UNRAVELING LINGUISTIC CHALLENGES: EXPLORING THE ROLE OF ARTIFICIAL INTELLIGENCE IN ONLINE HATE SPEECH AND HARASSMENT

BY

SHAISTA NOOR

NATIONAL UNIVERSITY OF MODERN LANGUAGES

ISLAMABAD

October, 2025

UNRAVELING LINGUISTIC CHALLENGES: EXPLORING THE ROLE OF ARTIFICIAL INTELLIGENCE IN ONLINE HATE SPEECH AND HARASSMENT

By

SHAISTA NOOR

Bachelor in English, Fatima Jinnah Women University (2021)

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF PHILOSOPHY
In **English**

To

FACULTY OF ARTS & HUMANITIES

NATIONAL UNIVERSITY OF MODERN LANGUAGES, ISLAMABAD

FACULTY OF ARTS & HUMANITIES
NATIONAL UNIVERSITY OF MODERN LANGUAGES

THESIS AND DEFENCE APPROVAL FORM

The undersigned certify that they have read the following thesis, examined the defense, are satisfied with the overall exam performance, and recommend the thesis to the Faculty of Arts & Humanities for acceptance.

Thesis Title: Unraveling Linguistic Challenges: Exploring the Role of Artificial Intelligence in Online Hate Speech and Harassment

Submitted By: Shaista Noor **Registration # :**214 MPhil/Eng/Ling/S22

Dr. Farheen Ahmed Hashmi
Name of Supervisor

____
Signature of Supervisor

Dr. Farheen Ahmed Hashmi
Name of Head (GS)

____
Signature of Head (GS)

Dr. Arshad Mahmood
Name of Dean (FAH)

____
Signature of Dean (FAH)

____
Date

AUTHOR'S DECLARATION

I, Shaista Noor

Daughter of Noor-Ud-Din

Registration # 214 MPhil/Eng/Ling/S22 Discipline English Linguistics

Candidate of **Master of Philosophy** at the National University of Modern Languages do hereby declare that the thesis **Unraveling Linguistic Challenges: Exploring the Role of Artificial Intelligence in Online Hate Speech and Harassment** submitted by me in partial fulfillment of MPhil degree, is my original work, and has not been submitted or published earlier. I also solemnly declare that it shall not, in future, be submitted by me for obtaining any other degree from this or any other university or institution.

I also understand that if evidence of plagiarism is found in my thesis/dissertation at any stage, even after the award of a degree, the work may be cancelled and the degree revoked.

____
Signature of Candidate

____
Name of Candidate

____
Date

ABSTRACT

Title: Unraveling Linguistic Challenges: Exploring the Role of Artificial Intelligence in Online Hate Speech and Harassment

The present research explores the role of Artificial Intelligence (AI) in the detection of online hate speech and the linguistic challenges encountered during the process. Grounded in Socio-Technical Systems Theory (STS) and Discourse Ethics Theory**,** the study investigates the linguistic challenges and ethical issues encountered by AI systems in identifying hate speech across diverse linguistic and cultural contexts. The research employed a mixed-method approach combining both quantitative and qualitative analyses. For the quantitative phase, the data was collected from online available datasets on websites such as Kaggle and Google Data Search. The analysis provided linguistic features and patterns of online hate speech on online platforms. It revealed that Twitter is the most widely used online platform for the spread of hate speech. Moreover, the analysis measured the frequency and percentage distribution of hate speech and confirmed that political hate speech is the most prevalent, followed by racism and religious hate speech. For the qualitative phase, interviews were conducted with 10 AI experts, working in different institutions. The interviews revealed several linguistic and ethical challenges faced by the AI models while detecting online hate speech. Some of these include the complexity of hate speech, lack of diversity of datasets on which the models are trained and the lack of contextual understanding. The present research contributes to the field of linguistics by advocating ethical AI systems and providing future recommendations for researchers and stakeholders. The findings underscore the significance of AI collaboration in ensuring transparency, and in tackling the evolving and complex nature of online hate speech. By analyzing the linguistic and ethical challenges, the research paves the way for more inclusive and effective AI systems, ultimately contributing to equitable and safer online environments.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

First and foremost, I am deeply thankful to Allah Almighty, who granted me the strength, patience and perseverance to complete this journey.

I am profoundly grateful to Prof. Dr. Arshad Mahmood, Dean Faculty of Languages, and Dr. Farheen Ahmed Hashmi, Head, Department of English (Graduate Studies), for their cooperation in the entire process.

I would also like to express my deepest thanks and appreciation to my supervisor Dr. Farheen Ahmed Hashmi who continuously guided me throughout this process. Her insightful feedback and expertise played a pivotal role in completion of this thesis. I owe immense gratitude to my family, especially my parents for their unwavering love, prayers and encouragement. Their support and encouragement have been my greatest strength. I would also like to thank my dear friends Hafsa Mateen and Rabia Saadullah who always encouraged me and helped me during this process. Their in depth discussions and feedback enriched my thesis. I also extend my thanks to the English department of NUML for creating an environment conductive to learning and growth.

I offer my sincere gratitude to everyone who supported me throughout this journey.

DEDICATION

This thesis is dedicated to my parents, especially to my dearest mother whose love, prayers, endless support and encouragement have been the foundation of this journey. Your faith in me has been my greatest strength.

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

The proliferation of online hate speech has become a pressing concern in recent years, with significant implications for individual and societal well-being. As the digital age continues to shape our social interactions and modes of communication, the need to develop effective and scalable solutions for the automatic detection of hate speech has become increasingly urgent. Researchers have explored various approaches to this challenge, including the use of machine learning algorithms and natural language processing techniques to identify and categorize hateful content (Asogwa et al., 2022). However, the inherent complexities of language, the context-dependent nature of hate speech, and the proliferation of novel forms of online expression pose significant barriers to the development of robust and generalizable hate speech detection systems (Anjum & Katarya, 2022).

The definitions of online hate speech are neither universally accepted nor are individual facets of the definition fully agreed upon. Ross, et al. believe that a clear definition of these terms can help the researchers in detecting them by making annotating hate speech an easier task, and thus, making the annotations more reliable. Online hate speech is broadly understood as any form of expression that targets individuals or groups based on their race, color, ethnicity, gender, sexual orientation, nationality, religion, or other defining characteristics. It has the potential to perpetuate harm, marginalize vulnerable populations, and erode the foundations of a just and inclusive digital landscape (Asogwa et al., 2022).

Linguistic challenges in hate speech detection refer to the complexities of human language that hinder AI systems from accurately identifying harmful content. These include semantic ambiguity, figurative language such as sarcasm and irony, rapidly evolving slang, code-switching, and culturally embedded expressions, all of which obscure speaker intent and complicate algorithmic interpretation (Fortuna & Nunes, 2018). Such challenges highlight the limitations of AI models trained on surface-level or monolingual datasets, which often fail to grasp the nuanced and context-dependent nature of online discourse.

One promising approach for addressing these challenges may involve leveraging the unique linguistic and contextual insights embedded within online communities themselves. Researchers have demonstrated the development of AI models that are in line with the context of online hate speech for its detection by drawing on an understanding of language use within various communities (Saleem et al., 2017). As this field of study continuously evolves, it is essential to acknowledge the potential for abuse and the need to balance the advantages of automated hate speech detection with the preservation of free speech and to protect the voices of minorities (Saleem et al., 2017). This research aims to explore the linguistic challenges that arise as a result of using AI for the identification of online hate speech. Its focus is to investigate such challenges in depth while also understanding the linguistic complexities that AI algorithms have to overcome for the effective identification of hate speech online. The research significantly contributes to the development of AI tools that are more efficient to tackle online hate speech by recognizing the challenges and developing strategies to address them.

Moreover, the study further investigates the ethical issues and potential biases associated with the application of AI for the identification and addressing of online hate that target individuals and communities. By addressing these challenges, this research aims to contribute to the effectiveness of existing AI tools to help curb the spread of online hate speech, making a safe and more inclusive digital space for all.

1.2 Statement of the Problem

Online hate speech is a pervasive and complex issue that targets individuals and communities online. One way to address this issue is through AI algorithms; however, there are various significant linguistic challenges involved in this process such as the evolving nature of language, ambiguity, and the use of non-standard language online. To ensure the effective use of AI tools, these linguistic and ethical challenges must be addressed. The problem is that the current AI tools for tackling online hate speech are limited due to linguistic biases, as well as difficulties in accurately detecting and interpreting the linguistic features and patterns of online speech. These challenges hinder the effectiveness of AI algorithms in finding ethical solutions for combating online hate speech. Therefore, linguistic research is needed to overcome linguistic challenges and to develop more accurate and sensitive AI algorithms that better identify and address online hate while mitigating potential biases and ethical concerns.

1.3 Objectives of the Study

1. To identify the main lexical and semantic features of online hate speech found on social media platforms like Twitter, Facebook, and Instagram.
2. To investigate the key linguistic challenges involved in developing AI models for the identification of online hate speech.
3. To evaluate the extent to which ethical considerations involved in using AI for the identification of online hate speech can be integrated into the development of AI systems.

1.4 Research Questions

1. What are the dominant lexical and semantic features that characterize online hate speech on platforms such as Twitter, Facebook and Instagram?
2. What are the key linguistic challenges faced in developing of AI models for the detection and identification of online hate speech?
3. To what extent can ethical considerations in the use of AI for detecting online hate speech be integrated into the design and development of AI systems?

1.5 Significance of the Study

Online hate speech has become a rampant feature of today’s digital age. As social media continues to expand, the task of finding a solution and addressing hate speech on online platforms have become an insurmountable challenge. One likely solution is to utilize the AI technologies, such as natural language processing, machine learning, and deep learning, to detect and eliminate hate speech in online media. However, the linguistic challenges involved in the accurate identification of online hate speech are complex, making it important to understand the biases and limitations of the AI tools. Linguistic challenges and existing limitations of the current AI tools have been reinforced by the previous research studies as well, such as the earlier research study by Cortiz and Zubiaga (2020), which argued that although AI might play a central role in addressing the issues online, this has a potential to create problems of freedom of expression.

The present research aims to explore these challenges in depth and discover ways on how to improve the accuracy and effectiveness of AI systems in recognizing and dealing with online hate speech. Some of the linguistic challenges that this research focuses on include ambiguity and figurative language, non-standard language and emerging language. All these characteristics of the language make it difficult for AI algorithms to determine and interpret such words and phrases. This means that AI algorithms need constant upgrading and modification to remain up to date with new trends and developments.

This study is significant as it highlights how linguistic features such as derogatory slurs, exclusionary language, and negative generalizations contribute to the spread of online hate speech. It underscores the importance of understanding these linguistic elements to enhance AI's ability to detect and interpret hate speech more accurately. The research also emphasizes the ethical dimension of AI applications, promoting fairness, inclusivity, and cultural sensitivity in automated detection systems. Moreover, it holds social significance by contributing to the creation of safer and more inclusive online environments through improved moderation tools.

The study contributes to both linguistic and AI research by linking linguistic analysis with the technological and ethical limitations of AI-based hate speech detection. It extends theoretical understanding through the application of Socio-Technical Systems Theory and Discourse Ethics, offering an interdisciplinary perspective on how language and technology interact. Practically, the research provides insights for AI developers, social media platforms, and policymakers to design context-aware systems that integrate human oversight. By revealing the limitations of AI in understanding linguistic complexities, it lays the foundation for future advancements in developing fair, transparent, and culturally responsive AI tools.

1.6 Delimitation

While conducting any research, it is essential to establish delimitations for the study which set the boundaries for the study. The study is delimited to AI systems that are currently in use rather than theoretical or speculative systems. The study is also delimited to the use of AI technology in identifying online hate speech specifically in the linguistic domain. This means that study does not delve into the technological aspects of AI, such as programing and implementation.

The population of the study is based on the speakers of English both native and non-native who are the users of social media platforms such as Twitter, Facebook and Instagram. The study is, however, delimited to the types of online hate speech, racism, sexism and Islamophobia. By analyzing the linguistic features of these different types of discourses, the researcher got an insight into the prevalence of each category and how they are different from each other.

The teachers of linguistics are not taken as participants because the study focuses on obtaining practical insights from experts directly engaged with AI systems. The aim is to explore the applied interaction between language and technology, rather than theoretical perspectives. In addition, the study aims to explore linguistic challenges in AI-based hate speech detection rather than resolve them from a purely linguistic perspective. This ensures the findings remained relevant to the real-world challenges of AI-based hate speech detection**.**

This study includes quantitative analysis to categorize and visualize patterns of hate speech but it does not rely solely on numerical data. The focus is delimited to exploring the linguistic and ethical aspects of AI's role in hate speech detection, something that numerical data alone could not fully capture. Lexical and semantic analyses were incorporated within the quantitative phase to identify how language features contribute to hate expressions. This scope allows for a more focused examination of linguistic challenges, supported by qualitative insights for deeper contextual understanding.

Furthermore, the research does not address the social and cultural factors that contribute to online hate speech, given the complex nature of the issue that goes beyond linguistic challenges. However, while focusing on the linguistic challenges, the research addressed the social and cultural context to some extent avoiding the complexity involved. Finally, the study considered only online hate speech that occur in public online spaces, such as social media platforms and online forums, and does not explore instances that occur in private messaging or other forms of communication. Summing it up, the delimitations of the study focus on the English language, current AI systems, linguistic challenges, and public online spaces. The study does not address social and cultural factors, other languages, theoretical AI systems, private communication, or other forms of online abuse.

1.7 Limitations of the Study

Although the research study investigated several insights into the linguistic challenges, AI and hate speech detection, several limitations should be acknowledged. One of the primary limitations for this study is to focus on online hate speech in the English language only. While there may be similar challenges in other languages, the study does not explore those due to limitations in the researcher's language abilities and resources.

Another limitation is the reliance upon datasets that are not diverse. The datasets used are not diverse enough to include various cultures and linguistic backgrounds. Although the researcher ensured a representative sample, most datasets existing online are in English and high-resource languages. As a result, the findings are not fully generalizable to low-resource languages and cultural contexts. Moreover, another limitation is the evolving nature of hate speech. Although the research included definitions of hate speech from various online platforms, however, the study cross-sectional approach captured the definitions of hate speech at that moment of time, but may not account for the evolving nature of hate speech. Therefore, the findings may not be generalizable to the linguistic features of future hate speech.

1.8 Organization of the Study

The present research is systematically structured to explore the linguistic challenges of detecting online hate speech. Each chapter is detailed aiming to provide a comprehensive understanding of the problem and propose linguistically informed solutions. The research study includes the following chapters:

i. **Chapter 1: Introduction**

   This chapter provides background and context for the study. It outlines some of the challenges of hate speech, AI role in its detection and the specific linguistic challenges it encounters. The chapter also presents research objectives, research questions, significance and delimitation of the study.

ii. **Chapter 2: Literature Review**

This chapter reviews existing literature on online hate speech, covering various definitions of hate speech. It also explores research studies on AI-based detection of hate speech, including various Natural Language Processing (NLP) and Machine Learning (ML) techniques. Further, the chapter identifies gaps and challenges in the current research that this study aims to address.

iii. **Chapter 3: Research Methodology**

This chapter details the research design, data sampling, data collection and data analysis methods. It also presents the theoretical frameworks used in the research study, and how various concepts are employed to carry out the analysis.

iv. **Chapter 4: Data Analysis**

This chapter provides the quantitative and qualitative data analyses. The quantitative analysis explores the linguistic features and patterns of online hate speech by lexically and semantically analyzing the hate comments collected from the datasets. The chapter also presents qualitative analysis of the interviews conducted with AI experts in order to explore the linguistic and ethical challenges that AI systems face. It presents the findings from both the analyses.

v. **Chapter 5: Conclusion**

This chapter summarizes the key findings and contributions of the study. It also presents limitations of the study and recommendations for future research and discussions, suggesting ways to improve AI models by incorporating deeper linguistic insights with collaborations of linguists and AI experts.

CHAPTER 2

LITERATURE REVIEW

The emergence of social media in recent times has presented novel prospects for individuals and groups to establish connections and participate in online communities. But this increased connectedness has also brought a significant increase in hate speech online, which can have adverse impacts on people personally as well as wider societal and political ramifications. Employing artificial intelligence to automatically identify and eliminate hate speech from internet platforms is one possible answer, while there are many other ways to tackle this issue. However, to create efficient AI-based hate speech detection systems, several language issues encountered in the use of this method need to be resolved. In the review below, the researcher explores the current state of research on the linguistic challenges of using AI for online hate speech detection, with a focus on the implications for developing effective and ethical solutions to this pressing problem. The review of the scholarly research works is based on four main themes. First, it investigates what online hate speech is, how it is defined, and the greater awareness of its online use. The second part addresses the role of AI in the detection and online hate speech. It shows how the AI models are trained and evaluated to discover online content. The third of these examines the linguistic difficulty of AI models while detecting online hate speech. It focuses on the intricacies of human language that AI models are unable to detect. The last section examines the potential ethical challenges of using AI for this purpose.

2.1 Key Themes in Existing Literature

2.1.1 Online Hate Speech

Deciding if a particular text contains hate speech can be a difficult task even for humans. Hate speech is a complex language that is inextricably linked to interpersonal relationships between individuals and groups. There is neither a consensus on the definitions of online hate speech nor is there agreement on specific aspects of these categories. According to Ross et al., a precise definition of these phrases can aid researchers in identifying them by simplifying the process of annotating hate speech and increasing the accuracy of the annotations.

Nobata et al. define hate speech as language that disparage or target a group because of their color, ethnicity, religion, handicap, gender, age, or sexual orientation/gender identity.

In the same way, Facebook defines hate speech to help its users understand the boundaries of acceptable communication and behavior. According to Facebook (2013), it does not allow content that attacks anyone based on their real or perceived color, ethnicity, national origin, religion, sex, gender identity, or sexual orientation, as well as their disability or illness. It does, however, allow overt attempts at satire or comedy that might otherwise be viewed as potentially dangerous or offensive (Fortuna and Nunes, 2018). Moreover, Facebook in 2019 adjusted and refined its hate speech policy six times that mostly included different forms of hate speech. According to Loebbecke et al. (2021), Facebook (2020a) defines hate speech as:

> A direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We protect against attacks on the basis of age when age is paired with another protected characteristic, and also provide certain protections for immigration status. We define attack as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, or calls for exclusion or segregation.

Similarly, online hate speech, according to YouTube (2017), is defined as any content that incites violence or hatred towards individuals or groups based on specific characteristics, including age, gender, veteran status, race or ethnic origin, religion, handicap, sexual orientation, or gender identity. What constitutes hate speech and what does not are two different things (Fortuna & Nunes, 2018). In addition to that, as per Loebbecke et al. (2021), YouTube (2020) defines hate speech as:

> Hate speech refers to content that promotes violence or hatred against individuals or groups based on certain attributes, such as race or ethnic origin, religion, disability, gender, age, veteran status and sexual orientation/gender identity. There is a fine line between what is and what is not considered to be hate speech. For instance, it is generally okay to criticize a nation-state, but not okay to post malicious hateful comments about a group of people solely based on their ethnicity.

Likewise, X (Twitter) (2020) defines hate speech as acts that" 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" (Loebbecke et al., 2021, p.3). Furthermore, Fortuna and Nunes (2018), also defines hate speech as any language that disparages or attacks a group of people based on their physical characteristics, religion, ethnicity, national or ethnic origin, sexual orientation, gender identity, or any other characteristic. It can take many different linguistic forms, even subtly, or when humor is used (Fortuna and Nunes, 2018). Hate speech" has long been prevalent in human interactions in the real world in a variety of forms, such as racism and bigotry, and it has since made a name for itself in the online world of social media (Thomas, 2011). It may be promptly disseminated to a huge number of people and is readily accessible, making the latter an obvious vehicle for" hate speech." While there's no consensus on what defines" hate speech," a definition similar to this has been put forth. Gitari et al. (2015) described this issue in light of the growing use of the term" hate speech" on social networks, noting that it is typically the result of hostile users who bias" others" because of particular benefits such as discrimination, creating fear, and instability. Erjavec & Kovai (2012) define hate speech as" any form of abuse, insult, intimidation, harassment, incitement to violence, hatred, or any other forms of violence." Hate speech" is defined by Awan (2016) as any discourse that aims to dehumanize somebody based on their race, gender, sexual orientation, religion, or any other attribute, such as physical or mental disability (Nazmine et al., 2021).

Any speech that is derogatory against an individual or a group on account of attributes including race, color, ethnicity, gender, sexual orientation, nationality, religion, or political affiliation constitutes hate speech (Zhang and Luo, 2018). Fortuna and Nunes (2018) have also defined cyberbullying which according to them is aggressive and deliberate act committed over time and repeatedly against a group or individual through electronic modes of connection.

In the same way, ElSherief et al. in their 2019 research, investigate the realm of hate speech on social media and offer a linguistic analysis of its target audience. By concentrating on the individual or group under attack, the researchers uncover intriguing indicators that differentiate focused hate speech from generalized hate speech. Based on who they were targeting, they examined two distinct forms of hate speech: directed and generalized. Directed hate speech, according to their analysis, is more aggressive, informal, and personal. It frequently uses terms that imply authority and influence as well as name-calling. On the other hand, terms that denote quantity, like “million” and “many,” and words that incite religious hatred, lethal words like “murder” and “kill,” are characteristics of generalized hate speech.

These psycholinguistic and linguistic indicators aid in differentiating between the two categories of hate speech. The research analyzes the particular individual or group being targeted adds significantly to one’s knowledge of hate speech on social media. The researchers provide insights into the distinctions between directed and generalized hate speech, which can help shape the creation of more focused tactics to counteract these two forms of hate speech. The researchers also draw attention to the issue of combating hate speech outside Twitter and the other sites that support it. The paper, however, would have benefitted more from thorough examination of the moral issues surrounding the use of language analysis to identify hate speech as well as any possible unforeseen repercussions.

Additionally, online hate speech has grown significantly in volume. Social media platforms support the freedom of expression, but they also encourage hate speech that is not well-considered because of the fast and anonymous publication options (Citron, 2014; CohenAlmagor, 2014; Mondal et al., 2018). Quick responses frequently encourage hate speech online to escalate (Coffey and Woolworth, 2004; Wheaton, 2019). Because of its international reach, internet hate speech defies national laws and necessitates collaboration across multiple jurisdictions (Gagliardone et al., 2015). The historical and cultural background also affects whether anything is considered hate speech. There is a growing recognition of the urgency of this issue (Gamback & Sikdar, 2017). According to Gagliardone et al. (2015), 40% of people in the European Union (EU) have felt intimidated or attacked by hate speech on social media, while 80% of people have come across it online.

On the other hand, referring to the increase in online hateful content, Nazmine et al. (2021) conducted a comprehensive research on hate speech and social media, synthesizing multiple studies and stating that increased audience response to online hate speech is a result of increased technology acceptability and higher communication skills. The research paper also emphasizes how important it is for mass media scholars to further explore this phenomenon and ensure that public dialogue on societal issues and media digitization is maintained. The research also reviewed a study that employed the categorization text approach to identify hate speech on social media, with the primary goal to differentiate hate speech from other types of social media text. A more profound analysis of the moral implications of hate speech on social media, particularly freedom of speech and the potential harm to people and groups, might have enhanced the evaluation, nevertheless. In addition, by looking into options like community driven moderation systems or increased platform accountability, the evaluation may have addressed measures to combat hate speech on social media.

2.1.2 Role of AI in Online Hate Speech

Over the years, addressing the harmful impacts of the rising issue of online hate speech has been the subject of discussion. For this, several initiatives have been taken some of which are the use of AI tools as a potential means to find, monitor and counter online hate speech. Recognizing hate speech in real-time and making online spaces safer, researchers and practitioners aim to develop automated systems. This is accomplished by applying AI algorithms and machine learning models.

