Cross-Platform NLP Framework for Detecting LGBTQIA Hate Speech: Evaluation on Reddit and Simulated Twitter Datasets

Alan Janbey
Department of Computer Technology, The London College, UCK
680 Bath Road, Cranford, London, TW5 9QX

ABSTRACT
Online hate speech targeting the LGBTQIA community presents a persistent challenge to social cohesion and individual well-being. This study proposes a computational approach to detecting and mitigating such content using Natural Language Processing (NLP) techniques. Data were collected from public Reddit forums, annotated into offensive and acceptable categories, and pre-processed using tokenisation, normalisation, and stopword removal. Both Count Vectorisation and TF-IDF Vectorisation were employed to generate features for training a Decision Tree Classifier. To enhance robustness and assess cross-platform applicability, a simulated evaluation was also conducted on a representative Twitter dataset. The Reddit dataset evaluation yielded an accuracy of 0.76, with strong precision for acceptable content but lower precision for offensive content due to vocabulary variability. The simulated Twitter dataset showed improved balance between precision and recall, achieving an accuracy of 0.81. High-resolution visualisations, including word clouds, class distribution charts, and an NLP workflow diagram, provide insights into data characteristics and model architecture. The results indicate that the proposed approach is effective for detecting offensive language in LGBTQIA-related discourse and adaptable to multiple social media platforms. Future research will explore multilingual extensions, multimodal content analysis, and real-time deployment for proactive content moderation.

General Terms
Artificial Intelligence, Natural Language processing.

Keywords
Natural Language Processing, Data analysis, online hate speech, LGBTQIA

1. INTRODUCTION
"Nothing is more lovely and powerful than someone who chooses to be themselves," said Maya Angelou. Coming out and being honest about who one is taking a lot of guts, especially whether one is gay, lesbian, transgender, or any other identity. Does the LGBTQIA community have the right to experience such hostility? Should they not accord the same level of respect as everyone else? This community has had many troubles gaining acceptance from a larger society, especially from people who are closest to them. They are frequently denied even the most fundamental human rights, such as the right to education, healthcare, and employment. Things absolutely need to shift to their current state The LGBTQIA community has been pushing back against oppression and fighting for everyone's acceptance while also working to secure their rights. It is necessary for them to do something to make the process significantly more welcoming and beneficial.

What exactly constitutes ‘hate speech’? Hate speech, in this definition, is an offensive language that targets a group or an individual based on the basis of their characteristics and has the potential to disrupt social harmony. It’s the use of insulting language or behaviour directed at a specific person or group because of their identity, whether in a written, spoken, or nonverbal form of communication. Communication can be verbal, written, or non-verbal. People can express their contempt for others through the use of memes, symbols, artefacts, gestures, and animations, whether they are communicating with them online or in person. It is estimated that 42% of LGBT individuals experience hostility at home. "Hate speech" is a term used by members of the LGBTQIA+ community to describe any kind of communication that promotes violence towards a particular group.

Hate speech includes comments on a person's or group's religion, nationality, descent race, gender, colour, origin, health status, or disability. There is an intolerable amount of vitriol directed against the LGBTIQ community and other marginalised communities worldwide [1] there is a hazardous environment for their survival because of the widespread use of pejorative phrases and slang terminology to describe them. Therefore, they avoid self-reflection and bury their emotions. As the use of social media platforms has grown, so has the amount of harassment and bullying directed toward this demographic. Consequently, identifying the language is crucial to ensure that people continue to live together peacefully and prosperously.

An abundance of hate speech, much of which targets members of the LGBTQIA community, is disseminated via social media. Those guilty for stoking divisions should show the error of their ways. The person(s) responsible for death threats should be held to the full extent of law. The fear and intimidation produced by hate crimes and hate speech prevent people who identify as lesbian, gay, bisexual, or transgender (often referred to as LGBT) from fully participating in society.

Natural language processing, NLP is a branch of AI that aims to provide machines with human-like language understanding abilities [2] Computer programmes that utilise natural language processing have allowed for rapid translation between languages, immediate responses to spoken requests, and rapid creation of huge quantities of text. These developments allow computers to "understand" the full meaning of human language, including intention and emotion, whether the language is written down, spoken, or collected as data. In natural language processing, the two most crucial steps are the "data pre-processing" stage and the "algorithm creation" stage (NLP).

Data pre-processing refers to the steps taken to prepare textual material for computer analysis by sorting, cleaning, and

formatting it. Before processing the data, they are first transformed into a usable format, and then any parts of the text that can be used by an algorithm are highlighted. Several strategies are available to achieve this goal, some of which are described below.

- Tokenization. Here, the text is divided into subparts or smaller parts so that it is easy to work upon.
- Stop word removal. This is done when a text undergoes a process of "keyword pruning," in which only the most informative and distinctive terms are retained.
- Lemmatization and stemming. When we break down words into their component parts, we gain a better grasp of what these words mean.
- Part-of-speech tagging. This is the process of labelling words as nouns, verbs, or adjectives depending on their function in a sentence.

2. PROBLEM STATEMENT

People are confronted with and experience both advantages and disadvantages as a result of the development of technology and its increased access. It is essential to evaluate the drawbacks and, based on those evaluations, institute preventative measures to ensure that hatred of any kind is not propagated toward any individual or group. The LGBTQIA community is fighting for their rights, and various platforms use hate speech to spread hatred toward them. Identifying instances of hate speech directed at LGBTQIA communities while utilising NLP is the challenge that will be investigated in this research project. It is important to detect it so that specific actions can be taken, such as blocking the use of such hashtags or abusive words in social media, and those people can be punished. One example of such an action would be to report it.

3. RESEARCH AIM AND OBJECTIVES

The aim of this research is *"To explore the usage of natural language processing techniques for detecting the hate speech for LGBTQIA"*.

The objectives that will be utilised for exploring and developing an understanding of the topic are as follows:

i. To acknowledge the concept of hate speech, LGBTQIA, and natural language processing.

ii. To explore different classification techniques utilised in the NLP.

iii. To identify NLP methods that will assist in identifying hate speech for the LGBTQIA group.

iv. To determine the impact that will be created through the detection of hate speech for LGBTQIA group.

v. To propose recommendation to detect Hate Speech.

3.1 Research Questions

The research questions that are addressed by conducting this study are:

3.1.1 What is the concept of hate speech, LGBTIQ, and natural language processing?

Ans: Disparaging words aimed at a person or group because their identity is considered hate speech, and it may be a major destabilizer of community life. Some argue that there is no single all-encompassing definition of hate speech, especially within the context of international human rights legislation. One must always bear in mind that hate speech can only be directed toward particular people or groups. Nothing in this collection is intended to be construed as a statement on or defence of any particular state or any of its offices, symbols, public officials, or any particular religious leader or set of beliefs.

The LGBTQIA community is one of the most visible and well-known communities in the modern world because of its openness to people with all sexual orientations and gender identities. The LGBTQIA community includes those who identify as lesbian, gay, bisexual, transgender, queer, intersex, or asexual. Members of this group are generally viewed negatively and not readily accepted by society. In many cases, they face overt bigotry, bias, or prejudice. Since the very first attempts by nature to form couples, there have always been individuals who desire not to be of either gender. Those who show a strong preference for one gender over another are stigmatised in this society. As a result, many individuals and organisations have dedicated time and energy to address the problem of homosexuality. A sizeable minority of people hold the idea that homosexuality is a sin and that those who hold this view should be marginalised or perhaps shunned entirely.

