preliminary investigation.

These will be subsequently used as the foundation for the project. Furthermore, the machine classifier used in this project was designed to learn from examples of classified text messages before using inductive inference to categorize fresh, unseen text messages. This inductive learning principle will be essential in the development automated machine learning that uses prior knowledge from specific examples to progress to greater generalization while keeping high performance (Shalev-Shwartz & Ben-David, 2014)

In addition, a full literature review and Internet search will be conducted for hate speech-related legislation, guidelines, and user policies from key social media networks and periodicals. This will be done in a descriptive manner, with an emphasis on the case of Kenya, where user-generated content bordering on hate speech on social media was triggered by the 2017 presidential elections.

These hate speech definitions will be subjected to content analysis to extract keywords.

To produce hate speech from the dataset, researchers used messages, a range of search strategies, including key hate phrases, pro-hate user profiles, and problematic hashtags.

After that, experiments for the text classification task was conducted, in which statistical machine learning models will be used to build classifiers that make inferences from sample data in classifying codeswitched messages from social media into three predefined classes: hate speech, offensive, or neither.

## 3.4 SYSTEM REQUIREMENT

For the proposed project, the memory required will be 8GB RAM and a storage of 2GB.

## 3.5 DATA PREPARATION

Convenient sampling was used to acquire data from the Twitter social media network. Unless the user indicates otherwise in their settings, Twitter makes every post public and programmatically available, unlike other social media networks. Furthermore, no account is required to view these tweets, and anyone can anonymously write, like, dislike, and quickly spread the words to a huge audience. The platform is vulnerable to the dissemination of hate speech because to these attributes and characteristics.

Ombui Edward Osoro's project on A Model for Classifying Hate Speech Text from Social Media Leveraging on Psycho-social Features and Machine Learning provided the dataset for this project. Codeswitched Dataset is the name of the dataset. According to the owner, the dataset is open to anyone interested in machine learning. The dataset is a .zip file with over forty thousand text posts from Twitter.

Unlike traditional research, big-data initiatives employ various sampling strategies to computationally capture all available online content (Kim et al., 2018), such as employing a web crawler or Twitter API to collect a large number of messages from social media based on specified key terms. Such methods are frequently free of the constraints that come with standard sampling methodologies (Kim et al., 2018b), such as the inefficiency and impracticality of collecting a large volume of hate speech data from many Kenyan social media users for machine learning reasons. Our work used simple random sampling to establish a study sample for annotation from the large volume of data collected. Previous research (Davidson et al., 2017) employed this sampling strategy to obtain study samples from social media.

Data from the Twitter social media network was collected via convenience sampling. Unlike other social media platforms, Twitter makes every post public and programmatically accessible unless the user specifies differently in their settings. Furthermore, accessing these tweets does not require an account, and anyone can anonymously publish, like, dislike, and rapidly transmit the messages to a large audience. Because of these traits and characteristics, the platform is vulnerable to the spread of hate speech.

This is a critical step of the research, the heart of the study, which verifies the findings (Seabrook et al., 2018).

The performance of the trained model is directly proportional to the quantity and quality of data collected. The desired data for acquisition included tweets from Kenya's presidential campaign in

August 2017, which includes a second election in October 2017. Previously, the Twitter API was used to create an app that gathered tweets during election days. A crawler based on Python programming was also used to supplement Twitter API's two-week data collection window in order to obtain a massive amount of archival tweets, which included tweets from the March 2013 general elections and the four months leading up to March, as well as two months after the results were announced. This time period and the events surrounding it have been the most prominent trigger events in the past, resulting in large spikes in online hate speech.

As a key data collecting strategy, the bootstrapping technique was adopted. To explore social media networks, seed words consisting of hate-related keywords (kw), phrase patterns (pp) with a connotation of hatred (Warner & Hirschberg, 2012), offensive hashtags (#), and pro-hate user account names (un) were used.

### 3.5.1 Data Cleaning

Data cleaning is an important part of the machine learning process because it removes distracting signals that would otherwise degrade the training and, as a result, the overall performance of a classifier model. To clean the data in this study, natural language processing techniques such as tokenization, stemming, and lemmatization were applied. HTML characters, non-ASCII and corrupted characters, empty rows, duplication, emoticons, stop words, and punctuations were all removed using regular expressions (regex). There were a lot of quotation marks, commas, apostrophes, exclamation marks, and other punctuation symbols. To normalize the data, all of the terms were also lowercased.