Natural Language Processing (NLP) and Machine Learning (ML) are two AI methods that are frequently utilized in the identification and control of online content. Natural language processing (NLP), an area of computer science, linguistics, and artificial intelligence, mainly focuses on how computers interact with natural human language (Reshamwala et al., 2013). Natural language processing (NLP) is a study of a set of methods that help computers comprehend natural language (Moy et al., 2021). These methods examine and understand the linguistic patterns connected to online hate speech. Among these methods are:

2.1.2.1 Word Embedding

One of the most prominent techniques of NLP is word embedding, that communicates the semantic meaning of a word (Bengio et al., 2003). It provides a helpful numerical description, depending on the context of the term (Saleh et al., 2023). N-dimensional dense vector which represents words, can be used to estimate how similar words are to one another in a given language (Liu, 2018; Mikolov et al., 2013).

2.1.2.2 Bidirectional Long Short-Term Memory (BiLSTM)

According to Hochreiter and Schmidhuber, (1997) Long Short-Term Memory is one of the deep learning models which is an enhaced version of neural network is designed to collect information from a sequence of data points. LSTM saves long sequences only from left to right, while bidirectional LSTM (BiLSTM) saves sequence data from both directions. In BiLSTM, two Long Short-Term Memory (LSTM) models are used; one model processes input in a left-to-right direction, while the other model processes data in the opposite direction. The resulting concatenated and flattened models enhance contextual knowledge (Saleh et al., 2023).

2.1.2.3 BERT Pre-Trained Language Model

A language model based on contextual representations and trained on a massive scale of data is called Bidirectional Encoder Representations from Transformers or BERT (Devlin et al., 2018). Comparing with the previous state-of-the-art in language, feature extraction layers of BERT consist of word embedding and model layers such as classification, question answering, and named entity recognition models, BERT is the newest model, delivering state-of-the-art results for many tasks in NLP. In the word embedding training process, BERT is distinct from the earlier word embedding models in that this generates a bidirectional representation of words, which can be learned in the left as well as in the right directions (Saleh et al., 2023).

Nobata et al. (2016) have developed a machine learning-based technique for detecting abusive language in the online comments of the users. The paper adopts a supervised classification methodology along with NLP features in order to nullify the drawbacks of classical methods including blacklists and regular expressions. Such NLP techniques are syntactic features and as well as distributional semantics for analyzing and classifying the comments of the user regarding the abusive language detection. The syntactic feature allows to understand the relationships between words and phrases in user comments whereas distributional semantics focuses on presentation of word’s meaning based upon their distribution patterns in a large corpus. These Features have improved the model’s semantic understanding of abusive language and that the research very much contributes to the field by creating an annotated corpus of comments users, providing a public dataset, curated for comparison of methods by researchers, and evaluation of a wide set of NLP features for detecting abusive language. The study confirms that the created methodology outperforms a deep learning approach, showcasing the effectiveness of the NLP-based classification model.

Moreover, the Fortuna and Nunes’ evaluation, in 2018, gives an idea about where the field stands today covering important techniques, attributes, and algorithms. In addition to the potential social impacts of improved automatic detection, it also highlights the difficulty of the definition of hate speech. This research indicates that low agreement in the classification of hate speech by humans, the requirement for cultural and social structure expertise, the difficulty of tracking all racial and minority insults due to the evolution of social phenomena and language, and the rapid evolution of language among young populations that communicate frequently in social networks are some of the challenges in defining and detecting hate speech. Especially for young people who are continually interacting on social media sites, language evolves quickly. The study also calls for knowledge about social structure as well as culture. It is hard to identify hate speech automatically due to these complexities.

In addition to the importance that a clear and precise definition of hate speech has toward improving its automatic identification, the work also gives room for an insightful analysis on how hate speech detection in text has come to be developed. For analysts and practitioners entrusted with the responsibility of identifying hate speech, their investigation of classification criteria and the merits and demerits they provide is sharply analytical. Still, the thorough discussion of at least potential costs and ethical dilemma the automated detection of hate speech poses as well as how better, more widespread social and cultural conditions impact the formation of hate speech, would have improved the research, nevertheless.

Similarly, Grondahl et al. (2018) conducted another research on avoiding online hate speech detection. The research underlines the vulnerability of existing hate speech detection techniques and offers a necessary perspective on the limits of datasets to evasion assaults. The proposed attacks and mitigations provide an understandable and realistic view of the challenges involved in the design of efficient systems for detecting hate speech. This work tries to address the limitations of the current approaches and points out the areas of future research to be undertaken for developing more reliable models and datasets, robust to attempts at evasion. This paper provides relevant information into efforts to suppress hate speech online and would be very resourceful to scholars interested in the development of a successful online hate speech detector.

2.1.3 Linguistic Challenges in Identifying Online Hateful Content

MacAvaney et al. (2019) while discussing the linguistic challenges of AI in detecting hate speech argued that the first challenge is the differences in the definitions of hate speech. This means that depending on how each definition is used, some content may be regarded as hate speech by some people but not by others. Conflicting definitions make it challenging to evaluate hate speech detection systems since there are several datasets that define hate speech differently, leading to datasets that not only come from diverse sources but also capture different information. This makes it challenging to determine which characteristics of hate speech to look for explicitly. Depending on the criteria, further difficulties in automatically identifying hate speech arise from linguistic nuances and ambiguities. Indeed, more society-influenced contexts would appear within this research since the development emphasizes that without a more thorough social and cultural perspective of context, the automated techniques have their limitations. However, it is important to note that the focus of the research on technical aspects may overlook other important considerations detecting hate speech, such as the broader social and cultural factors that contribute to online hate speech.

Likewise, Kiritchenko, Nejadgholi, and Fraser (2021) identify a number of the most common hurdles in the detection and solving problems of abusive language in cyber space. According to the report, abusive language may take various forms and that, to recognize it precisely, a specific linguistic technique needs to be used for each form. This is crucial because it highlights how complex AI systems that can recognize a wide range of abusive words are needed. It also establishes the need for context while identifying abusive language on the web. This is again important issue, because it accepts that the meaning of offending language depends on historical background, cultural standards, and many other contextual elements.

Researchers are looking for automatic methods to detect hateful tweets because of the non-scalable nature of manual method. However, the complex nature of language which include various targets, forms of hatred, and ways to portray the same meaning in numerous ways (Badjatiya et al., 2017), makes the task extremely challenging. Moreover, the challenges be either because of the intricacies of language or about the ethical use of language. Park et al. (2018) expressed concerns about the possibility that words that appear frequently in the data set could be over fitted into hate speech recognition algorithms, thus causing biases in the model's detection process.

Additionally, Sap et al. (2019) highlighted that some commonly used corpora by researchers have been proven to perpetuate racial bias against African American English (AAE). AAE is being classified to be offensive twice as likely (Moy et al., 2021). Moreover, Zhang and Luo (2019) have explored the issues in the identification of hate speech on Twitter when hate speech is not clearly evident or relies on subtle language cues. The authors suggest a new deep neural network to detect hate speech online. The authors also conducted data analysis to identify the linguistic features of hate speech on Twitter. In addition, they evaluate their technics and discuss what may be directions in the near future for advanced hate speech identification. One key contribution of this work to the field is in the area of Deep Hate, which refers to a new form of deep neural network designed for detecting hate speech with much sensitivity. The suggestion of focusing on highly sophisticated models agrees with the general trend in hate speech detection research to consider more advanced machine learning methods. Among other recommendations is the recommendation of including user data and network architecture for consideration in analysis. This places the benefits of the use of contextual data as key to raising hate speech detection accuracy. However, with this one, it would be important to reflect on how that might impede privacy, consequences of collecting and analyzing user data. Since it brings to light the inadequacies of the current datasets in capturing the diversity of hate speech and its circumstances, the proposal to work towards the development of larger datasets for training and testing hate speech detection systems is equally important. In order to ensure that such datasets represent the greater population, it is important to consider any bias in them. According to Kovacs´ et al. (2021), certain linguistic challenges are closely linked to the limitations of keyword-based methods. Words can be obfuscated through two ways, either through intentional attempts to evade automatic content moderation or using social media for communication; for instance, some posts tend to replace letters with similar-looking numbers, such as “E”s for 3s or ”I”s for 1s, and so on.

Additionally, a lot of expressions are not intrinsically offensive, but they can be in the correct situation. However, even when it comes to slurs, different slurs not only have varying degrees of offensiveness, the offensiveness can also change depending on the time, the audience, the users, and the way the same word is used. They further found out that although the usage of the “f word” significantly raises the likelihood that a tweet would be classified as hateful or offensive, tweets that include it in a hashtag are not more likely to be classified as such than any other tweet.

According to Davidson et al. (2017) a major problem with a lot of earlier research is that offensive language is mistakenly classified as hate speech because the definition is too wide. The researchers are able to reduce these mistakes because of their multiclass framework; only 5% of their actual offensive language is classified as hate speech. The tweets that have been appropriately classified as offensive typically contain slang and frequently sexist language. It seems that human coders view terms that are homophobic or racist as hateful, but only regard words that are sexist and demeaning of women to be offensive, which is in line with previous research (Hovy and Waseem, 2016).

In a similar way, a study by Ullmann and Tomalin (2019) points out that hate speech often uses coded language, latent connotations, and subtle undertones, which are hard for automated systems to detect. The correct identification of hate speech is contingent upon the ability to understand context and intent in language usage. Furthermore, it becomes challenging for hate speech detection systems to identify sarcasm, irony, or other figurative language. Such linguistic tricks may be used to hide hate speech or subtly spread destructive ideas, so the interpretation of language by systems beyond its literal meaning is necessary. Hate speech can contain ambiguous or interpretive words and phrases that can have more than one meaning, leading to polysemy. Systems that successfully disambiguate language are necessary to distinguish between instances of hate speech and legitimate uses of the language.

The research paper is innovative in the use of a conceptual framework that is different from existing ones. This conceptual framework draws parallels between hate speech detection and malware containment; it is an approach which can be termed proactive in curtailing the spread of offensive content. Using ethical analysis, the paper underlines the need to strike a balance between freedom of expression and responsible content moderation. Technical perspectives are also included in the work, like that of automated hate speech detection systems, in a forward-looking methodology. Interdisciplinary of methodology allows for a holistic study of the challenges around hate speech moderation online and contributes to the discussions on safety technology online and ethical content management with great importance.

2.1.4 Ethical Challenges of AI in Identifying Online Hate Speech

The role of Natural Language Processing (NLP) in the identification of online hate speech has been investigated by Kiritchenko, Nejadgholi, and Fraser (2021). To determine the extent of harm by the online harmful content, modern technologies can help, but they are not able to identify the hidden meaning of human-written statements. The research paper emphasizes the importance of ethical considerations in NLP technology to avoid silencing underrepresented groups and other unexpected results. The research also discusses eight established ethical principles that should be applied when using NLP technology to detect abusive language on the internet: privacy, accountability, safety and security, openness and explainability, impartiality and non-discrimination, human control over technology, professionalism and the advancement of human values.

Another work by Kumarage et al. (2024) is on the application of Large Language Models to hate speech detection. These models have a number of challenges regarding reliability and interpretability, especially with such complex models, potential biases from the training data, issues in deployment at scale, requirement for diverse and high- quality data, vulnerability to adversarial attacks, and ethical issues around privacy and free speech. Overcoming these challenges requires efforts to improve transparency, mitigate biases, ensure efficient deployment, improve data quality, strengthen model security, and establish ethical guidelines for the responsible use of LLMs in combating online hate. The research therefore fully examines the role of LLMs in hate speech detection and evaluates the proficiency of these models, but would have been more useful with a greater variety of LLMs for the comparison of their performance in hate speech detection.

Furthermore, Field et al. (2020) explore the issue of accountability for the decisions made automatically. It deals with the problems of verification and replication, evaluation of impacts, review, audit standards for fault and legal accountability. Moreover, the responsibility principle has much to do with safety and security and explainability transparency and human control of technology. Kiritchenko, Nejadgholi, and Fraser (2021) states that it has become widely recognized that responsibility for the outcomes and impact of AI systems should be entrusted to the organization developing and utilizing them. The fact that AI systems are not juridical personalities by themselves does not make a legal sense for holding them liable because problematic (Bryson et al., 2017). There are ethical rules regarding AI that make distinctions between the liability of the implementing organizations and the liability of the developers of the AI system. Transparency and professional responsibility norms are often considered to be the most acceptable forms of accountability at the development level.

Additionally, according to a research conducted by Cortiz and Zubiaga (2020), when it comes to ethical challenges, representativeness is the first one. To ensure a more representative and diverse dataset, the AI experts need to be careful with the filters they apply during the collection process. The model developers may obtain data that only accurately depicts the realities of a small subset of people if they employ a certain set of keywords. If they decide to gather data from a specific forum (some Chan, for example) or a narrow subnet of a social network (such as gathering data only from a particular group of users in Twitter), they may encounter bias and produce an unrepresentative data set.

Privacy is yet another essential issue. Real users' posts, who may not wish to be identified are used in the data collecting process. It can be argued that since the data is accessible to the public on a social network, privacy is not an issue. One counterargument, though, is that the person simply shared anything on social media to express their ideas, beliefs, and feelings; he or she did not consent to having their content utilized to develop an artificial intelligence model. Given the volume of laws and regulations being considered and drafted globally, privacy is a major issue on the agenda (Cortiz, Zubiaga, 2020).

In their 2019 study, Sap et al. examined how racial prejudice in automatic hate speech detection models can result from annotators' insensitivity to dialectal distinctions, thereby causing even more harm to minority communities. Cortiz and Zubiaga (2020) argued that any AI project must include the annotation process, but it becomes even more important when employing the supervised learning method. Annotators have some degree of influence over an AI system's future actions. It is necessary that, in order to prevent issues with bias and representativeness, the group of annotators must be numerous.

The question of the disparity of the population between hate speech and non-hate speech was a concern raised by the study conducted by Arango et al. (2020). As a result, a data set had numerous hate tweets coming from the same individual. Consequently, user overfitting is prevalent in the detection models for hate speech. In addition, Grondahl et al. (2018) also highlighted the fact that the data sets do not contain the reason behind the low generalizability of models is the whole range of hate speech that can be encountered on social media. There are several topics that fall under hate speech, such as sexual orientation, gender, religion, and race. In most of the data sets, certain domains of hate speech are preferred over others (Moy et al., 2021).

In addition, another study by Ullmann and Tomalin (2019) notes that one of the chief ethical dilemmas is a balance between the right of free speech and the duty of stopping hate speech from causing harm. Hate speech might be a fundamental right; however, it may pose negative impacts, for example, inciting violence and discrimination. Quarantine measures for hate speech raise issues of censorship and the risk of an overreach by authorities or platforms. The thin line between deleting offending content and stifling free speech and criticism exists here. Algorithmic bias and accuracy is another ethical question that arises. Systems for automatic hate speech identification could be biased, misclassifying some information as hate speech. Questions of accuracy, impartiality, and a potential for discriminatory results are then raised.

This means that the automated hate speech detection systems may be biased, and some content might end up being incorrectly flagged as hate speech. This has brought concerns of unfairness, accuracy, and even discrimination. Not much is known about how platforms moderate content and how they determine what speech is hateful. To maintain credibility and self-esteem, accountability for these decisions and practices should be given. The marginalized groups rely on the online spaces to make their voices heard and to fight for their rights, they might be disproportionally affected by the prohibition of hate speech. It is a huge ethical dilemma to find a that delicate balance between harm prevention protection of those populations. Another factor is the diversity of culture and community around the world and their possible distinct understanding of hate speech. Any general approach for curbing hate speech may totally neglect these differences in culture and create some untimely reactions (Ullmann and Tomalin, 2019).

2.2 Pakistani Context and Contributions

In the context of Pakistan, a study by Rizwan, Shakeel, & Karim (2020) contributed to hate speech detection by focusing on Roman Urdu, an under-resourced and colloquially complex language, addressing a notable gap in existing research predominantly centered on English. The study mainly focuses on developing models without thoroughly exploring the nuanced semantic and contextual challenges inherent in Roman Urdu's informal and code-switched usage, which are critical for real-world application. Moreover, the lack of classification and cross-lingual validation suggest that further work is needed to handle the cultural and linguistic variability of hate speech in diverse social media contexts. Overall, this research underscores the importance of language-specific resources and modeling approaches but highlights the ongoing challenge of capturing the complex semantics of hate speech in less-resourced, informal dialects.

In addition, another study by Abro et al. (2020) makes a valuable contribution to hate speech detection. Their study focuses on systematically comparing multiple techniques for hate speech detection. The study relies on a single, publicly available dataset consisting of Twitter messages, which, although standardized, may not fully capture the diversity of hate speech across different platforms or languages. This limits the generalizability of the findings, especially considering the evolving nature of hate speech language and tactics. Moreover, the exclusion of deep learning models, such as contextual embeddings like BERT, constrains the scope of comparison, especially given recent advancements that often outperform traditional classifiers. The three-class classification scheme fails to address the nuanced severity levels of hate speech or account for potential class imbalance issues, which are critical for practical deployment. Moreover, short-text limitations and the absence of consideration for ethical implications such as bias and false positives underscore areas for future enhancement, rendering the study a foundational, yet somewhat narrow, step towards more robust and context-aware hate speech detection systems.

Ali et al. (2021) makes a significant contribution by addressing critical challenges for detecting Urdu tweets through systematic techniques such as feature selection and oversampling. This resulted in the performance improvement of classifiers such as SVM. However, its reliance on relatively simple models and feature engineering methods limits its capacity to capture the nuanced linguistic features inherent in social media language, especially given the brevity and informal nature of tweets. The dataset, although pioneering for Urdu, remains modest in size, which may restrict generalizability, and the evaluation primarily relies on micro F1 scores without exploration of deeper linguistic or contextual errors. Future research could benefit from incorporating advanced deep learning models, rich semantic features like word embeddings, and more comprehensive error analysis to better handle complex language phenomena such as sarcasm or coded hate speech. Despite these limitations, this work lays an important foundation for low-resource language hate speech detection and highlights the ongoing need for resource development and methodological innovation in this area.

Additionally, Bilal et al. (2022) provides a comprehensive study on the detection of hate speech in low resource language like Urdu. They proposed a new model which is context-aware deep learning model, and developed a comprehensive dataset. Their work focuses primarily on the lexical diversity, normalization challenges, and contextual nuances specific to Roman Urdu, demonstrating that incorporating lexical normalization and attention mechanisms enhances detection accuracy (Bilal et al., 2022). They compared a deep learning model with a traditional model which revealed the superiority of deep learning model over traditional learning model. However, there are some limitations such as reliance on manual annotations and potential bias in dataset construction. Overall, this research provides a valuable foundation for developing more robust and culturally sensitive hate speech detection systems in resource-low languages (Bilal et al., 2022).

Kamal and Shibli et al. (2023) also contributed significantly to hate speech detection in Urdu language. They utilized a large, region-specific dataset and hybrid deep learning models, achieving moderate to high accuracy. However, challenges inherent to low-resource languages including data sparsity, class imbalance, and linguistic complexity due to dialectal variations and code-switching are not fully addressed, raising concerns about the models' robustness and generalizability. Additionally, limited details on annotation reliability and evolving hate speech patterns suggest that static datasets and models may quickly become outdated, emphasizing the need for adaptive, context-aware approaches like contextual embeddings and transfer learning. While the work demonstrates the potential of deep architectures in this domain, further exploration into demographic factors, linguistic nuances, and continual dataset updates is essential to enhance the effectiveness and ethical deployment of hate speech detection systems in low-resource languages like Urdu.

Furthermore, Kousar et al. (2024) present a research study introduces a new way to detect hate speech in different language. Their study introduces an efficient model that can detect hate speech across twelve different languages. It combines different methods to understand the meaning of texts in 12 languages, making it useful for more places than just English. The use of special techniques helps the model pick up on the hidden details in messages, making it better at catching hate speech. However, since the model learns from examples created by humans, it might struggle to keep up with new ways people express hate or meet different online platforms. Overall, the model performs very well even when the data is unbalanced, making it a helpful tool for stopping hate speech quickly and efficiently, an important step forward in this area (Kousar et al., 2024).

2.3 Linguistic Frameworks for Understanding Online Hate Speech

It is crucial to ground the discussion in fundamental linguistic frameworks, specifically pragmatics and discourse analysis, in order to completely comprehend the intricacy of online hate speech and the limitations of AI detection algorithms. These approaches provide valuable tools for understanding how meaning, intent, and ideology are communicated in subtle and context-dependent ways. Recent studies have emphasized the importance of these frameworks in computational models. For example, Hüsünbeyi et al. (2022) integrated discourse analysis techniques with neural networks to improve hate speech detection in Turkish-language news content. By incorporating features such as ideological framing and power dynamics, their model achieved greater accuracy than baseline systems reliant on lexical data alone. Similarly, Yu et al. (2022) explored the role of pragmatic context in identifying hate and counter-speech on Reddit. Their study found that when models were trained to recognize conversational context**,** performance improved significantly, demonstrating that pragmatic cues such as intent and tone are critical in distinguishing between offensive and non-offensive speech.

Another study *A Pragmatic and Discourse Analysis of Hate Words on Social Media* provides valuable insight into the contextual and ideological dimensions of hate speech. The researchers use pragmatic principles to analyze how hate words go beyond their literal meanings, frequently carrying irony, sarcasm, or implicit aggressiveness that makes simple categorization difficult. The study also employs discourse analysis in order to show how such language is woven into larger narratives that uphold group-based prejudice, especially along religious, gender, and racial lines. The study draws attention to the limits of AI systems that only use surface-level information by concentrating on how meaning is created through speaker purpose, context, and social power relations. It highlights the value of linguistically trained models and is in line with current research that focusses on detecting the nuanced and coded types of hate speech that automated detection techniques frequently overlook.

A notable contribution to the methodological advancement in hate speech detection is the study by Wanniarachchi et al. (2023), which investigated fat stigma on social media through a combined use of sentiment analysis**,** topic modelling, and discourse analysis. This mixed-methods approach provided a layered understanding of how stigmatizing language is both emotionally charged and ideologically constructed. The integration of computational tools with linguistic frameworks allowed for a deeper exploration of implicit bias and community-level discourse patterns. Moreover, the use of sentiment analysis underscores its potential utility in enhancing

AI’s ability to detect subtle or emotionally nuanced hate speech, pointing toward valuable directions for future research.

Conclusively, the research works that so far have been critically reviewed within this paper discuss a multitude of issues that cut across hate speech detection, such as the different definitions of hate speech as can be seen from Loebbecke et al. 2021. They define different meanings of hate speech on social media platforms such as Twitter, Facebook and YouTube are in some ways similar but are not the same. Secondly, all of these sites have been refining these definitions from time to time. Similarly, various research works have defined hate speech, such as Zhang & Luo, (2018) defined it as a derogatory speech against an individual or group. According to Fortuna and Nunes (2018), hate speech is a deliberate and aggressive act committed over and over, and it can take several linguistic forms. In addition to this, hate speech and misinformation are found on social media sites, linguistic issues, deep learning techniques, technical issues, and dataset limitations. The AI algorithms applied for the two main technologies in automatic online hate speech detection are Natural Language Processing (NLP) and Machine Learning (ML). Some of the NLP techniques include Word Embedding, Bi-directional Long Short-Term Memory (BiLSTM) and pre-trained BERT language model. The linguistic challenges of online hate speech detection are multifaceted, considering the complex interplay of various factors. The first challenge is that diverse communication platforms online, as well as the enormous quantity of user-generated content, make it difficult to detect and effectively address hate speech. Another challenge is that different types of hate speech demand very advanced algorithms for their detection can analyze multimodal content. The global nature of the internet also presents a difficulty in terms of cultural differences, legal frameworks and linguistic nuances in its effort to develop universal technique to detect hate speech. Also, ethical considerations based on user privacy, freedom of speech and censorship add another complexity to the detection process. Finally, the speed and the rate at which technology evolves together with online behaviors requires consistent adjustments in detection strategies before the new threats emerge. In short, the complex challenges of online hate speech detection require a holistic and subtle approach that takes into account linguistic, technical, ethical, legal, and cultural considerations.

The research works highlights the significance of developing effective models for identifying online hate speech that can account for contextual and linguistic subtleties in the speech. However, for systems to perform better and generalize, the researchers admit that only AI techniques are not enough; instead, they require an understanding of the linguistic and social context, for systems to perform better and generalize. The present research has important ramifications since hate speech on the internet can be harmful to the people and organizations that it is targeted toward. Safer online environments can be realized by minimizing the spread of hate speech and using efficient online hate speech detection techniques. The weaknesses of the current approaches, such as data accessibility, linguistic complexity, and evasion strategies, point out the need for further study and development in this area to enhance the accuracy of hate speech identification.