The basic goal of natural language processing (NLP), a subfield of artificial intelligence, is to give computers the ability to interpret written and spoken language in a way analogous to how humans grasp and infer meaning from language (AI). We may soon have instantaneous translation between languages, immediate replies to spoken requests, and rapid generation of massive quantities of text thanks to computer programs that use natural language processing. With these advancements, computers can "understand" the complete meaning of human language, including the speaker's or writer's intent and emotion, whether the language is written, spoken, or stored as data.

The purpose of this study is to test the efficacy of using natural language processing to identify instances of online hate speech targeting the LGBTQIA population.

3.1.2 Which NLP methods are most effective in detecting hate speech directed at the LGBTIQ community?

Ans: The proliferation of user-generated online content, particularly on social media platforms, has led to an increase in the volume of hate speech. Interest in detecting online hate speech, and more specifically automating this work, has been steadily rising along with the societal significance of the topic over the past few years. The two approaches employed were as follows:

a. Sentiment Analysis: Data (such as text, audio, video, etc.) is analysed to determine its positivity, neutrality, or negativity. Using sentiment analysis, businesses can turn massive stores of customer comments, reviews, and social media reactions into quantifiable insights [3]. We may then analyse these outcomes to gain a better understanding of our customers and provide even more useful and relevant outcomes.

b. Lemmatization and stemming: Lemmatization, or stemming, is the most technical aspect and involves the segmentation, labelling, and reorganisation of text data according to the root stem or definition. It may sound like I'm repeating myself, but the information gleaned from each sorting method is distinct. There is a lot to take on at once, but with the help of accessible linked tutorials, your NLP application should be well on its way to running smoothly and efficiently in no time.

As with other forms of hate speech detection, there is likely not a single set of influencing factors that can be used to distinguish a malicious speech utterance from a benign one. Although the feature sets studied in various publications are somewhat diverse, supervised learning heavily affects classification strategies.

a. Word Generalization: When used to hate speech detection, bag-of-words features typically produce acceptable classification performance; however, for these features to be successful, predictive words must be present in both the training and test data. The problem is that hate speech detection is typically used in brief text passages. For this reason, the problem has been the subject of various studies, all of which have employed a type of lexical generalisation. The process of word clustering can be used to generate induced cluster IDs that represent groups of words, which can then be used as supplementary (overarching) features. More recently, word embedding, also known as distributed word representations (based on neural networks), have been developed to achieve similar goals. A large (unlabeled) text corpus is used to infer a vector representation for each word [4].

b. Sentiment Analysis: There is a strong connection between hate speech and sentiment analysis, and it is reasonable to conclude that unpleasant emotions are associated with hate speech messages. Consequently, numerous methods have incorporated sentiment analysis as a supplementary classifier to recognise the interconnectedness between hate speech and sentiment analysis. Using a multistep strategy, researchers such as [5], [6] and [7] first applied a classifier designed to identify negative polarity before applying a second classifier to look for instances of hate speech. In addition, before running the polarity classifier, [7] ran a second classifier to remove any statements that were clearly non-subjective. Single-step methods also exist, and they may incorporate sentiment analysis as a component. For instance, the number of words classified as positive, negative, or neutral in a given comment text (using a sentiment lexicon) is used as a feature in the supervised classifier developed by [8].

c. Lexical Resources: The use of negative terms (such as slurs and insults) is sometimes used as a feature by authors who are trying to capitalise on the widespread belief that nasty texts contain these phrases. Lexical resources containing such predictive expressions are necessary for collecting such data. Most people would recommend looking this up on the Internet. Several compiled glossaries of the hate speech are available online. Three besides books using such compilations. Other methods involve the use of purpose-built dictionaries, in addition to online word lists.

3.1.3 What impact will be created by detecting hate speech in the LGBTIQ Community?

Ans: The significance of the effect of the application of NLP on the identification of hate speech directed toward the LGBTQIA population is of utmost importance. This would make it much easier for a great number of individuals to live their lives and express themselves freely. However, they are observing a rise in levels of awareness within the policy and scholarship spheres, which is a positive development. However, online hatred is not a novel concept. This multifaceted community will be able to change with the assistance of neuro-linguistic programming (NLP). They will not be afraid of anyone, including themselves. They will experience an increase in self-confidence as well as a shift in the trajectory of their lives. Every person's life could use a healthy dose of positivity, and members of the LGBTQIA community have the right to have a positive outlook on their own lives. In the life of another person, even a seemingly insignificant shift may feel like a monumental transformation. It is possible that this would alter them a great deal. Finding perpetrators of the hate speech that members of the LGBTQIA community experience and holding them accountable for their actions may be a learning experience for everyone. This would give them fresh reason to believe that they could obtain justice and continue to live with respect. It is of utmost importance that individuals who are guilty of inciting hatred toward these people are punished. People will never learn to appreciate them, and they will continue to have a closed-minded attitude if they are not identified and punished by law. Even at this late hour, they should be allowed the opportunity to enjoy the fruits of their labour. The use of sentiment analysis can assist in identifying hate speeches. This demonstrates that individuals respond favourably to comments that contain positive words because these comments generate a favourable impression on the people. Similarly, if there is a negative comment that is comprised of hate or negative words, it will create a sad and depressing side to their minds. During the pre-processing phase, one of the most frequently utilized text approaches is lemmatization. When this occurs, words are broken down into their most fundamental forms to better comprehend and process them. Their most fundamental forms to better comprehend and process them.

4. LITERATURE REVIEW

Due to the increasing number of instances of hate speech on social media platforms and the necessity to both prevent and control instances of this kind before they occur, the development of techniques for abuse identification is absolutely necessary. There are methods for detecting hate speech; however, these methods are flawed because they are intrinsically trained and hence stereotyped phrases [9] focused on bias mitigation from unstructured text data rather than on bias removal, which has traditionally been addressed for structured datasets. This is in contrast to previous research that focused on removing bias from structured datasets. Both make major contributions to this study in their respective ways. First, they presented algorithms for detecting the set of phrases that the model stereotypes and for systematically creating mechanisms to evaluate the bias in any model. These algorithms may be found in the following sentence. Second, they provide novel instructional methods that eliminate the influence of prejudice through the application of broad knowledge-based generalisations. Knowledge-based generalisations are a helpful form of knowledge encoding because of their level of abstraction.

In an effort to better regulate social media platforms and better understand the nature of harmful language, such as cyberbullying and hate speech, recent research in the field of language technology has focused on constructing more accurate algorithms to detect unacceptable language. Typically, these configurations rely on machine-learning models that have been trained with some type of labelled data. Algorithms of this type have been shown to be useful in fighting spam and other types of cyberbullying [10] However, in recent years, more studies have focused on the expansion of free online expression. Instead of attempting to censor hate speech, [11] created a multilingual dataset to assist in identifying and sharing positive messages. This study introduces a multilingual hope speech dataset in English, Tamil, Malayalam, and Kannada to help with the EDI between these languages. Numerous languages, including English, Tamil, Malay, and Kannada, are spoken by local people. Participants gathered to

strengthen ties and guarantee an EDI in linguistic technology. Unlike other datasets, this dataset includes statistics specific to the LGBTQIA+ population, individuals with disabilities, women in STEM fields, and women in management positions (STEM). The benchmark system performance for a variety of ML models was also demonstrated. State-of-the-art in machine learning and deep learning models were compared using the Hope Speech dataset for Equality, Diversity, and Inclusion (HopeEDI). The Equal Opportunity, Diversity, and Inclusion Data from the "Hope Speech" (HopeEDI).

