# CHAPTER ONE

## INTRODUCTION

## 1.1 Background to the Study

In recent years, hate speech has grown to be a major social issue with profound consequences on individuals, groups, and society as a whole. The growing number of online communities and the emergence of social media have opened up new channels for the spread of hate speech, multiplying its effects and presenting new difficulties surrounding its regulations. The need to comprehend the dynamics and effects of hate speech has been acknowledged by researchers from a variety of professions, which has resulted in an expanding body of research on the subject.

According to a report by Statista in earlier 2023, Nigeria social media users represents about 0.66 percent of the total social media users in the world[[1]](#footnote-1). These users freely share their opinions on events that occur within and outside the country which include comments that are discriminatory, humiliating, derogatory, demeaning, and threating to individuals or groups based on some protected characteristics like religion, race, gender, disability and others. Therefore, there is an urgent need for regulation of what people share on the Internet, especially the social media platforms. In 2016, the European Commission signed an agreement with Facebook (Meta), Twitter, Youtube and Microsoft on countering hate speech which requires reviewing and removing hateful contents within 24 hours[[2]](#footnote-2).

Traditionally, a user reports a post as offensive or hate speech, these platforms manually review it and act accordingly. Some other platforms maintain a list of offensive words and expressions for detecting and removing offensive posts. However, these approaches of relying on a user report and the use of a lexicon of words and expressions in tasking and not efficient considering the amount of data generated daily by on these platforms (Saleem et al., 2017). Furthermore, users tend to incorporate images, videos and user abbreviations to evade detection (Schmidt & Wiegand, 2017). These have been identified as one of the major setbacks of the above approaches.

Despite the efforts by social media platforms and the Government in combating hate speech, more needs to be done to develop AI solutions that will perform real-time detection, especially for words that users intentionally change the spellings by omitting or adding a letter or code-mixing and multimodal posts, to avoid being detected (Gröndahl et al., 2018).

Natural Language Processing (NLP)

Natural language processing (NLP) is a branch of Artificial Intelligence (AI) that focuses on how computers interact with human languages. It is a combination of various fields like Computer Science, Linguistics, Statistics and Artificial Intelligence that uses ‘computational techniques’ to create algorithms and models for generation and interpretation of human languages (Thanaki, 2017). In recent years, NLP has gained significant attention and prominence, particularly in the domain of text classification. Text classification is a fundamental task in NLP which involves the categorization of text documents into predefined classes or labels. NLP has a variety of applications which include:

* Information Retrieval: The techniques of using query to search and obtain results from documents or websites (Mei & Radev, 2016).
* Machine Translation: Automatically translating text from one language to another using Natural Language techniques (Stahlberg, 2020).
* Text Summarization: The process of automatically creating a text summary (Aries et al., 2019).
* Questions and Answering: The process of automatically answering a posed question in natural language (Rogers et al., 2023).
* Text Classification: Classifying and labelling text based on its context. This includes tasks such as Sentiment Analysis, Spam detection, Hate speech detection, an many more (Alabbas et al., 2016).

Offensive Language

Offensive speech is any form of speech that hurts, annoys, or offends individuals (Sai & Sharma, 2020). It include the use of obscene words to dehumanize, insults and belittle someone (Mallikarjun et al., 2009). Today, social media has become a breeding ground for offensive contents which can be detrimental to the society. This necessitates the research community to develop various methods to detect and remove these contents within the shortest possible time. There are areas that are being researched under this task including hate speech (Ombui et al., 2019), abusive language (Nobata et al., 2016), cyberbullying (Haidar et al., 2018), and more. Hate speech is one of the most researched areas alongside offensive speech (Fortuna & Nunes, 2019; Husain & Uzuner, 2021; Pradhan et al., 2020).

Hate speech

Hate speech is a complex phenomenon and does not have an agreed definition (Alkiviadou, 2019). This complexity made various organizations, platforms and the researchers to come up with different definitions. Among the definitions by social media platforms are:

1. Meta (Facebook): “We define hate speech as a direct attack against people – rather than concepts or institutions – on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease[[3]](#footnote-3)”
2. YouTube: “content promoting violence or hatred against individuals or groups based on any of the following attributes: age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin, and veteran status[[4]](#footnote-4)”.
3. Twitter: “Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow inciting harm towards others on the basis of these categories[[5]](#footnote-5)”.
4. Instagram: defined hate speech as “attacks on people based on their protected characteristics, including race or religion[[6]](#footnote-6)”.

Researchers (Davidson et al., 2017; Frenda et al., 2019; Gitari et al., 2015; Mansourifar et al., 2021; Modi, 2018; Pereira-Kohatsu et al., 2019) have given different definitions of hate speech and we summarize it as any form of communication that disparaging an individual or a group of people based on some protected characteristics like religion, ethnicity, gender, sexual orientation, age and disability.

The definitions above show that the target for hate speech are individuals or groups based on certain characteristics like race, gender and nationality. Other concepts related to hate speech include:

1. Abusive language: Any form of communication intended to offend or hurt the feelings of an individual or group. This includes subtle form of abuse (Waseem et al., 2017)
2. Cyberbullying: This is an act of electronically insulting, harassing or bullying an individual. It usually involves repeatedly sending threatening or abusive messages that may have negative effect on the victim (Cheng et al., 2019).
3. Sexism: This is a hate against one’s sex or gender. It is based on the notion that one gender is to the other. Most victims of sexism are females (Jiang et al., 2022).
4. Racism: This is a form of discrimination based on person’s skin color, nationality or tribe. It also entails treating on race as superior to other races (Al-Hassan & Al-Dossari, 2019).
5. Radicalization: Araque & Iglesias (2020) defined Radicalization as “emotional transformation” where a group feels it is being unjustly treated by another outside-group. Hence, this group will try to end, the out-side group.

## Hate Speech Detection

Hate speech detection is the process of identifying hateful contents within a communication. That is, given a set of posts X = {x1, x2, …, xn}where n represents the number of posts. The goal is to classify these posts as either hate or non-hate based on a target attribute. It has proven to be a difficult task especially with the proliferation of social media platforms. These platforms generate huge amount of data which makes the traditional method of manually reviewing post almost in practicable (Plaza-Del-Arco et al., 2021). Researchers have applied different approaches to solve this complex problem including lexicon-based, ruled-based, machine learning, Deep learning and Hybrid.

### 2. Lexicon-based approach

The lexicon-based approach relies on a set of predefined lists of words or phrases associated with hate or offensive speech that are assigned subjectivity scores. The sentiments of words in a text are summed up or averaged to compute its subjectivity. Lexicons are created by collecting potentially offensive or hate words and give experts to evaluate them based on contextual meaning (Vargas, de Góes, et al., 2022). Alternatively, subjectivity analysis can be used in conjunction with semantic features (Gitari et al., 2015). One major disadvantage of this approach is its inability to capture semantics in text as they rely on the presence or absence of certain words in a text.

### 2.2 Rule-based approach

Rule-based is one of the earliest techniques for analyzing and processing textual information in NLP. This approach uses a predefined set of linguistic rules to identify hate speech. The rules are usually in the form IF-Then conditions used to make informed decisions (Schneider & Xhafa, 2022). This approach falls short of detecting contextual nuances such as sarcasm, humor and irony (MacAvaney et al., 2019).

2.6.3 Machine learning approach

This approach uses algorithms and statistical models to learn from a given data and make prediction on an unseen data based on the learned patterns. The algorithms require little human intervention and their performances increase with more training on the input data (Mullah & Zainon, 2021). With good data at hand, these algorithms can be used to build robust prediction models. Building a machine learning model entails creating a corpus – which involves collecting written or spoken documents from various sources like newspapers, social media platforms, books and so on, to study linguistics features and patterns in the data (de Gibert et al., 2018). This data is then cleaned by removing duplicates, replacing mentions, and more to prepare it for annotation. The annotation can be done by group of experts, crowdsourcing or through training some individuals (Alshalan & Al-Khalifa, 2020; Nobata et al., 2016). Depending on the researchers’ choice, at least 2 people are used for the annotation (Rahman et al., 2021). The annotated data is further processed into a form that will be suitable for the machine learning algorithms to process. Examples of these algorithms include – support vector machine (SVM), naïve bayee (NB), Decision tree (DT), Logistic regression (LR), and K-Means. These are the supervised and the unsupervised algorithms (Fortuna & Nunes, 2019). The supervised models are the most preferred (Ali et al., 2021; Davidson et al., 2017; Waseem & Hovy, 2016). The data is used to train these models to learn patterns from the data and be able to make predictions. The models’ parameter are tuned until a good prediction level is attained (Klubička & Fernández, 2018). Classical machine learning models have proved to work well even with small data (Pereira-Kohatsu et al., 2019).