CHAPTER 3

RESEARCH METHODOLOGY

This chapter presents details about the overall approach that was adopted to carry out the present research. It begins by focusing on the research design, followed by research sampling, data collection and methods of data analysis used to achieve the research objectives. Lastly, the chapter describes the theoretical framework employed in this study.

3.1 Research Design

The present research follows explanatory sequential design. This is a type of mixed-method design that involves the collection of both quantitative and qualitative data in a sequential manner. This design usually consists of an initial quantitative phase followed by the subsequent qualitative phase in elucidating the findings further arising from the first phase of the research (Abutabenjeh & Jaradat, 2018).

This research design is chosen to elaborate further on the explanation of the research problem. This design is also referred to as a two-phase model (Creswell & Clark, 2011). The justification for this design is that the qualitative data can elaborate, or provide deeper insights into the statistical results obtained in the first phase (Creswell, 2014). This is relevant to this research as the findings obtained from the qualitative phase help explain and provide deeper insights into the initial quantitative results. The qualitative findings helped the researcher gain in-depth knowledge about the existence of certain linguistic challenges faced by the AI systems.

Following the two-phase process, the researcher first collected large datasets of hate speech already existing online. The analysis of this data provided the linguistic features and patterns of online hate speech. In the second phase, the researcher collected qualitative data through interviews with AI experts, which gave insights into the linguistic challenges, the ethical implications and the contextual problems faced by the AI models while detecting online hate speech. Thus, following explanatory sequential design the researcher was able to first identify the characteristics of online hate speech and then the experts' insights helped the researcher to explore linguistic challenges and hurdles faced by AI algorithms in order to detect online hate speech. The results from the qualitative analysis helped the researcher gain deeper insights into why certain linguistic challenges occur in AI detection of online hate speech.

3.2 Research Sample and Sampling

The research study employed two different sampling methods for the quantitative and qualitative phases. For the quantitative phase, random sampling was used since random sampling is considered statistically more accurate and efficient (Alvi, 2016). This type of sampling is often considered as the most rigorous form of sampling, as it reduces the potential for bias and enables the generalization of results to the broader population (Naseri & Rahmiati, 2022). In the present study, it was used to select data from much larger datasets of online hate speech. The sample was drawn from various online platforms including Twitter, Facebook and Instagram that have been identified as having instances of hate speech. The online available datasets were categorized as Multilingual Hate Speech Dataset, Dynamically Generated Hate Speech Dataset, Labeled Hate Speech Detection Dataset (Cooke & Shane, 2022) and others. These datasets provided a diverse range of textual data, comprising of English language and containing various aspects of online hate speech. Given the vast amount of content available on platforms like Twitter, the researcher cleaned and filtered the data and created a new dataset containing only the labeled hate comments from the larger datasets containing both labeled and non-labeled hate comments. The initial sample was a total of 153,426 comments, including both labeled and non-labeled hate comments present on Twitter, Facebook and Instagram. The new dataset was selected through a combination of key word filtering ensuring the dataset to be representative of both overt and subtle forms of hate speech. Then the final sample was a total of 96,942 labeled hate comments including both overt and subtle forms of hate speech.

Initial Sample Collected

Table 1

Total Number of Collected Comments (Initial Sample)

Sr. No.	**Platforms**	**Frequency**	**Percentage**
1.	Facebook	42022	27%
2.	Instagram	43562	28%
3.	Twitter	67842	44%
	Total	153426	100%

Final Sample

Table 2

Total Number of Hate Comments Analyzed (Final Sample)

Sr. No.	**Platforms**	**Frequency**	**Percentage**
1.	Facebook	28813	30%
2.	Instagram	21942	23%
3.	Twitter	46187	48%
	Total	96942	100%

For the qualitative phase, highly purposive sampling technique was employed to select AI experts for interviews. According to Etikan (2016), purposive sampling allows the researchers to target subgroups or explore phenomena in depth. Purposive sampling was selected to ensure the inclusion of participants with specific expertise and relevant experience in the field of AI. Given the study's focus on exploring the linguistic challenges in AI-based hate speech detection, it was essential to obtain insights from individuals capable of providing informed and contextually rich perspectives. This sampling method allowed the researcher to intentionally target subject matter experts who could contribute depth to the data. Thus, purposive sampling was most appropriate for achieving the study's objectives, as it prioritized quality and relevance of information over representativeness.

The interviewees were chosen based on their expertise in the field of AI. The inclusion criteria required participants to have a direct experience in the field of AI such as academic researchers who have conducted research in the same field and are teaching at various universities. These participants possess a minimum of one year of professional or academic experience in artificial intelligence, machine learning, or natural language processing. They were selected based on their familiarity with linguistic and ethical challenges in online communication and their ability to provide informed perspectives on hate speech detection systems. This targeted approach allowed for the inclusion of participants whose knowledge directly aligned with the study’s objectives, ensuring the collection of meaningful and contextually rich data. A total of 10 participants were selected as a sample based on their expertise in the field and teaching at prominent universities in Islamabad.

Participants of the Interviews

Table 3

Participants of the Interviews

Sr. No.	**Name**	**Qualification/Area of Interest**	**Currently Working**
1	Dr. Fahad Ahmed Satti	PhD (CS), Department of Artificial Intelligence and Data Science (AI&DS-DoC)	Assistant Professor at NUST, Islamabad
2	Dr. Hassan Mujtaba	PhD (CS), Data Science and Artificial Intelligence, Founding Member AIM (Artificial Intelligence and Machine Learning) Lab.	Head of Department, Professor at FAST, Islamabad
3.	Adil Majeed	MS (CS), Artificial Intelligence and Data Science	Lecturer at FAST, Islamabad
4.	Ayesha Safdar	MS(CS), Artificial Intelligence, Machine Learning, Deep Learning	Lecturer, NUML Islamabad
5.	Sadia Ashraf	PhD (CS), Artificial Intelligence, Software Testing and Application of AI in Testing	Lecturer, NUML Islamabad

6.	Anam Taskeen	MS(CS), Artificial Intelligence	Lecturer, Department of Software Engineering, NUML, Islamabad
7.	Hashim Ayub	MS (CS), Artificial Intelligence	Associate Professor at CUST, Islamabad
8.	Muhammad Kashif	MS (CS), Artificial Intelligence and Deep Learning	National Engineering & Scientific Commission Mission (NESCOM),
9.	Ali Raza	MS (CS), Artificial Intelligence Deep Learning	National Engineering & Scientific Commission Mission (NESCOM),
10.	Saad Ur Rehman	MS (CS), Artificial Intelligence and Deep Learning	Lecturer at Air University

3.3 Data Collection Methods

3.3.1 Google Data Search and Kaggle

The study used a combination of data collection techniques to address the research questions. For the quantitative data, online datasets were accessed through websites such as Google Data Search and Kaggle. Google Data Search is a search engine designed specifically for datasets on a wide range of topics. The datasets are available to the researchers from various organizations and research institutions. Kaggle is also an online platform that contains thousands of datasets for data scientists and machine learning practitioners. The datasets contain a large number of hate comments were then filtered and cleaned to include only those comments that were instances of hate speech and to remove the irrelevant data.

3.3.2 Interviews

The qualitative data was collected by conducting semi-structured interviews with subject matter experts. Interviews allow the researchers to examine the depth of experience gained by participants and the elicitation of their opinions (Flick, 2014). The AI experts were identified and approached by the researcher based on their expertise and experience. The interviews were conducted both in person or via phone depending on the availability of the participants, and each interview was 30-40 minutes long. The interviews were conducted in a manner that allowed for open-ended responses, enabling the participants to provide rich, detailed information about their experiences and perspectives. The interview questions were designed to elicit responses that shed light on the key factors influencing data collection methods. The interviews were mostly in English and were recorded and transcribed verbatim to ensure the accuracy of the data. For interviews guide, see Appendices.

3.4 Data Analysis Methods

The research study employed different data analysis methods for the quantitative and qualitative data analyses. The statistical analysis of the quantitative data was carried out using SPSS as a software tool. This software helped in the organization and management of the data which facilitated the analysis. Then, the researcher classified the data based on the lexical and semantic features which resulted in the identification of several different categories of hate speech such as racism, sexism and political hate speech. The comments from the different categories were first analyzed lexically and semantically which highlighted the linguistic patterns and features of online hate speech. Moreover, descriptive statistics method was used to summarize the frequency and distribution of hate speech across various online platforms like Twitter, Instagram and Facebook. The frequency distribution indicated the number of different categories of hate comments that appeared in the dataset. The mean, median, and mode values represented the characteristics of data such as the average frequencies of hate speech occurrences across Twitter, Instagram and Facebook. The standard deviation measured the variation or dispersion within the dataset. These descriptive statistics provided an overview of the dataset, the key characteristics of hate speech and its frequency and percentages across the three main online platforms including Twitter, Instagram and Facebook.

The qualitative data collected from the interviews was analyzed using thematic analysis. Thematic analysis followed the six-step process outlined by Braun and Clarke (2006). The interviews were transcribed in order to have text-based data for analysis. After transcription, the researcher generated initial codes based on the key issues discussed by the participants such as challenges in hate speech, AI limitations and ethical issues. Following this, the researcher grouped the similar codes in order to generate broader themes. Key themes were extracted from the codes which were refined and then analyzed in order to get insights into the major areas of concern. Some of the themes included linguistic features of hate speech, challenges in AI detection of hate speech and Ethical challenges. The thematic analysis allowed for the identification of the linguistic and technical challenges that AI systems face and some future regulations suggested by the experts.

3.5 Theoretical Framework

The research study is grounded in two theoretical frameworks: Socio-Technical Systems Theory and Discourse Ethics Theory. Both the frameworks are employed to guide the quantitative and qualitative analysis. Socio-Technical Systems Theory is applied to explore the interaction between the AI systems and the social contexts ensuring that the technology aligns with the social and linguistic contexts in which it operates. Similarly, Discourse Ethics Theory provides ethical guidelines for AI in facilitating ethical and inclusive online discourse. By combining these frameworks, the research study evaluates both the technical effectiveness of AI and the ethical implications of AI tools, aiming to develop socially responsible and context sensitive approaches to online hate speech detection.

3.5.1 Socio-Technical Systems Theory (STS)

Socio-Technical Systems Theory emerged in the 1950s and 1960s from the work of Eric Trist and Fred Emery. It was later developed by the work of several other researchers. The two pioneers were social scientists who were interested in how the social and organizational structures were impacted by the changes in technology. In the present study, the organizational structure such as the people who spread hate speech online and their linguistic norms are impacted by the change in technology. Due to the availability and advancement of technology, it has become much easier for users to target individuals and communities for hate speech online. Moreover, the theory emphasizes that technology is not just a tool that can be added to a social system, rather technology shapes and is shaped by the social systems. The technological systems such as Twitter, Instagram and Facebook shape and are shaped by the social systems such as people and their culture. According to Clegg (2000), it recognizes that the performance of any organization or system depends on both its technical components and the social system that shape its design and implementation.

3.5.1.1 Recent Developments in Socio-Technical Systems Theory

Socio-Technical Grounded Theory (STGT) proposed by professor Rashina Hoda in 2022, is a methodological advancement to adapt Traditional Grounded Theory for recent socio-technical contexts. This approach provides clearer methodological guidelines, accommodating the dynamic nature of human-technology interactions in fields like AI and human-computer interaction. STGT offers a robust framework for exploring the co-evolution of social and technical systems, making it particularly suited for analyzing how linguistic challenges arise in AI-mediated online environments. By focusing on the emergent patterns of interaction between users, moderators, and algorithmic systems, this approach enables a grounded understanding of the interpretive ambiguities and socio-ethical implications inherent in hate speech moderation.

Individualization at work is another revised socio-technical perspective for the 21st century proposed by Roger Fischer & Richard Baskerville in 2020. It highlights a shift towards individualization in socio-technical systems. The theory was updated to address modern digital work environments emphasizing challenges like decentralized decision-making, remote work and information overload. In addition, Enid Mumford’s 2019 research ‘Continuous Socio-Technical Design’ advocated for ongoing design and redesign of socio-technical systems, recognizing the evolving nature of human-technology interactions. This perspective underscores the importance of human-centered design and systemic adaptability.

Another one is Game-Theoretic Control in Socio-Technical Networks by Zhu and Başar (2024), which explores the application of Game Theory to manage socio-technical systems. It addresses like misinformation and infrastructure resilience. Their work emphasizes aligning individual behaviors with system-wide objectives through frameworks like Stackelberg games and mechanism design.

3.5.1.2 Key Principles of STS and Relevance to the Research

I. **Interdependence of Social and Technical Systems**

One of the core principles of STS is the interconnectedness of social and technical systems. Organizations consist of social systems such as people and culture and technical systems such as tools, processes and technology, which are interconnected in a way that changes in one area affect the other. Therefore, it is crucial to have a holistic approach to designing and understanding complex systems that involve both social and technical components (Whetton & Georgiou, 2010).

In the present research study, the social systems are the individuals, groups or communities that engage in online hate speech and are targeted by it. The technical systems are the online platforms that are used for the spread of hate speech and the moderation systems such as the AI algorithms used to detect online hate speech. Both of these systems are highly interdependent on each other, such that the outcomes of one influence the performance of the other. For instance, if a particular community or group of people are targeted for hate speech in the social environment as a result of any prejudice towards them, then that particular community faces prejudice online as well through platforms like Twitter where they are targeted by the users. This indicates that any changes in the social system cause changes in the technical system and vice versa.

II. **Joint Optimization**

Another key aspect of this theory is joint optimization of socio-technical systems, which means that the social and technical systems cannot be optimized independently, rather both the systems need to be designed and optimized together (Clegg, 2000). This indicates that if the systems are not designed in harmony, then changes in one system such as technology can lead to potential harm or unintended consequences in the other system such as the human factor.

This is particularly relevant to this research as the AI systems are developed by keeping in view the broader social factors. As explained by Zowghi and Rimini (2023), the AI systems are not developed and deployed solely in a technical vacuum but they are designed and shaped by the values, behaviors and interactions of humans, teams and organizations in the social context. The AI-based decision making is greatly influenced by interconnectedness of the social and technical components (Dolata et al., 2021). This theoretical lens acknowledges that the AI algorithms in hate speech detection are designed and developed in integrating the changing linguistic norms as well as social contexts, which helps to find online hate speech.

III. **Adaptability and Flexibility**

This theory has another core principle, which speaks about adaptation to changing conditions. The socio-technical systems should be designed such that flexibility and modifications are introduced whenever needed without causing much disruption. This is relevant to this research as it focuses on the need for the AI systems to be flexible and adaptable to the constantly changing nature of language. The online hate speech changes rapidly, with new slang words and emerging languages, making it challenging to be detected by the AI systems. In addition, the AI systems need to be flexible to adapt to the various social and contextual interpretation of hate speech ensuring the accuracy of detection. If the AI models are flexible enough to be trained on data from diverse cultures and contexts, this enables them to detect hate speech from diverse contexts, and hence more effective identification of online hate speech.

3.5.2 Discourse Ethics Theory

Discourse Ethics Theory was primarily developed by German philosopher Jurgen Habermas in the 1980s. This theory mainly focuses on the role of rational discourse and communicative action in the process of establishing ethical norms and resolving conflicts. The central focus of this theory is that ethical norms should be validated through a process of mutual communication and dialogue where each party have equal chance of participation and sharing their perspectives (Murphy, 1994). In the present study, the ethical discourse is mutually decided by the online platforms, AI developers who are involved in training the models and the users.

3.5.2.1 Recent Development in Discourse Ethics Theory

Digital Discourse Ethics theory was introduced by Rafael Capurro, Charles Ess in 2009. Their main focus was to adapt Habermas’ principles to digital communication. The theory emphasizes that online discourse must focus on transparency, inclusion and accountability. Digital platforms have responsibility of fair and ethical online discourse. AI algorithms must support democratic dialogue instead of suppressing it. In addition, Mark Coeckelbergh, Thomas Metzinger, and recent European AI ethics bodies proposed AI and Discourse Ethics focusing on how the AI systems must follow discursive norms. They mainly focused on how AI must enable fair participation in public discourse. The online systems responsible for decision-making should be inclusive, explainable, and contestable. The automated decisions should be legitimate only if they could be justified in a discourse involving those affected.

3.5.2.2 Key Principles and Relevance to the Research

I. **Communicative Rationality:**

Discourse Ethics emphasizes the notion of communicative rationality where participants take part in the process of rational discussion and dialogue without any type of force, self-interest and manipulation (Metselaar & Widdershovan, 2016). This process requires that all the individuals share their perspectives about the standard moral norms and can express their views openly ensuring transparency and mutual understanding.

In context of this research, hate speech is a very complex phenomenon and there are varying conceptions of what hate speech is. Besides, there are challenges on the side of AI detection of hate speech in regards to the issues of free speech, bias, and cultural sensitivity. Thus, in determining what term falls under hate speech and what doesn't, the AI models must be trained based on data which is labeled by mutual discussion and decision involving all the AI regulators, developers, and users so that all the members are equally involved in decision making.

II. **Principle of Universality:**

The moral norms that the parties agree on must be universal in nature. In other words, they ought to be applicable universally, and they should be true for all the people regardless of their differences or social conditions (Bohnet, 1997). Ethical norms should only be relevant if they are accepted by all the affected parties through their participation in the discourse. The principle ensures moral norms are subjective but universal in the claim to the validity (Metselaar & Widdershovan, 2016).

This is relevant to this study because the principle of universality demands that the ethical standard required for AI systems in order to detect hate speech online should be both fair and applicable to everyone, without regard to their social, cultural, or personal identity. This means that the rules AI uses to discern hate speech should be based upon norms that are equally binding for all users, guaranteeing fairness and impartiality. Therefore, universalization principle provides that the parties concerned should agree equally to the detection process of hate speech through AI. It is the implication that all the actions that AI may take regarding online hate speech must be justifiable to all parties and must not contain any bias.

III. **Normative Ground for Ethical Action:**

This principle of Discourse Ethics provides a framework for moral behaviors, decisions, and actions. It is concerned with what people ought to do and the standards that should govern ethical decision-making. It provides guidance on how individuals in a society should act, what should be the moral principles and what should be considered ethical and unethical, just and unjust and right and wrong.

This is relevant to the present research as it deals with the ethical challenges of AI systems. In this case, the normative question of whether AI systems should identify certain comment as hate comment even if it might be part of free speech. In this situation, the AI developers and policy makers rely on the normative principle to justify their actions. Furthermore, when it comes to drawing the line between free speech and hate speech, it is specifically crucial as some may argue that hate speech should be flagged as hate speech as it is part of the ethical norm to prevent harm. On the other hand, some may believe that protecting free speech is equally important, so AI should be more lenient towards free speech. In such situation, the AI developers and policy makers mutually decide upon the ethical norms that are justifiable to all the parties involved and balance free speech and hate speech.

CHAPTER 4

DATA ANALYSIS

This chapter presents both the quantitative and qualitative analysis. Section one presents the quantitative data analysis which is followed by qualitative data analysis in section two. Both the analyses are grounded in the theoretical frameworks of Socio-Technical Systems Theory and Discourse Ethics Theory.

The quantitative analysis analyzes the data collected from various online sources. It examines the numerical data to explore the linguistic features and patterns of online hate speech. The researcher analyzes the most salient lexical and semantic features of hate speech as well as the presence of different categories of hate speech across different online platforms. On the other hand, the qualitative analysis explores the relationship between AI and linguistics in the detection of online hate speech, highlighting the linguistic challenges and ethical issues that arise in the process. This analysis primarily focuses on how AI models can accurately identify hate speech while navigating the intricacies of language and upholding ethical standards related to privacy and free speech.

4.1 Section 1: Quantitative Analysis

The analysis further explores the integration of ethical principles in AI systems are integrated into AI systems to identify hostile content. The use of descriptive statistics allows for the thorough understanding of online hate speech patterns and its characteristics, indicating its frequency and distribution. The findings in this section are then discussed in terms of their implications for AI-driven solutions aimed at combating hate speech online.

4.1.1 Data Collection

The researcher gathered data from multiple datasets available on Google Data Search and Kaggle. The data comprised of hate speech available on online platforms. Some of the datasets have combined hate speech with non-hate speech and have included clear labels marking indications which comments are hateful and which are not. In some datasets, only hate speech comments were found, and very few had labels. The researcher ensured data integrity by cleaning the data, so the redundant information or the one that did not have relevance was removed, and only hate speech content processed. After the cleaning process, the researcher came up with a new list in which statements of hate speech were compiled from the original datasets. After completing the dataset, the researcher classified the comments according to their lexical and semantic features. Thus, different categories of hate speech were identified including racism, sexism, and religious intolerance. The process of categorization allowed for a much more sensitive understanding of the different types of hate speech found in the given dataset. This in turn then provided strong ground for further analyses in patterns and characteristics behind online hate speech.

4.1.2 Data Analysis

After collecting hate speech comments, they were categorized based on lexical and semantic features, with a focus on the language used to target individuals or groups. By examining specific words, phrases, and the context of these comments, distinct categories of hate speech were identified, providing insights into the linguistic patterns and features on online hate speech. These categories are presented below. The data is analyzed at three different levels for each category, i.e. lexical, semantic, and theoretical to gain deeper insights into the linguistic features of hate speech.

4.1.2.1 Racism

To categorize comments as racial based on their lexical and semantic features, specific language elements that target or demean individuals based on their race are analyzed. For example:

a. “All these niggas are ruining this country. Send them back where they belong.” (Twitter)
b. “I can’t believe they let coons like him get into universities.” (Twitter)
c. “These brown people are stealing all the jobs. They don’t belong here.” (Facebook)
d. “Black people are lazy and always looking for handouts.” (Instagram)
e. “They should just deport all Mexicans; they are nothing but criminals.” (Facebook)

A. Lexical Analysis

i. **Slurs and Derogatory Terms:** Phrases like the words such, “niggas”, “coons”, “brown people” and “Mexicans” are either racial epithets or a sweep offensive generalization. Even young boys calling others, “niggas,” are using a word still steeped in the legacy of the oppression of Black people “Who are coons”. This is a derogatory referring to black people particularly in a disrespectable context. In its turn, “brown people” and “Mexicans” are used here in a condensing and dehumanizing manner.

ii. **Verbs and Actions Words:** “Ruining”, “stealing”, “send”, “deport”, “belong”, “criminals” etc. are the words that build a negative image of the racial groups that were mentioned earlier. The verbs “ruining” “stealing”, and “deport” suggest that there is some problem that these people are notorious for, that is they are involved in crimes.

iii. **Polarized Language:** The use of sweeping generalizations (“all Mexicans,” “all the jobs”) makes the language more divisive, promoting broad and unjust conclusions about entire groups based on limited stereotypes.

iv. **Exclusion and Belonging:** Such sentences like “should just deport” and “Send them back where they belong” are most emphatic as though to remove people from that place since they do not belong there.

v. **Possessive Pronouns:** It is also clear from phrases such as “this country” and “our jobs” that the people who make these claims feel that certain ethnic communities do not belong to or have a right in the country or that certain ethnic groups dominate employment in the country.

B. Semantic Analysis

i. **Cultural and Racial Superiority:** By asserting that certain groups “don’t belong,” the speaker implies a form of racial hierarchy, positioning white individuals or those not targeted as the rightful inhabitants or dominant class. The comments suggest a clear division.

ii. **Stereotyping:** These comments activate common racial stereotypes. For instance: “Black people are lazy and always looking for handouts” refers to a stereotype historically used to justify inequality and oppression. “Brown people are stealing all the jobs” invokes the stereotype of immigrants as job-stealers and economic threats. “Mexicans are nothing but criminals” perpetuates the false stereotype of Mexican people as criminals, playing into xenophobic tropes.

iii. **Violent or Hostile Language:** While not overtly violent, there is an under-current of hostility that suggests potential violence through deportation or exclusion (“deport,” “send them back”).

iv. **Fear-mongering:** These comments create a narrative of fear and threat where minority groups are presented as responsible for societal decline (“ruining this country,” “stealing all the jobs”), perpetuating racial tension and xenophobia.