### 3.5.2 Tokenization

Tokenization is a way of separating raw text communications into phrases and then into a list of individual words, also known as tokens, using whitespaces, newlines, tabs, and other delimiters. This was significant because computers can digest token units significantly more easily and swiftly than the original corpus documents in terms of machine learning. Tokenization was also used to handle standard abbreviations and hyphenated terms, the bulk of which were in English.

Two tokens were created from the hyphenated words. The NLTK (Natural Language Toolkit) word tokenizer function was altered by adding to the list numerous frequent codeswitched abbreviations, in addition to the default, which is in the Standard English list.

### 3.5.3 Removing Stop Words

The NLTK corpus stop words library was used to delete all English stop words, with the exception of third-person pronouns, which, according to the study's conceptual model, would be highly indicative of hate speech. The English stop word list now includes Swahili corresponding stop words like "ni," which is the Swahili equivalent of the English stop word "is." Stop words are usually filtered away since they do not provide meaning to a phrase's deeper meaning.

### 3.5.4 Filtering Out Punctuations

Regular expression methods in Python's natural language tool kit (NLTK) module were used to eliminate duplicate messages, single letters, non-alphanumeric data, HTML components, dates, emoji, and URLs, and punctuations and other non-ASCII characters were replaced with a space.

The Python isalpha () function was used to cycle over all tokens and filter out the single punctuation. Advertisements based on popular hashtags were also eliminated from spam mails.

### 3.5.5 Stemming

The process of stemming reduces a word's inflectional forms to its base form, resulting in a smaller vocabulary. For example, hate, hated, hating, and hates are all reduced to the stem word hate. The data for this inquiry was stemmed using the Porter Stemmer method in the NLTK.

### 3.5.6 Case Normalization

Another preparation operation was case normalization, which involved changing all of the text messages to a single case, which was lowercase. This was performed by applying the lower () function to each word in Python.

### 3.5.7 Splitting Into Train and Testing Datasets.

Data preparation requires splitting the dataset into training and testing data samples.

The training dataset is typically provided a bigger fraction than the testing dataset in order for the classifier to comprehend the underlying data distribution. The magnitude of the data available typically determines the proportions. For example, if the original data set is massive, just a small portion of it will be required for examination. Two typical training ratios are 80:20 and 70:30. The dataset was utilized for training 80% of the time, while the remaining 20% was used to test and assess the classifier model. To split the data, the train test split package in Scikit learn will be utilized.

Overfitting happens when a model gets deceivingly good by "regurgitating answers" from memory. Separating the two data sets is a machine learning strategy for avoiding overfitting. This is because the data samples that were viewed during training are sent to the model again during testing. As a result, the model does well during training but fails miserably when faced with new or unknown data samples.

### 3.5.8 Exploration Data Analysis

The data exploration process involves examining and comprehending the dataset in general by looking for patterns in data types, class distribution, word frequencies, missing values, and other factors using quantitative and visual methodologies. Following an in-depth analysis that determines the interpretation and accuracy of the conclusions regarding the machine classification of hate speech text, these are frequently visually plotted using word clouds for text data, pie charts or bar graphs, or any other statistical chart to provide some high-level insight into data patterns and other characteristics that will give confidence to the kind of expected results.

The Scikit-Matplot program was used to make the chart visualizations. Displaying the class distribution, describing the counts, mean standard, max, and min of the data, and explaining the counts, mean standard, max, and min of the data are all made easier using the quantitative method. Furthermore, the researcher will be able to ask the appropriate questions without introducing bias into the study through incorrect data assumptions. The two primary columns of relevance in the dataset used in this investigation were the class and tweet message columns.

## Deployment

The project will be employed on a web based interactive interface. The web based version will be on heroku an online based site for hosting Machine learning and data science projects

CHAPTER FOUR

SYSTEM ANALYSIS

Introduction

This chapter show the current state of the system .It’s a detailed analysis of the current system using flow charts.