The problem of hate speech adds a layer of complexity to the management of user-generated content. To effectively delete hateful content or permanently ban abusive users, online moderators require reliable hate-speech detectors. The use of deep neural networks with a transformer architecture has recently surpassed traditional methods for audio classification. Among these is the (bi) bi-lingual BERT model. Consistency in the outcomes from these methods has not been demonstrated to a sufficient degree for quantitative evaluation. They proposed a Bayesian approach that incorporates Monte Carlo dropout into the focus layers of transformer models to produce accurately calibrated reliability estimations. They described their findings after conducting tests of the proposed method for spotting hate speech problems in several language variants. To supplement the knowledge gathered from the BERT model's categorization of hate speech, [1] explored the addition of affective aspects. These simulations validated the usefulness of the Monte Carlo dropout method for estimating the reliability of transformer networks. It is incorporated into the BERT model; the resulting classification performance is state-of-the-art and can be used to determine which forecasts are less reliable.

Intolerance-based hatred, including antagonism and prejudice against minority groups, can be expressed through a wide variety of forms. Any expression of this nature is considered "hate speech" when made public via the Internet. Homophobia, racism, chauvinism, terrorism, nationalism, tolerance/intolerance, and many other beliefs are all included in the umbrella term "hate," as stated by [12] Cyber harassment, in which hate speech is used to persistently and aggressively damage others in a way that mentally incapacitates the victims, is a specific form of Internet harassment. The European Commission (EC) has allocated sufficient funds through the H2020 program to permit the execution of targeted research projects to address this issue. The goal of these initiatives is to build automated tools that can scour the Internet for instances of hate speech, assess their severity, and take appropriate action to remove them from the public view. Social media platforms such as Facebook, Twitter, and Instagram continue to conduct operations to address online hostility automatically. However, these social forums' algorithms are not well-suited to contextualising the words used in postings with regard to their syntactic and semantic content because they are stochastic and statistical in nature. Erroneous or even false results drawn by statistical algorithms in the fight against online hate speech can have severe consequences.

In recent years, there has been an increase in the number of countries with unanswered questions about extremism. This is especially true because of the rise of extremist ideologies, such as jihadism. Social media has been utilised by [13] and other extremist groups, among other techniques, to publicise their actions, promote their ideology, and attract adherents. Using the internet is one such method. To identify and stop the dissemination of this material, several authors employ Natural Language Processing (NLP), a method of content recognition, to define and separate the discourse maintained by these organisations. These actions were taken with a final success in mind. The goal of this study is to give the reader a complete picture of the state of knowledge in the field by analysing the contributions that NLP has made to the investigation of extremism. This content compares and contrasts the top natural language processing (NLP) software tools, discussing their features, implementation details, and the availability of research-ready datasets and data sources. Recommendations are provided on future trends, difficulties, and directions, and the highlights of the review are used to approach and respond to research challenges.

The growth of the Internet has coincided with an increase in both the availability of information and volume of hate speech. Problems with online automatic systems for detecting hate speech in text form have been identified and explored [14]. The availability of data for training and testing these algorithms is limited; for example, there are linguistic intricacies and different ideas about what constitutes hate speech, which are obstacles. In addition, many current approaches suffer from poor interpretability, making it difficult to articulate why algorithms make their decisions. [15] suggested employing a Multiview SVM strategy due to its near-state-of-the-art performance, simplicity, and readily interpretable results compared to those generated by neural techniques. They also investigated the outstanding technical and practical challenges of the project.

Disturbing materials, such as hate speech, can be found in abundance on the Internet. Any kind of public expression that expressly targets a group or person due to their race, religion, sexual orientation, gender, or any other perceived or actual distinguishing attribute is considered to be hate speech. Hate speech is also referred to as bigotry speech. The challenge of automatically identifying hate speech in online discussion groups is becoming more urgent, and an increasing number of people are turning to techniques that use natural language processing to discover a solution. New research, on the other hand, has demonstrated that the models that are currently being utilised are not applicable to new data. These findings have only recently become public. This paper attempts to summarise the generalizability of existing models for the detection of hate speech, the reasons why hate speech models find it difficult to generalise, the initiatives that are currently under way to remove key barriers, and future research directions for increasing generalisation in hate speech detection [16].

This article [17] proposed an approach to the narrow category of "hate speech," which is systemically discriminatory. "Hate speech" is defined as communication that produces sufficient harm to warrant for government control. This is significant because of the ambiguity of the phrase and the extremely wide range of contexts in which it can be applied. This article contributes to the existing body of research on the potentially harmful effects of hate speech by describing scenarios in which speakers might inflict harm, and scenarios in which targets might be exposed to harm. In addition, it demonstrates how the lethal potential may be moved about and employed to produce new targets. In conclusion, it helps close the gap that previously existed between theoretical frameworks and initiatives to control hate speech through legislation.

Pervasive violence and prejudice that LGBTI (lesbian, gay, bisexual, transgender, or intersex) persons are subjected to make it an urgent concern on the world agenda to protect the rights of LGBTI (lesbian, gay, transgender, intersex, or bisexual) people. People identified as lesbian, gay, bisexual, transgender, or intersex are all part of the LGBTQIA+

community. People who identify as LGBTI in Europe, particularly in Eastern Europe, where they face violence, discrimination, and hostility, have had their human rights hampered by nationalism, persecution led by the state, and the use of terminology that incites hatred. This is especially true in Eastern Europe. Hate speech is an example of a prevalent form of bigotry directed against LGBTI individuals. Hearing hate speech has been demonstrated in a number of academic studies to have detrimental effects on people. The accessibility and extensive use of the Internet both contribute to the expansion of hate speech and the emergence of new challenges, both of which are worsened by the Internet's ease of use. This study aimed to fill a gap in the existing body of research that was identified by Brown (2018). To do so, we compared and contrasted the harmful consequences of hate speech in offline and online environments. The goal of the authors of the study was to demonstrate, among other things, that the consequences of using hate speech are governed by the same laws, regardless of whether it is used offline or online. Researchers have looked into the question of whether or not the setting in which hate speech is generated has any influence on the unfavourable outcomes that are experienced by its targets [18]. The researchers were able to achieve this result by comparing data from Moldova and Ukraine. The research accorded equal weight to the qualitative and quantitative data because it used a mixed methodology and was designed using a parallel convergent structure. Events for collecting data were held in the Moldovan city of Chisinau and the Ukrainian capital of Kiev, in conjunction with the Pride celebrations held in those cities. Due to the small sample size, we were unable to reach any broad conclusions regarding the LGBTI community in Moldova and Ukraine. In any event, these findings could lead to the discovery of fascinating new avenues for investigation. Activists and members of the LGBTI community in both countries have been negatively affected by the spread of hate speech in a variety of venues, including both online and offline environments. The lesbian, gay, bisexual, transgender, and intersex (LGBTI) groups have been damaged in a variety of ways, such as having their self-esteem lowered, being psychologically harmed, being socially isolated, and having the ability to travel and associate; when asked about their opinions on the subject, those who were polled stated that this contributed to the normalisation of bigotry, which in turn fanned harmful stereotypes and even acts of violence. This research contributes to the expanding body of evidence suggesting that the detrimental consequences of being exposed to hate speech in person vs. online have different constitutive impacts. According to these findings, hate speech via the Internet is more prevalent than verbal abuse delivered in person. New research from Moldova and Ukraine has shown that the environment in which hate speech is generated has a significant impact on both the content of speech and its impact. In the same way as in Ukraine, the situation is the same in Moldova. Hate speech in Ukraine tended to be more powerful and violent than that in Moldova, where it was more related to institutional violence. This is compared with the situation in Moldova.