C. Theoretical Analysis

STS suggests that technology and society are intertwined, with the social aspects (culture, norms, language) and the technical aspects (tools, systems) co-shaping each other. In the context of online hate speech, this framework helps analyze how technology facilitates, amplifies, or constrains certain social behaviors, like the spread of hate speech. Discourse Ethics, as articulated by Jurgen Habermas, revolves around the principle that communication should follow specific ethical guidelines fostering mutual respect, truthfulness, and inclusivity. It underscores that moral norms arise from rational dialogue, where every participant has an equal opportunity to express their views and acknowledges the perspective of others. Various key tenets of Discourse Ethics are violated by hate comments.

i. **Technological Affordances and Amplification of Hate Speech:** From a Socio-Technical Systems Theory perspective, digital platforms function as interconnected systems**,** where technological capabilities interact with social behaviors, sometimes with unintended consequences. The lexical use of racial slurs and generalizations in the comments (“niggas,” “coons,” “brown people”) mirrors how digital platforms enable people to propagate harmful words fast and anonymously. Technology enables hate speech that was to be amplified through algorithms that did not differentiate between harmful and neutral content, promoting posts based on engagement, regardless of ethical considerations.

ii. **Social Structures and Exclusion:** The language analyzed promotes social exclusion “Send them back where they belong,”. They should just deport all Mexicans”), which ties into how socio-technical systems can reproduce societal inequalities. Online platforms are the reflection of offline power dynamics, with hateful content targeting minorities. This indicates the reinforcement of systemic social exclusion and prejudice, indicating that digital spaces are not neutral; rather they are motivated by and perpetuate already existing social norms of prejudice. As Baskerville and Fischer (2020) argue, socio-technical systems are “adaptive, evolving continuously as users and machines co-shape each other’s behavior” (p. 14), highlighting how user actions and technological systems reinforce one another in amplifying harm.

iii. **Automation and Dehumanization:** The expressions “lazy,” “criminals,” or “stealing all the jobs” represents an inclination to dehumanize people, reducing them to harmful stereotypes. Technology then can further amplify this thinking because of the lack of contextual understanding. Socio-technical systems are often unable to identify the complex social identities of individuals, breaking down their relationships into hurtful categories. To conclude, STS shows how these comments and the mediums in which they exist evolve together showing that technology can be both an enhancer of social stratification and an arena in which harmful discussion thrives.

iv. **Exclusion from Rational Discourse:** The comments reinforce exclusion of the participants in the discourse. The self-ascribed rights of legitimacy for certain individuals’ existence in society puts a vociferous closure on rational debate, core to discourse ethics. The silenced or labeled “outsider” status given to such marginalized groups negates the very kernel of discourse ethics: equality and inclusion in the conversation.

v. **Lack of Mutual Respect:** Hate speech inherently violates the mutual respect principle. Words such as “coons,” or “lazy” are specifically meant to degrade and dehumanize. Such speech for Habermas must always be aimed to gain understanding and to build consent. These comments foster hate and division while creating an environment that can never lead to constructive exchange.

vi. **Lack of Truth and Fairness:** Discourse Ethics also puts much emphasis on truth and fairness, as argued “in digital ethics, it is not enough to ensure procedural neutrality; we must strive for communicative justice grounded in shared understanding and transparency” (Capurro & Ess, 2009, p. 31). False generalizations (“Brown people are taking all our jobs,” “Black people are lazy”) appear in the statements spreading poisonous stereotyping. Truth or reason does not serve as their basis but biased social yarns that bear distorted truths regarding the real situation. Honest expression is no longer maintained over such unethical requirements.

vii. **Denial of Reciprocity:** The Discourse Ethics calls for openness and understanding of the other side when approaching a discourse. As Habermas, (1990), claims "The telos of human speech is understanding, not success" (p. 288). The comments lack the mutual understanding and ethical grounds and reflect prejudices towards the other person. There is lack of understanding the perspective of others.

4.1.2.2 Islamophobia

The comments showing hatred towards Islam reflect an alarming trend of the use of this language. The Islamophobic discourse is marked by narrative driven fear and exclusionary behavior. Such rhetoric is grounded in stereotypes related to Muslims, with an objective of strengthening the notion that they are outsiders who are not entitled to acceptance within society.

a. "Those towel heads are ruining our neighborhoods. They don't belong here". (Facebook)
b. "All these sand niggers are terrorists; we need to keep them out of our country." (Twitter)
c. "You are just another shitlord Muslim trying to impose your beliefs on us." (Twitter)
d. "Muslims are nothing but extremists waiting for their chance to attack us." (Twitter)
e. "Every mosque is a breeding ground for terrorists; they should all be shut down." (Instagram)

A. Lexical Analysis

i. **Derogatory Terms:** The remarks used to dehumanize people based on their religion include; "towel heads" and "sand niggers". "Towel heads" mock those who wear traditional head coverings trivializing a vital cultural or religious activity. In the same manner, "sand niggers" is a racial slur used against Arabs and Muslims with racial and religious hate blended together to demean a person based on his race and religion.

ii. **Negative Verbs of Action Words:** Through the expressions “ruining” and “terrorists,” Muslims are portrayed as being harmful or dangerous by nature. Phrases like “ruining our neighborhoods” and “terrorists” indicate that Muslims are a threat to the safety of society, instill fear, and frame them as a negative presence. This narrative is built upon long-standing stereotypes that have linked Muslims with violence or destruction.

iii. **Exclusion and Alienation:** The slogans “They don’t belong here” and “keep them out of our country” support the notion that the Muslim are strangers thus fostering social exclusion. The words promote “us against them” mentality that Muslims are aliens to the society involved and must be kept out. Such rhetoric paves the way for an environment in which Muslims are considered unwanted furthering their marginalization process. These include statements such as “all these sand niggers” and “keep them out.” Such statements, if taken to extreme meanings, would mean all Muslims are terrorists or threats. Such general claims provide grounds for dividing lines by instilling fears about a specific group because of the acts of only some individuals. This overall results in exclusion and alienation.

B. Semantic Analysis

i. **Social Exclusion:** Phrases like “They don’t belong here” contain exclusionist language that propounds Muslims should not be allowed to occupy specific geographies or cultures. The language forms the vision of “eternal alien” and then touts that Muslims are anti-norm forever. The remarks made on portraying Muslims as aliens who “destroy” neighborhoods or are dangerous grow the feeling of hate towards the Muslim community.

ii. **Cultural and Religious Superiority:** The definition of the act of Muslims “trying to impose their beliefs” or “don’t belong” reflects the speaker’s perception that his cultural or religious identity is superior and that what the Muslims are practicing or believing poses a threat. This kind of discourse relies on a belief in cultural superiority with Islam occupying the threatened, dangerous, or inferior category that needs to be held at bay or rejected.

iii. **Stereotyping and Fear-Mongering:** These remarks feed on stereotypes that portray Muslims as violent extremists or terrorists, especially when such words are used as “All these sand niggers are terrorists.” Such frightening creates a feeling of fear by feeding into negative views of Muslims because they are considered dangerous.

iv. **Hostility and Threats:** Although not violent, the tone of these remarks is hostile, indicating Muslims ought to be feared and out of society. The claim that Muslims “ruin neighborhoods” or “force their beliefs” carries within it an implied threat that Muslims are to be excluded or mistreated based on the belief that they pose a societal risk.

C. Theoretical Analysis

STS explains how society is impacted by the development of technology, particularly social media; it provides a lense to analyze the ways in which hate speech against Muslims is spread online. It describes some of the effects the online environment has on the use of harmful speech against Islam.

i. **Technical Amplification of Hate Speech:** In the current era of technical growth, a number of ethnic slurs can be disseminated extensively and anonymously. People are encouraged to interact with the contents of magazines and social media platforms, but unfortunately, doing so reinforces the negative sentiments that these outlets hold about Muslims. Consequently, this amplifies hate and intensifies the victims’ experiences of being subjected to hate speech in the media.

ii. **Reinforcement of Social Exclusion:** Such systems display social relegation processes which are facilitated by the linguistic subsets under discussion. The Internet, along with the deafening silence, is a reflection of the hierarchies that already exist in society, where certain groups like Muslims are more vulnerable to the acts of aggression than others. Hate speech is made easier in the virtual world by combative inclusions conveyed through comments that say “They don’t belong here,” these in turn, reach a larger audience and are supported by a community.

iii. **Dehumanization Through Technology:** Muslims being portrayed through such harmfully stereotyped words as “terrorists” is an example of how new media perpetuates rhetoric that dehumanizes people. The notion that identity problems could ever be reduced to an overly generalized or simplistic classification as something negative, Islamophobic discourse itself reduces complexity, and technology only serves to expedite the spread of ideas without taking into account their subtleties. Socio-technical systems often fail to bring dehumanization accompanying hate speech to the forefront and instead allow the damaging language to seed itself.

iv. **Exclusion from Rational Discourse:** In the form of comments, such as "They don't belong here" and "keep them out of our country," there is exclusion of Muslims from equal participation in social discussions. Such exclusion goes in contradiction to the value of inclusion found at the heart of Discourse Ethics which propounds equal voices in the discussion. Such language further suppresses all avenues of rational discussion by excluding Muslims from participating in societal, cultural, or political discussions.

v. **Violation of Mutual Respect:** According to Discourse Ethics Theory, everyone should actively practice empathy and consider the viewpoints of others as well. As Habermas (1990) argues that "only those norms can claim to be valid that meet with the approval of all affected" (p. 66), The remarks exhibit no such inclination. Instead, they propagate entrenched, intolerant attitudes that ignore the dignity and opinions of Muslims. Such one-way interaction increases the chances of the audience comprehension even more, thereby violating the tenets of Discourse Ethics even further.

4.1.2.3 Ethnicity

Since statements about ethnic hate contain hateful content towards a particular race, such comments are classified under ethnic hate speech as they are expressed in disparaging words advanced to people of certain ethnic groups. For example:

a. "Throw these chinks out of our country, they are aliens" (Twitter)

b. "All these beaners are stealing our jobs and ruining the economy." (Facebook)

c. "All Indians smell bad and don't know how to live in a civilized country." (Facebook)

d. "These Africans are thieves and scammers." (Instagram)

e. "These Asians are taking over our schools and businesses; they should stay in their own countries." (Twitter)

A. Lexical Analysis

i. **Discrimination and Devaluation:** Using words like “steal” and “aliens” implies that these ethnic minorities are negative or even intolerable in certain sphere. Some expressions like “these chinks should be out of our country”, “stealing all the jobs” and “taking over our schools and businesses” support the idea that this group does not deserve of any place in that society encouraging their ostracization and marginalization.

ii. **Exclusion and Belonging:** Such phrases as “all Indians are flatulent” or “ethnic toilet travelers are thieves” and “scoundrels” irrevocably ruin the dignity of a certain ethnic group and irrespective of existence of positive instances work towards creating the disharmony among the people.

iii. **Boundary Setting and Group Inclusion:** The statements “they are aliens” and “throw” cut across and move towards ethnic intolerance as the referents of these phrases are ethnic alien people or invaders. It asserts a hierarchical identity to the in-group and justifies the exclusion of the out-group by ethnic distinction.

iv. **Slurs and Derogatory Terms:** These include the use of the words “chinks,” “beaners,” “African” and “Indian” which in the ethnic context are intended to insult a particular group of ethnic backgrounds. The term “chink” is a pejorative term for people of East Asian descent, “beaner” is directed toward individuals of Hispanic or Latino descent, while “Gypsies” is often intentionally used in a derogatory sense as a stereotype of the Roma community as thieves or outsiders.

B. Semantic Analysis

i. **Ethnic Exclusion and Marginalization:** Quotations like “they are aliens” and “throw them” tell a story of exclusion where certain ethnic groups remain outside as if not belonging rightfully to the society. Such rhetoric does indeed bring about more social and cultural exclusion and marginalizes ethnic minorities.

ii. **Stereotyping and Dehumanization:** Using slurs such as “beaners” or “Africans” takes the entire ethnic group and reduces it to a stereotypical image that is less than desirable which might be that of a criminal or economic threat. Comments made with regards to “Africans are nothing but thieves and scammers” dehumanize the group as if they do not have any trusting essence and are dangerous in intent.

iii. **Cultural or Racial Superiority:** Conceit sustains belief in cultural or racial superiority, the belief that ethnic minorities” do not belong” and are “stealing” resources. The words describe that it is proper for the race of the speaker to be the rightful owners of resources, jobs, or a piece of cultural space while ethnic minorities are seen as an inappropriate intruder in such a space.

iv. **Economic and Social Threat:** Such language as “stealing all the jobs” galvanizes an economic and social fear that again launches the ethnic minorities as threats to the security of jobs or of social stability.

C. Theoretical Analysis

i. **Amplification of Ethnic Hate Speech:** Online platforms enable the rapid creation and circulation of ethnic hate speech. Algorithms on these sites function on the principle of amplification of engagement attached to the content; inflammatory posts are therefore amplified as they go viral to normalize ethnic hate speech. As Habermas (1990) maintains, ethical discourse must allow all participants an equal voice under conditions free from coercion or marginalization. When online platforms permit or ignore dehumanizing language, they violate this principle, allowing dominant groups to reinforce exclusionary narratives.

ii. **Normalizing Ethnic Exclusion:** The normalization of ethnic prejudice on digital platforms is a product of both technological affordances and sociocultural dynamics, as explained by Socio-Technical Systems Theory. Online communities begin sharing prejudice and hate speech against any ethnic group normalized on such digital platforms. Comments that tend to categorize “all Indians as smelling bad” or “send them back” circulate freely and repeatedly, they not only reflect existing societal biases but also become entrenched through algorithmic reinforcement**.**

iii. **Systematic Ethnic Marginalization:** STS shows how technology exemplifies the next modes of social organization offline, further marginalizing ethnic minorities. Social media sites facilitate more participation that consolidate exclusionary practice as hate speech motivated by ethnicity is condoned or encouraged by many within the particular community. The rhetoric that ethnic groups are “stealing” jobs or resources offline in comments is a reflection of exclusionary patterns denying groups equal access to opportunities offline.

iv. **Disdain and Disparity:** The terms “chinks” and “beaners” breach the core ethical principle that underpins Discourse Ethics. There is a need to understand that such words have only been coined to belittle, insult, and obliterate all forms of courtesy and sensibility.

v. **Inability to Participate in Rational Discourse:** Certain verbal assaults associated with ethnic bias normatively forbids individuals from being treated as equal in public dialogues, exemplified by phrases such as “they do not belong in this place” and “deport them all.”

4.1.2.4 Sexism

The comments analyzed here trace how sexist language perpetuates damaging stereotypes, keep certain genders out of a social and professional space. For example:

a. “All these faggots are destroying the sanctity of marriage. They don’t deserve equal rights.” (Twitter)

b. “You’re just another tranny trying to pretend you’re a real woman. No one will ever accept you.” (Twitter)

c. “Women are too emotional to be leaders. They should stick to raising kids and leave important jobs to men.” (Instagram)

d. “All these dykes are disgusting, acting like they’re men. They need to be put in their place.” (Facebook)

e. “Nobody takes a ladyboy like you seriously; stop pretending you’re something you are not.” (Instagram)

A. Lexical Analysis

i. **Derogatory Terms:** Words like “faggots,” “tranny,” and “dykes” are slurring that target to insult and belittle a person because of his or her sexual orientation or gender identity. “Faggots” is for gay men, “tranny” and “ladyboy” are used in reference to transgendered individuals, while “dykes” have been used towards lesbians. All these make people into names or labels tagged to their gender or sexual identity leaving behind the whole element of human beings. Other terms used for women are “slut”, “simp”, and “bimbo” that reflect women as unintelligent and silly.

ii. **Action Words:** The action words involved in such expressions are: “act like you are a real woman”, “women should stay in the kitchen”, etc. The message is that gender identity or expression does not exist; therefore, there are certain roles one has to play according to his/her gender. Language that exhibits traditional gender identities reflects the message that people who do not live like the traditional gender identity roles need to be marginalized.

iii. **Polarizing Language:** Use of phrases such as “women are too emotional to be leaders” and “faggots destroying the sanctity of marriage” generalize groups based on stereotype, depicting them as lesser or evil rather than ordinary individual contributors or behavior.

iv. **Exclusion and Belonging:** Phrases like “no one will ever accept you” and “stay in the kitchen where you belong” encourage the notion that specific genders or identities need to be excluded from public or professional life. Such language creates exclusion based on gender roles, thus substantiating social inequalities.

B. Semantic Analysis

i. **Gender-Based Exclusion:** Phrases such as “women should stay in the kitchen” and “too emotional to be leaders” perpetuate the notion of a natural lack of qualification of women to take on responsibility or assume exposure in society. This rhetoric aims at including less women participation and has secluded them into imprinted traditional roles lodged in the old stereotypes.

ii. **Dehumanization and Stereotyping:** Terms such as “tranny” and “faggots” strip people down to the most essential aspect of sexual orientation or gender identity; it advances even further in removing human aspects. Such terms press in the reinforcement of debilitating stereotypes for the function of labeling identities as deviant.

C. Theoretical Analysis

i. **Amplification of Digital Spread:** The anonymity of the digital platforms coupled with the tendency of posts to be viral tends to fuel sexist language spread. Sexist comments, such as “faggots” or “women should stay in the kitchen,” get shared and liked or commented on in a short time frame and can become part of a viral methodology through which misogynistic language or homophobic language can traverse instantly. Socio-Technical Systems Theory posits that these phenomena are a result of the co-evolution of technical design and user behavior, with platforms that prioritize visibility over accountability amplifying the social reproduction of detrimental gender norms (Baxter & Sommerville, 2011). Traditional gender inequalities are reinforced by the viral design of these systems, particularly in digital areas where such viewpoints are consistent with prevailing cultural values.

ii. **Exclusion of the Marginalized Genders:** STS reveals ways in which technology reflects the offline biases the electronic sphere that can enhance gender differences within the digital world. The exclusionary saying of "no one will ever accept you" mirrors broader societal patterns that marginalize those who do not fit within rigid mainstream gender norms making digital spaces another place where such practices are heightened and reinforced.

iii. **Violation of Mutual Respect:** Such words as "faggots" or "dykes" are in fact a part of sexist language which directly violates a principle so fundamental to ethical discussion that cannot be elicited from anybody using such words, dehumanizing and insulting people in terms of gender or sexual orientation, leaving no room for respect.

iv. **Exclusion from Rational Discourse:** Comments such as "women are too emotional to be leaders" and "real women don't act like that" bar individual exclusion from being seen as rational and capable participants in discourse. From a Discourse Ethics perspective, this violates the foundational principle that public communication must uphold mutual respect, dignity, and moral accountability (Habermas, 1990). When misogynistic language circulates without challenge, it creates an environment where certain identities are excluded from ethical dialogue**,** undermining the possibility of equal participation in discourse.

v. **Reaffirmation of Gender Hierarchy:** Sexist remarks reaffirm the systematic hierarchy that insists on declaring certain genders, either male or female, to be inferior in capacity or legitimacy. Such a person who praises discourse ethics propounds equity and equal rights; however, a remark that reduces a person to his or her gender or sexual identity defies these notions by emphasizing inequality and exclusion.

4.1.2.5 Political Hate Speech

The comments analyzed here demonstrate how political hate speech reinforces division and escalates hostility, particularly in online environments. For example:

a. "All fascists like you should be locked up. You're a threat to democracy." (Facebook)

b. "Anyone who votes for this traitor is an enemy of the state." (Twitter)

c. "These left-wing snowflakes are destroying our country with their idiocy." (Instagram)

d. "All liberals are traitors to the nation and should be silenced." (Twitter)

e. "Conservatives are brain-dead idiots who only know how to ruin the country." (Facebook)

A. Lexical Analysis

i. **Using Pejorative Labels:** The words "fascist", "brain-dead" and "traitor" are used against political opponents to make them look bad or less important. This is a term associated with extreme right-wing ideologies where certain political figures or political beliefs are compared to fascism. Mainly the term "traitor" carries with it the sticky connotation of treason which suggests that the people who have been targeted here are undermining the very state or cause for which they seem to be claiming allegiance.

ii. **Verbs and Action Words:** Expressions such as "be locked up", "destroying", and "idiocy" put an aggressive spin to the discourse. They present a desire to go beyond normal civil disagreement suggesting punishment or even grievance of political ideologies. This style of language serves to give credence to the notion that there is a real threat posed by political enemies and it is the only appropriate way to deal with them.

iii. **Polarizing Language:** Using broad statements like, "all of those left-wing snowflakes", "everyone who votes for this traitor!" and "conservatives are brain-dead idiots" tend to stereotype whole political parties in terms of their caricatures and leads to further aggravation of the factionalism and political intolerance.

iv. **Inclusion and Exclusion:** Phrasings such as "should be locked up" depict a tendency to stigmatize as unacceptable within a given political or societal sphere, certain categories of individuals defined by specific political ideas.

Such comments encourage the view that political rivals must be excluded and eliminate any chance for cohabitation within a pluralistic political framework.

B. Semantic Analysis

i. **Social and Political Exclusion:** Statement of the kind, “everyone who votes for this traitor is a state enemy”, implies that there are certain people who must be kept out. This narrative allows the representation of the rival in politics as a threat and consistency is maintained by using the idea of good and evil politics.

ii. **The Fighters are Illegitimizing their Opponents:** If the speaker is calling a certain person a fascist or a traitor then that person is made to look even worse from a political standpoint in that this position is apparently not just wrong, it is also dangerous.

iii. **Stereotyping and Fear-Mongering:** Here is how the political nickname “left-wing snowflakes” describes an entire ideology, giving the impression that it is weak or irrational. Expressions such as “will destroy our country” infuse urgency and instill fear. Such language makes audiences tap into their anxieties to perceive opponents of politics as instantaneous threats.

iv. **Hostility and Threats:** Words, though not aggressive, like “should be locked up,” convey serious punishment and even imprisonment for political opponents. Such use of words propagates hostile atmosphere where political differences are received with hostility and exclusion.

C. Theoretical Analysis

i. **Technological Amplification of Political Hate Speech:** Political hate speech spreads rapidly over the internet platforms, mainly when political tension is at its peak. Once the people use algorithms that follow sensationalized contents and say “fascists” or “traitor,” they amplify similar comments. According to the Socio-Technical Systems Theory, technological structures are not neutral; rather, user behaviours and discourses both influence and are influenced by them (Baskerville & Fischer, 2020). In this case, the platform's architecture of virality and low accountability makes it easier for political hate speech to spread and thrive with no repercussion for the speaker.

ii. **Reinforcing Political Exclusion:** STS claims that online environments amplify, often exaggerate, political cleavages. In this way, the ability to provide echo chambers for individuals, political hate speech such as "left-wing snowflakes are destroying the country" is further legitimated within enclosed online communities. Such a space perpetuates exclusionary discourse, pushing existing wedges between parties deeper.

iii. **Exclusion from Rational Discourse:** Exclusion from the point of conversation arises by hate speech political talk such as "Anyone who votes for this traitor is an enemy of the state". With this, basic tenets of Discourse Ethics are violated, as such language fosters the objectives of inclusivity as well as respect by silencing and shunning political adversaries rather than engaging in rational discourse.

iv. **Violation of Mutual Respect:** Political hate speech violates the very respect that should be given to everybody involved in discourse. Words such as "fascist" and "traitor" dehumanize political opponents thus does not contribute to constructive dialogue and promote division instead of understanding. From a Discourse Ethics standpoint, such amplification contradicts the ethical demands of rational, inclusive, and respectful public communication (Habermas, 1990).