This chapter focuses on data gathering and facts interpretation , diagonosis problem using the information to reccomemnd to recommend and improve the system . System analysis is a problem solving activity that requires communication between users and the project developer.

Here, the system is viewed as a whole, the inputs are identified and the model is subjected to close study to identify the problem areas.The solution are given in proposal.The model is tunned if it did a misclassification . The model will be retrained with new data to improve its accuracy

4.2: REQUIREMENT ANALYSIS

A complete understanding of the problem to be solved will require a good and accurate model to be developed .This will have the developer learn that there is new data generated every minute thus creating new room for improvement . The improvement is the terms of collecting new data and retraining the model again.This will improve the model’s ability to classify tweets accurately.

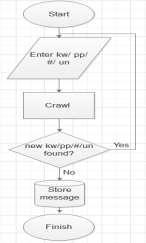
The model is work with changing trends,this will be challenging as there are new things trending on daily basis. The solution to this is to collect the data and retrain the model to improve its future prediction in the case the same trends are repeated.

4.2.1 Fuctional requirements

This project aims at making content annotation possible and eliminating the human aspect.The model to learn new features from previous patterns to improve its accuracy in the future predictions. This will take less time or more time as the data generated is more and its classification gets harder as there no build yet datasets to help in the classification. In the near future yhe model is expected to work well with big datasets to eliminate hate speech .

If large datasets are classified, it will be easy to retrain models that are more accurate over a period of time.

4.2.2 Non-functional requirements

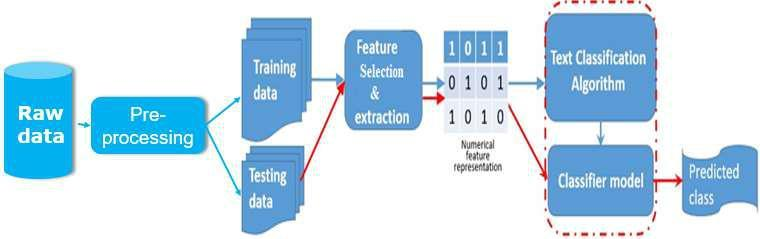


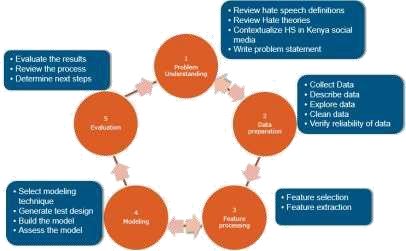
# Chapter 5: System Design Introduction

Detailed design of the proposed system using tools such as ERDs,

DFDs, UML, etc.

5.1 Architectural design

 *The Supervised Machine learning model*



**Modelling**

Modeling comprises three main steps: the selection of the model, the training of the model, and the tuning of the model’s parameters.