## 2.6.4 Deep learning

Deep learning algorithms are a subset of machine learning algorithms that mimics the function of human brain. They are used in identifying patterns from data just like the classical machine learning algorithms. Deep learning model can be seen as a neural network model with a minimum of three layers: input, hidden and output. The Input layers take data and pass it to the hidden layers where to learn different features from the data and used it to make predictions (Polson & Sokolov, 2018).

Nigeria is a country that is facing serious security challenges like the Boko Haram Islamic Sec, the Indigenous People of Biafra (IPOB), Niger Delta Militants, The Fulani Herdsmen and the Unknown Gunmen. In addition, there are communal crises that happens in some part of the country that are religiously or ethnically motivated. The social media platforms have been observed as a key actor in fueling some of these crises through the spread of online hate speech against some communities, religion believers or ethnic groups and are mostly politically motivated (Pate & Ibrahim, 2020). Christian Chukwuebuka Ezeibe & Okey Marcellus Ikeanyibe (2017) conducted research on hate speech in Nigeria between 2010 and 2015. Their results showed a tremendous increase in hate speech during the election period and a decrease after. Similarly, they reported that politicians’ statements during campaign periods triggers physical violence and killings. In a related report of pre-election violence, a score of people were killed in some parts of the country as a result of violence incited through hate speech during political campaigns in 2015 (Ezeibe, 2021).

The Nigerian Government proposed a hate speech bill in an effort to foster national unity. The bill proposed a capital punishment (death) against any individual or group that propagates hate speech, ethnic, religious or racial discrimination. All though there was a serious outcry from the public that necessitated the removal of the death amongst the punishment (Eka, 2019).

Hausa Language

Hausa language is the second most spoken language in Africa after Swahili. It is an Afroasiatic language with more than 40 million native speakers and 15 million people using it as second or third language (Inuwa-Dutse, 2021). Most of the speakers of the language are found in the Northern part of Nigeria, in states like, Kano, Sokoto, Kaduna, and Katsina. Niger, Chad, Cameroon, Sudan, Benin, Togo, Burkina faso and Ghana also speak the language. Kananci, katsinanci and sakwkwatanci are the three most popular Hausa dialects with kananci being the standard dialect used by both local (e.g Freedom Radio News) and international media (e.g BBC news) (Adelani et al., 2020). Hausa language uses different writing styles including the Ajami which is Arabic-based was pre-dominant before the colonization. The Hausa language is one of the well-documented sub-saharan African languages. Yet, the language lacks enough resources for Natural Language Processing (NLP) related tasks, thus, it is considered as low resource languages. Hence, the need to enrich the language with more resources to support the conduct of research in the language.

## 1.2 Statement of the Problem

Hate speech detection has proven not to be an easy task (Ousidhoum et al., 2019a). A lot of research has been conducted to detect hate speech in social media using Natural Language Processing techniques (NLP). According to the literature, hate speech detection is treated as a text classification task using approaches like lexicon-based, classical machine learning and deep learning (Petrocchi & Tesconi, 2017).

Nigeria, just like other countries is not free of the menace of hate speech. Coupled with its multicultural and religious differences, its social media space is full of hateful and offensive speeches. In Nigeria, English is considered the official language of communication. However, there are three major languages spoken in the country, Hausa, Yoruba and Igbo. Hence, some Nigerians prefer communicating on social media using their native languages and some code-mixed. There are a lot of hate speech propagated in the social media space and some of which led to physical crises (Alakali et al., 2017). Currently, there are a number of research conducted to address the issues of hate speech propagation in social media. Several datasets and models have been developed for the automatic task of hate speech detection in languages like English (Davidson et al., 2017; Gröndahl et al., 2018), Arabic (Alkomah & Ma, 2022; Alshalan & Al-Khalifa, 2020), French (Battistelli et al., 2020), Chinese (Jiang et al., 2022) and India (Sreelakshmi et al., 2020). Both datasets and models for hate speech detection in the three major Nigerian languages are scarce as there are only a few research (Udanor & Anyanwu, 2019; Wilson, 2019) studies conducted for Nigerian social media space.

The importance of combating hate speech in social media can not be overemphasize and there is a great need to have datasets and models to be used to automatically detect hate speech in Hausa texts. Hence, we identify the following problems:

1. Absence of labelled hate speech dataset.
2. Low-resource languages do not have sufficient data to train state-of-the-art models.
3. In reality, social media posts are not monolingual, they contain texts in more than one language.
4. No existing tool for detecting hate speech in Hausa texts.

## 1.3 Aim of the study

The aim of this research is to develop a framework for the automatic detection of hate speech in Hausa social media text.

### 1.4 Objectives of the study

The research objectives are to:

1. Develop a manually annotated benchmark dataset for detecting hate speech in Hausa texts
2. Develop a transfer learning-based hate speech classifier for Hausa texts.
3. Develop a multilingual model for hate speech detection of Hausa code-mixed texts.
4. Develop a tool for automatic detection of hate speech text in Hausa

## 1.5 Significance of the Study

This research will provide the necessary resources like lexicon, dataset and models required to help control the activities of individuals on social media. The models can be used by both the social media companies and the government to control the spread of hate speech texts written in Hausa language, that has resulted in many physical harms to some individuals and group people. There by restoring the victims’ confidence in using the social media. Furthermore, these resources will foster research into the area of hate speech detection for Nigerian languages and more models will be available to combat hateful posts written in other languages in the country.

## 1.6 Scope of the Study

The study will look at the research done in hate speech detection, identify the progress and challenges of the field and propose ways to make improvements. Hate speech post written in Hausa language is the main focus of the research, but the model will also be tested on two more Nigerian languages, Yoruba and Igbo with the intent of developing a multilingual model.

## 1.7 Operational definition of terms

# CHAPTER TWO

## LITERATURE REVIEW

## 2.1 Introduction

The advancement in digital technology has given rise to development of various social media platforms, and has revolutionized communication among people. This has come with a lot of advantages as well as challenges, of which hate speech in one of the most disturbing among them. This chapter will look into what has been done and what needs to be done in terms of detecting and combating online hate speech. The review concentrates on the classical machine learning and deep learning approaches used for the task of hate speech detection.

Hate speech detection

Hate speech detection is the task of identifying hate in hateful contents in textual and multi-model data. The task entails the use of dataset and machine learning classifiers to train models to learn some patterns from the training data and be able to generalize on unseen data. A number of approaches have been used for this task, including the use of classical machine learning algorithms like Support Vector Machine (Defersha & Tune, 2021; Raj et al., 2021; Swaminathan et al., 2022; Tazeze & R, 2021), Naïve Bayes (Gurmessa et al., 2020; Ibrohim & Budi, 2019; Mossie & Wang, 2020), Logistic Regression (Hossain et al., 2022; Zia et al., 2021), Decision Tree (Defersha & Tune, 2021), and Random Forest (Arega, 2022; Rajalakshmi et al., 2022), Deep learning models such as in (Corazza et al., 2020; Dirkson & Verberne, 2019; Garcia-Diaz et al., 2022; Komiya & Shinnou, 2018, 2018; Kyriakakis et al., 2019; Mathew et al., 2021; Pouran Ben Veyseh et al., 2022; Shvets et al., 2021; Zhang et al., 2018) and Transformers (Caselli, Basile, et al., 2021; Das et al., 2022; M. M. Khan et al., 2021; Manerba & Guidotti, 2021; Salawu et al., 2020; Wu & Dredze, 2020; Zhou & Srikumar, 2022).

Recently, survey papers have shown significance increase in the number of publications on hate speech detection. It is observed that supervised learning is the most common approach used for the task. However, deep learning approaches are gaining popularity because of high performance.

Lexicon approach review

Vargas et al., (2022) used a lexicon of implicit and explicit offensive words to detect offensive language and hate speech in Brazilian Portuguese. The lexicons were annotated with contextual information for effectiveness. Their approach achieved a high performance than the baseline. In a related study, Jiang & Zubiaga, (2021) investigated the effect of combining hate-related lexicons with pre-trained embeddings on a Cross-lingual Capsule Network model. This model was found to perform better that all the baseline models used for comparison. A study by Hoang et al., (2022) combines Center Loss and Lexicon Attention Pooling with PhoBERT to improve hate speech detection in Vietnamese texts. Results shows a significant increase in computational cost as opposed to using just the PhoBERT model.

Rule-based

Hutto & Gilbert, (2014) created a model for sentiment analysis that uses a pre-defined set of rules to and lexical features to identify the sentiment of a given text. The model, A Parsimonious Rule-based for Sentiment Analysis of Social Media Text (VADER) was found to perform better than human ratings and has done much better than existing baselines. (Gitari et al., 2015) created a lexicon of hate speech and followed a similar approach using sentiment analysis techniques to classify text. Results of their experiment shows a promising future for this approach.