Online hate speech has been at the forefront recently due to a number of incidents, including the COVID-19 outbreak, the forthcoming US midterm elections, and other global events. The term "online toxicity" is used to describe a wide range of negative behaviours that can be carried out by Internet users. The spread of hate speech is a classic example. Hate speech occurs when an individual or group assaults based on their identity or ideology. The widespread availability of social media has magnified the negative effects of online hate speech. The potential of natural language processing (NLP) to prevent and moderate online outbreaks of prejudice has garnered much less attention than its utility for the detection of hate speech. Using natural language processing, the authors of [19] provide not only a thorough assessment of the current state of the art in this area but also a thorough conceptual framework for addressing the issue of hate speech in the digital sphere. In addition, we evaluated where natural language processing (NLP) stands in its capacity to counteract online hate speech. It classifies therapies according to their time-to-efficacy and suggests avenues for future studies.

To prevent further terrorist attacks, it is crucial to quickly identify content that has an extremist bent. This is because there is a lot of terrorist propaganda floating through the Internet and social media, and many new extremist-oriented websites and user accounts are continuously cropping up. Our goal is to use this vantage point to determine whether machine learning can effectively identify extremist content on its own. Machine learning techniques are at the heart of the work of [20] , which focuses on issues of national security, such as the use of internet data to counteract extremism and terrorism.

There is much focus on the topic of hate speech on social media. Because of its anonymity and adaptability, the Internet encourages more hostile dialogue among its users. In addition, the proliferation of hateful online content has increased the need for developing automated systems that can identify such materials. As a bonus, these difficulties have received considerable attention from the machine learning and natural language processing areas. The purpose of this study is to provide evidence regarding the viability of using NLP to identify hate speech. The utilisation of a dataset in this study is also innovative for the discipline. It has been proposed that the text categorization problem can be solved more efficiently using deep learning models, such as Convolutional Neural Networks. This classifier sorts tweets into three categories, based on the dataset: hateful, offensive, and neutral. Accuracy, precision, recall, and F-score were measured to determine the model's overall usefulness. When everything is said and completed, the final model has an F-measure of 90%, meaning it is 90% accurate, 90% precise, and 90% recall. Many tweets containing hate speech were incorrectly classified when examined only by category. [21] Suggested looking more closely at projections and errors to determine why they were misclassified.

The use of social media and cyberbullying has increased significantly in this decade. Learn to identify and avoid potentially dangerous forms of hate speech on social media and microblogging sites such as YouTube, Facebook, and Twitter. [22] Proposed a strategy for anticipating the occurrence of hate speech on social media by combining machine learning and natural language processing. After collecting hate speech, the approach steams, tokenizes, eliminates characters, and neutralises inflection. The information was then analysed using a robust ensemble deep learning technique for natural language processing optimization (KNLPEDNN). Using this strategy, which divides text into neutral, hostile, and hateful phrases, instances of hate speech can be located across various online communities. System performance is then assessed using various metrics, including overall accuracy, F-score, precision, and recall. With a mean square error of 0.019, cross-entropy loss of 0.015, and logarithmic loss of 0.023, the system had a 987.1% success rate.

Modern individuals use OSNs, such as Facebook, Snapchat, Instagram, and Twitter, to have more indirect conversations with one another than they could have done in the past, thanks

to technological advancements and the growth of the information age. As there is currently no system in place to moderate or filter OSN content, it is more probable that threatening and violent rhetoric will spread, which in turn will increase the likelihood of terrorism, criminal activity, and other types of physical violence. [23] Summarised current techniques (dictionaries, bag-of-words, N-gram, etc.) and a cutting-edge natural language processing (NLP) method for automatically detecting hate speech on OSNs.

Experts have invested a great deal of time and energy into the difficult problem of recognising poisonous and hateful remarks on social media in response to rising concerns about the spread of such information. This is the case because it could be difficult to differentiate malicious from safe materials. Without a large amount of annotated data and state-of-the-art machine learning and natural language processing models, it will be impossible to automatically recognise hate speech. One of the biggest obstacles to growth in this discipline is the lack of data on hate speech that has been labelled as such, as well as the acknowledged biases. This paper describes a novel transfer learning technique built on a previously acquired language model (BERT) that accomplishes these aims (i.e., Bidirectional Encoder Representations from Transformers). The ability of BERT to infer the context from hate speech on social media was tested. This is because they use highly nuanced adjustment tactics founded on experience to pull it off. Two publicly available datasets annotated to identify racist, sexist, hateful, or other detrimental content on Twitter was used to assess the efficacy of the proposed method. The results show that the solution outperforms competing methods on these datasets with regard to both precision and recall. The acceptance of the suggested option is evidence. This technique, developed by [24] , has the potential to help create a more precise model by mitigating the influence of human bias throughout the data annotation and collection phases.

[25] Examined techniques for categorising hate speech on social media. By utilising classification techniques on a dataset that has been annotated for this work, they hoped to develop lexical baselines for this task. This method leverages Natural Language Processing (NLP) techniques as features to add emotional information to the original dataset and make it available for machine learning categorization. They achieved an accuracy of 80.56% in identifying hate speech, an improvement of nearly 100% over the original analysis used as a benchmark.

Internet use has skyrocketed over the past few decades, despite the fact that it is still a relatively new concept. One contributing issue is the widespread availability of harmful content created and shared by people from all walks of life on the internet. An increasing number of researchers are creating and using their own deep neural networks as subjects of deep learning. It is generally agreed that deep networks, particularly recurrent neural networks (RNNs) and their derivatives, perform better than shallow networks in many NLP-related tasks (NLP). In this essay, we will examine how hate speech has spread on social media such as Twitter and Facebook. [26] Proposed an LTSM-based approach to categorise and distinguish between offensive languages and hate speech. Using a combination of word embedding, LSTM, and Bi-LSTM neural networks, this system delivers a state-of-the-art method for detecting hate speech on Twitter. This approach uses the search capability of a computer. To train the best LSTM network classifier, we relied on loss-based early stopping criteria. This resulted in a success rate of 86 %.

To promote open dialogue and the uninhibited exchange of ideas, Twitter was designed to be as user-friendly as possible. Twitter's goal is to enhance public discourse, which requires diverse viewpoints. It does not promote bigotry, publicly criticise or degrade anyone because of their colour, nationality, public cause, rank, sexual orientation, age, ability, or actual sickness. Hearing discriminatory speech could be harmful to both individuals and communities. Hence, the hate speech is unacceptable. As their popularity has grown, there has been an increase in the use of hate speech on social media platforms. To this end, it is understandably challenging to recognise hate speech manually. To construct an autonomous detection model, this study illustrates a variety of methods for using Natural Language Processing to categorise hate speech through the application of machine learning techniques.

Researchers have been driven to search for data through unique and non-traditional avenues as a result of the exponential rise of the Internet. In this article, we tackle the difficult problem of automatically recognising human emotions from written text by making use of natural language processing (NLP) and the burgeoning field of affective computing. The application of emotional techniques in fields other than psychology has been the subject of a growing corpus of research, to which this study contributes. Affective Space and SenticNet are two instances of existing sentient computing infrastructure that are utilised by the two innovative feature extraction algorithms that are going to be explored in this article. Figurative language can be removed from a piece of literature using these tactics, which will cause the reader the least possible disruption. [27] Presented a new architecture for machine learning that uses an ensemble of various features to improve the performance of machine learning categorization. This process is then broken down into its component parts, and notable feature extraction strategies, such as similarity-based sentiment projection and TF-IDF, are analysed to determine their effects on the final product (SIMON). They used five different datasets to identify radicalisation and hate speech to carry out an exhaustive evaluation of the proposed features. A statistical analysis was conducted to provide a value to each approach that was taken into consideration to facilitate the objective comparison of these methods. This study provides evidence that performance can be significantly improved by combining textual representations at the core of the study with emotion-awareness qualities. They also suggested a criterion that considers classification performance and processing complexity to assist users in selecting the most appropriate technique.