4.1.2.6 Abusive Language

The comments analyzed here show how abusive language undermines respect and civility, particularly in online interactions where anonymity emboldens people to use hurtful language. Some of such comments are:

a. "You're a complete moron. No one cares about your stupid opinions. "(Twitter)

b. "Only an idiot like you would think this way. Go crawl back into your hole." (Facebook)

c. "Shut up, you worthless loser. You don't matter to anyone." Instagram)

d. "You're a pathetic waste of space, and no one would care if you disappeared." (Twitter)

e. "You're just a filthy piece of shit that no one wants around." (Facebook)

A. Lexical Analysis

i. **Abusive Speech:** Expressions such as "idiot", "moron" "pathetic" and "filthy piece of shit" are purposely used as an attempt to insult a person. Such language reduces the very person they are speaking to, to one negative word cut intended to damage his/her esteem or place in society.

ii. **Word Types:** Action words such as the examples “crawl back in your hole”, “shut up” and “no one wants around” are disrespectful comments with the intent to subdue an individual. Such expression is powered with a tendency to suppress the other person, where the second person is led to submission or is made obscure.

iii. **Exclusion:** The comments use radical statements in order to categorize a person as worthless or unintelligent. Statements to the effect that “You don’t matter to anyone” and “no one wants around” serve to insult the person who is targeted and seek to reject their voice from the debate.

B. Semantic Analysis

i. **Abuse and Devaluation:** Abusive lexicon aims at degrading all the right. ‘You’re a complete moron” and “filthy piece of shit” sends the message that the target has low IQ, and psychologically undermines the victim. The intention here is to lower the person’s value to nothing besides what was used against him or her.

ii. **Social or Interpersonal Assaulting Remarks:** These types of comments aim for emotional damage. The attacker adopts labels like “loser”, “idiot” and “filthy” in order to hurt the individual behind them. It is rhetoric that displays the aggressor’s platform. And such aggression usually promotes the alienation of the victim and discourages active participation – including verbal aggression.

C. Theoretical Analysis

i. **The Issue of Anonymity and its Effect on Abusive Language:** On digital platforms, anonymity functions as a socio-technical affordance that profoundly changes how people communicate, frequently reducing social accountability and permitting more dehumanizing or abusive rhetoric. Statements such as "Crawl back into your hole" demonstrate how people who are encouraged by anonymity may use harsh language that they would normally refrain from using in person. Socio-Technical Systems Theory explains this phenomenon as the result of interactions between technical design (anonymous profiles, weak moderation) and social behavior, where technology not only reflects but reinforces social dynamics.

ii. **Reinforcing Abusive Dynamics:** Not interfering on the user's part, indirectly reinforces power dynamics where conversation is dominated by the abusive users. According to Socio-Technical Systems Theory, this reflects a misalignment between user behavior and technological structures, where platform algorithms prioritize engagement over ethical considerations.

iii. **Violation of Respect and Equality:** Abusive talk violates the principle of mutual respect in discourse directly. Within the framework of Discourse Ethics, mutual respect and equality are non-negotiable conditions for ethical communication. Abusive language, such as "shut up" or "you're a complete moron," directly violates these norms by aiming not to engage in dialogue, but to silence, insult, or humiliate the other party. As Habermas (1990) argues, valid moral discourse depends on the equal inclusion of all participants**,** where each speaker is treated as a rational and autonomous individual deserving of dignity.

iv. **Denial of Reciprocity:** Reciprocity refers to the expectation that each participant listens, responds, and seeks mutual understanding which is the core of communicative rationality in Discourse Ethics. As Habermas (1990) notes, genuine dialogue presumes a "symmetrical distribution of communicative rights," where all speakers are open to being challenged and to justifying their positions. Abusive language disrupts this balance by shutting down engagement and refusing the ethical obligations of listening and understanding**.**

4.1.2.7 Vulgarity

Vulgarity refers to the use of indecent language, often intended to shock or to offend. The comments analyzed here reflect how vulgar language disrupts respectful discourse by prioritizing shock value over meaningful interaction.

a. "You're a total motherfucker, always spewing garbage online." (Twitter)

b. "You're nothing but a cunt who loves to stir up trouble." (Instagram)

c. "I can't stand these shit-eating idiots who ruin everything." (Facebook)

d. "She's a whore who sleeps her way to the top; no real talent at all." (Instagram)

e. "You cunt; you look absolutely disgusting." (Twitter)

A. Lexical Analysis

i. **Obscene and Crude Terms:** Phrases such as “motherfucker”, “cunt”, “whore” and ’shit-eating’ are extremely offensive and have been invented for shock value. They do not serve the purpose of the dialogue, but aim at irritating and offending the other party.

ii. **Verbs and Action Words:** The swear words are mostly used together with the assertive phrases like “you’re a total motherfucker,” and “you cunt” which serve as direct offensive language. The use of such foul language with insults increases the degree of aggressiveness of the statement.

B. Semantic Analysis

i. **Shock Value and Offense:** The chief aim of obscenities is to offend and shock others. For words like “cunt” and “motherfucker”, the speaker shows no care for the social etiquette and the respect that one is supposed to show to others. The expressed language is aimed to disrupt the acceptable mode of communication and to incite anger.

ii. **Dehumanization and Disrespect:** The concept of vulgarity, in most instances, reduces people to beastly or uncouth expressions without respect for their status in society. Expressions such as “shit-eating idiot” are demeaning and reduce the individual to the lowest levels of humanity as well as rob the target of their respect and dignity.

C. Theoretical Analysis

i. **Amplification of Vulgarity:** In some instances, the course of events can lead to the increase in the deployment of vulgarity in the content due to the nature of content engagement, which is shortly driven in this case by the word “engagement” and content that is disrespectful or shocking. Expletives such as “motherfucker” or “cunt” tend to spread in cyberspace at an obnoxious rate where such obnoxiousness seems to be more valued over civility. The ethical basis of speech is undermined when digital communication turns into personal attacks, which prevents comprehension or consensus. The democratic potential of digital platforms, which depend on free and rational communication, is undermined by such speech in addition to degrading the interpersonal space.

ii. **Lack of Decency and Respect:** Vulgarity lacks decency and respect, two standards for which ethical discourse must provide. Usages of terms such as "cunt" and "motherfucker" simply throw away the dignity of others; there is no possibility of a reason or respectable address able to be found. This directly violates the core principle of Discourse Ethics, As Capurro and Ess (2009) emphasize, digital communication must not only transmit information but also uphold communicative justice.

iii. **Vile Discourse Destroys Constructive Communication:** According to Discourse Ethics**,** the goal of communication is to foster understanding through rational, respectful, and inclusive dialogue**.** Rather than contributing to rational discourse, vile discourse insults or belittles and prohibits any meaningful interaction or mutual understanding.

4.1.2.8 Harassment

The comments analyzed here illustrate how harassment undermines a person's sense of security and can escalate into real-life consequences, particularly when perpetuated in online environments where anonymity often shields perpetrators from accountability.

a. "I'll expose your private life to everyone if you don't shut up. You're finished." (Facebook)
b. "I'm going to find you and rape you. You better watch your back." (Twitter)
c. "You deserve to be killed for saying that. I'll make sure it happens." (Facebook)
d. "I'll ruin your life you piece of garbage. You're dead." (Instagram)
e. "I know where you live, and I'll make sure you regret everything you've ever said." (Twitter)

A. Lexical Analysis

i. **Threatening and Violent Language:** "expose" "rape," "killed," and "ruin your life" are direct threats of physical and emotional harm. These comments instill fear and are meant to convey that severe consequences will occur for the victim. Verbs like "find," "killed" "expose," and "make sure it happens" therefore, once again stress the fact that these violent threats will be carried out, raising the alarm on this language even further.

ii. **Hate Words:** Phrases like “watch your back” and “you’re dead,” these phrases are used to suggest the violence that the individual was targeted for. These words comprise warnings or ultimatums which shows how one must get severely harmed unless he or she does something to prevent the violence.

iii. **Dehumanizing and Polarizing Language:** The commenter in “You piece of garbage,” dehumanizes the victim reducing that person to something lesser than human. In this type of language, one is encouraged to find an aggressive power dynamic where the commenter believes he is justified in perpetrating violence because he perceives the other as worthless or deserving of harm.

B. Semantic Analysis

i. **The Immediate Threat to Harm:** Taunting statements such as “You deserve to be killed” and “You’re dead” are immediately threatening to the safety and well-being of the person. This repeated use of language that is violent in nature produces an atmosphere of threat, and it makes the victim feel she has little recourse or protection.

ii. **Psychological and Emotional Terror:** “I am going to ruin your life” and “I will find you” include not only physical harm but also emotional manipulation and control. These are threats of personal, social, or professional long-term damage. The harasser attempts to create an atmosphere where the victim is never secure on any level: physically and emotionally.

iii. **Coercion and Fear-Mongering:** The harasser says, “You better watch your back,” and “I’ll make sure it happens.” In doing so, the harasser is using coercion, forcing the individual to comply through fear. The statements imply that bad things will happen to him or her as a result of his or her actions not taking action or failing to act in order to continue the harassment.

iv. **Dehumanization:** “You piece of garbage” is used to dehumanize the victim. The speaker degrades the person to rationalize his intention of hurting them, thereby making it easy to commit harm. In dehumanizing the victim, the harasser dismisses any possible empathy or moral responsibility in committing the harassment.

C. Theoretical Analysis

i. **Amplification of Harassment in Digital Spaces:** Users can harass others with relative anonymity found online that emboldens some to threaten as violently as saying "I'm going to find you and rape you" type of statements without consequence. The design of platforms facilitates such fast amplification of harassment by often not adequately moderating or providing consequences for it. The identity of such perpetrators is never known, as anonymous accounts or fake profiles are used to make threats. That contributes to the systemic nature of cyberbullying.

ii. **Violation of Mutual Respect and Dignity:** Abuse fundamentally contravenes the point of mutual respect that Discourse Ethics esteems. Threats like "I'm going to find you and rape you" or "I'll ruin your life" take away the dignity of the victim and his or her ability to act. This kind of speech does not allow one to engage in respectful or rational discourse.

iii. **Exclusion from Ethical Communication:** The remarks "You better watch your back" and "You're dead" exclude the victim from reasonable, ethical communication by making the interaction into one of fear and domination. Discourse Ethics demands a free dialogue between people with respect and fairness towards each other, but harassment denies any such possibility reducing it to intimidation.

iv. **Dehumanizing Vocabulary:** Of course, saying the word "you piece of garbage" is completely dehumanizing and lacks full ethical communication. The malefactor destroys the two real pillars of Discourse Ethics - the principle of empathy and the principle of reciprocity in abusing his victim. Therefore, the harasser is totally guided by the goal of causing harm instead of being bothered by the ethical or moral value of his speech.

v. **Coercion and Threats:** Comments like "You deserve to be killed for saying that" are coercive, trying to make the person submit or comply with a threat. Discourse Ethics rejects coercion but instead allows discourse to be open, respectful, and rational. Such threatened comments betray ethical standards to be placed in presumptions made in communication.

4.1.2.9 Body Shaming

The comments analyzed here highlight how body shaming reinforces harmful societal standards and perpetuates the idea that worth is tied to appearance.

a. “Look at that fat cow. No wonder no one wants to date her.” (Instagram)

b. “You’re so disgustingly skinny, you look like a walking skeleton.” (Facebook)

c. “With that flabby body, you should be ashamed to wear anything tight.” (Twitter)

d. “Your double chin is gross, ever thought about losing some weight?” (Instagram)

e. “No one likes ugly people like you. Fix your face before you go out in public.” (Twitter)

A. Lexical Analysis

i. **Derogatory Terms and Insults:** Expressions like “fat cow,” “flabby,” “disgustingly skinny,” and “double chin,” meant to slight and humiliate someone for their body shape or body weight. Such terms degrade the identity of the person in light of their physical appearance and encourage negative body stereotypes.

ii. **Action Words:** Phrases like you should be ashamed” and “fix your face” refer to something being wrong with how you look. Such talk perpetuates social oppression toward unattainable beauty.

iii. **Polarized language:** When somebody says “no one wants to date her” or “you look like a walking skeleton” in making general judgments over everyone’s look as though, it is indeed bad. Such language divides the world into “acceptable” and “unacceptable” categories based solely on the way people look.

iv. **Exclusion and Belonging:** Expressions like, “no one wants to date her” and that “you should be ashamed to wear anything tight”, so it depicts the utter exclusion from social circles or romantic opportunities because of what you might look like. In this line of speech, exclusion forms a certain meaning where people who do not fit into societal beauty standards should not belong in spaces

B. Semantic Analysis

i. **Body Devaluation:** Such comments dictate how people’s lives would be devalued based on body appearance. Comments like “flabby body” and “fat cow” reduce an individual to negative body descriptors, continually asserting the false narrative that some body types are undesirable or shameful.

ii. **Reinforcing Beauty Standards:** The targeting of individuals who are tagged as either “fat” or “disgustingly skinny,” suggests that only certain body types are acceptable, while others warrant ridicule. This reinforces societal beauty standards to a damaging level, promoting unrealistic and unhealthy ideals.

iii. **Social Exclusion and Stigmatization:** Body shaming remarks take it a step further to give people the impression that those who do not fit the expected physical specifications are not good enough to love or be respected. Saying “no one wants to date her” or “fix your face before you go out in public” are statements emphasizing that only those who are aesthetically pleasing according to certain standards are granted social validation.

iv. **Emotional and Psychological Harm:** Language is intended to cause emotional harm in the body-shaming comments. Through insults to personal appearance and indicated shame over the body, the comments do not only cause psychological harm but are also being used to reinforce harmful narratives on worth or value through the use of physical appearance.

C. Theoretical Analysis

i. **Amplification of Body Shaming on Digital Platforms:** Social media amplifies body-shaming discourse by allowing the people to spread anonymous posts, memes, or viral content. This includes sharing comments such as “fat cow” or “disgustingly skinny” which can fast-track like, share, and comment momentum, having brutal impacts on those targeted. The problem is worsened by the socio-technical design of the platforms, where images and especially physical features are often strongly emphasized.

ii. **Normalizing Destructive Standards of Beauty:** On the internet, users enjoy endorsing and promoting in society’s beauty standards by using photo-shopped images, endorsed influencers, and gym guides. In these spaces of the internet, the norm of shame-ridden comments about a person’s body - remarks such as “fix your face” or “nobody wants to date her” is normalized. Instead of expressing outrage and rejection over this commentary, it is normalized and enacted in these online spaces where appearance-based judgments are considered normative.

iii. **Systemic Exclusion:** The offline beauty standards that create an exclusionary space for those whose bodies do not fit the ideal mold is, in many ways, reflected in this digital space. Comments that body shaming is a reflection of the way technology reinforces social hierarchies by excluding a person simply because they don't fit the ideal mold of what one likes, thus propagating a culture of only desiring that few bodies.

iv. **Disrespecting the Individuals:** Body shaming goes against the mutual respect principle which forms the basis of communication. Comments like "you look like a walking skeleton" or 'you should be ashamed' are made with the intention of being insulting and degrading rather than respectful dialogue. Such language makes inclusive and ethical discourse impossible.

v. **Stigmatization:** Such comments are not only offensive but also promote negative stigma on body size. For instance, who should respect or love someone defined as "fat" or "skinny"? Discourse Ethics emphasizes equity and understanding in any discourse. Utterances that reflect body discrimination are based on stereotypes which are unduly prejudicial and counterproductive to the objectives of engagement.

vi. **Devaluation of Human Life:** The use of body-shaming language subjectifies individuals by portraying them as merely their bodies. Expressions such as "fat cow" or "don't be flabby" pertain exclusively to one's physical body, thereby denying the person-nation and assigning a value towards a person based on such superficialities. Discourse Ethics principles are that every engagement should uplift the dignity of the human person in contrast those comments stripped the dignity of the target individuals.

4.1.2.10 Intrigator

These comments often use seemingly benign or less overtly offensive language but still serve to alienate, demean, or marginalize individuals based on their identity, beliefs, or actions. While less aggressive than other forms of hate speech, Intrigator comments perpetuate division and exclusion by masking hostility with faint praise or backhanded compliments.

a. “All spazzes are useless and should be kicked out of society.” (Facebook)

b. “These queers are ruining everything with their agenda. They should just disappear.” (Twitter)

c. “You’re such a batty boy, but at least you’re trying to fit in.” (Twitter)

d. “We may have different views, but you’re still a bit of a freak to me.” (Facebook)

e. “It’s good you’re trying, but stop acting like such a weirdo.” (Instagram)

A. Lexical Analysis

i. **Derogatory or Backhanded Terms:** The term “spazzes” or “queers” for ex- ample, other terms that are not acceptable in the current society, such as “freaks”, “weirdos”, and “batty boys” are all derogatory in nature for every single one of these words is directed towards a specific type of person’s physical characteristics, sexual orientation, or social standing. They are all negative and are used to portray people as abnormal or unwanted individuals. Certain expressions may seem neutral or even inclusive (“a batty boy” or “freak”), but there is an implicit judgment that it carries which is detrimental to people and groups and their concepts.

ii. **Backhanded Compliments and Patronizing Language:** Common phrases include such statements as “but at least you’re trying to fit in”, “it’s good you’re trying” which is a typical case of backhanded compliment. Although statements of these types may seem particularly supportive, they serve to belittle the target by further endorsing the fact that he or she is different or inferior.

iii. **Polarized Language:** The phrases stating “All the spazzes are useless and should be kicked out from the society” or “They should just disappear” are clear instances of exclusionary language that seeks to expel, shun, or suppress certain people regardless of their actions.

B. Semantic Analysis

i. **Implied Superiority:** Most of the comments go beyond the intended notions and have an element of superiority in them on the part of the speaker. For instance, “You’re such a batty boy, but at least you’re trying to fit in” indicates the tension of acceptance as the speaker classifies him or herself in the exclusive group while the subject is attempting to blend in. This social ordering glorifies and embeds the concept of a hierarchy whereby the subject is viewed as “less than”.

ii. **Subtle Marginalization:** While the comments above are not as out-and-out degrading as some others, they are peripheral at base; they comment on a difference in inherent fault or inferiority that lies at the basis of the attack. A fairly standard declaration such as “We may have different views, but you’re still a bit of a freak to me” recognizes the presence of differing views while degrading the validity of the other person’s identity or perspective with the word freak.

iii. **Stereotypes Reinforced:** The phrases like, “These queers are ruining everything with their agenda,” and” I hate how feminists and social justice warriors are turning the world into a mess,” rely on stereotypes to make this point. Feminists somehow come to be seen as inherent in disrupting society; these groups have harmful agendas and go against the stability of society. The very use of terms like “queers” or “social justice warriors” points to the inevitable nature of bias and disrespect for such groups.

iv. **Social Exclusion and Denigration:** The remark “All spazzes are worthless and should be kicked out of society” is a great example of social exclusion, very clearly communicating that persons with disabilities deserve to be thrown out of society altogether. In this language, what little persons with disabilities contribute is devalued, making their “worth” so small that people suggest that they should be excluded from societal participation altogether.

C. Theoretical Analysis

i. **Amplification of Hostile Stereotypes:** Intrigator comments often trigger amplification through social media; the less virulent forms of bias are unlikely to be detected by an automated moderation system. For instance, phrases like “but at least you’re trying” or “you’re still a bit of a freak” would not be kept back by content filters but represent the kinds of utterances that feed into a culture of exclusion and veiled aggression. The socio-technical architecture makes precisely these kinds of comments travel and are empowered by like-minded others who reinforce harmful norms.

ii. **Normalization of Passive-Aggressive Discourse:** The nature of STS, with the tendencies to normalize subtle manifestations of aggressiveness, as can be seen on the comments of Intrigator, evidences the backhand compliments and veiled jabs that are tolerated and in turn receive likes, shares, and even comments. This normalized low-level hostility and exclusion as acceptable forms of communication spread within cultures.

iii. **Reinforcement of Social Exclusion:** The electronic environment may propagate social exclusion by allowing some groups of people to marginalize others through mild forms of insult. In a comment that is quite amusing but inciting towards superiority for the dominant social group, the intrigator remarks, “We may have varying opinions, but you are still somewhat of a freak to me” promotes the dominant social groups towards a form of superiority while peripherally advancing the exclusion of those individuals who do not form a part of the “norm.”.

iv. **The Detriment of Mutual Respect:** Intrigator’s comments are aimed to destroy an integral part of any ethical argument – that of mutual respect. The soft bigotry of “You’re such a batty boy, but at least you’re trying” or “It’s good you’re trying, but stop acting like such a weirdo” is far removed from hate speech but still disparages the target. Such comments do not encourage exchange of ideas; instead they serve to belittle the target thereby making any fruitful engagement between the parties impossible.

v. **Lack of Reciprocity and Inclusivity:** According to Discourse Ethics, there should be inclusion and reciprocity such that every person’s voice is regarded as having the same legitimacy. “We may have different views, but you’re still a bit of a freak to me” and “stop acting like such a weirdo” are intrigator comments which do not allow genuine dialogue to happen as they label the other person’s viewpoint, as freakish. Such language shows that the speaker is not willing to understand the other side, thus blocking reciprocal communication.

4.1.3 Distribution of Hate Speech on Different Social Media Platforms

The tables below represent the hate speech distribution on the three main online platforms including Twitter, Facebook, and Instagram.

4.1.3.1 Facebook Hate Speech

Table 4 presents the hate speech distributed over Facebook. The total number of hate speech comments collected was 28,813.

Table 4

Hate Speech Categories on Facebook

Sr. No.	**Categories**	**Frequency**	**Percentage**
1.	Racism	6051	21%
2.	Islamophobia	5186	18%
3.	Political	4033	14%
4.	Body Shaming	2881	10%
5.	Religious	2305	8%
6.	Vulgarity	2018	7%
7.	Harassment	1729	6%
8.	Ethnicity	1729	6%
9.	Intrigator	1153	4%
10.	Sexism	864	3%
11.	Abusive	864	3%
Total		**28813**	**100%**
Mean		4.8200	
Median		5.0000	
Std. Deviation		3.26328	
Variance		10.649	
Skewness		0.338	
Std. Error		0.014	

Table 4 shows that majority of hate comments on Facebook consist of racism and that accounts for 6,051 comments, which is 21% of the total. Next in line is Islamophobia which constitutes 5,186 comments approximating to 18% of the hate speech. Political targeted hate speech is also a huge portion of the comments with 4033 comments (14%) indicative of the polarized politics and political conversations that take place on social media platforms such as Facebook.

There are also hate comments targeting an individual’s ethnic background as well as comments about bullying, both representing 6% apiece with 1729 comments noted un- der each category. Religious based hate speech comprises 8% of the pool noted with 2305 occurrences whereas the element of profanity in the pool consists of 7% comprising of 2018 comments. Body shaming, an increasingly prominent issue on social media, comprises 10% of Facebook’s hate speech, with 2,881 instances.

Sexism and abusive language both represent smaller portions of the data, each contributing 3%, with 864 instances in both categories. Intrigator hate speech, which involves the instigation or provocation of harmful behavior, is the least prevalent category, making up just 4% with 1,153 instances.

Measures of Central Tendency

The mean number of comments across all categories is 4.8200, with a median of 5.0000, indicating a balanced distribution of hate speech categories. The standard deviation of 3.26328 suggests moderate variability in the distribution of hate speech across different categories, while the variance of 10.649 shows that the dispersion is significant but not excessive. The skewness value of 0.338 indicates a slight rightward skew, meaning there are more categories with lower frequencies of hate speech and fewer categories with very high frequencies. The low standard error of 0.014 suggests a high level of precision in the data analysis, indicating that the sample is a reliable representation of the hate speech categories on Facebook.

4.1.3.2 Instagram Hate Speech

Table 5 presents the breakdown of hate speech categories on Instagram, where a total of 21,942 hate speech comments were analyzed.

Table 5

Hate Speech Categories on Instagram

Sr. No.	**Categories**	**Frequency**	**Percentage**
1.	Islamophobia	4388	20%
2.	Racism	3511	16%
3.	Body Shaming	2633	12%
4.	Abusive	2414	11%
5.	Religious	2194	10%
6.	Ethnicity	1755	8%
7.	Sexism	1536	7%
8.	Political	1536	7%
9.	Harassment	1097	5%
10.	Intrigator	878	4%
Total		**21942**	**100%**

Mean	4.7901
Median	4.0000
Std. Deviation	3.13478
Variance	9.827
Skewness	0.434
Std. Error	0.017

According to table 5, Islamophobia represents the largest portion of hate speech on the platform, with 4,388 comments accounting for 20% of the total. Racism follows, comprising 3,511 comments, or 16% of the hate speech data. Body shaming also features prominently, contributing 12% of the total with 2,633 comments, indicating that this form of hate speech is relatively common on Instagram.