2.1 Dataset

Training a machine learning model requires a dataset. A dataset is a collection of related data points. It is an integral part of machine learning task and one of the challenging aspects of it. Some datasets have been created by researchers and made it available to the public. Most of these publicly available datasets are in high-resource languages like English, Portuguese and Spanish. Datasets in low-resourced languages are scarce and research have to create one when training machine learning models.

Vadesara & Tanna, (2023) used keywords related to hate speech to crawl tweets from twitter. They used trained annotators to classify 10,000 tweets as hate or non-hate. The inter-annotator agreement was computed using Fleiss’ kappa as 0.86. The final corpus contains about 69% of hateful tweets. Nobata et al., (2016) collected comments from Yahoo! Finance and News and annotated it into clean and offensive. A study by Tuarob & Mitrpanont, (2017), created a dataset of offensive comments curated from Facebook. Five annotators were trained to label each comment as dirty, rude, figurative, offensive or non-offensive. Alfina et al., (2017) created a dataset for hate speech detection in Indonesia by crawling tweets related to politics. The tweets were labelled into hate or non-hate by a group of 30 students, with each tweet annotated by at least three people. Only tweets with 100% agreements are considered. In a related study, Ibrohim & Budi (2018) used offensive keywords to collect Indonesian tweets. The dataset contains about 2000 tweets which were annotated by 20 different persons into three classes. The annotators were provided with a guideline and the final dataset contains only tweets that have 100% agreement. Chopra et al., (2023) used a combination of keywords and hashtags related to politics to crawl tweets in code-mixed Hindi and English. The tweets were annotated by experts in both English and Hindu languages. They uses a binary classification of hate and non-hate for the labeling and an inter annotator agreement score of 0.97 using Cohen;s kappa. A Serbian dataset for the detection of offensive language in twitter was created by Jokić et al., (2021). Tweets were labelled by 10 trained annotators on two levels: first, each tweet was label as either offensive or not offensive and those marked as offensive were further categorize as profanity, hate, derogatory and others. They reported a Cohens kappa of 0.513 as the inter-annotator agreement attained for the binary classification. Chiril et al. (2020) created a dataset of sexism in French. They used 5 experts to label sexist tweets into direct, descriptive and reporting. Majority vote was used to determine the label of a tweet. They reported an inter-annotator agreement of 0.72 Cohens’ kappa. In another study, pooling technique was used to sample tweets that are likely hateful from a large number of comments and posts collected from twitter. Active learning was used to select tweets to annotate. The inter-annotator agreement was reported to be 0.12 using Fleiss kappa (Rahman et al., 2021). Sreelakshmi et al., (2020) created a dataset for hate speech detection of code-mixed English-Indian tweets by merging three existing datasets and labelled. The tweets were labelled as hate or non-hate. Petrocchi & Tesconi, (2017) collected Italian posts and comments from Facebook and were annotated by 5 people. Each comment and post were first labelled as either, strong, weak or no hate. The hate category was further divided into 7 more classes for a fine-grained classification. Fleiss kappa of 0.19 was obtained for the first level annotation and a 0.26 for the second level annotation. Salawu et al., (2021) labeled 62, 587 tweets to detected abusive contents. They used a multi-label annotation with 17 people to label the tweets as profane, spam, porn, insult, bullying, sarcarsm, threat, exclusion or none.

Another study created a dataset of hate speech by collecting data from Stormfrost. Three annotators conducted a pilot annotation using the drafted annotation guide, and they updated the annotation guide and before the main annotation. Each sentence was annotated as hate, no-hate, relation and skip. Inter-annotator agreement was measured using both Fleiss and Cohens kappa (de Gibert et al., 2018). (Ali et al., 2021) developed the first dataset for hate speech detection in Urdu language. They used a combination of hate lexicons and target keywords to crawl posts and comments from twitter. The keywords were all related to national security, religion or ethnicity. Experts were used to annotated the tweets on two levels; type and score. They computed inter-annotator agreements for both level with Cohens kappa with the first level having the highest agreement (0.799). Davidson et al. (2017) used lexicon of hate words to collect tweet from twitter. The authors utilized crowdsourcing to annotate the tweets into hateful, offensive or neither. The annotators were provided with a clear guideline for the task and advise to consider the contextual meaning of words as used in the tweets not just the appearance of the word in a tweet. The agreement between the annotators was found to be 0.92. A total of 24802 tweets were labelled and only 0.5 percent were found to be hate speech. (Trye et al., 2022) developed a hybrid method for annotating tweets in Maori-English language. Tweets were annotated at both word and sentence level. Vargas et al. (2022) collected comments from the Instagram accounts of some politicians to detect offensive languages in Brazilian Portuguese. They used 3 esperts to annotate the comments at three level. The first level was a binary classification of the comments into offensive or non-offensive. The intensity of the offensive comments was determined at the second level of the annotation. At the third level, offensive comments that have target were further annotated to determine category of the target. A good inter-annotator agreement was reported for the binary classification. In a related study, (Zampieri et al., 2019a) used a combination of politics related offensive and general offensive keywords to collected tweets form twitter. They used crowdsourcing to annotate these tweets on three levels; offensive language detection, categorization and target identification. Majority voting was used to select tweets for the corpus. A dataset of offensive language detection in Greek was created by (Pitenis et al., 2020). Using the approach in (Zampieri et al., 2019a), tweets were classified as offensive or non-offensive by a 3 volunteer annotators. The authors used Cohens kappa to measure Inter annotator agreement. Similarly, (Ajvazi & Hardmeier, 2022) created a dataset from Facebook, and YouTube comments in Albanian language. They followed the OLID annotation guide by (Zampieri et al., 2019a) to classify the comments as offensive or not at the first stage. In stage two, comments were labeled as either targeted or untargeted, and the last stage involves identifying the comments as targeting an individual or a group. They measure interannotator agreement using Cohens’ kappa. (Waseem & Hovy, 2016) created a dataset of 16914 tweets that are potentially offensive. The authors manually annotated the tweets into sexist, racist and neither. A third annotated was involved to mitigate annotation bias. (Jha & Mamidi, 2017) extended the dataset of (Waseem & Hovy, 2016) by crawling comments from twitter using sexism related keywords. They manually annotated these tweets with the assistance of a feminist into benevolent sexism or not. The benevolent class was then merged with the sexist and neither class of the Waseem & Hovy (2006) to form a new dataset with three categories; benevolent, hostile and others category.

Chung et al., (2019) used a counter-narratives approach to create a multilingual dataset of combat hate speech in English, French and Italian language. They generated some sentences that contain hate and distributed two forms to respondents; one to respond to the pre-generated questions and the other to provide their counter hate responses. These sentences were generated by a set of 2 experts from each language. The authors employed 3 annotators to labelled the sentences and the counter-hate responses according to target and type respectively. (Alshalan & Al-Khalifa, 2020) created a dataset for hate speech detection in Arabic language. They used a combination of keywords and hashtags to crawl tweets related to racism, religion and Ideological hate. 10,000 tweets were sampled from the pool of tweets collected for annotation. These tweets were annotated as hateful, abusive or normal by crowd workers, trained annotators and experts. A multilingual dataset of English, French and Arabic language was created by (Ousidhoum et al., 2019b). It consists of tweets annotated by Amazon Mechanical Turk workers into five different levels, with each level having at least two labels. The final tweets for the corpora were selected based on majority vote. Agreement between the annotators was measured with Krippendorff alpha. In another study, three datasets were developed to detect offensive language targeted at foreigners and refugees in Germany. The datasets consist of posts and comments collected from three different Facebook pages that spread hate against these groups. For each dataset, 3 experts annotated the posts and comments into offensive and non-offensive, with the offensive class further annotated based on severity and target of the post. They recorded a good inter-annotator agreement using Cohens kappa (Bretschneider & Peters, 2017). (Fortuna et al., 2019) created a dataset with 5,668 tweets scrapped from twitter using hate related keywords and users’ profile. The tweets were annotated first by amateurs into a binary category of hate or non-hate. The hate categories were further annotated by a team of experts for a fine-grained classification. They obtained a very low inter annotator agreement using Fleiss kappa, which the authors attributes to the use of “non-expert” annotators. (Albadi et al., 2018) developed a dataset of religious hate in the middle East social media space. They collected tweets and used crowdsourcing to annotate them on a two-level scheme. First, to label the tweets as hate, non-hate and unrelated. At the second level, the target of the religious hate is selected from seven religious categories provided by the authors. The authors observed a high agreement score at the first level classification and a low agreement at the target identification level which they attribute to the number of categories at this level. (Çöltekin, 2020) collected a dataset from tweets in Turkish language. They used a multi-level annotation scheme similar to (Zampieri et al., 2019c) and a team of native speakers as annotators to categorize the tweets as offensive or not. Next, the offensive tweets were labeled as targeted or untargeted. The tweets labeled as targeted were further annotated as targeting an individual, group or others. (Ljubesic et al., 2022) conducted a two-round annotation on comments collected from videos related to Covid-19. In the first round, in-context annotation, full threads of conversations for a video were given to annotators to label as acceptable, inappropriate, offensive and violent. In the out-context annotation, the comments were mixed-up and given to annotators to label into same four categories. Krippendorf’s alpha was used to compute the agreement between annotators and a higher value was recorded for the in-context annotation. This suggests that in-context improves the quality of manual data annotation. See table 2.1 for a summary of some datasets created by authors for different languages. Guellil et al., (2020) used Arabic keywords related to politics to collect comments from Youtube. They sampled 5000 comments from their collections and employed 3 native speakers to annotate them into hate or non-hate. The annotated comments were used to create two datasets, one with an unbalance hate class and the other with a balanced hate class. Mollas et al., (2022) proposed a protocol for curating datasets. Using data collected from Reddit, they created two different hate speech datasets. The forst one contains 998 comments labeled as hate or not-hate. The second one contains 433 hateful comments that were labeled on three steps. Step 1 involves labeling the comments as communicating violence or not. Step two identifies whether the comment is directed to an individual or a group. Lastly, each comment was categorized as either race, gender, sexual orientation, nationality, disability or religion. The annotation was conducted on figure-eight, a crowdsourcing platform using 3 annotators. They used kappa coefficient to compute inter-annotator agreement for each annotation label