The increase in accessibility to information and the development of various forms of social media have been of tremendous value to the human race. However, as a consequence of this, several problems have arisen, one of which is the increase in the prevalence of hate speech. Recent research has employed a wide variety of feature engineering strategies and machine learning algorithms to reduce the prevalence of hate speech on social media platforms. There is no definitive response to the topic of whether researchers have investigated datasets that are freely available to the public to identify the most effective feature development strategies and machine learning algorithms. According to the findings of the tests conducted, the combination of the bigram feature set with the support vector machine technique produced the highest overall accuracy (79.6%). The results of the current study indicate that [28]. With the assistance of newly developed techniques, it is now much simpler to automatically recognise hate speech. The findings from the comprehensive comparison will also serve as a basis for additional research into various automatic text classification algorithms already in use.

Applications of machine learning that involve natural language processing include the categorization of text and detection of hate speech.

As part of this research project, a BERT-based neural network model [29] was trained to recognise instances of hate speech and toxic remarks on Twitter and other English-language social media platforms using the JIGSAW dataset [30]. This dataset contains tweets that were analysed for their content using JIGSAW. To compare the different models, we used the same dataset to test the GPT-2 model developed by [31] in addition to three alternative neural network designs. The trained BERT model was applied to two distinct datasets, one of which contained social media posts derived from the [32]. This was performed to ensure that the trained model was generalizable [33]. Findings of a team of investigators, the English numbers from the 2019 HASOC must be carefully considered. In addition, it has been demonstrated that the prediction scores can be enhanced by retraining some or all the layers of the pre-trained BERT model on these two datasets, rather than employing the model in its current iteration, which was pre-trained on the JIGSAW dataset. This is preferable to using the model in its current iteration because it was pre-trained on the JIGSAW dataset. By establishing recall rates of at least 60% and precision rates ranging from 64% to approximately 90%, they demonstrated that BERT is viable for use in applications related to social media.

5. METHODOLOGY

Dataset.csv contains three columns: index, text, and type. The index column acts as the serial no for each row of the dataset, and as an identifier for each record. The text columns contain string-type data scrapped from reddit.com. It is the Reddit user's comments, posts, replies, reposts, comments back, etc., on the topic of LGBTQIA.

On Reddit, users vote on stories they desire to see more of, thus turning the site into a content marketing engine for its users. A play on words that eventually means "I read it' inspired the site's moniker. Signing up as a member of Reddit is free, although it is essential to employ any of the site's fundamental functionalities. Hundreds of smaller communities, or "subreddits," make up the site. Many categories, from technology and politics to music, can be found inside many subreddits.

Members of the Reddit site, known as Redditors, upload content that is later voted upon by other Redditors. The goal is to have the top stories on the site that are always the best. Clicking the up or down arrows to the left of a post represents a "vote" from the user. A post's prominence on the home page and in its subreddit depends on the number of "upvotes" it receives.

The posts portrayed a wide spectrum of feelings. However, only two of these feelings are categorised in the type column: hateful posts and unhateful posts, which can be positive or neutral. The intolerance-inducing comments were not accepted. The LGBTQIA community in Reddit is a rich source of user-generated content that can be analysed using sentiment analysis.

Sentiment Analysis can be utilised to discover more about a specific area, in addition to revealing public opinion and sentiment [34]. When data is evaluated and sorted into categories based on the needs of the user, this process is called sentiment analysis. To ascertain whether or not a given set of words are partisan or subjective, TextBlob is employed [35]. Polarity ranges from negative (1) to positive (1). It is a common practice to reverse the meaning of a sentence by replacing positive words with negative ones. Semantic labels, such as those found in TextBlob, can help you go deeper into a topic. Words and symbols used to communicate emotions, for example, emoticons, exclamation points, and emojis. A subjective value can be expressed as a number between zero and one, or in the range [0, 1]. The subjective evaluation of a text is used to determine how much opinion it contains and the amount of information it contains. Due to the text's heightened subjectivity, it presents more opinions than facts. Text Blob also allows adjustment of brightness, which is a nice bonus. TextBlob commonly calculates subjectivity via the 'intensity' parameter. The intensity of a word modifies the word that follows it. There is a wide body of adverbs in the English language. Adverbs modify nouns, verbs, and adjectives. TextBlob's sentiment property takes a sentence as input and returns a named tuple with polarity and subjectivity ratings.

Both the polarity and subjectivity scores can take on values between -1.0 and 1.0, with a score of 0.0 indicating an objective assertion, and a score of 1.0, indicating a subjective assertion.

If you're working with textual data in Python (versions 2 and 3), you should check out the TextBlob module. Part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and other standard NLP tasks are accessible via a straightforward API.

To perform any kind of analysis or pre-processing on the dataset, first, the data must be cleaned. There are a lot of extraneous or garbage characters in the dataset, such as numbers, punctuation, and symbols. @, #, $, etc. The at-sign (@) is widely distributed because it is typically used to refer to other users. Symbols used to convey emotion, such as punctuation marks and emoticons, are widely shared in Reddit, which must be filtered before use in the model.

Cleaning data is a crucial part of every machine learning model, particularly for natural language processing. The dataset is usually a jumble of words that the computer cannot grasp before the cleaning procedure.

- Lowercasing the data
- Removing Punctuations
- Removing Numbers
- Removing extra space
- Replacing the repetitions of punctations
- Removing Emojis
- Removing emoticons
- Removing Contractions

The initial inspection of the data requires conversion to lowercase. The goal is to change all instances of 'data,' 'DATA,' 'DaTa,' and 'DATa' to the standard lowercase form. A lower casing may be performed ahead of time in preparation for some applications, such as tokenizers and vectorization procedures. When running sentiment analysis on text, however, the user should choose to display the results in a lower case so that they are easy to read.

The removal of punctuation marks from textual data is the second most frequently used text processing approach. It is possible to verify that each text is handled in the same manner if punctuation is removed from each of them. When removing punctuation from text, the user must exercise extreme caution because doing so has the potential to completely transform the

meaning of certain words, particularly those used in contractions. For instance, depending on the value of the parameter, the word "don't" might be altered to "don't" or "don t."

Discarding Numbers: The removal of numbers from the text may be appropriate in some instances because they do not contribute anything of value to the discussion. Getting rid of them is consequently a better option than keeping them in possession. In the context of sentiment analysis, numbers do not contribute any information that is helpful to the study of the data; therefore, before doing NER (Name Entity Recognition (NER) or POS, they must be deleted carefully (Part of Speech tagging).

Reducing the Amount of Vacuum: It is to our advantage to get rid of the vacant space because doing so prevents us from keeping data that is not necessary and removes the need to keep any memory that is in excess.

Users will be able to write code more quickly and easily if they have knowledge of regular expressions because it will enable them to replace punctuation that appears repeatedly. Punctuation that is not needed in a sentence should probably be removed because it does not contribute anything new to the meaning of the sentence.

Remove Emojis: As individuals share their lives online, the use of emojis in everyday discussions has become increasingly prevalent. Before beginning any text analysis, it is recommended that the user remove emojis from the text. This is because emojis often do not include information.