5.2 Database design

5.3 User interface design

# Chapter 6: Implementation and testing

Introduction

6.1 Development environment

6.2 System components

6.3 Test Plan (test data, test cases, test results)

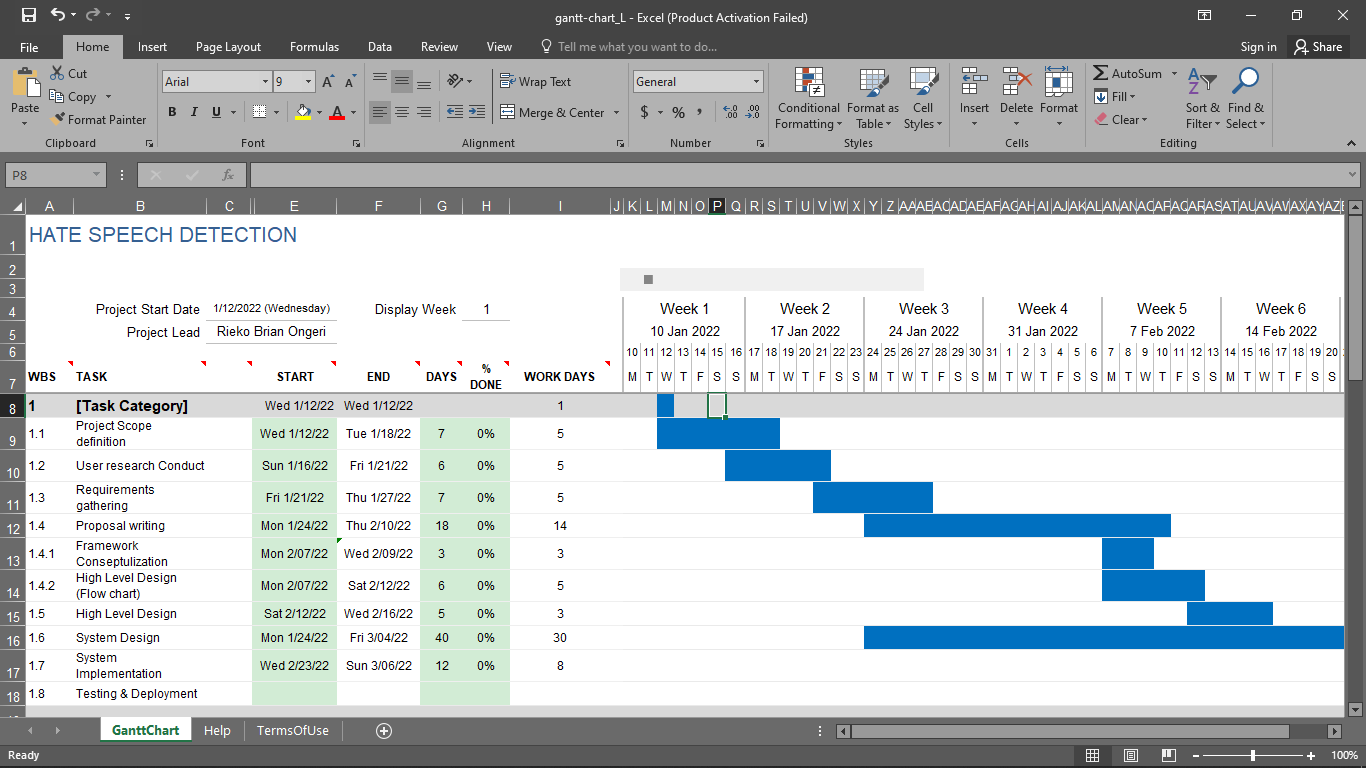
6.4 System Testing

# Chapter 7: Conclusions and Recommendations

## Budget

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S. No.** | **Items** | **Specifications** | **Quantity** | **@ Ksh** | **Amount, Ksh** |
| **1.** | Printing costs  (Proposal, Report) | - | 2 | 500 | **1000** |
| **3.** | Writing pad | A4, high quality | 1 | 100 | **500** |
| **4.** | Laptop |  | 1 | 50,000 | **50,000** |
| **5** | Internet subscription | 5mbps |  | 4000 | **4000** |
| **5.** | Equipment |  |  |  | **2,000** |
| **6.** | Subsistence |  | 15 days | 1,500 | **5,000** |
| **TOTAL** | | | | | **58,900** |

## Implementation Schedule



# REFERENCES

Ajulu, R. (2002). Politicised Ethnicity, Competitive Politics and Conflict in Kenya: A Historical Perspective. *African Studies*, *61*(2), 251–268. https://doi.org/10.1080/0002018022000032947

Atiah Alsolamy, A., Ahmed Siddiqui, M., & Hussain Khan, I. (2019). A Corpus Based Approach to Build Arabic Sentiment Lexicon. *International Journal of Information Engineering and Electronic Business*, *11*(6), 16–23. https://doi.org/10.5815/ijieeb.2019.06.03

Burnap, P., & Williams, M. L. (2015). Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet*, *7*(2), 223–242. https://doi.org/10.1002/poi3.85

Burnap, P., & Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(1). https://doi.org/10.1140/epjds/s13688-016-0072-6

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1). https://ojs.aaai.org/index.php/ICWSM/article/view/14955

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media. *ArXiv: 1804.04257 [Cs]*, *1*(56466446). https://arxiv.org/abs/1804.04257