| Author(s) | Language(s) | Task(s) | Size | % of  Offensive  /Hate | Annotation | Annotators | Platform/ Source |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Qian et al., (2019) | English | Hate | 33,776 | 0.43 | Binary | Crowdsource/Amazon Mechanical Turk | Gab, Reddit |
| (Vadesara & Tanna, 2023) | Gujarati | Hate | 10,000 | 69.3 | Binary | Trained annotators | Twitter |
| (Aljuhani et al., 2022) | Arabic | Offensive | 29901 | 49.0 | Binary | Authors/Crowdsourcing | Twitter |
| Ganfure, (2022) | Afaan Oromo | Hate | 42,100 | - | Multiclass | Native speekers | Facebook/ Twitter |
| Çöltekin, (2020) | Turkish | Offensive | 36,232 | 0.19 | Multi-label | Trained volunteers | Twitter |
| Guellil et al., (2020) | Arabic | Hate | 5000 | 27.16 | Binary | Native speakers | Youtube |
| Albadi et al.,(2018) | Arabic | Hate | 6,136 | 0.45 | Binary | Crowdsource/CrowdFlower | Twitter |
| Nurçe & Keci, (2019) | Albanian | Abusive | 11,874 | 13.2 | Multi-label | Authors/Trained annotators | YouTube |
| Ousidhoum et al., (2019) | Arabic | Hate | 3,353 | 0.64 | Multi-label | Crowdsource/Amazon Mechanical Turk | Twitter |
| Romim et al., (2021) | Bengali | Hate | 30,000 | 0.33 | Binary | Trained annotators | YouTube & Facebook |
| Caselli et al., (2021) | Dutch | Abusive | 8,156 | 15/explicit  8/implicit | Multi-label | Trained annotators | Twitter |
| Mulki et al., (2019) | Arabic | Hate | 5,846 | 0.48 | Multi-label | Native speakers | Twitter |
| Abdelhakim et al., (2023) | Arabic | Offensive | 24,071 | 48.2 | Multi-label | Native speakers | YouTube |
| Albanyan & Blanco, (2022) | English | Hate | 5652 | 100 | Multi-label | Trained annotators | Twitter |
| Toraman et al., (2022) | English/ Turkish | Hate | 100,000 | 58.3 | Multi-label | Trained annotators | Twitter |
| (Mubarak et al., 2017) | Arabic | Hate | 1,100 | 0.59 | Multi-label | Crowdsource/ CrowdFlower | Twitter |
| Shekhar et al., (2022) | Croatian | Abusive | 2,240 | 100 | Multi-label | Trained annotators | S4sata Newspaper |
| Cercas Curry et al., (2021) | English | Hate | 20,710 | 20.4 | Multi-label | Experts | Chatbots |
| Vidgen et al., (2021) | English | Hate | 41,255 | 54 | Multi-label | Trained annotators | Synthetic-Data |
| Grimminger & Klinger, (2021) | English | Hate | 3,000 | 12 | Multi-label | Experts | Twitter |
| Kirk et al., (2022) | English | Hate | HatemojiCheck  = 3,930  HatemojiBuild  = 5,912 | 69  50 | Multi-label | Trained annotators | Synthetic-Data |
| Karayigit et al., (2022) | Turkish | Hate | HATC - 31,290  resHATC – 27,642 | 32.7 | Multi-label | - | Instagram |
| Fanton et al., (2021) | English | Hate | 5,003 | 1 | Multi-label | Experts | Synthetic-Data |
| (Mathew et al., 2021) | English | Hate | 20,148 | 0.57 | Multi-label | Crowdsource/ Amazon Mechanical Turk | Twitter & Gab |
| Zampieri et al., (2019) | English | Hate | 14,000 | 0.33 | Multi-label | Experts | Twitter |
| Vidgen, Nguyen, et al., (2021) | English | Abusive | 25,000 | 24 | Multi-label | Experts | Reddit |
| Pamungkas et al., (2020) | English | Abusive | 1,511 | 0.41 | Binary | Experts (Authors) | Twitter |
| Vidgen, Nguyen, et al., (2021) | English | Abusive | 25,000 | 24 | Multi-label | Experts | Reddit |
| Davidson et al.,( 2017) | English | Hate/ Offensive | 24,802 | 5.8/ 77.4 | Multi-label | Crowdsource/ CrowdFlower | Twitter |
| De Gibert et al., (2018) | English | Hate | 9,916 | 0.11 | Multi-label | Authors | Stormfront |
| (Waseem & Hovy, 2016) | English | Hate | 16,914 | 0.32 | Multi-label | Annotators/ Expert | Twitter |
| Waseem, (2016) | English | Hate | 4,033 | 0.16 | Multi-label | Crowdsource | Twitter |
| Founta et al., (2018) | English | Hate/ Abusive | 80,000 | 5.7/ 11 | Multi-label | Crowdsource/ CrowdFlower | Twitter |
| Mollas et al., (2022) | English | Hate | 988 | 0.43 | Multi-label | Crowdsource/ Figure-Eight | Reddit |
| Chung et al., (2019) | French/English/  Italian | Hate/ Counter-Hate | 1,719 | 1 | Multi-label | Non-experts | Synthetic/ Facebook |
| Demus & Pitz, (2022) | German | Offensive | 10,278 | 10.85 | Multi-label | Trained annotators | Twitter |
| Bretschneider & Peters, (2017) | German | Offensive | 5,836 | 0.11 | Multi-label | Experts | Facebook |
| Pitenis et al., (2020) | Greek | Offensive | 4,779 | 0.29 | Multi-label | Volunteers | Twitter |
|  |  |  |  |  |  |  |  |
| Mathur et al., (2018) | Hindi/English | Offensive | 3,189 | 0.65 | Multi-label | Trained annotators/ Experts | Twitter |
| Bohra et al., (2018) | Hindi/English | Hate | 4,575 | 0.35 | Binary | Trained annotators | Twitter |
| Mandl et al., (2019) | Hindi/ German/ English | Hate/ Offensive | English = 7,005  Hindi = 5,983  German = 5,137 | 0.36  0.51  0.24 | Multi-label | Trained annotators | Twitter & Facebook |
| Alfina et al., (2017) | Indonesian | Hate | 712 | 0.36 | Binary | Volunteers | Twitter |
| Ibrohim & Budi, (2019) | Indonesian | Hate/ Abusive | 13,169 | 0.42 | Multi-label | Crowdsource | Twitter |
| Ibrohim & Budi, (2018) | Indonesian | Abusive | 2,016 | 0.54 | Multi-label | Volunteers | Twitter |
| (Sanguinetti et al., 2018) | Italian | Hate | 1,837 | 0.13 | Multi-label | Crowdsource/ CrowdFlower & Experts | Twitter |
| Fortuna et al., (2019) | Portuguese | Hate | 5,668 | 0.32 | Multi-label | Experts/ non-experts | Twitter |
| Andrusyak et al., (2018) | Russian & Ukrainian | Abusive | 2,000 | 0.33 | Binary | Native speakers | YouTube |
| Toraman et al., (2022) | English & Turkish | Hate | English = 100k  Turkish = 100k | 58.3  34.5 | Multi-label | Trained annotators | Twitter |
| Khan et al., (2021) | Roman Urdu | Hate | 5000 | 71 | Multi-label | Author $ trained annotator | Twitter |

2.2 Machine learning approach

Machine learning algorithms have become very popular in the task of hate speech detection. The approach involves the use of data to train models to learn patterns from the data and be able to make predictions on unseen instances (Pyingkodi et al., 2023). In most cases, authors use supervised approach (labeled data) to train these models, while others use unsupervised (unlabeled data) or semi-supervised (labeled and unlabeled data) to train these models (Alrehili, 2019).