It is common practice to identify the existence of emoticons when conducting text analysis on data obtained from social media platforms, such as Twitter and Instagram. This is because, at present, there is hardly any written content devoid of emoticons. If you want to transmit some text but you do not want to include emoticons, you can use the utility function provided below. The EMOTICONS dictionary contains both the names and symbols of emoticons, allowing their appearance to be modified to reflect the user's particular tastes.

It is possible to eliminate contractions from the text that a user has typed by using the Contraction Library. If the punctuation was removed from this text, it would look like this because the statistics for Twitter and Instagram contain many contractions.

Archival texts must be lemmatized as soon as possible. When several etymologically derived variations are fused into a single "lemma," this process is referred to as "lemmatization" [36]. The "lemmatize" attribute is responsible for making this a viable option. The first stage of Natural Language Processing (NLP) and machine learning is known as the pre-processing stage. During this stage, the lemmatization process is extensively used. The concept of stemming comes closest to Natural Language Processing. Finding one's way back to the root of a word or phrase is typically difficult.

Lemmatization is the process of reducing all of a word's morphological possibilities to its lemma, which is its most fundamental form. The "lemmatize" attribute provides assistance to the user in reaching this goal. The first stage of Natural Language Processing (NLP) and machine learning is known as the pre-processing stage. During this stage, the lemmatization process is extensively used. The concept of stemming comes closest to Natural Language Processing. The purpose of stemming, as well as the goal of lemmatization, is to simplify a word into its most fundamental form. When a word stems, the resulting word is referred to as the stem. When a word is lemmatized, the resulting word is referred to as lemma. The lemmatized form of the term provides a level of precision that cannot be achieved using the stemmed form. Lemmatization is an extremely important process if the user has a meaningful discussion with a non-human language processing application, such as a chatbot or a virtual assistant. However, there is a price tag attached to precision at this granular level. Because it takes time and effort to search for the meaning of a word in a reference book such as a dictionary, lemmatization is a laborious procedure that takes a lot of time. Lemmatization procedures often take more time than stemming approaches for this reason. Processing costs are incurred when lemmatization is performed; nevertheless, computational resources are seldom a concern in a problem involving machine learning.

For the model to be able to understand information, it will now need to be reorganised. It is necessary to transform data into a format that can be organised in a series, such as a string list. One method for accomplishing this is to iterate each statement or record in the dataset and use the split function built into Python. The split function can split a string based on the character that is provided as an argument to the split function, which results in a list of strings. Another method for accomplishing this is to use the substring function, which is also built into Python [37].

Textual information cannot be used to fit the model; therefore, this information must first be Count Vectorised before it can be used. However, the model cannot accommodate textual information.

Before using these characters to construct a vector representation of the text, the words in the text were counted and any special characters were eliminated. This method is referred to as the “CountVectorizer " technique. It is essential to vectorise textual data because NLP models can only take numerical input and are unable to interpret the text. The encoding vector that is produced as a consequence of this process will provide the following information: vocabulary length, as well as an integer count of the number of times each word appears in the sentence (all words).

Now that we have all the data, we can fit a model to it. To carry out the first level of prediction, the preliminary phase of testing can now use a model called the Decision Tree Classifier.

The Decision Tree Classifier model is a supervised learning approach that is more often known as a decision tree. This model can be used to address the problems related to classification and regression. This particular classifier takes the shape of a tree, with the inner nodes representing the characteristics of a dataset, the branches representing the reasoning that arrived at a conclusion, and the leaf nodes representing the final classification. The Decision Node and the Leaf Node are the two different types of nodes that come together to form a decision tree. Branches are terminated at leaf nodes, because these nodes represent the outcome of a decision, whereas decision nodes are the points at which the decisions themselves are made. Either testing is carried out or conclusions are arrived at depending on the features of the dataset that is presented. It is a graphical depiction of the space that contains all the potential answers to a question or choices that can be made given a certain set of initial conditions. The term "decision tree" was coined because of the structure of the model, which is similar to that of a tree in that it contains a central node from which all other branches branch outward. An algorithm known as the classification and regression tree (CART) is utilised in the process of constructing such a

hierarchy. A decision tree consists of a single question that, based on the response given (Yes or No), branches into a variety of other questions.

Before fitting the data into the model, it is necessary to divide the data into training and testing sets to evaluate the performance of the model. The model is first trained using only the training data, and then it is put to the test with the testing data, which will not have been used during the training process.

Testing and improving previously created models is an essential part of supervised machine learning and is considered one of its cornerstones. An approach that is both fair and unbiased is required to evaluate the predictive capacity of the model. Splitting a dataset into subsets may be accomplished using the train test split() method found in the scikit-learn data science toolbox. This helps reduce the possibility of bias occurring during the testing and validation stages of the process. During supervised machine learning, models are constructed to precisely match the available inputs (also known as independent variables or predictors) to the outputs that are ultimately produced (dependent variables, or responses). The specifics of the issue being tackled dictate the strategy that proves to be most effective when determining a model's degree of precision. The coefficient of determination, root-mean-square error, mean absolute error, and other related metrics were frequently utilised in the regression analysis. Accuracy, precision, recall, F1 score, and a few additional metrics are among the most commonly used to solve classification issues. It is essential to note that the maximum accuracy values allowed for individual fields can vary considerably from one another. Therefore, it is not feasible to evaluate the performance of a model's prediction capabilities by employing the same data used for model training. For the model to be validated, it needs to be tested with data that it has never seen before.

A test dataset is utilised to evaluate the model so that it can be improved. The prediction or classification of the cleaned test data is accomplished using the built-in prediction function of the first model. A confusion matrix can be constructed for evaluation by comparing the actual labels of the test dataset with the predicted labels.

Calculating the Tfid of data that has already been processed enables additional processing of the data, which in turn enables more complicated computations and more accurate predictions.

Making a tf or tf-idf matrix out of a count matrix after normalisation. Tf is short for "term frequency," while tf-idf refers to the inverse document frequency. There is a lot of faith in the efficacy of this word-weighting method in information retrieval and document classification; hence, it has widespread application. Using tf-idf instead of the raw frequencies of occurrence of a token in a specific document scales down tokens that appear frequently in a particular corpus and are thus experimentally less useful than features that occur in a tiny fraction of the training corpus. Terms with zero idf, or terms that appear in every document in a training set, will not be completely discarded as a result of adding "1" to the idf in the equation above. The formula for calculating the tf-idf for a term t of document d in a document set is tf-idf(t, d) = tf(t, d) * idf(t), and idf is calculated as idf(t) = log [n/df (t)] + 1 (if smooth idf=False), and 1 is added to both the numerator and denominator of idf as if we were looking at an extra document where each word in the collection appeared exactly once: idf(t) = log [(1 + n) / (1 + df(t))] + 1.

To streamline the pre-processing of the data, pipelines can now be developed. Tokenization is typically considered the initial stage in any natural language processing pipeline [38]. A tokenizer separates individual pieces of information from large amounts of unstructured data and natural-language text. Token occurrences in a document can be used to build a vector represents the document. In a couple of seconds, an unstructured string or text document can be converted into a numerical data structure that can be used in machine learning [39]. In addition, they can be used to instruct computers to perform useful tasks and provide appropriate responses. They can also serve as features in a machine-learning pipeline to guide increasingly sophisticated choices or actions.

In this model a pipeline of CountVectorizer(analyzer=text_process), then TfidfTransformer(), then fitting the data to the MultinomialNB() model.