Fersini, E., Rosso, P., & Anzovino, M. (2018). *Overview of the Task on Automatic Misogyny Identification at IberEval 2018*. Semantic Scholar. https://www.semanticscholar.org/paper/Overview-of-the-Task-on-Automatic-Misogyny-at-2018-Fersini-Rosso/b550270ea54715b5f587b8692c91d4b36432da13

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., & Nunes, S. (2019, August 1). *A Hierarchically-Labeled Portuguese Hate Speech Dataset*. ACLWeb; Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3510

Gupta, P., Kamra, A., Thakra, R., Aggarwal, M., Bhatti, S., & Jain, V. (2018). A Proposed Framework to Analyze Abusive Tweets on the Social Networks. *International Journal of Modern Education and Computer Science*, *10*(1), 46–56. https://doi.org/10.5815/ijmecs.2018.01.05

Kim, H., Jang, S. M., Kim, S.-H., & Wan, A. (2018a). Evaluating Sampling Methods for Content Analysis of Twitter Data. *Social Media + Society*, *4*(2), 205630511877283. https://doi.org/10.1177/2056305118772836

Kim, H., Jang, S. M., Kim, S.-H., & Wan, A. (2018b). Evaluating Sampling Methods for Content Analysis of Twitter Data. *Social Media + Society*, *4*(2), 205630511877283. https://doi.org/10.1177/2056305118772836

Kumar, R., Ojha, A. Kr., Malmasi, S., & Zampieri, M. (2018, August 1). *Benchmarking Aggression Identification in Social Media*. ACLWeb; Association for Computational Linguistics. https://www.aclweb.org/anthology/W18-4401/.

Kumilachew Tegegnie, A., Nega Tarekegn, A., & Anteneh Alemu, T. (2017). A Comparative Study of Flat and Hierarchical Classification for Amharic News Text Using SVM. *International Journal of Information Engineering and Electronic Business*, *9*(3), 36–42. https://doi.org/10.5815/ijieeb.2017.03.05

Ombui, E., Karani, M., & Muchemi, L. (2019, May 1). *Annotation Framework for Hate Speech Identification in Tweets: Case Study of Tweets during Kenyan Elections*. IEEE Xplore. https://doi.org/10.23919/ISTAFRICA.2019.8764868

Ombui, E., Muchemi, L., & Wagacha, P. (2021a). Building and Annotating a Codeswitched Hate Speech Corpora. *International Journal of Information Technology and Computer Science*, *13*(3), 33–52. https://doi.org/10.5815/ijitcs.2021.03.03

Ombui, E., Muchemi, L., & Wagacha, P. (2021b). International Journal of Information Technology and Computer Science (IJITCS). *International Journal of Information Technology and Computer Science (IJITCS)*, *13*(6), 29. https://www.mecs-press.org/ijitcs/ijitcs-v13-n6/v13n6-3.html

Orehek, E., & Human, L. J. (2016). Self-Expression on Social Media. *Personality and Social Psychology Bulletin*, *43*(1), 60–70. https://doi.org/10.1177/0146167216675332

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. *ArXiv: 1701.08118 [Cs]*, *DD* (4545). https://doi.org/10.17185/duepublico/42132

Seabrook, E. M., Kern, M. L., Fulcher, B. D., & Rickard, N. S. (2018). Predicting Depression from Language-Based Emotion Dynamics: Longitudinal Analysis of Facebook and Twitter Status Updates. *Journal of Medical Internet Research*, *20*(5), e168. https://doi.org/10.2196/jmir.9267

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf

Sternberg, R. J. (2003). A Duplex Theory of Hate: Development and Application to Terrorism, Massacres, and Genocide. *Review of General Psychology*, *7*(3), 299–328. https://doi.org/10.1037/1089-2680.7.3.299

Waldron, J. (2014). *The harm in hate speech*. Harvard University Press.

Warner, W., & Hirschberg, J. (2012). *Detecting Hate Speech on the World Wide Web* (pp. 19–26). Association for Computational Linguistics. https://aclanthology.org/W12-2103.pdf

Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *Proceedings of the NAACL Student Research Workshop*, *545*(44545). https://doi.org/10.18653/v1/n16-2013

YouTube, Y. (n.d.). *Hate speech policy - YouTube Help*. Support.google.com. https://support.google.com/youtube/answer/2801939?hl=en