A study by (Çöltekin, 2020) used a supervised approach to train a linear support vector model (SVC) to detect offensive language in Turkish. Using word and character n-gram features, the model was able to successfully classify the comments as offensive or not and also identify the target of the offensive tweet. The model was evaluated using recall, precision and f1-score. (Joksimovic et al., 2019) collected comments from an online course to train a machine learning model to classify abusive comments. They utilize linguistic and contextual features, LIWC among other features. They experiment SVM in conjunction with each feature, and then combine all the features together. The later combination produced the best AUC ROC. (Bhattacharyya, 2022) experimented with various machine learning models to classify offensive comments in Tamil language. The author performed three different experiments using both classical and transformer-based models. The gradient boost classifier performed well on sampled data and the transformer model outperformed the classical models in the two other setups. In a related study, (Defersha & Tune, 2021) and (Kanessa & Tulu, 2021) used data collected from Facebook to detect hate speech in Afan Oromo language. While the former experimented with five classical machine learning models and reported Linear Support Vector as the best classifier with an F1-score of 0.64, the later combines SVM with Tf-IDF, word2vec and ngram features and achieved an accuracy of 0.96. (Raj et al., 2021) conducted a similar experiment for Tamil language and achieved the f1-score with Support Vector classifier. (Swaminathan et al., 2022) evaluated the performance of machine learning and transformer-based models on the task of detecting homophobia and transphobia in English, Tamil and code-mixed English-Tamil languages. Three different datasets were used and Precision, Recall and micro F1-score were reported. The models performed differently on the three datasets.

(Ibrohim & Budi, 2019) conducted a research to detect hate speech in Indonesian tweets. They experiment Naïve Bayes, Random Forest Decision Tree and Support Vector machine models with char and word n-grams to classify tweets into offensive, abusive and normal. The best result was achieved using a combination of Naives Bayes classifier with word uni-gram and bi-gram. The same authors conducted a multilabel hate speech detection by combining their datasets with datasets from other related research. The datasets were re-annotated as hate, offensive or normal. In addition, the hate category was further labeled to identify target, category and level. They compared the performance of same algorithms used in previous study with additional features and recorded the best performance from Random Forest Decision Tree classifier. (Mossie & Wang, 2018) used tf-idf and word2vec features with Naïve Bayes and Random Fores classifiers to detect hateful comments from Facebook data. They obtained a promising result from Naïve Bayes classifier with word2vec feature. (Rajalakshmi et al., 2022) evaluated the performance of eight classical machine learning classifies on the task of detecting abusive comments on two datasets, one containing abusive comments in Tamil only and another containing a mix of Tamil-English comments. Random Forest classifiers performed best with 0.78 f1-score. (Zia et al., 2021) proposed a pipeline for detecting hateful meme in multimodal comments. They used a Logistic Regression classifier to identify the target and type of hate in a meme. The model was found to perform very well with an AUROC of over 0.90 for both tasks.

2.3 Deep learning

A study by Alshalan & Al-Khalifa, (2020) evaluates the performances of CNN, GRU, CNN+GRU and BERT models on the detection of hate speech using two different dataset. The dataset set were annotated into hateful, abusive and normal, but the authors dropped the abusive class during the experiment. Using word2vec embedding trained with CBOW algorithms, these algorithms were used to classify Arabic tweets as hateful or normal. The CNN model performed best with an F1-score of 0.79 and AUROC of 0.89.

Meedin et al., (2022) proposed a crowdsource annotation platform. They used the platform annotate comments from Facebook, Twitter and YouTube at different levels. Both male and females of different age groups and religion were selected for the annotation. The first level annotation categorized the comments as ‘offensive content’, ‘not offensive content’ and ‘cannot tell’. The second level identifies hate content from the offensive comments. The third level is to identify the target and the fourth level identifies the hate category. The inter-annotator agreement was computed using Cohen Kappa coefficient, Fleiss’ kappa and Krippendorff’s alpha. The final corpus contains 52,646 Comments from Facebook and 45,000 comments from each from Twitter and YouTube.

Almaliki et al., (2023) proposed a Mini-BERT model for detecting hate speech in Arabic tweets. In order to evaluate the performance of their model, they developed several baseline models using Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Random Forest (RF), Gradient Boosting (GB)., Convolutional Neural Network (CNN) and Long Short-Term Memory network (LSTM) and a combination of CNN + LSTM. The tweets were classified into normal, hate and abusive. The Arabic BERT-mini model was found to perform better that the baseline models with an accuracy of about 0.99.

Rathnayake et al., (2022) investigated how adapter based pre-trained language models (PMLM) fine-tuning can improve the classification task of code-mixed and code-switch text. They proposed some novel techniques of adapters stacking, fine-tuning and training of base model “without freezing” for a single code-mixed and code-switch task.

Stankovic & Mladenovic, (2023) used a combination of lexicons and keywords to collect comments from YouTube, blic.rs and b92.net news channels. The comments from YouTube relates to general hate while those from the two other sites were related to sports in Serbian. These comments were annotated into yes, no, or neutral as it relates to hate. The comments from the YouTube were used to create a general hate dataset while those from the news site were used to create a sport related hate dataset. Krippendorff’s alpha of 0.67 was obtained as the inter-annotator agreement. A BiLSTM model was trained on both datasets. They obtained a good precision score with a low recall.

Markov et al., (2022) proposed an ensemble approach to hate speech detection in Dutch text. They used two existing Dutch hate speech datasets to evaluate the performances of BERTja and RobBERT, pre-trained Dutch language models and also a classical model, SVM. The ensemble model was formed by combining these models with a gradient boosting and other features like number of hate words, personal pronouns, length of comments and more. The BERTja performs better for the individual models for both datasets and the ensemble performed best in the overall experiment.

Zhou et al., (2020) proposed a fusing method to improve the performance of hate speech classification. They used the dataset of SemEval 2019 Task 5 to train Embeddings from Language Models (ELMo), Bidirectional Encoder from Representations Transformers (BERT) and Convolutional Neural Network (CNN) classifiers to categorize tweets as hate or non-hate. The results of the individual classifiers were combined through a set of rules like “voting, meaning, maximizing and production” to obtained a better result. They recorded an improvement in the F1-score of about 13% when compared to the best F1-score obtained in the individual first case.

Mollas et al., (2022) investigated the task of hate speech detection on two different datasets. The study uses both machine and deep learning models to perform a binary classification of comments as hateful or not. The machine learning models were implemented without feature extraction with the deep learning models were implemented with FastText and GloVe embeddings. In addition, they used BERT and BistilBERT models. The DistilBERT recorded the highest f1-score.

Badjatiya et al., (2017) used the dataset of Waseem & Hovy, (2016) to explore the use potentials of neural networks in hate speech detection. The dataset contains about 16,000 tweets labeled as racist, sexist or neither. They used Logistic regression, support vector machine and Gradeint boosted decision tree with features like TF-IDF and Bag of Words Vectors to train some baseline models. Next, they combined CNN, FastText, LSTM and GBDT with GloVe and random embeddings. The result of LSTM + random embeddings + GBDT was the best.

Baydogan & Alatas, (2022) created two datasets for Covid-19 hate related tweets. They used both machine learning and deep learning algorithms to detect hateful tweets towards the Asians, who a seen as originators of the outbreak. In all the two experiments, the deep learning models outperform the machine learning models.

Karim et al., (2021) proposed a hate speech detection model for Bangla language to provide an understanding of the black-box transformer models using the Sensitivity analysis (SA) and layer-wise relevance propagation approach. They developed both machine and deep learning baseline models and compared them with an ensembled approach of four monolingual and multilingual transformer models. The best performance was obtained from the ensemble model with an F1 – score of 88%.

Aljuhani et al., (2022) used a blended model of AraVec and a domain-specific AraOffW2V models with BiLSTM to detect offensive comments in Arabic tweets. The proposed model recorded a promising performance when compared to existing models for offensive speech detection in Arabic.

Boulouard et al., (2022) used the dataset of Alakrot et al., (2018) to compare the performances of four different BERT-based models; BERT and mBERT models for English translated version of the comments, and mBERT and AraBERT for the original Arabic comments. The models were trained on the datasets and evaluated using the accuracy, precision, recall and f1-score metrics. The BERT trained on the translated English comments performed better in terms of all the four metrics (precison, recall, F1, and accuracy).

Emon et al., (2022) investigated the performance of Bangla BERT, Bengali DistilBERT and XLM-RoBERTa models in classifying Facebook comments written in Bangla as sexual, threat, religious, troll and not-bully. The Begali DistilBERT model performed best with an F1-score of 84%.

Ganfure, (2022) utilized five deep learning algorithms to detect hate speech in Afan Oromo comments collected from Facebook and Twitter. The comments were categorized as hate, offensive, both hate & offensive or neutral. Using CBOW embeddings, the BiLSTM performed best with an F1 – score of 91%.