Abusive language makes up 11% of the total, with 2,414 instances, while religious hate speech accounts for 10%, with 2,194 comments. Ethnicity-based hate speech is present in 1,755 comments, representing 8% of the total. Sexism and political hate speech both contribute 7% of the dataset, with 1,536 comments each.

Harassment accounts for 5% of the hate speech on Instagram, with 1,097 comments. Intrigator hate speech, involving provocation or encouragement of harmful behavior, represents 4% of the dataset, with 878 instances.

Central Tendencies of Mean, Median and Mode

The mean of comments per category is 4.7901, with a median of 4.0000, indicating that the majority of hate speech categories have relatively balanced frequencies. The standard deviation of 3.13478 suggests moderate variability across the categories, while the variance of 9.827 further illustrates the spread of hate speech data. The skewness value of 0.434 indicates a slight rightward skew, meaning there are more categories with lower frequencies and fewer categories with high frequencies of hate speech. The standard error of 0.017 reflects high precision in the data, ensuring that the analysis provides a reliable depiction of hate speech distribution on Instagram.

4.1.3.3 Twitter Hate Speech

Table 6 provides the distribution of hate speech categories on Twitter, where a total of 46,187 hate speech comments were analyzed.

Table 6

Hate Speech Categories on Twitter

Sr. No.	**Categories**	**Frequency**	**Percentage**
1.	Political	13394	29%
2.	Racism	8776	19%
3.	Islamophobia	4619	10%
4.	Religious	4619	10%
5.	Harassment	4157	9%
6.	Ethnicity	2771	6%
7.	Intrigator	2308	5%
8.	Abusive	1847	4%
9.	Body shaming	1386	3%
10.	Sexism	1386	3%
11.	Vulgarity	924	2%
Total		**46187**	**100%**
Mean		5.0898	
Median		6.0000	
Std. Deviation		3.05966	
Variance		9.362	
Skewness		0.252	
Std. Error		0.011	

Table 6 shows political hate speech is the most prevalent category, accounting for 13,394 comments, or 29% of the total. This in itself speaks to the relevance of political rhetoric within hate speech on Twitter, indicating how overly politicized Twitter is with regard to its environment.

The second largest category is racism which represents 19% of the comments at 8,776. This means that racial discrimination is one of the significant issues on Twitter, just like its reach in every corner of the world and users it has accumulated. Incidence of Islamophobia constitutes 10% of the hate speech with an incidence count of 4,619, and religious hate speech also represents 10% of the hate speech with an incidence count of 4,619. These suggest that religious intolerance, especially against Muslims, constitutes a considerable proportion of the problems it has on the platform.

The ethnicity-based hate speech is 6 percent of the total number, standing at 2,771 comments while abusive language was at 4 percent of the total with 1,847 instances. Harassment is also significant as harassment accounted to 9 percent of hate speech with 4,157 comments. Finally, intrigator hate speech that has inciting harmful behavior accounts for 5 percent of the total, coming in at 2,308 instances.

Sexism and body shaming are at par in the lead for hate speech on Twitter at 3% of the total with 1,386 comments on the platform in both categories. Vulgarity is the smallest part of hate speech on the platform, taking up 2% with 924 comments.

Central Tendencies

In table 6, the mean value is 5.0898 with a median of 6.0000 indicating a relatively balanced tendency. The standard deviation of 3.05966 explains a moderate variability across the categories. Moreover, the variance of 9.362 further illustrates that the dispersion is significant. The skewness of 0.252 indicates a slight rightward skew. The standard deviation of 0.011 highlights high precision in the data, indicating that the analysis represents the correct depiction of data.

4.1.3.4 Hate Categories Derived from Total Hate Comments

Table 7 provides the spread of hate speech comments categorized by type, with a total of 96,942 comments analyzed across 10 categories. The categories range from racism, Islamophobia, and ethnicity-based hate speech to body shaming, vulgarity, and harassment.

Table 7

Hate Categories Derived from Total Hate Comments

Sr. No.	**Categories**	**Frequency**	**Valid Percent**
1.	Political	18963	20%
2.	Racism	18338	19%
3.	Islamophobia	14193	15%
4.	Religious	9118	10%
5.	Ethnicity	6255	7%
6.	Body Shaming	6900	7%
7.	Harassment	6983	7%
8.	Abusive	5125	5%
9.	Intrigator	4339	5%
10.	Sexism	3786	4%
11.	Vulgarity	2942	3%

Total	**96942**	**100%**
Mean	4.9418	
Median	5.0000	
Std. Deviation	3.14155	
Variance	9.869	
Skewness	0.315	
Std. Error	0.008	

This data reveals a pattern in the types of hate speech found within the dataset as well as showing the scale of different types of hate speech. Political hate speech constitutes majority of this and it answers for 20% of the votes cast. This demonstrates the depth of the social divide within the framework of political communication as people tend to try and undermine and insult other people on the grounds of their political opinion and their political orientation. The abundance of remarks belonging to this category further implies that people have strong opinions about politics and that politics is one of the most common areas where hatred and real hostility respect speech.

The second clearly defined category is racism which comes within the scope of this dataset and comprises 19% of all comments. This high percentage indicates another wide problem in language-based hate speech happening online, and that is racism which encompasses name-calling and negative representations of certain races. The next most prevalent category is Islamophobia which accounts to 15% of all comments. 10% Religious hate speech intolerance towards individuals from a certain faith was detected by the researcher. This would include hate speech targeting other religious groups besides Muslims, and thus proves religion continues to be a divisive force in online environments.

Body shaming, harassment, and ethnicity-based hate speech each accounts for 7% of hate comments in the dataset. This presents the idea that all of these three share the same prevalence, so online harassment is not only pointed towards physical appearance, namely body shaming, but harassment and discrimination targeting ethnic identity. Although their percentages may appear to be the same, the four subcategories portray variations in how each of them independently contributes to the hate speech landscape and impacts people differently through various adverse effects.

It is abusive language, the remarks combining insults and simply derogatory which does not fall under any other category accounted for 5% of the total. Some commentators termed these comments “intrigator” subtle or not so obvious forms of exclusion and marginalization. While less aggressive they still can be seen to input to stereotype reinforcement and disease of injurious social norms.

Sexism, which includes derogatory language aimed at demeaning individuals based on their gender or gender identity, makes up 4% of the comments. This shows a smaller but still significant presence of gender-based discrimination in the dataset. Vulgarity, representing only 3% of the comments, involves the use of obscene or offensive language.

4.1.3.5 Distribution of Hate Speech Across Different Categories

The mean value (4.9418) represents the average frequency of comments across all categories. This suggests that, on average, the categories are relatively well-distributed, though some, like political hate speech and racism, are more prevalent than others. The median of 5 indicates that half of the hate speech categories have frequencies equal to or less than 5% of the total comments. This suggests a skew in the distribution, where a few categories dominate (e.g., political hate speech and racism. A value of 3.14155 of standard deviation suggests moderate variability, meaning there are significant differences between the prevalence of certain hate categories (e.g., political hate speech and vulgarity).

Variance (9.869), the square of the standard deviation, further illustrates the dispersion of data. The relatively high variance indicates that some hate categories have much higher frequencies than others. The positive skewness of 0.315 suggests that the distribution is slightly skewed to the right. This means that while most categories fall around the average, there are a few categories with higher frequencies (e.g., political hate speech and racism), pulling the distribution to the right.

The distribution of hate speech across these categories highlights the complex nature of online hostility, with different forms of prejudice intersecting and reinforcing one another. These findings suggest that political and racial issues are often at the center of online hate, though other forms of identity-based discrimination remain persistent. The data further emphasizes the need for comprehensive strategies to address the multifaceted nature of online hate speech.

4.1.4 Findings of Quantitative Analysis

4.1.4.1 Total Number of Collected Comments

Table 8 shows the distribution of total comments gathered from three social media platforms, Facebook, Instagram, and Twitter, using datasets obtained from Kaggle and Google Data Search. A total of 153,426 comments were collected.

Table 8

Total Comments Gathered from Facebook, Instagram, and Twitter

Sr. No.	**Platforms**	**Frequency**	**Percentage**
1.	Twitter	67842	44%
2.	Instagram	43562	28%
3.	Facebook	42022	27%
Total		**153426**	**100%**
Mean		2.1683	
Median		2.0000	
Std. Deviation		0.82931	
Variance		0.688	
Skewness		-0.322	
Std. Error		0.006	

According to table 8, Twitter contributes the highest percentage of comments (44%), followed by Instagram (28%) and Facebook (27%).

Summary of Statistics

The mean value represents the average number of hate speech comments per platform. A mean of 2.1683 suggests that, on average, the comments are moderately spread across the platforms. A standard deviation of 0.82931 suggests that the number of hate speech comments across platforms is moderately dispersed around the mean. The low standard error value shows that the estimate of the mean is precise and reliable, reflecting minimal uncertainty in the calculated mean value.

4.1.4.2 Total Number of Hate Comments

Table 9 presents the distribution of hate speech comments collected from three social media platforms—Facebook, Instagram, and Twitter. A total of 96,942 hate comments were collected, out of total 152,426 comments.

Table 9

Total Comments Gathered from Facebook, Instagram, and Twitter

Sr. No.	**Platforms**	**Frequency**	**Percentage**
1.	Twitter	46187	48%
2.	Facebook	28813	30%
3.	Instagram	21942	23%
Total		**96942**	**100%**
Mean		2.1792	
Median		2.0000	
Std. Deviation		0.86113	
Variance		0.742	
Skewness		-0.353	
Std. Error		0.008	

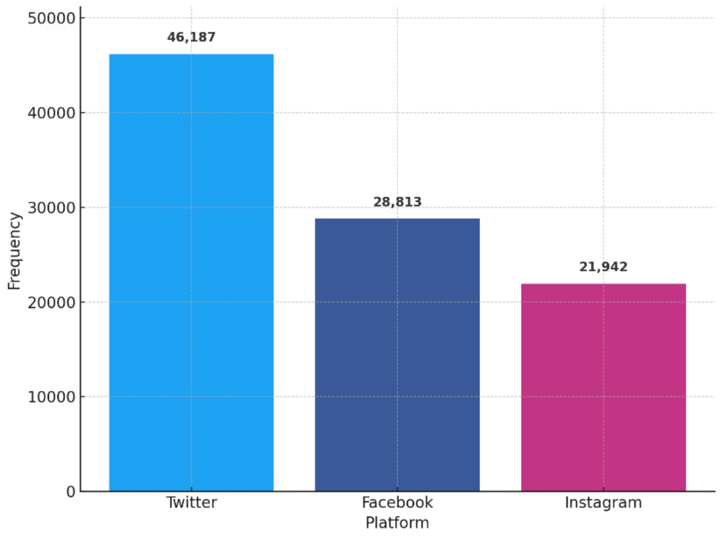
Out of the total 96,942 hate comments, 28,813 (30%) were sourced from Facebook. Instagram contributed 21,942 hate comments, representing 23% of the total. Twitter accounted for 46,187 hate comments, which is 48% of the total.

The dominance of Twitter in hate speech comments suggests that the platform may have structural features or policies that make it more conducive to hate speech, compared to Facebook and Instagram. Twitter's higher engagement levels, ease of information dissemination, and public nature may contribute to this outcome.

In summary, the lexical and semantic features revealed that hate speech on online platforms is characterized by derogatory slurs, negative generalizations, and exclusionary language particularly targeting race, ethnicity, and gender of individuals, groups or communities. Further, the analysis measured the frequency and percentage distribution of various categories and confirmed that political hate speech is the most prevalent, followed by racism, and religious hate speech. Descriptive statistics further highlights Twitter as the platform with the highest concentration of hate speech, amplifying harmful narratives more than other platforms. These findings underscore the urgent need for enhanced moderation policies and discourse ethics to combat the spread of online hate speech.

Figure 1

Distribution of hate speech instances across social media platforms based on collected data.



The quantitative results revealed that Twitter hosted the highest volume of hate speech instances among the platforms analyzed, with political hate speech emerging as particularly dominant. This supports earlier observations by Warner and Hirschberg (2012), who identified Twitter’s open-access structure and brevity-driven discourse as contributing factors to the proliferation of hateful content. Similar trends were observed by Cortiz and Zubiaga (2020), who emphasized the role of platform architecture in enabling the spread of toxic content, especially within political and ideological contexts. These findings reinforce the argument that social media platforms are not neutral spaces but actively shape the visibility and proliferation of hate speech.

4.2 Section 2: Qualitative Analysis

The qualitative analysis primarily focuses on the accurate detection of hate speech through AI while navigating the intricacies of language and not violating ethical standards related to privacy and free speech. The analysis is based on the themes extracted from the interviews conducted. The thematic analysis follows the six-step process outlined by Braun and Clarke (2006). First, the researcher read the text in order to get familiarized with the text of the interviews, then the initial codes were developed based on the key issues. Similar codes were then grouped into more general themes for analysis. These themes were analyzed on the basis of theoretical frameworks that helped in the exploration of the linguistic intricacies and ethical challenges that AI faces in the identification of online hate speech. Further, the analysis is grounded in two theoretical frameworks: Socio-Technical Systems Theory (STS) and Discourse Ethics Theory.

4.2.1 Analysis of Trends Emerged from Interviews

4.2.1.1 The Linguistic Complexity of Hate Speech

One of the major issues highlighted by the interviews is the complex nature of hate speech online. Hate speech can be both direct and indirect in different forms and situations. It may be simple, such as just insults and slurs, or very complex, like euphemisms, sarcasm, and dehumanizing analogies. Such diversity in hate speech makes AI detection more challenging since algorithms must be able to detect not only overt hate speech but also its more covert forms, which may be masked by sarcasm, humor, or seemingly neutral language. For example, while some hate speech cases are straightforward, many are complex and can be masked by sarcasm or culturally sensitive language, making them hard for AI systems to identify (Clarke et al., 2023). The interviews suggest that hate speech is represented by specific linguistic features. “Any form of communication. where it fuels or advocates for hatred, discrimination, or hostility towards people or groups” (Participant 1) is how participant from the first interview defined hate speech. This definition shows that the language used online is a reflection of social attitudes toward specific groups.

4.2.1.2 Challenges in AI Detection of Hate Speech

There are several challenges faced by the AI systems in the accurate detection of hate speech, mainly due to the complex nature of language, cultural contexts and technological limitations.

i. **Linguistic Complexity and Nuance**: AI systems have extreme difficulty in understanding the subtleties of language, such as sarcasm, slang, and changing linguistic patterns. During an interview, an interviewee said, “AI misses the subtext when the subject is not from the Western culture... it falls short when dealing with Eastern languages” (Participant 3). This shows that mainly the AI models trained based on the Western datasets would not be able to correctly identify hate speech in many linguistic contexts.

ii. **Cultural Context Variability**: Cultural background has a deep impact on the functioning of AI systems. The statement, “pre-determined models do not really work in other social contexts...we have not differentiated between free speech and hate speech,” was made by Participant 1. This claim signifies how cultural definitions of hate speech can be vastly different, therefore making it much harder to function in AI systems that had been trained in one particular set of cultural standards as opposed to another.

iii. **Technological Limitations**: Technical limitations are another challenge faced by AI in identifying hate speech. According to Participant 2: “The first one is the limitation of hardware availability... only limited models are trained because of the non-availability of hardware (Participant 2). In that regard, the scarcity of resources may hinder the formulation of sustainable AI systems, which can understand complicated patterns in speech.

4.2.1.3 Role of Linguistics in AI Detection

This theme brings to attention the linguistic analysis and how it may optimize the function of AI technologies designed towards hate speech recognition online. The interviews exhibit a wide range of several important concerns on this aspect, some of which encompass linguistic patterns recognition, recognition of differences in cultures and overcoming specific limitations that restrict AI present technologies. What linguistics contributes to detection in AI is the awareness of distinctive language patterns associated with hate speech. A participant commented, “The role of a linguist is to identify the linguistic patterns on the basis of which we can identify hate speech and free speech” (participant 4). This shows how linguists can be used in developing algorithms that can detect more subtly expressed forms of anger and contempt in addition to overt hate speech. In addition, cultural context plays a very significant role in the manifestation of hate speech. According to one of the respondents, “AI needs to be continuously updated with cultural insights to remain relevant” (Participant 1). Linguists offer critical knowledge regarding cultural references, idioms, and social norms that inform how language is used in different communities.

The current limitations of AI technology also involve the understanding subtle linguistic expressions such as sarcasm and emotional tone. As an interviewee stated, “AI cannot understand the subtle meanings of words and phrases in different contexts” (Participant 2). Analysis of language can help mitigate these limitations by providing a sense of how language works in different contexts.

4.2.1.4 Ethical Challenges and AI Oversight

This theme focuses on the ethical implications of AI and the monitoring process required to ensure its ethical use. The interviews highlighted a number of concerns about the dangers of AI spreading harmful content if it is not given proper instructions. Participant 3 said, “AI will replicate whatever it is trained on; if AI is trained on data that has bad language, it will reproduce bad language”. This shows that a bias may exist in a training dataset, which might exacerbate societal differences by reinforcing pre-existing biases and stereotypes.

Additionally, the balance between hate speech detection and protecting privacy as well as free speech is another dilemma. One participant underscored the importance of caution in labeling data, stating, “For free speech and hate speech differentiation, we have to be careful while labeling the data. applying pre-trained models do not work in other social contexts” (Participant 1). This shows the importance of context in defining hate speech and that AI systems should not infringe on Legitimate expressions of free speech.

In order to address the ethical problems in detecting hate speech by AI, a number of oversight mechanisms have to be implemented. To start with, there has to be clear rules indicating what hate speech is. These rules should conform to the human rights principles so as to ensure equity and the protection of individual rights. “The training data fed to the AI model has to be so diverse that it also embraces cultural aspects outside the West” (Participant 1). This highlights the need for context and indicates that without contextual meanings, one would struggle to moderate content over various cultural contexts effectively. The other critical oversight mechanism is transparency. AI algorithm development and content assessment need to be explained in detail by platforms and AI developers. According to a participant, “There should be machines for the appeal of users’ decisions, and users should be enlightened on why such a thing happened” (Participant 5). This aspect is helpful in building users’ confidence in AI systems by them feeling secure about their contents being moderated.

In addition, bias in the AI system must be detected to rectify them, therefore, regular bias audits are required. Monitoring the systems regularly may help prevent reinforcing negative stereotypes and ensures effective treatment of all groups. In the context of the current changing linguistic norms, as one participant insisted owning the need for adaptable techniques, he mentioned: “ongoing evaluation of AI systems is critical. to recognize and correct biases” (Participant 6). The AI systems must also ensure privacy by eliminating redundant data gathering.

4.2.1.5 Balancing Free Speech and Hate Speech

Differentiating hate speech from free speech was one of the most important concerns for the interviewees. As mentioned by a participant, “For free speech and hate speech differentiation, we have to be careful while labeling the data… applying pre-trained models do not really work in other social contexts” (Participant 1). This means that the definition of hate speech and free speech varies with culture and society. What may be taken as hate speech by someone in one region: may not be hate speech but free speech in another. Thus, for AI models to get the correct representation of diverse cultures and situations, it is important to train the system on culturally relevant data that is representative of the local norms and values.

In addition, there are other more critical concerns of privacy issues for the individual. According to one of the participants, protecting users’ privacy should be addressed during the implementation of the effective measures of hate speech detection. As indicated by one of the participants, “AI should be trained to understand the context in which speech occurs... this helps prevent overreach in censorship while effectively targeting harmful behavior” Participant 7. This implies that AI systems should be focused more on recognizing the linguistic patterns and context rather than recognizing individuals without sufficient context. This means that AI systems must focus on linguistic patterns and context rather than recognizing individuals without a clear reason. Platforms can offer a safer online environment without violating the right to free speech, content analysis should ensure that it respects user privacy.

To achieve this balance, it requires a few strategies. First, guidelines of what hate speech entails have to be well put forward. These guidelines ought to conform to the human rights principles that uphold justice and the protection of individuals’ rights. Transparency of AI systems is required as AI developers and online service providers should make clear and transparent what they do to the users and how those systems are developed and in what ways the content is evaluated. As one of the interviewees pointed out: “There should be provisions for the users to seek appeal of decisions made on them and understand the decision-making process” (Participant 9). That shows transparency which leads to being accountable and building confidence into AI systems.

4.2.1.6 Future of AI in Hate Speech Detection

An important imperative for this is the more excellent contextual comprehension in AI models. According to the participants, the current models often have problems with linguistic nuances especially about the understanding of numerous cultural contexts. For instance, participant 7 said that: “AI should be trained to understand the context in which speech occurs... this helps prevent overreach in censorship while effectively targeting harmful behavior” (participant 7). Such AI models need to be developed which can identify tone, intent, and cultural context better to identify hate speech in the online platforms.

Another important area that needs enhancement in the future is training diverse datasets. As stated by one of the interviewees, “The training data fed to the AI model needs to be diverse enough to include cultural aspects other than the West” (Participant 1). This shows the significance of inclusively representative datasets covering a wide range of linguistic and cultural contexts, thereby making the AI-based detection of hate speech is feasible without feeding into biases and prejudices.

Moreover, there is an urgent need for real-time intervention by AI systems. The interviewees pointed out that the sophisticated AI models can identify harmful content and send alerts immediately. In one of the participant's statements, "AI can provide personalized feedback to users, educating them about the proper online behavior and the consequences of hate speech." (Participant 10). It could potentially decrease the diffusion of hate speech. Another important thing is to always evaluate and update AI systems. Regular audits may reveal biases in detection algorithms that may be addressed. An interviewee said that "evaluation of AI systems on a continuous basis is important to identify and address biases" (Participant 8), emphasizing that continuous monitoring helps ensure fair treatment across different demographics while preventing the reinforcement of harmful stereotypes.