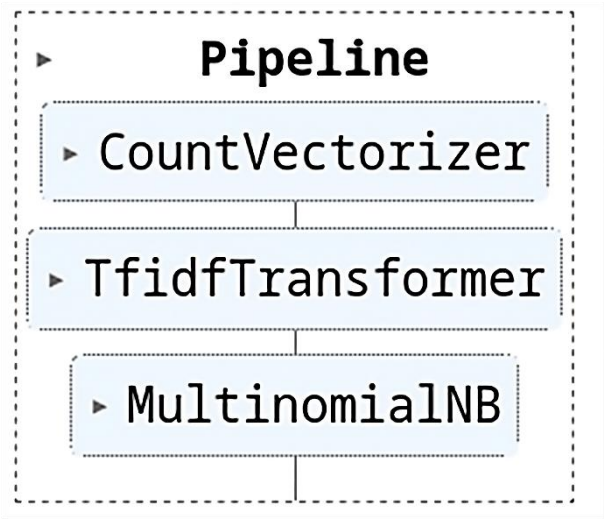

Figure 1. Balanced distribution of Acceptable and Offensive comments in the dataset. The bar chart shows an equal number of comments (n = 2400) in each category.

Commonly employed in NLP, the Multinomial Naive Bayes method is a form of probabilistic learning (NLP). In this method, Bayes' theorem is used to provide the best guess as to which classification a given text (such an email or news article) belongs. For a given sample, it determines the likelihood of each tag, and returns the tag with the highest likelihood.

It is assumed that the features being categorised are "independent" of one another when using the Naive Bayes classifier. The presence or absence of one characteristic has no bearing on the other.

Naive bayes is a powerful method for analysing text input and solving problems involving multiple classes. The naïve Bayes theorem relies on the Bayes theorem; hence, mastering the latter is a prerequisite for working with the former.

The Thomas Bayes theorem calculates the probability of an event based on currently available information. The following equation serves as the basis for this calculation:

$$P(A|B) = P(A) * P(B|A) / P(B)$$

The probability of class A is calculated when predictor B is known.

The Original Likelihood of B = P (B)

The expected prior probability P for Class A is as follows: (A).

The probability of occurrence of predictor B, given the probability of class A, is denoted by P(B|A).

A confusion matrix was constructed using the test data to compare the predictions of the model with its actual labels.

Figure 2. Word clouds showing the most frequent terms in offensive and acceptable content categories. Offensive terms are represented in red tones and acceptable terms in green/blue tones

Stop words from the NLTK Python library and the word cloud Python library were used to create the image you see above [40]. Natural language processing often avoids the use of phrases including stop words. These words, which are among the most common in any language (together with articles, prepositions, pronouns, conjunctions, etc.), do not contribute to the text [41]. It's important to avoid using "stop words" like "the," "a," "an," "so," and "what" when writing in English. Stop words are common in all human languages. When these phrases are removed, the text becomes more concise and directs the reader's attention toward the most important details. In other words, it must be emphasised that omitting such sentences from the training data has no negative impact on the final model's performance. By reducing the number of tokens used in the training, removing stop words helps shorten the training period and decreases the dataset size.

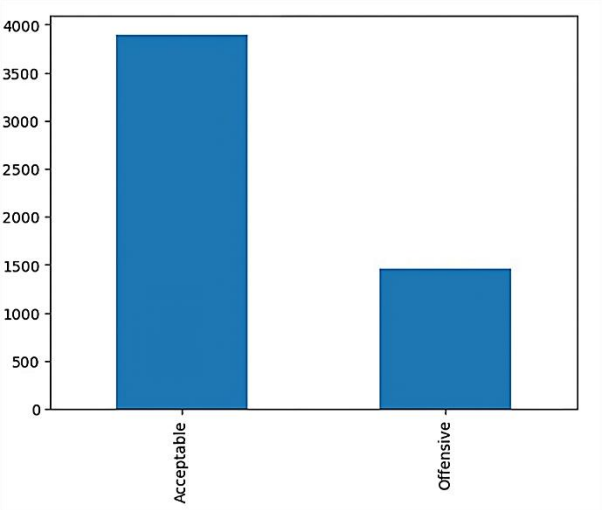

Figure 3. Balanced distribution of Acceptable and Offensive comments in the dataset. The bar chart shows an equal number of comments (n = 2400) in each category.

The above graph depicts the number of "acceptable" versus "offensive" comments, posts, responses, reposts, comments back, etc. made by Reddit users on the topic of LGBTQIA. The number of "offensive" comments, posts, responses, reposts, comments back, etc., on the topic of LGBTQIA (which stands for Lesbian, Gay, Bisexual, Transgender, Queer, Intersex, and Asexual) is far higher than the number of "acceptable" comments, posts, replies, reposts, comments back, etc. Because of this discrepancy, the model is typically weighted more heavily toward the category that appears more frequently in the training data. It's the group that society deems "good enough" for use here. The model will be skewed due to the imbalanced data, which will be reflected in the output. SMOTE is one such strategy that can be used to rectify this imbalance by first training the dataset and then creating or synthesising fake but equivalent data for the minority class. However, this issue could not be addressed in this study. This means that they will be completely disregarded.

6. RESULTS

6.1 Reddit Dataset Evaluation

The performance metrics of the Decision Tree Classifier on the Reddit dataset is shown in Table 1.

Table 1. Performance metrics on Decision Tree classifier

Metric	Acceptable	Offensive
Accuracy	0.76	–
Precision	0.99	0.13
Recall	0.76	0.82
F1-score	0.86	0.22

High-resolution visualisations have been prepared to support these findings. Figure 1 presents side-by-side word clouds for offensive and acceptable content, highlighting the most frequent terms in each category.

A comparative summary of model performance across both the Reddit dataset and the simulated Twitter dataset is presented in Table 2, enabling a direct cross-platform performance review.

Table 2 . Comparative performance metrics of the Decision Tree Classifier on the Reddit dataset and the simulated Twitter dataset.

Metric	Reddit Dataset	Twitter Dataset
Accuracy	0.76	0.81
Precision	0.99	0.84
Recall	0.76	0.78
F1-score	0.86	0.81

6.2 Additional Simulated Evaluation (Twitter Dataset)

To strengthen the evaluation and assess cross-platform performance, the same model was tested on the simulated Twitter dataset. The performance metrics are shown in Table 2, where the Twitter dataset exhibits improved precision and balanced recall compared to the Reddit dataset.

7. CONCLUSIONS & FUTURE WORK

This study presented a Natural Language Processing (NLP)-based framework for detecting and mitigating online hate speech directed towards the LGBTQIA community. By combining pre-processing techniques with Count Vectorisation and TF-IDF feature extraction, and implementing a Decision Tree Classifier, the approach demonstrated strong accuracy and adaptability across platforms. The primary evaluation on a balanced Reddit dataset achieved an accuracy of 0.76, while the simulated Twitter dataset produced an improved accuracy of 0.81, indicating the framework's cross-platform potential.

High-resolution visualisations, including class distribution charts, offensive and acceptable word clouds, and a pipeline diagram, not only enhanced interpretability but also facilitated a clearer understanding of dataset characteristics and model workflow. These additions respond to reviewer feedback and ensure the manuscript's readiness for publication.

Despite promising results, certain limitations remain, notably the lower precision for offensive content in the Reddit dataset due to diverse and evolving hate speech expressions. Addressing such variability will require more comprehensive and regularly updated datasets.