Saleh et al., (2023) used about five existing datasets and used two different methods to detect hate speech. At first, they combine domain-specific and domain agnostic embeddings with Bidirectional Long Short-Term Memory (BiLSTM). The domain-specific embeddings outperformed the domain agnostic embedding because they were trained using hate speech terms. The second approach used a Bidirectional Encoder Representations from Transformers (BERT). BERT-base and BERT-large were then fine-tuned on these datasets and their performances evaluated using precision, recall, f1-score and AUC for a binary classification. The best result was obtained from the BERT experimental setup.

Guellil et al., (2020) created two datasets of hateful comments related to politics in Algeria and train both classical and deep learning models to classify these comments as hate or non-hate. They used Word2Vec and FastText models to extract features. Experimental results show that hate speech detection models perform better when trained with a balanced dataset.

Ketsbaia et al., (2020) explored the performances of various machine learning models in the task of hate speech classification using two existing datasets. The models were both implemented with unigrams, bigrams, and trigrams and their performances were measured using accuracy, precision, recall, and F1-score. The Linear SVC with unigram, bigrams and trigrams recorded the best performance in the second dataset. Next, they investigated the use of CNN model with CBOW and Skip-gram embedding models for the same task. In these setup, CNN + Skip-gram had the best AUC ROC. Lastly, the authors investigated the performance of two vector representation models, Distributed memory mean (DMM), Distributed bags-of-words (DBOW). These two models were trained individually, and combined together. The best performance resulted from the combined setup.

Chiril et al., (2022) used seven existing datasets to develop a model to identify the target and topic of hate in a comment. These datasets where trained on state-of-the-art models to identify similar characteristics of hate speech. The aim of the training was to see if the learned characteristics can be transferred from topic-generic datasets to topic-specific datasets. They create a baseline model using Linear Support Vector Classifier (LSVC). The result of the state-of-the-art models where compared against the baseline and a little increase was obtained which the authors consider as negligible and concluded that topic-generic datasets are not good at detecting topic-specific hate speech. Furthermore, they assessed the potential of topic-related datasets for multi-target hate speech detection using same models. High performances where observed for most of the models which indicates the potential of using topical datasets in detecting multi-target hate speech.

Halawani et al., (2022) proposed a hate speech detection model, Enhanced Seagull Optimization with Natural Language Processing Based Hate Speech Detection and Classification (ESGONLP-HSC). The model uses attention-based bidirectional long short-term memory (ABLSTM) and glove for feature extraction while the enhance seagull optimization algorithms was used for hyperparameter optimization. They trained their model of two datasets, one curated from Storm front and the other from Crowd-flower to classify comments as hateful, offensive or neutral. The model was compared against some existing models and performed best with an accuracy of 99.24%.

Khan et al., (2021) studied hate speech detection in Roman Urdu by creating a dataset of 5000 tweets labeled on three stages. After preprocessing the tweets, an iterative guideline was developed for annotation. The first step in the annotation involves labeling tweets as either neutral or hostile. Next, the hostile tweets were annotated as simple or complex. Lastly, the simple or complex tweets were labeled as either offensive or hate. They combined some classical and deep learning models with features like word level TF-IDF, count vectors, and word embeddings to established a baseline model. The best performance setup was obtained from Logistic regression and count vector with an f1-score of 0.932.

Pandey et al., (2022) employed sentiment analysis approach to detect hate in social media posts. They used a dataset from Kaggle with comments labeled as insulting or not insulting. The authors preprocessed the data and extracted features using Bag of words model. This model was used together with naïve bayes classifier for the task on 80 – 20% train – test split. They obtained an accuracy of 99.7%. Furthermore, they compared the performance of other feature extraction models, TF-IDF and word2vec models and implemented each with Logistic regression and Random forest classifier. The best accuracy was obtained from the word2vec models.

M. U. S. Khan et al., (2021) proposed a generic framework for multi-label hate speech detection for social media sites. They used n-grams and sequential convolutional neural network (SCNN) to implement the framework. The authors implemented a baseline model using n-gram features and support vecto machine (SVM). There results show a significant increase in performance when compared to some state-of-the-art models. The authors also argue that modeling hate speech detection task as multi-label problem gives a better performance that as a multi-class problem.

A study by Roy et al., (2023) used real-time tweets to detect the presence of hate speech. They developed a deep learning model using Long-short term memory (LSTM) model with TF-IDF feature extraction model. Using the same feature extraction model, they experimented with other classifiers like SVM, LR, NB, XGB, BERT, KNN and ANN and compare their performance with the proposed LSTM model using Aurea under the curve (AUC), precision, recall, accuracy and f1-score metrics in which the LSTM model outperformed all the other models in terms of accuracy.

Karayigit et al., (2022) proposed a technique for the detection of hateful and homophobic comments from Instagram. Out of the collected comments, 1226 were classified as homophobic, 10237 as hateful and 19827 as neutral. They used random oversampling and under-sampling techniques to create a new dataset with balanced classes of homophobic, hatful and neutral comments. These two datasets were used to train a multilingual BERT and compared the performance with other traditional and deep learning models. Accuracy, precision, recall and F1-score metrics were used to evaluate the models’ performances. The best performance was obtained from mBERT with an F1-score of 91.75%.

Oriola & Kotzé, (2020) applied a semi-supervised approach to detect abusive language from South African tweets. Unlabeled data were assigned labels based on the features extracted form both labeled and unlabeled clusters through majority voting as against the self-learning and active learning employed by several studies. The authors used Logistic Regression, Support Vector Machine and Neural Networks with the proposed method and compared it with the state-of-the-art semi-supervised methods. Experimental results show that their proposed approach performed better.

Ali et al., (2021) used a combination of keywords and lexicon to collect tweets in Urdu. The tweets were preprocessed and a final corpus of 16,000 tweets was obtained. Two annotators were used to label the tweets. Firstly, each tweet was categorized under one of the three domains; National security, Religious or Ethnic distinctions. Next, each tweet was labeled as containing offensive language or not. This was done by assigning a score of 1or 0 for offensive and non-offensive tweets respectively. They used a TF-IDF and two classifiers; Support Vector Machine (SVM) and Multinomial Naïve Bayes (MNB) to create baselines. The best performance was recorded from SVM. The authors also addressed the problem of sparsity, high dimensionality and class imbalance using “dynamic stop word filtering”, variable global feature selection scheme (VGFSS) and synthetic minority over-sampling technique (SMOTE) respectively. They performed another experiment and compare it with the baseline. Just like in the first case, SVM performed best with an F1-score of 0.98.

Pamungkas et al., (2023) argued that high performance may be attained by using cross-lingual word embeddings to train some models on the source language and apply them to the target language, which is devoid of labeled examples. Next, by bootstrapping labels using an ensemble of various model architectures, the authors used unlabeled target language data for additional model improvements. Additionally, they addressed the problem of class imbalance in hate speech datasets.

Shishah & Fajri, (2022) compared the performance of classical machine learning, deep learning and transformer model using some renowned datasets. Logistic Regression and Support Vector Machines with Bag-Of-Words and TF-IDF feature representation for the classical models. For the deep learning models, CNN, LSTM, BiLSTM with GloVe and Word2Vec. BERT, RoBERTa and XLNet as the transformer models. The authors used G-Means, F1-score, Accuracy and AUC score to evaluate performance. The transformer models were observed to perform better although it requires high computational resources.

Velankar et al., (2023) compared the performance of monolingual and multilingual transformer models in classifying hate speech text in Marathi language. They used a Marathi hate speech dataset to train mahaBERT, mahaALBERT, mahaRoBERTa as the monolingual models and used the mBERT, indicBERT and xlm-RoBERTa as the multilingual models. The result of their analysis shows that monolingual models performed better. Furthermore, the authors concluded that the monolingual models give a better sentence representation.

A study by Kar & Debbarma, (2023) presented a new approach to hate speech detection in multilingual text. They proposed an improved seagull optimization (ISO) algorithm for feature extraction and applied a quantum search optimization algorithm on the features to minimize dimensionality. The enhanced features were then combined with a hybrid diagonal gated recurrent neural network and experimented on HASOC 2019 dataset which contain comments in English, Hindi and German. The authors tested on all the three Tasks of the HASOC competition. They compared their approach with existing models like SVM, LR, k-NN, and it performed best in all the three tasks.