4.2.2 Analysis of Data Using Theoretical Perspectives

4.2.2.1 Socio-Technical Systems Theory

The Socio-Technical Systems Theory emphasized the interdependence of social and technological structures. This theory draws attention to how social standards and technological structures affect linguistic elements in the case of online hate speech.

i. **Interdependence of Language and Technology:** Based on the interviews, hate speech is characterized by specific linguistic elements such as threats, dehumanization, and offensive words. As one participant puts it, "Hate speech entails any form of communication. where it fuels or advocates for hatred, discrimination or hostility towards people or groups" (Participant 2). This definition emphasizes how attitudes in society are reflected in language. Within the framework of Socio-Technical Systems Theory, this highlights how AI systems, being products of both technical design and social input, must be contextually informed to function responsibly. A participant in the interview pointed out, "For free speech and hate speech differentiation, we have to be careful while labelling the data… pre-determined models do not really work in other social contexts" (Participant 1). This underscores a core STS principle that technology is not value-neutral; its performance is shaped by the social environments and assumptions embedded in its development (Baxter & Sommerville, 2011). A comment from participant 3, “AI will repeat whatever it is trained on. if AI is trained on data that has bad language, it will reproduce bad language” further illustrates the co-dependence between human agency and machine behavior**,** reinforcing that biased inputs or culturally unrepresentative datasets can reproduce harm unless socio-technical alignment is intentionally achieved.

ii. **Interdependence of Social and Technical Elements:** The theory highlights the connections between social elements (linguistic variety and cultural environment) and technological abilities (AI training approaches). As Trist (1981) originally conceptualized, “any organizational work system is made up of two interdependent systems: the social and the technical” (p. 3). Applying this to hate speech detection, it becomes evident that effective moderation cannot be achieved through algorithmic precision alone, but must also account for linguistic diversity, user behavior, and cultural context. Mumford (2000) similarly stressed that ignoring the social dimension in technology design is “not only ineffective, it is ethically questionable” (p. 126), which aligns with the present study’s findings on the ethical risks of biased or culturally uninformed AI training data. “Models are trained on the data we provide... we need to train models based on our data,” as participant 1 stated. It emphasizes how social elements, like cultural background and privacy issues, and the technological capability of AI algorithms are interdependent. According to the interviews, there might be substantial cultural variations in what constitutes hate speech, hence it is crucial that AI models be trained with data that represents these many settings. A participant 5 said, “For free speech and hate speech differentiation, we have to be careful while labelling the data... applying pre-trained models do not really work in other social contexts”. This indicates how linguists, technologists, and social scientists must work together to develop models that accurately capture regional norms and values while guaranteeing the protection of free speech. This emphasizes how important it is for linguists and technologists to work together.

iii. **Adaptation and Flexibility:** STS asserts that social norms and technology must develop together. Reliance on small datasets may result in misunderstandings and inefficient anti hate speech interventions. As a participant said, “Different real-world models are trained in developed countries... therefore, they already have identified what is free speech and what is hate speech” (participant 4). This indicates that AI systems must adapt to local contexts to remain effective. Modern socio-technical systems are "adaptive, evolving continuously as users and machines co-shape each other’s behaviour," as argued by Baskerville and Fischer (2020). AI systems must therefore be adaptable and context-aware, updating often to take into account regional language customs, regulatory requirements, and cultural values. In order to prevent systems educated in one environment from imposing improper categories in another, this calls for a design strategy based on ongoing socio-technical alignment.

iv. **Human-Centered Design:** The notion of human-centered design is a key principle in updated interpretations of STS. Therefore, it puts more onus on creating technology based on social circumstances and understanding human needs. Since definitions vary from culture to culture on the concept of hate speech, there is a need to have an ethical framework guiding the discussion on harmful languages. AI systems run the risk of ignoring offensive material or stifling free speech in the absence of a well-grounded ethical framework. According to Hoda (2022), in order to guarantee equity, flexibility, and contextual awareness, "AI systems, as socio-technical artefacts, must be iteratively developed with sustained engagement from diverse stakeholders to ensure fairness, adaptability, and contextual sensitivity" (p. 7). This underscores the importance of collaboration with linguists and cultural specialists, which is critical to the development of AI systems that appropriately detect hate speech without violating regional cultures.

v. **Feedback Loop:** STS recognizes the bidirectional relationship between society and technology, where social practices shape technological development and in turn, technology reshapes those very practices (Trist, 1981; Baxter & Sommerville, 2011).

The feedback loop is especially relevant in the context of hate speech detection, where cultural norms and language use are constantly evolving. The challenge here is the cultural context; as definitions change in society, similarly the technology frameworks used for detection procedures should also evolve. Participant 3 said, “AI needs to be continuously updated with cultural insights to remain relevant”, which calls for constant adaptability. Zhu and Başar (2024) similarly note that “understanding the social dependencies of technical performance is central to the responsible design of intelligent systems” (p. 4), reinforcing the STS perspective that socio-cultural input is not a one-time concern but an ongoing requirement**.** Therefore, the effectiveness and ethical soundness of AI technologies depend on maintaining an open feedback loop between users, designers, and evolving social realities.

vi. **Adaptation to Environment:** Technological constraints are a reflection of the need for social behaviors and technology to co-adapt. The ability to create complex models that can precisely analyze language is limited by a lack of hardware resources. In order to overcome these constraints, technological investment is necessary, and social requirements must be met by providing sufficient assistance for linguistically varied societies. These constraints reflect the broader STS insight that technical progress must be supported by social investment, both in terms of infrastructure and inclusive research practices (Baxter & Sommerville, 2011).

vii. **Resource Limitations:** Socio-Technical Systems Theory highlights how social structures, such as concerns about equity and resource distribution, are intricately entwined with technical functionality and access. Fair and inclusive access to technical tools must be in line with user expectations, especially when it comes to hate speech identification, where certain communities may experience more frequent or severe cases of harm. However, these communities cannot benefit from preventive measures without the required resources, such as computational equipment, reliable internet connectivity, or localized AI technologies. This highlights a major STS concern: systems run the risk of escalating rather than reducing existing imbalances when social requirements are not met by sufficient technical capacity (Mumford, 2000).

vii. **Adaptation to Linguistic Norms:** Socio-Technical Systems Theory underlines the need for flexibility and adaptability in the development of technology, particularly in AI systems, play a significant role intended to identify hate speech. Since language and social norms are constantly changing, it is important that AI adapt as well. As one interviewee noted, “Linguists can help in identifying ever-evolving patterns in languages. so machines can identify linguistic patterns more effectively”. This adaptability is important to ensure that AI systems stay relevant and accurate in the fast-changing world and rapidly changing digital landscape. As Hoda (2022) points out, socio-technical systems must be designed with the flexibility to adapt to underrepresented communities, requiring not only innovation in hardware and software**,** but also institutional support for marginalized languages and regions**.**

viii. **Diverse Linguistic Contexts:** Socio-Technical Systems Theory foregrounds the importance of designing technology that reflects the diverse social, cultural, and linguistic environments in which it operates. In the domain of hate speech detection, cultural understanding and linguistic variation are not peripheral, they are central to the system’s effectiveness and fairness. As one interviewee observed “The training data fed to the AI model needs to be diverse enough to include cultural aspects other than the West”. This reinforces the STS principle that technology must be socially embedded**,** meaning it should be informed by the realities, languages, and communication norms of varied populations (Trist, 1981; Hoda, 2022). This indicates that AI systems must be developed by collaboration, taking insights from linguists, sociologists, and cultural experts to ensure that models accurately reflect diverse linguistic realities.

4.2.2.2 Discourse Ethics Theory

Discourse ethics theory stresses mutual respect and dialogue as fundamental features of communication in a democratic society.

i. **Ethical Communication Standards:** The interviews revealed that determining hate speech presents moral questions over whether or not speech is appropriate. Discourse Ethics deals with standards for just communication in democratic societies. AI is incapable of perceiving linguistic contexts that vary and is thus raise a question of fairness and representation. AI systems are likely to unintentionally restrict marginalized voices if they are unable to identify culturally specific hate speech. One participant pointed out that “If AI is used to control online hate speech… it needs to be trained well” (participant 3) which under-scored the important role moral training procedures hold. In another case, “There is already work being done on automatically analyzing online text to find the writer’s intent… even the writer’s emotional state can be guessed” (Participant 3). This strategy is in line with Discourse Ethics, which emphasizes inclusive discussion to reach a consensus on what constitutes hate speech. This accountability is crucial for building users’ confidence and trust on online platforms.

Discourse ethics also holds that there should be moral responsibility in deciding on language use. Respondent 2 was of the opinion, “The role of a linguist is to identify the linguistic patterns, they can also help to add more context and meaning to pre-existing data”. Such responsibility makes it that even when AI systems want to identify hate speech with greater efficiency, it is done in such a way that human dignity is preserved.

ii. **Balancing Free Speech and Hate Speech:** The interviews revealed how crucial it is to differentiate between identifying hate speech and protecting the right to free expression. A participant remarked, “If AI is used to control online hate speech. AI has the capability to detect hate speech pretty well” (participant 4). However, as Discourse Ethics argues, automated moderation must also interpret intent, social meaning, and possible harm buried in language, rather than merely depending on surface-level lexical clues. Without endorsing discourse that dehumanizes or encourages violence, the principle of communicative rationality demands that all voices, including those who criticize prevailing power structures, be heard and respected.

iii. **Contextual Understanding:** Discourse Ethics mainly focuses on the importance of context in language interpretation as it states “The telos of human speech is understanding, not success” (Habermas, 1990, p. 288). The complexity of language demands an understanding of cultural nuances in order to effectively identify hate speech. According to one interviewee, “Understanding context is very crucial; it can distinguish between harmless discussions and harmful language” (participant 3). This underscores the requirement of ethical considerations in AI training processes that value diverse linguistic contexts.

iv. **Inclusivity:** The need for transparency in AI systems aligns with this theory's focus on accountability and inclusivity. "There should be mechanisms for users to appeal decisions," interviewee 4 said. This aligns with Discourse Ethics which emphasizes that "Discourse ethics in a digital context must account for asymmetries in power, access, and cultural interpretation of meaning."

(Capurro & Ess, 2009, p. 23). This underscores that users should be able to raise their voices about the nature of monitoring of their content. Stakeholders representing different cultures should be engaged in elaborating common terms to have an understanding of hate speech by all. Diverse opinions on definition can lead to more equitable AI systems that respect different cultural contexts and give room for constructive discussion.

v. **Moral Responsibility:** Discourse Ethics holds that people and organizations are morally obligated to engage in ethical conversations, that is inclusive, just, and transparent. This moral responsibility is underscored by the problem of cultural diversity; if some populations are deprived of advanced AI tools because of financial constraint, they can be grossly affected by hate speech online without sufficient assistance to identify and remove it. Habermas (1990) argues that "only those norms can claim to be valid that meet with the approval of all affected" (p. 66), reinforcing the view that equitable access to ethical AI is not optional rather it is a shared moral obligation rooted in dialogic justice and the universal right to participate in shaping communication norms.

vi. **Transparency in Decision Making:** In Discourse Ethics Theory, one core principle is the standard of the norm for just communication and moral responsibility in discourse related to hate speech. It also argues for being transparent with the AI machines. As illustrated by a participant of the discussion, "There should be mechanisms for users to appeal decisions and under- stand reasoning behind it" (participant 4). This also ensures trust and accountability in the process of hate speech detection. The concept emphasizes the need for openness and transparency of the platforms and developers about the training and design of the algorithms used for content moderation. Users could not comprehend why certain content was ignored or why certain moderation steps were taken if the technological restrictions were not clearly communicated.

vii. **Accountability Mechanisms:** The discourse ethics supports processes in which people are granted ability to question judgments made by AI systems. According to Habermas (1990), when communicative acts are subject to examination and rational justification, ethical legitimacy emerges. In the context of AI-based hate speech identification, when opaque algorithms have the potential to silence or mislead users, especially those from marginalized communities, this principle becomes even more important. Trust erodes and feelings of exclusion are worse when users can't question AI judgements or see the logic behind moderation decisions.

viii. **Mutual Decision-Making:** This further focuses on how much context is necessary while understanding language in Discourse Ethics. The theory holds that understanding language ethically requires careful attention to context, intent, and the plurality of meanings**,** which cannot be determined unilaterally by technologists or external institutions. The difficulty to find hate speech in varied cultural contexts has required ethical practice that involved all relevant stakeholders in definition talks. This approach assures that the AI systems not only easily identify harmful content but respect various cultural nuances as well.

ix. **Sensitivity to Cultural Nuances:** The contextual understanding, emphasized in Discourse Ethics further relates to interpreting language. It also emphasizes on involving community members in this process of determining hate speech, which has a different definition across cultures. For example, one participant observed on the issue of guidelines: “Clearly, establishing guidelines and policies that define what constitutes hate speech is vital” (participant 5). This approach ensures that AI systems are sensitive to cultural nuances while effectively identifying harmful content.

Conclusively, both the theoretical frameworks emphasize that AI systems should be evaluated and improved continuously with the changing and evolving linguistic norms. To ascertain fairness and Identify biases, regular audits can help. One of the interviewees said that continuous monitoring of AI systems is crucial for the identification of evolving forms of online hate speech.

4.2.2 Findings of Qualitative Analysis

The findings of this section reveal several linguistic and ethical challenges faced by the AI algorithms in the process of online hate speech detection. The interviews with the experts reveal the linguistic and ethical challenges, the limitations of AI systems and the future opportunities. The discussion reflects on these challenges and limitations connecting them to the theoretical frameworks.

The findings of the analysis show that one of the significant challenges to AI systems lies in the complexity of the language. Hate speech proves to be a complex issue and is mostly expressed by using ambiguous language, sarcasm, and emerging languages, which makes it tougher for AI to detect. Moreover, the socio-cultural barriers are also involved in language complexity. Hate speech differs across cultures and languages, and it needs to be understood by AI systems. This makes AI systems to be limited only to the already trained models on language limited to one community or culture. Therefore, technology must, as explained by Socio-Technical Systems theory, be integrated with the social and linguistic context to be more effective (Bijker, Hughes, & Pinch, 2012). The improved contextual understanding is crucial for AI systems in order to detect hate speech which is only possible if the AI models are trained on diverse data.

Another significant finding revealed by the interviews is the bias in AI models due to its training on biased data. If the data for training AI systems is good and not biased, then the AI systems are not biased and vice versa. The interviewees emphasized the need for the training of data on diverse datasets. The AI models can inherit prejudices from the data they learn on (Sap et al., 2019). It means that inclusive and representative datasets that showcase a broad range of different cultural and linguistic contexts become necessary. According to the STS, this illustrates how social contexts shape technology together with linguistic environments in which this technology operates. Thus, AI models must be trained on datasets that are diverse in including various languages, dialects, and cultural contexts (Waseem & Hovy, 2016).

The interviews further highlighted the issue of ethical use of AI systems. The most challenging task for AI to find a balance between hate speech and free speech. This corresponds to the principle of Discourse Ethics, which requires a balance between freedom of expression and the prevention of harm in online discourse. Furthermore, the results also highlighted the risk of misinterpretation of text particularly those expressions of hate speech that are obvious. Thus, there is a need for the development of AI systems which are transparent and with accountability mechanisms.

Moreover, the findings highlighted the need for the collaboration between linguists and AI experts for the future of hate speech and AI. To overcome the linguistic challenge and the difficulty of contextual understanding, linguists can play a significant role. Besides, the interviewees underscored continuous evaluation and adjustments of AI systems for future developments. In summary, to overcome the linguistic and ethical challenges, there is a need for continuous monitoring of AI systems along with human oversight in order to accurately detect online hate speech.

The qualitative findings underscored the difficulty AI systems face in detecting hate speech that is context-dependent**,** sarcastic**,** or implicitly encoded. This reflects Fortuna and Nunes' (2018) concern about the limitations of models that rely primarily on overt lexical features. MacAvaney et al. (2019) similarly argue that conventional classification models fail to capture the semantic subtleties and contextual cues needed to identify nuanced hate speech. Further, Kiritchenko, Nejadgholi, and Fraser (2021) emphasize the need for incorporating pragmatic and discourse-level understanding in hate speech detection, as such expressions often rely on implied meanings and cultural references. Yu et al. (2022) also affirm the value of conversational context in distinguishing between hate and counter-speech. Together, these insights support the current study's findings that purely algorithmic approaches, without linguistic enrichment, remain insufficient for tackling the complexity of online hate speech.

4.3 Critical Discussion

This study investigated the role of artificial intelligence (AI) in identifying and addressing online hate speech using a mixed-method approach that combined quantitative and qualitative analyses. The results show a multifaceted and evolving phenomenon influenced by linguistic complexity, cultural and technological constraints. The quantitative results show a notable prevalence of hate speech on the internet, with Twitter standing out as a major source for the spread of politically motivated hateful discourse. Based on statistical trends, people and communities are regularly the victim of disparaging slurs, exclusionary statements, and unfavorable generalizations because of their political affiliations, ethnicity, gender, and religion. This trend reinforces the idea that social polarization can be amplified by online platforms, particularly if they are not monitored or are not sufficiently controlled by algorithmic filters.

However, the underlying nuances and latent forms of hate speech could not be explained by quantitative data alone. Qualitative analysis was used to close this gap and found that hate speech frequently goes undetected because of its subtle linguistic composition. AI systems trained on surface-level patterns are unable to detect hateful expressions since they are often imbedded in irony, sarcasm, or culturally coded language. Furthermore, the strict frames of the existing hate speech detection algorithms are challenged by the usage of developing slang, memes, and implicit language. This qualitative insight reveals that hate speech is not always a matter of overt hostility but often takes the form of covert discursive strategies aimed at masking hateful intent while maintaining plausible deniability.

The synthesis of both analytical strands points to a critical tension in the current technological landscape: while AI has shown measurable success in detecting explicit and clearly defined hate speech, it struggles significantly with identifying implicit, context-dependent, or culturally embedded forms. This shortfall is exacerbated by the over-reliance on static datasets that do not represent the full spectrum of global linguistic diversity. AI models trained on such datasets are ill-equipped to handle the fluidity of online discourse, leading to either missing real hate speech or mislabeling non-hateful content.

In addition, the research study underscores a pressing ethical issue: AI systems run the risk of reinforcing the biases they are designed to counteract when they are taught on biased or non-representative data. For example, the AI might not recognize hate speech directed at minority communities or, worse, might mistakenly label their vernacular as hostile if a dataset under-represents the speech patterns of those people.

This presents significant questions about privacy, algorithmic fairness, and the ethical duty of platform authorities and tech developers.

The implications of these findings are both practical and theoretical. Practically, they suggest the need for continuous retraining of AI models using datasets that are not only linguistically and culturally inclusive but also reflective of the rapidly changing nature of online language. There is a pressing need for hybrid moderation systems that combine AI detection with human oversight, especially in complex or borderline cases. Theoretically, the findings call for a reevaluation of the assumptions underpinning computational linguistics and AI ethics, particularly the belief that technological solutions alone can resolve deeply social and contextual issues.

This study concludes demonstrating that although AI has potential for addressing hate speech online, its effectiveness is limited by its shortcomings in linguistic understanding, contextual interpretation, and ethical foundation. Designing AI systems that are both technically sound and socially conscious requires a more thorough, multidisciplinary approach that incorporates knowledge from sociolinguistics, discourse analysis, computer science, and digital ethics. Therefore, to guarantee that AI is a useful instrument in the battle against online hate and harassment, future research and policy initiatives must place a high priority on inclusion, openness, and contextual awareness.

CHAPTER 5

CONCLUSION

The primary aim of this research study has been to explore the growing concern of online hate speech and the linguistic challenges it presents to the AI technology used for its detection. The central problem addressed in this research is the linguistic challenges faced by AI systems while detecting online hate speech. The three primary objectives of this research were: first to identify the linguistic features and patterns of online hate speech, second to explore the significant linguistic challenges encountered by the AI systems in the detection of online hate speech and third to highlight the ethical challenges involved in the detection of online hate speech through AI. By achieving these objectives, the study aimed to contribute to both the field of computational linguistics and Natural Language Processing (NLP), providing a clear understanding of how AI can be used to effectively detect and address hate speech on online platforms.

5.1 Key Findings

5.1.1 Lexical and Semantic Features of Hate Speech

The quantitative findings reveal that online hate speech being characterized by derogatory slurs, exclusionary remarks and negative generalizations targeting individuals and groups. These hateful statements were commonly observed on several platforms, with Twitter becoming as a focal point for hateful speech that is politically tinged. Such posts frequently used dehumanizing analogies, stereotyped generalizations, and language that was loaded with racial or ideological connotations. These manifestations went beyond personal insults to include stigmatization of groups on the basis of political affiliation, gender, race, religion, and ethnicity.

The qualitative analysis also revealed that implicit meanings and semantic complexity are important characteristics of hate speech in the modern day. Contextually elusive, hate speech is sometimes coated in irony, sarcasm, euphemisms, memes, and coded language. By disguising a hostile tone, these linguistic patterns enable users to get over content control systems and nonetheless accomplish the rhetorical objective of offence or marginalization. Furthermore, culturally unique references and changing terminology add more semantic ambiguity that is frequently missed by current AI systems.

5.1.2 Linguistic Challenges in AI-Based Detection

The findings underscore that although AI has made significant progress in identifying hate speech that is explicit and based on keywords, it continues to face challenges in recognizing sentiments that are nuanced and contextually embedded. The pragmatic functions of language, including speaker intent or implicit meaning, are difficult for machine learning algorithms to comprehend because they are usually trained on massive text corpora and mostly rely on surface-level lexical data. One significant drawback is the training data itself. When faced with under-represented kinds of hate speech, AI models frequently rely on datasets that are linguistically and culturally constrained, leading to blind spots. Due to the model's incapacity to comprehend subtleties beyond literal interpretation, expressions that are culturally or community-specific often go undetected.

5.1.3 Ethical Considerations in AI Moderation Systems

This study highlights a number of ethical issues with AI-based moderation, with a particular emphasis on privacy, fairness, and bias. The existence of systemic bias in training datasets is one important problem. In addition to failing to identify targeted hate against specific communities, the AI model may incorrectly label their expressions as offensive if the input data reflects prevailing cultural stereotypes or under-represents particular groups.

Such misclassifications have significant repercussions that go beyond accuracy; they also affect equity and freedom of expression, especially when automated systems silence or misinterpret the views of marginalized groups. Concerns around over-surveillance are also becoming more prevalent, particularly as AI technologies are used widely without clear accountability frameworks.

Ethical integration, therefore, requires more than technical adjustments; it calls for inclusive design practices and hybrid moderation frameworks that combine automated detection with human review**.** This approach would not only improve contextual accuracy but also help mitigate the ethical risks associated with unchecked automation.

5.1.4 Commonalities and Differences with Previous Studies

The findings of this study align with previous research in recognizing that online hate speech is linguistically complex, context-dependent, and deeply embedded in social and cultural discourse (Fortuna & Nunes, 2018; Kiritchenko et al., 2021; Gamback & Sikdar, 2017). Similar to earlier studies, this research confirms that AI models often struggle to detect hate speech effectively due to semantic ambiguity, figurative language, and evolving online expressions. Consistent with the work of Cortiz and Zubiaga (2020), and Field et al. (2020), the present study also identifies ethical challenges, including bias in training datasets, accountability and the potential suppression of minority voices.

However, this study differs from prior research by adopting a socio-technical and ethical lens, integrating linguistic, cultural, and ethical dimensions in analyzing AI’s role in hate speech detection. While previous studies often focused on either the linguistic or technological aspects in isolation, this research emphasizes the interdependence between language complexity, AI limitations, and ethical fairness**.** Moreover, the inclusion of qualitative insights from expert interviews adds a human perspective often absent in earlier computational studies, thereby contributing to a more holistic understanding of the challenges in AI-based hate speech detection.

5.2 Recommendations for Future Research

Future research may prioritize the development of models that can effectively detect hate speech across different cultures as the research revealed the need for the development of AI models that can effectively incorporate deeper linguistic understanding beyond just detecting key words. The models may analyze tone, context and cultural references unlike the current models that are often restricted to surface- level detection.

As AI models struggle to detect subtle forms of hate speech such as sarcasm, coded language and evolving language, future research can focus on developing models that can accurately interpret and detect cultural contexts and linguistic norms. This could involve a hybrid approach where human oversight is needed for the machine learning models to be trained in order to identify linguistic markers of overt forms of hate speech. Additionally, future studies can seek out collaborative frameworks that allow humans to provide input which enhances the AI processing power. The context-based detection of subtle, human-specific hate speech, requires the input of the human in understanding linguist and cultural contexts. Therefore, human moderators can have real-time feedback about their AI models to better understand them as well as adapting and interpreting the complex linguist nuances.

Datasets training for AI tends to generate biases, leading sometimes even to an aggravation of hate speech due to being disproportional on a specific group. Future research should focus on identifying biases within the datasets of hate speech and mitigating those biases through techniques such as auditing, diversifying datasets and data augmentation. This enables AI systems to make accurate and fairer identifications across a wide range of social contexts. Moreover, AI systems also raise other ethical concerns such as the privacy of users. Future research can investigate these ethical challenges by exploring ways to ensure accountability, transparency and privacy of users. In addition, development of AI models with accountability and transparency help the researchers to understand the difference between free speech and hate speech. Hence, future research may also explore the legal frameworks and policies to draw a line between free speech and hate speech.

In addition, future studies should also investigate the incorporation of multilingual AI detection systems, especially as hate speech is increasingly being communicated on international platforms in non-English languages, regional dialects, and code-switching. Although the current study was restricted to English-language data, cross-linguistic hate speech identification is still a field that needs more research and has important practical applications. Furthermore, the interpretive depth of AI models may be enhanced by combining sentiment analysis methods with hate speech classification. Sentiment analysis can improve the model's capacity to differentiate between hostile, neutral, and sarcastic content by examining emotional tone and polarity, particularly in situations that are unclear or culturally complex. Therefore, extending future studies in these areas will aid in the creation of AI systems that are more ethically conscious and context-aware.