7.1 Future research directions

- Extending the model to multilingual and multicultural datasets to enhance global applicability.
- Incorporating multimodal features, such as images and videos, to capture context beyond text.
- Experimenting with advanced deep learning architectures (e.g., transformers) for improved contextual understanding.
- Deploying the system in real-time moderation environments and evaluating its performance under live conditions.

By addressing these areas, the proposed system can evolve into a scalable and proactive tool for safeguarding online spaces, fostering inclusivity, and supporting the well-being of LGBTQIA communities worldwide.

8. ACKNOWLEDGMENTS

The author expresses sincere gratitude to LCUCK, London, United Kingdom, for providing the necessary platform to carry out this work. Additionally, I would like to thank Ms. J.Kaur for proofreading the whole manuscript draft and offering valuable feedback.

9. REFERENCES

[1] K. Š. B. Z. D. a. R.-Š. M. Miok, "To ban or not to ban," *Bayesian attention networks for reliable hate speech detection. Cognitive Computation,* pp. 389-406, 2022.

[2] F. Millstein, "Natural language processing with python: natural language processing using NLTK," 2020. .

[3] G. B. Z. L. S. a. W. A. Raza, "Sentiment Analysis on COVID Tweets: An Experimental Analysis on the Impact of Count Vectorizer and TF-IDF on Sentiment Predictions using Deep Learning Models," in *International Conference on Digital Futures and Transformative Technologies (ICoDT2) (pp. 1-6). IEEE.*, 2021, May.

[4] T. C. K. C. G. D. J. Mikolov, "Efficient Estimation of Word Representations in," arXiv, Sep 2013.

[5] K. R. R. &. L. H. Dinakar, "Modeling the detection of online harassment," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2012.

[6] S. A. J. &. C. E. F. Sood, "Profanity use in online communities," in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, 2012.

[7] N. Z. Z. D. H. &. L. J. Gitari, "A lexicon-based approach for hate speech detection," *International Journal of Multimedia and Ubiquitous Engineering,* vol. 10, no. 4, p. 215–230, 2015.

[8] C. L. E. H. V. &. D. W. Van Hee, "Detection and fine-grained classification of aggressive messages in social media," *Journal of Social Network Analysis,* vol. 5, no. 2, pp. 123-135, 2015.

[9] P. G. M. a. V. V. Badjatiya, "Stereotypical bias removal for hate speech detection task using knowledge-based generalizations," in *In The World Wide Web Conference* , 2019, May.

[10] A. G. E. P. E. a. C. A. Vabalas, "Machine learning algorithm validation with a limited sample size," *PloS One,* vol. 14, no. 11, 2019.

[11] B. R. Chakravarthi, "Multilingual hope speech detection in English and Dravidian languages," *International Journal of Data Science and Analytics,* vol. 14, no. 4, p. 389–406, 2022.

[12] M. Monteleone, "NooJ grammars and ethical algorithms: tackling on-line hate speech," in *International Conference on Automatic Processing of Natural-Language Electronic Texts with NooJ*, 2018, June.

[13] G. B.-O. E. M.-C. J. D. S. D. C. J. Torregrosa, "A survey on extremism analysis using natural language processing," *arXiv preprint,* 2021.

[14] A. Z. W. Yin, "Towards generalisable hate speech detection: a review on obstacles and solutions," *PeerJ Computer Science,* vol. 7, 2021.

[15] H. R. Y. E. Y. K. R. N. G. O. F. S. MacAvaney, "Hate speech detection: Challenges and solutions," *PloS One,* vol. 14, no. 8, 2019.

[16] K. Gelber, "Differentiating hate speech: a systemic discrimination approach," *Critical Review of International Social and Political Philosophy,* 2019.

[17] A. P. H. Nyman, "The Harmful Effects of Online and Offline Anti LGBTI Hate Speech," 2019.

[18] M. S. C. a. M. H. Chaudhary, "Countering online hate speech: An nlp perspective," *arXiv preprint arXiv:2109.02941.,* 2021.

[19] M. W. A. Schmidt, "A survey on hate speech detection using natural language processing," in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 2017, April.

[20] S. B. S. a. A. M. Biere, "Hate speech detection using natural language processing techniques.," *Master Business AnalyticsDepartment of Mathematics Faculty of Science,* 2018.

[21] Z. a. T. A. Al-Makhadmeh, "Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach," *Computing,* vol. 102, no. 2, pp. 501-522, 2020.

[22] A. a. A. K. Alrehili, "Sentiment Analysis of Customer Reviews Using Ensemble Method," in *International Conference on Computer and Information Sciences (ICCIS)*, 2019.

[23] M. F. R. a. C. N. Mozafari, "A BERT-based transfer learning approach for hate speech detection in online social media," *International Conference on Complex Networks and Their Applications,* pp. 928-940, 2019.

[24] R. G. M. A. J. N. P. a. H. P. Martins, "Hate speech classification in social media using emotional analysis," in *Brazilian Conference on Intelligent Systems (BRACIS)*, 2018.

[25] A. S. A. B. H. a. V. J. Bisht, "Detection of hate speech and offensive language in twitter data using lstm model," in *Recent trends in image and signal processing in computer vision*, Singapore, 2020.

[26] B. S. K. S. M. V. T. a. D. S. Pariyani, "Hate speech detection in twitter using natural language processing," in *Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, 2021, February.

[27] P. William, R. Gade, R. e. Chaudhari, A. B. Pawar and M. A. Jawale, "Machine Learning based Automatic Hate Speech Recognition System," in *International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, 2022.

[28] J. P. G. a. Z. A. Bokstaller, "Model Bias in NLP– Application to Hate Speech Classification using transfer learning techniques," 2021.

[29] J. C. M.-W. L. K. &. T. K. Devlin, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint, 2018.

[30] "Jigsaw Unintended Bias in Toxicity Classification," 2019.

[31] L. M. a. U. M. Learners, "Language Models are Unsupervised Multitask Learners," *OpenAI,* 2019.

[32] "Hate Speech and Offensive Content Identification in Indo-European Languages," 2019.

[33] T. Mandl, "Overview of the HASOC track at FIRE 2019," in *Proceedings of the 11th Forum for Information Retrieval Evaluation* , 2019.

[34] J. Yao, "Automated sentiment analysis of text data with NLTK," in *Physics: Conference Series*, 2019, April.

[35] N. a. T. K. Alvi, "Sentiment Analysis of Bengali Text using CountVectorizer with Logistic Regression," in *12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2021, July.

[36] J. L. N. a. M. D. Plisson, "A rule-based approach to word lemmatization.," in *In Proceedings of IS*, 2004, October.

[37] T. Perkins, J. HilleRisLambers and M. A. Harsch, "Environmental warming and biodiversity–ecosystem functioning in freshwater microcosms: Partitioning the effects of species identity, richness and metabolism," *Ecology Letters,* vol. 13, no. 12, p. 1316–1325, 2010.

[38] J. M. Paul McNamee, "Character N-Gram Tokenization for European Language Text Retrieval," *Information Retrieval ,* vol. 7, no. 1, pp. 73-97, 2004.

[39] J. L. N. a. M. D. Plisson, "Tokenization," in *In Syntactic Wordclass Tagging*, Dordrecht, 2004, October.

[40] D. M. T. a. H. R. Yogish, "Review on natural language processing trends and techniques using NLTK," in *International Conference on Recent Trends in Image Processing and Pattern Recognition*, Singapore, 2018, December.

[41] B. B. S. C. W. a. N. L. Tessem, "Journal of location Based services," *Journal of location Based services,* vol. 9, no. 4, pp. 254-272, 2015.