.1 Transformers

2.3.2 Language models

| S/N | Authors | Categories/Levels(L) | Algorithms | Features/Embeddings | Performance |
| --- | --- | --- | --- | --- | --- |
| 1 | Almaliki et al., (2023) | L1: Hate, Abusive, Normal | SVM, DT, RF, GB, KNN, BERT-mini, AraBERT. | N-Gram, BERT embeddings | Accuracy – 0.986  Recall – 0.986  Precision – 0.986 |
| 2 | Stankovic & Mladenovic, (2023) | L1 – Hate: Yes, No, Neu | BiLSTM | CBOW, one-hot | Precision – 0.96  Recall – 0.32  F1 – 0.48  Accuracy – 0.46 |
| 3 | Saleh et al., (2023) | L1: Hate, non-hate | BiLSTM, BERT | Word2Vec, GloVe, CBOW | F1-0.98 |
| 4 | Roy et al., (2023) | L1: Hate, non-hate | LSTM | TF-IDF, N-gram | Accuracy – 0.7 |
| 5 | Pamungkas et al., (2023) | L1: Presence of swear word or not  L2: Swear word is offensive or not | BERT, LSVC, LR, RF | n-gram, syntactic, stylistic | F1 – 0.68 |
| 6 | Kar & Debbarma, (2023) | L1: Non-offensive, hate speech  L2: Hateful, offensive, propane word  L3: Targetted, untargetted |  | Sentiment base – e.g word polarity  Semantic based – e.g capitalization, punctuation, interjections.  Unigram  Pattern based | L1 accuracy – 87.74  L2 accuracy – 88.98  L3 accuracy – 84.74 |
| 6 | Halawani et al., (2022) | L1: Hate, offensive, neutral | ABLSTM | GloVe | Accuracy – 0.99 |
| 7 | Chiril et al., (2022) | L1: Hate, not hate | LSVC, BERT, ELMo, CNN, LSTM | FastText | F1 – 08.4 |
| 8 | Emon et al., (2022) | L1: bully, not-bully  L2: Sexual, threat, troll, religious | Bangla BERT, Bangali DistilBERT, XLM-ROBERTa | - | F1 – 0.86 |
| 9 | Aljuhani et al., (2022) | L1: Offensive, non-offensive | BiLSTM | AraVec, Domain-specific AraOffW2V | F1 – 0.93 |
| 10 | Baydogan & Alatas, (2022) | D1: Counterhate, non-Asian Aggression, Hate, Neutral  D2: Counter-speech, Entity-Directed-criticism, Discussion-of-Eastasian-Prejudice, Entity-Directed-Hostility | SVM, NB, LR, RF, KNN,  ANN, RNN, CNN, LSTM, GRUs | Document matrix, TF-IDF, Bags-of-Words | Accuracy – 0.79 Precision – 0.60 Sensitivity - 0.55 F1 – 0.51 |
| 11 | Mollas et al., (2022) | L1: Hate, not-hate | MNB, BNB, LR, SVM, RF, GB, CNN, LSTM, BiLSTM, BERT, DistilBERT | FastText, GloVe | F – 79.92 |
| 11 | Markov et al., (2022) | L1: Hate, non-hate | Gradient boosting, SVM, BERTja, RobBERT | - | Precision – 68.1  Recall – 75.2  F1 – 66.9 |
| 12 | Boulouard et al., (2022) | L1: Abusive, neutral | BERT, mBERT, AraBERT, LR, NB, LSTM, RF, SVM | BERT | F1. Accuracy, Recall, Prescision – 0.98 |
| 12 | Karayigit et al., (2022) | L1: Hateful, homophobic, neutral | SVM, NB, RF, AdaBoost, XGB, GB, LSTM, BiLSTM, GRU, mBERT | n-gram, TF-IDF, GloVe | F1 – 91.75 |
| 13 | Ajvazi & Hardmeier, (2022) | L1: Not-offensive, offensive. L2: Targeted/untargeted. L3: Individual/Group/ Others | Bi-LSTM | - | F1-0.86 Task A |
|  | Ali et al., (2021) | L1: Field – National security, Religion, Ethnic  L2: Offensive, not-offensive | SVM, MNB | TF-IDF | F1 – 0.98 |
| 15 | (Karim et al., 2021) | L1: Political, personal, geopolitical, religious, gender | Bangla BERT-base, Multilingual-BERT-cased/uncased, XLM-RoBERTa, | FastText | F1 – 0.88 |
| 14 | Alshalan & Al-Khalifa, 2020) | L1: Hateful, Normal | CNN, GRU, CNN+GRU, BERT | Word2vec embedding | F1 – 0.79  AUROC – 0.89 |
| 15 | Ketsbaia et al., (2020) | L1: Hate, non-hate | LR, LSVC, MNB, BNB, CNN | Unigram,bigram, trigram, Doc2vec | Accuracy – 0.91  Precision – 0.92  Recall – 0.91  F1 – 0.91 |
| 15 | Guellil et al., (2020) | L1: Binary | GNB, LR, RF, SGD, LSVC, CNN, MLP, LSTM, BiLSTM | Word2Vec,FastText | F1 – 0.91 |
| 15 | Zhou et al., (2020) | L1: Hate, Non-hate | ELMo, CNN, BERT | ELMo | F1 – 0.698  Accuracy – 0.732 |
| 16 | Badjatiya et al., (2017) | L1: Racist, Sexist, Neither | LR, SVM, CNN, GBDTs | TF-IDF, GloVe, BoWV | Precision, Recall, F1 – 0.930 |
|  |  |  |  |  |  |
| 17 | Khan et al., (2021) | L1: Neutral, hostile  L2: Simple, complex  L3: Hate speech, offensive | NB, LR, NF, SVM, CNN, | CLV, NGV, WLTF, WE | F1 – 0.93 |
|  |  |  |  |  |  |

CLV = Character level vectors, NGV = N-gram vector, CV = Count vector, WLTF = Word level TF-IDF, WE = Word embeddings, MNB = Multinomial Naïve Bayes,

* 1. **Review of Related literature**

T et al., (2020) used the dataset of SemEval-2019 Task 6 to investigates the usefulness of three features namely Google sentence encoder, Fast-text, Dynamic mode decomposition, alongside the Random kitchen sink (RKS) approach for detecting offensiveness in a post. The result of the experiment was evaluated using four metrices and the best performance was recorded from the combination of fast-text and the Random kitchen sink.

* 1. **Summary of Literature Review**

**2.12. Research Gap**

**CHAPTER THREE**

**Dataset**

3.1 Introduction

This chapter explains the details of steps taken to collected data from twitter using keywords related to hate speech and offensive language, the approach used in cleaning the data before and after annotation, the entire annotation process and the final corpus.

3.2 Data source

The choice of data source is crucial in dataset creation. Data should be collected from reliable sources and be of good quality. Different sources have been used by different researchers for hate speech detection tasks. Most common among these sources are the Twitter, Facebook, Yahoo!, Wikipedia, and YouTube (Fortuna & Nunes, 2019; Jahan & Oussalah, 2023). In this research, we chose to use twitter as our dataset because 1) it is one of the most used social media platforms by Nigerians, 2) the availability public API to crawl data freely, 3) it is good for studying real time events.

Seed keywords



Twitter

Crowdsourced keywords

Crawled tweets

Initial pre-processing

Annotation

Dataset

Social media and conflict in Nigeria, doc

Figure 3.1: Dataset creation flowchart

Keywords selection

One of the approaches to crawling tweets is the use of keywords, hashtags, and identifying users that a known to post offensive contents (Hoefels et al., 2022; Ibrohim & Budi, 2018; Zampieri et al., 2019a). Researchers employ this technique to increase the chances of getting offensive tweets from twitter (Schmidt & Wiegand, 2017). We adapted this approach and begin by collecting keywords from a document of hate speech lexicon terms in Nigeria by (Ferroggiaro, 2017) as shown in figure 3.1. We extracted terms used for Hausa language and gave it to two Hausa language experts for validation. Six of the terms, arne (non-muslim), kafir (non-believer), inyamuri (Igbo), almajiri (student of traditional qur’anic school), boko haram, and karuwa (prostitute) were unanimously agreed by the two experts as conveying hate depending on context used in the sentence.

Obviously, the six terms will not be enough for the collection, hence, we resorted to crowdsourcing. We drafted a google document (see appendix ..) and a shared it on Facebook, Twitter and through emails requesting people to provide use with examples of hateful and offensive words in Hausa language. Interestingly, people responded positively, and we collated these words and gave it to our language experts for validation. Those words that passed the validation were retained and used for the tweet crawling. We maintained two lists, one for hate speech and the other for offensive words. For each word in the list, we updated the masculine, feminine and plural of the word. Table 3.1 shows the total number of hate and offensive words used to collect tweets.

Table 3.1: Keywords statistics

|  |  |
| --- | --- |
| Hate | Offensive |
| 51 | 149 |

Tweets collection

These keywords were used to collect tweets using the twitter academic API which gives access to historic data. Two separate searches were conducted for hate and offensive keywords. A total of 10345 and 119648 tweets were collected for the hate and offensive keywords respectively.

Table 3.2: Crawled tweets

|  |  |
| --- | --- |
| Keywords | No. of tweets |
| Offensive  Hate | 119648  10345 |

Data cleaning

We notice the possibility of having many non-Hausa tweets from our collection. This is so because Nigeria is a multilingual country. There are terms that have similar spellings but different meanings in different languages. Example of a word in Hausa with different meaning in Igbo is *“biya”* which means *“to pay”* in Hausa language but means *“beer”* in Igbo language. Therefore, we used location, longitude, latitude and radius parameters to crawl the tweets. Additionally, we used Google CLD3[[7]](#footnote-7) and Natural Language API[[8]](#footnote-8) to detect and remove these tweets not written in Hausa language. We removed duplicates, retweets, URLs, mentions and redundant whitespaces. All tweets were equally lowercased.

Data sampling

Since we used keywords for the collection, there is high chances of bias as tweets related to some keywords may dominate the data (Çöltekin, 2020). Therefore, to mitigate the bias, we sampled tweets per keyword from the data. The data sampling strategy involves looking at time span of the tweets, in which case we consider 2022 and 2023. Again, we look at events that occurred within this periods that may likely promote offensive and hate speech. Table 3.3 gives a list of events that were considered.

Table 3.3: Sampling strategy events

|  |  |
| --- | --- |
| Event | Date |
| Zamfara massacre where over 200 people were killed  ASUU strike  Abuja-Kaduna train attack  Owo church attack  Kuje Prison attack  Naira redesign  General Election  Removal of fuel subsidy | Jan, 2022  Feb-Oct, 2022  March, 2022  June, 2022  July, 2022  2023  Feb-Mar, 2023  May, 2023 |

Data annotation

Data annotation is the act of labeling data into a pre-defined class or category. We started the processing by developing an annotation guideline adapted from Petrocchi & Tesconi, (2017). The process involves classifying a tweet into offensive (OF), hate (HT), normal (NL), and indeterminate (IN), at the first task. In addition, annotators were instructed to select all the offensive/hate words in the tweet. The next task identifies the target of hate tweets as gender, religion, ethnicity, political affiliation and others. If an annotator chooses the “others” target, he/she is required to mention the specific target.

**Definitions:**

* **Offensive language** is any form of bad language expressions including r*ude, impolite, insulting or belittling utterance intended to offend or harm an individual.*
* **Hate speech** is language content that expresses **hatred towards a particular *group or individual*** based on their **race**, **ethnicity**, **religion**, **gender**, **sexual orientation, or other characteristics. It also includes threats of violence**
* **Normal** is any form of expression that does not contain any bad language belonging to any of the above classifications.
* **Indeterminate** is any tweet that is not **readable** or is **completely** written in another language other than your language of annotation.

The annotation guideline

***Task Description:***

***Task 1****: Given a tweet, select the option that best describes it:*

* *Offensive language*
* *Hate speech*
* *Normal*
* *Indeterminate*

*Select all words/terms that are hateful or offensive.*

***Task 2:*** *If the label in task 1 is hate, select the target of the hate speech*

* *Ethnicity*
* *Religion*
* *Disability*
* *Gender*
* *Political affiliation*
* *Others*

*For the “others” target, mention the target.*

Annotators training

Three native Hausa language speakers and one language annotator were recruited and trained for the annotation exercise. All of them have a minimum of first-degree qualification and have background in Computer Science or Linguistics. These annotators have previously participated in one data annotation or the other. We used LabelStudio to train the annotators and asked them to label a sample of 100 tweets as a pilot study. The annotators were given google doc to record their annotation experience and suggestions. The feedback from the annotators was used to the annotation guide. For example, in the first draft of the guidelines, task 1 did not include the *“indeterminate”* class. But the annotators suggest for another category to label tweets that do not have any meaning and those in another language, so we added the indeterminate. At the end of the pilot annotation, an adjudication was conducted under the supervision of the language coordinator. This was done to give the annotators a better understanding of the task. Annotation is an iterative process, as such, after updating the guideline, they were given another set of 200 sampled tweets for a second pilot study. At this round, we observed a significant increase in inter-annotator agreement score among the annotators. In this round, another target class was added because there were some tweets that contain political hate/offensive related terms as observed by the annotators. This resulted in having the *“political affiliation”* target label. We iteratively update the guidelines based on the adjudication report.

Main annotation

The main annotation was divided into two phases. In the first phase, a sample of 1000 randomly sampled tweets were annotated by two experts and all agreements were resolved by a third expert annotator. These labels were used as gold standards for the quality control in the second phase annotation. In the second phase, our trained annotators were given a total of 5000 tweets in a batch of 1000 samples to annotate. After each round of annotation, disagreements were between annotators were resolved through adjudication with the coordinator.

|  |  |
| --- | --- |
| Sentence | Translation |
|  | Jan, 2022  Feb-Oct, 2022 |

The gold standard label

One of the common approaches to choosing a gold label is the simple majority vote (Çöltekin, 2020; Hoefels et al., 2022; Jokić et al., 2021; Nobata et al., 2016). Simple majority vote considers tweets with same label all through, in our case, where all three annotators agreed on a label and where two annotators agreed on a label. For every annotated tweet, we considered one of the following:

Three-way agreement: This is when all the three annotators agree on the label of a tweet. In this case, we consider the agreed label as gold standard.

Two-way agreement: This happens when two annotators agreed on a label and one disagree. For example, two annotators label a tweet as hate (HT) and the third label it as offensive (OF). This is when adjudication happens and the final label is determined by the coordinator, who doubled as fourth annotator.

Three-way disagreement: Here, all the three annotators have different label for the tweet. For example, the first annotator label it as HT, the second as OF and the third as NL. Similar to the two-way agreement, the adjudication is done and the final label determined by the fourth annotator.

Table 3. 5: Task 1 labels

|  |  |
| --- | --- |
| label | No. of tweets |
| Offensive  Hate  Normal  Indeterminate |  |

Inter-annotator agreement (IAA)

Inter-annotator agreement (IAA) measures the degree of agreement between two or more annotators. It is used to access the reliability of the annotation. Some of the commonly used metrics for computing inter-annotator agreement include Cohen’s Kappa, κ (Mulki et al., 2019), Fliess’ kappa, κ (Akhtar et al., 2020), and Krippendorff’s alpha, α (Meedin et al., 2022). In our case, we used the Fliess’ kappa to measure the agreement between the three annotators. Table 3.6 shows the IAA for task 1 with 4-classes and task 2 with 5 classes.

Table 3.6: Inter-annotator agreement for task 1

|  |  |
| --- | --- |
| label | No. of tweets |
| Offensive  Hate  Normal  Indeterminate |  |

Furthermore, we computed the IAA of each class with other classes to determine the classes with difficulties during the annotation. *To complete after annotation*

Annotation observations

We made some observations during the annotation. This include a smaller number of hateful tweets. Most of the tweets are either normal or offensive. Words like ‘Almajiri’ which means a Quranic school student, is mostly being used as a hate term towards individuals, especially, Hausas. In addition, some tweets contain offensive or hateful words but do not convey offense or hate as used in that context. Examples include:

*“@username amma fa mu ba dolaye bane wallahi”* the word *“dolaye”* means *“fools”* in Hausa, which is offensive. But the tweet when translated means *“we are not fools”.* Hence, the tweet is labeled “normal”.

*“@username ku guntaye”.* Normally, the word *"gajere"* means *"a short person"* which can be classified as normal but in this tweet the word *"guntaye"* is used which means *“short people”* in an offensive way which is why the tweet is classified as “offensive”.

*“in dr congo: opposition leader urges citizens to reject kabila ##url”*

The above sentence contains the word *“kabila”* which in means a person from another tribe, and usually used in an offensive way. But in this context, it is a name of an individual, hence, it is completely in English language, therefore, marked as ‘indeterminate”.

*“@username shege mutumina, ashe ka nan”.* The word shege in Hausa means *“bastard”* which is offensive, but used here in a complementary way, as such it is labeled as *“normal”.*

Proverbs or wise saying in Hausa were challenging for the annotators because most of them do have more than one interpretation. Another challenge was annotating tweets that are prayers. After a through consultations with experts, it was resolved that all prayers should be classified as ‘*Indeterminate*’ to reduce bias. It was also observed that some tweets appear to be normal but the meaning in real sense is offensive.

.

1. https://www.statista.com/statistics/1176096/number-of-social-media-users-nigeria/ [↑](#footnote-ref-1)
2. https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online\_en [↑](#footnote-ref-2)
3. https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/ [↑](#footnote-ref-3)
4. https://support.google.com/youtube/answer/2801939?ref\_topic=9282436 [↑](#footnote-ref-4)
5. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy [↑](#footnote-ref-5)
6. https://about.instagram.com/blog/announcements/an-update-on-our-work-to-tackle-abuse-on-instagram [↑](#footnote-ref-6)
7. <https://github.com/google/cld3> [↑](#footnote-ref-7)
8. https://cloud.google.com/natural-language/ [↑](#footnote-ref-8)