5.3 Final Remarks

The present research study concludes that the issue of online hate speech is complex and multifaceted. It has explored the interaction of linguistics and

Artificial Intelligence and how these disciplines can collaborate to address the complex issue of evolving online hate speech and its detection. The study emphasizes that in order to address this issue, there is no single solution. It requires continuous adaptation, collaboration, and inclusivity in AI design and development to address the evolving and complex nature of language and the diverse cultural contexts in the detection of online hate speech. The study concludes as a stepping stone for future research and practical initiatives inspiring efforts to harness the potential of AI in fostering safer and more inclusive online spaces.

REFERENCES

Abro, S., Shaikh, S., Khand, Z. H., Zafar, A., Khan, S., & Mujtaba, G. (2020). Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, *11*(8), 484–491. https://doi.org/10.14569/IJACSA.2020.0110861

Abutabenjeh, S., & Jaradat, R. (2018). Clarification of research design, research methods, and research methodology: A guide for public administration researchers and practitioners. *Teaching Public Administration, 36*(3), 237–258.

Ali, M. Z., Rauf, S., Javed, K., & Hussain, S. (2021). Improving hate speech detection of Urdu tweets using sentiment analysis. *IEEE Access, 9*, 84296–84305. https://doi.org/10.1109/ACCESS.2021.3087827

Alvi, M. (2016). *A manual for selecting sampling techniques in research* (MPRA Paper No. 70218). University Library of Munich, Germany.

Anjum, R., & Katarya, R. (2022, April). *Analysis of online toxicity detection using machine learning approaches*. In G. Sanyal, C. M. Travieso-González, S. Awasthi, C. M. A. Pinto, & B. R. Purushothama (Eds.), *International Conference on Artificial Intelligence and Sustainable Engineering: Select Proceedings of AISE 2020* (Vol. 836, pp. 381–392). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-8542-2_29

Asogwa, D. C., Efozia, F. N., Chukwuneke, C. I., & Nnaekwe, K. U. (2022). Automatic text classification on blogs using support vector machines (SVM). *International Journal of Research Publication and Reviews*, *3*(4), 805–809.

Awan, I. (2016). Islamophobia on social media: A qualitative analysis of the Facebook’s walls of hate. *International Journal of Cyber Criminology, 10*(1), 1–20. https://doi.org/10.5281/zenodo.58517

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017, April). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion* (pp. 759-760).

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers*, *23*(1), 4-17.

Bengio, Y. (2012, June). *Deep learning of representations for unsupervised and transfer learning.* In Proceedings of the ICML Workshop on Unsupervised and Transfer Learning (pp. 17–36). JMLR Workshop and Conference Proceedings.

Bijker, W. E., Hughes, T. P., & Pinch, T. J. (Eds.). (2012). *The social construction of technological systems: New Directions in the sociology and history of technology*. MIT Press.

Bilal, M., Khan, A., Jan, S., & Musa, S. (2022). Context-aware deep learning model for detection of Roman Urdu hate speech on social media platforms. *IEEE Access, 10*, 121133–121151. https://doi.org/10.1109/ACCESS.2022.3215672

Binns, R. (2018). *Fairness in machine learning: Lessons from political philosophy.* In *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency* (pp. 149–159). ACM.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Capurro, R. (2009). *Digital ethics.* In The Academy of Korean Studies (Ed.), *2009 Global Forum on Civilization and Peace* (pp. 203–214). Korean National Commission for UNESCO / The Academy of Korean Studies.

Clarke, C., Hall, M., Mittal, G., Yu, Y., Sajeev, S., Mars, J., & Chen, M. (2023). *Rule by example: Harnessing logical rules for explainable hate speech detection* (arXiv:2307.12935). arXiv. https://arxiv.org/abs/2307.12935

Clarke, V., & Braun, V. (2017). Thematic analysis. *The Journal of Positive Psychology*, *12*(3), 297-298. https://doi.org/10.1080/17439760.2016.1262613

Clegg, S. R., Pitsis, T. S., & Kornberger, M. (2019). *Managing and organizations: An introduction to theory and practice* (4th ed.). SAGE Publications.

Coffey, B., & Woolworth, S. (2004). “Destroy the scum, and then neuter their families:” The web forum as a vehicle for community discourse? *The Social Science Journal*, *41*(1), 1–14. https://doi.org/10.1016/j.soscij.2003.11.003

Cohen-Almagor, R. (2014). Countering hate on the Internet. *Annual Review of Law and Ethics*, *22*, 431–443. https://doi.org/10.1146/annurev-lawsocsci-102013-030358

Cortiz, D., & Zubiaga, A. (2021). Ethical and technical challenges of AI in tackling hate speech. *The International Review of Information Ethics*, *29*(3), 1–10. https://doi.org/10.29173/irie416

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 11, No. 1, pp. 512–515). AAAI. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15584

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Elsafoury, F. (2020). *Cyberbullying datasets* [Data set]. Mendeley Data. https://data.mendeley.com/datasets/jf4pzyvnpj/1

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018, June). Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 12, No. 1). AAAI. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17878

Erjavec, K., & Kovačič, M. P. (2012). “You don't understand, this is a new war!” Analysis of hate speech in news web sites' comments. *Mass Communication and Society, 15*(6), 899–920. https://doi.org/10.1080/15205436.2011.624706

Fischer, L. H., & Baskerville, R. (2020). Revising the socio-technical perspective for the 21st century: New mechanisms at work. *Journal of Database Management (JDM), 31*(4), 69–87. https://doi.org/10.4018/JDM.2020100104

Flick, U. (2014). Mapping the field. In U. Flick (Ed.), *The SAGE handbook of qualitative data analysis* (pp. 3–18). SAGE Publications.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR), 51*(4), 1–30. https://doi.org/10.1145/3232676

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech.* UNESCO Publishing.

Gambäck, B., & Sikdar, U. K. (2017, August). Using convolutional neural networks to classify hate speech. In *Proceedings of the First Workshop on Abusive Language Online* (pp. 85–90). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3012

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering, 10*(4), 215–230. https://doi.org/10.14257/ijmue.2015.10.4.20

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018, January). All you need is “love”: Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security* (pp. 2–12). Association for Computing Machinery. https://doi.org/10.1145/3270104.3270106

Gu, J., Hassan, H., Devlin, J., & Li, V. O. (2018). Universal neural machine translation for extremely low resource languages. *arXiv preprint* arXiv:1802.05368. https://arxiv.org/abs/1802.05368

Habermas, J. (1990). *Moral consciousness and communicative action*. MIT Press.

Hüsünbeyi, Z. M., Akar, D., & Özgür, A. (2022, June). Identifying hate speech using neural networks and discourse analysis techniques. In *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference* (pp. 32–41). European Language Resources Association.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Ishtiaq, M. (2019). Book review: Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches.* Thousand Oaks, CA: Sage. *English Language Teaching, 12*(5), 40.

Kiritchenko, S., Nejadgholi, I., & Fraser, K. C. (2021). Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research, 71*, 431–478. https://doi.org/10.1613/jair.1.12590

Kousar, A., Ahmad, J., Ijaz, K., Yousef, A., Shaikh, Z. A., Khosa, I., … & Anjum, M. (2024). MLHS-CGCapNet: A lightweight model for multilingual hate speech detection. *IEEE Access, 12*, 106631–106644.

Kovács, G., Alonso, P., & Saini, R. (2021). *Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources.* SN Computer Science, 2(2), 95. https://doi.org/10.1007/s42979-021-00457-3

Kumarage, T., Bhattacharjee, A., & Garland, J. (2024). Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection. In B. Di Fátima (Ed.), *Methods, techniques and AI solutions in the age of hostilities: Online Hate Speech Trilogy* (Vol. 3, pp. 111–134). LabCom – University of Beira Interior & Universidad Icesi. https://doi.org/10.18046/EUI/ohst.v3

Liu, R., Lehman, J., Molino, P., Petroski Such, F., Frank, E., Sergeev, A., & Yosinski, J. (2018). An intriguing failing of convolutional neural networks and the CoordConv solution. *Advances in Neural Information Processing Systems, 31*.

https://papers.nips.cc/paper/8169-an-intriguing-failing-of-convolutional-neural-networks-and-the-coordconv-solution

Loebbecke, C., Luong, A. C., & Obeng-Antwi, A. (2021). AI for tackling hate speech. In *ECIS 2021 Research-in-Progress Papers* (Paper No. 1205). AIS eLibrary. https://aisel.aisnet.org/ecis2021_rip/10/

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLOS ONE*, *14*(8), e0221152. https://doi.org/10.1371/journal.pone.0221152

Martínez, I. M. P. (2020). Methods of data collection in English empirical linguistics research: Results of a recent survey. *Language Sciences, 78*, 101263. https://doi.org/10.1016/j.langsci.2019.101263

Masood, Z., Hoda, R., & Blincoe, K. (2020). Real world scrum: A grounded theory of variations in practice. *IEEE Transactions on Software Engineering, 48*(5), 1579–1591.

Metselaar, S., Meynen, G., & Widdershoven, G. (2016). Reconsidering bias: A hermeneutic perspective. *The American Journal of Bioethics*, *16*(5), 33–35.

Mikolov, T., Yih, W.-T., & Zweig, G. (2013, June). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746–751).

Mondal, M., Silva, L. A., Correa, D., & Benevenuto, F. (2018). Characterizing usage of explicit hate expressions in social media. *New Review of Hypermedia and Multimedia*, *24*(2), 110–130.

Mumford, M. D. (2000). Managing creative people: Strategies and tactics for innovation. *Human Resource Management Review*, *10*(3), 313–351.

Naseri, R. N. N., & Rahmiati, F. (2022). What is a population and sampling technique used in intention towards online halal cosmetic purchasing research.

International Journal of Academic Research in Business and Social Sciences, 12(5), 707–712.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016, April). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 145–153). ACM. https://doi.org/10.1145/2872427.2883062

Oberholzer-Gee, F., Bohnet, I., & Frey, B. S. (1997). Fairness and competence in democratic decisions. *Public Choice, 91*(1), 89–105.

Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). Hate speech: A systematized review. *Sage Open, 10*(4), 2158244020973022.

Pérez, J. M., Arango, A., & Luque, F. (2020). ANDES at SemEval-2020 Task 12: A jointly-trained BERT multilingual model for offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 1524–1531). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.199

Pinch, T. (2012). The social construction of technology: A review. In W. E. Bijker, T. P. Hughes, & T. J. Pinch (Eds.), *Technological change* (pp. 17–35). MIT Press.

Plano Clark, V. L. (2017). Mixed methods research. *The Journal of Positive Psychology*, *12*(3), 305–306.

Rehman, A. A., & Alharthi, K. (2016). An introduction to research paradigms. *International Journal of Educational Investigations, 3*(8), 51–59.

Reshamwala, A., Mishra, D., & Pawar, P. (2013). Review on natural language processing. *IRACST Engineering Science and Technology: An International Journal (ESTIJ), 3*(1), 113–116.

Retta, M. (2023). A pragmatic and discourse analysis of hate words on social media. *Internet Pragmatics*, *6*(2), 197–218.

Rizwan, H., Shakeel, M. H., & Karim, A. (2020). Hate-speech and offensive language detection in Roman Urdu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2512–2522). https://doi.org/10.18653/v1/2020.emnlp-main.197

Safdar, K., Nisar, S., Iqbal, W., Ahmad, A., & Bangash, Y. A. (2023). Demographical based sentiment analysis for detection of hate speech tweets for low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(4), Article 1. https://doi.org/10.1145/3616867

Saleem, H. M., Dillon, K. P., Benesch, S., & Ruths, D. (2017). A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint*. arXiv:1709.10159.

Saleh, H., Alhothali, A., & Moria, K. (2023). Detection of hate speech using BERT and hate speech word embedding with deep model. *Applied Artificial Intelligence*, *37*(1), 2166719. https://doi.org/10.1080/08839514.2023.2166719

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019, July). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668–1678). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1163

Tareen, M. K., Tareen, H. K., Noreen, S., & Tariq, M. (2021). Hate speech and social media: A systematic review. *Turkish Online Journal of Qualitative Inquiry, 12*(8), 6143–6156.

Teddlie, C., & Yu, F. (2007). Mixed methods sampling: A typology with examples. *International Journal of Multiple Research Approaches*, 1(1), 77–100.

Theofanidis, D., & Fountouki, A. (2018). Limitations and delimitations in the research process. *Perioperative Nursing-Quarterly*, *7*(3), 155–163.

Thomas, P. (2011). 'Mate crime': Ridicule, hostility and targeted attacks against disabled people. *Disability & Society*, *26*(1), 107–111.

Ullmann, S., & Tomalin, M. (2020). Quarantining online hate speech: Technical and ethical perspectives. *Ethics and Information Technology, 22*(1), 69–80. https://doi.org/10.1007/s10676-019-09516-z

Wanniarachchi, V. U., Scogings, C., Susnjak, T., & Mathrani, A. (2023). Hate speech patterns in social media: A methodological framework and fat stigma investigation incorporating sentiment analysis, topic modelling and discourse analysis. *Australasian Journal of Information Systems, 27*. https://doi.org/10.3127/ajis.v27i0.3929

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop* (pp. 88–93). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-2013

Wheaton, E. M. (2019). *The economics of human rights*. Routledge.

Young, J., Swammy, P., & Danks, D. (2018). *Beyond AI: Responses to hate speech and disinformation* [Unpublished manuscript, Carnegie Mellon University]. Retrieved from http://jessicayoung.com/research/Beyond-AI-Responses-to-Hate-Speech-and-Disinformation.pdf

Yu, X., Blanco, E., & Hong, L. (2022). Hate speech and counter speech detection: Conversational context does matter. *arXiv preprint* arXiv:2206.06423. https://arxiv.org/abs/2206.06423

Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? The challenging case of long tail on Twitter. *Semantic Web, 10*(5), 925–945. https://doi.org/10.3233/SW-190378

Zhu, Q., & Başar, T. (2024). Revisiting game-theoretic control in socio-technical networks: Emerging design frameworks and contemporary applications. *arXiv preprint* arXiv:2411.01794. https://arxiv.org/abs/2411.01794

Zowghi, D., & da Rimini, F. (2023). Diversity and inclusion in artificial intelligence. *arXiv preprint* arXiv:2305.12728. https://arxiv.org/abs/2305.12728

APPENDIX A

INTERVIEW GUIDE

Instructions for Interview Participants

The purpose of this study is to explore the role of AI in detecting online hate speech and the linguistic challenges involved. Please read the following instructions carefully:

i. The interview is semi structured, there are guiding questions but the participant is free to elaborate, share examples and raise points that he/she finds important.

ii. The interview will be approximately 30 minutes long.

iii. The responses will be kept confidential and will be used only for academic purposes.

iv. Participation is voluntary**.** Any question can be skipped if the participant is uncomfortable with.

v. The participant may withdraw anytime if he/she does not want to continue further.

vi. Feel free to ask for clarification if a question is unclear.

Interview Questions

Linguistic Aspects:

1. How do you define online hate speech and harassment?
2. What are some of the linguistic features and patterns that characterize online hateful content?
3. In what ways can linguistic analysis techniques be adapted to better identify different types of online harassment?
4. How do you see the role of language and linguistics in improving the accuracy of AI models for identifying online hate speech and harassment?

AI Improvement:

5. What are the most significant linguistic challenges you have encountered while working with AI to detect and address online hate speech and harassment?
6. In what ways do you think AI could potentially replicate or perpetuate harmful language patterns or biases if not properly trained or developed?

7. In your opinion, what are the limitations of current AI technology when it comes to understanding the complexities of language and detecting nuances in meaning?
8. How do you see linguistic analysis and AI being used in the future to address the problem of online hate speech and harassment?

Ethical Challenges:

9. How do you balance the need for effective detection of online hate speech and harassment with concerns around privacy and free speech?
10. In your opinion, what kind of oversight or regulation is necessary to ensure that AI is being used ethically and responsibly in detecting online hate speech and harassment?

APPENDIX B

SAMPLE INTERVIEWS

Interview 1

1. **How do you see linguistic analysis and AI being used in the future to address the problem of online hate speech?**

"General distributions of words are defined within English language. Identifying the frequency of alphabets is very important. But again, this is only done in the English language, not in Urdu or in any other language. This partly because words are formed of combinations. In Urdu for example, the entire writing style is different. Similarly, Chinese is different, in Chinese characters represent words. So it differs language to language."

2. **In what ways do you think AI could potentially replicate or perpetuate harmful language patterns or biases if not properly trained or developed?**

"Machines recognize patterns in the language. Some are supervised learning, some are unsupervised learning, and semi supervised learning. Labelling the data and the artifacts with in it. Unsupervised learning is not labelling the data, just providing raw data to the machine. So, if the models are not properly trained on the labeled data set, then harmful language patterns can be perpetuated through it."

3. **How do you balance the need for effective detection of online hate speech and with concerns around privacy and free speech?**

"For free speech and hate speech differentiation, we have to be careful while labelling the data, it's a human task to identify certain linguistic patterns as free speech and as hate speech. Models are trained on the data we provide. Different real world models are trained in the developed countries where they want to get rid of hate speech and promote free speech. Therefore, they already have identified what is free speech and what is hate speech. So, the pre-determined models do not really work in other social contexts. For instance, in Pakistan we have not differentiated between free speech and hate speech. In Pakistan the definition of hate speech and free speech are different from that of Europe. So applying the pre-trained models do not work in Pakistani context effectively. For that, we need to train models based on our data, it requires effort from institutions as well as individuals. The AI models recognize patterns, a particular set of words used in a particular way identified as hate speech. A little flexibility is that if not an exact sentence but close sentence. Accountability is not as such the job of a linguist, his job is to identify and detect the text. Privacy, accountability and other ethical concerns do not fall under the domain of linguistics. For free speech and hate speech differentiation, we have to be careful while labelling the data, it's a human task to identify certain linguistic patterns as free speech and as hate speech. Models are trained on the data we provide. Different real world models are trained in the developed countries where they want to get rid of hate speech and promote free speech. Therefore, they already have identified what is free speech and what is hate speech. So, the pre-determined models do not really work in other social contexts. For instance, in Pakistan we have not differentiated between free speech and hate speech. In Pakistan the definition of hate speech and free speech are different from that of Europe. So applying the pre-trained models do not work in Pakistani context effectively. For that, we need to train models based on our data, it requires effort from institutions as well as individuals. The AI models recognize patterns, a particular set of words used in a particular way identified as hate speech. A little flexibility is that if not an exact sentence but close sentence. Accountability is not as such the job of a linguist, his job is to identify and detect the text. Privacy, accountability and other ethical concerns do not fall under the domain of linguistics."

4. **How do you see the role of language and linguistics in improving the accuracy of AI models for identifying online hate speech?**

"The role of a linguist is to identify the linguistic patterns on the basis of which we can identify hate speech and free speech. Linguists can also help in the identification of ever-evolving patterns in languages, new sequences and the safe results of those changes. They can also help to add more and more context and meaning to pre-existing data, so the machines are able to identify the linguistic patterns more effectively, especially in the multilingual societies."

5. **In your opinion, what are the limitations of current AI technology when it comes to understanding the complexities of language and detecting nuances in meaning?**

"The first one is the limitation of hardware availability, it's not available in many parts of the world. So, only limited models are trained because of the non-availability of hardware. Another limitation is lack of separating words in different languages. For instance, in the English language, the most commonly used one is wide space detection, to identify words separately, we the experts look at the wide space. Such challenges are faced in languages other than English. Even within the English language, this word separation technique and tokenization techniques are limited."

6. **In your opinion, what kind of oversight or regulation is necessary to ensure that AI is being used ethically and responsibly in detecting online hate speech and harassment?**

1. "Labelling the data is the most important thing. Identify the artefacts within it. Recognizing the actual word again every linguistic pattern.
2. Further specify the grammatical category, whether a word is a noun, adverb or an adjective etc. so that it is simpler to extract meaning from it.
3. On social media, it is important to identify the meaning of non-verbal emoticons so that the intended meaning can be extracted from the context.
4. Overall, it is important to train the models in such a way that they are able to identify the meaning and context out of a particular data, and for that collaboration of linguists and AI experts is needed."

Interview 2

1. **How would you define online hate speech?**

"Hate speech entails any form of communication, expression or speech in the online platform where it fuels or advocates for hatred, discrimination or hostility towards people or groups in society on grounds of race, colour, nationality, religion, sex, gen- der, disability, sexual orientation, or political affiliation. It comprises text, images, videos, or audio that in one way or another seeks to offend others, threaten or harm them, and it can be an outright attack, threat or use of slurs, other examples include rejection or encouragement to embrace stereotypical notions that have negative implications on individuals, groups, and communities."

2. **What linguistic features and patterns are commonly found in online hateful content?**

"The different kinds of prejudice displayed on the internet is often characterized using degrading terms, threatening language as well as encouraging people to commit acts of violence. It often entails general negative attributions and stereotyping as well as the use of verbal abuse with slurs and/ or epithets. Lies and exaggerations are generally used to enhance the extent of fear and hatred, in addition to falsehood and conspiracy theories. The content of such a letter also involves partisanship and a kind of language that pits 'us' against 'them' and which profits from outrageously emotionalized phrases. Profanity and the aggressive message may be buried under some sort of code or slang so that only people in on the joke will get the message. Repetition of specific sections or words also helps to spread hatred to the audience as well. All these are the linguistic features and patterns that assist in the fight against online hate speech."

3. **How can linguistic analysis techniques be adapted to more effectively identify various forms of online hate speech?**

"General extended techniques of linguistic analysis in the context where different forms of hate speech in social media can be detected with higher efficiency by incorporating enhanced models based on Natural Language Processing (NLP) like BERT or GPT forest-born from the data itself. Sentiment analysis is used to detect negative emotional states and invectives, contextual embeddings deal with subtle variations and colloquialisms. Operating in close cooperation with the text analysis, the multimodal analysis combining both text and image, video, and audio content, is aimed to analyse hate speech in various formats. Preprocessing also involves text analysis which includes the identification of targets as well as the meanings of content that is considered as being hateful. Feature engineering from linguistic properties and contextual filtering in which the messages are analysed with respect to the preceding and following messages and user interactions takes detection to the next level. Machine learning and deep learning classifiers increase their accuracy with time, while understanding user's activity and their interaction with content lets to identify possible sources of hate speech."

4. **What role do language and linguistics play in enhancing the accuracy of AI models for detecting online hate speech?**

"Language and linguistics play a crucial role in enhancing the accuracy of AI models for detecting online hate speech by providing a deeper understanding of the nuances and complexities of human communication. Linguistic comprehension enables one to determine how hate speech is less overt, depends on context and is often couched in euphemisms, slangs and code words. With the help of the syntax, semantics, and pragmatic vocabularies of NLP, AI detection models' accuracy increases, subsequently overcoming the hate speech problem. Furthermore, features such as part of speech tagging, parse, sentiment analysis can be formulated from the linguistic analysis to aid models to distinguish between the dangerous and harmless content. contextual embeddings and Named Entity Recognition help in the identification of the relations between these words and entities as well. All in all, application of linguistic knowledge helps AI-based algorithms better recognize and prevent different types of cyber aggression."

5. **What are the most significant linguistic challenges you have faced while working with AI to detect and address online hate speech?**

"The two prominent problems of language when it comes to AI to predicting and combating hate speech relate to ambiguity and context sensitivity, where words and phrases may have more than one interpretation depending on the surrounding and purpose. This is a problem because the creation of new words, phrases and forms of expression in the media is bewilderingly fast, and the models built have to be kept up to date. Also, the pictures and relocation of irony, which involve rather complex perception, proved to be rather difficult. Trying to find the hidden bias that calls for prejudice in an oblique way and advocates for stereotyping people while avoiding improper language is challenging semantically and pragmatically. Diversity in cultures and languages also poses another dimension that makes the hate speech expressions complicated since the use of language differs depending on the culture. Lastly, false positive and false negative problem to censor extremely excessive speech but let go of hate speech take a lot of effort of fine-tuning and calibrating the AI models."

6. **How might AI unintentionally replicate or reinforce harmful language patterns and biases if it is not properly trained or developed?**

"That is why if AI is not trained or developed, it can indirectly reproduce or retransmit the language and even biased patterns in several ways. Biased training results in AI having prejudices of the society hence it flags contents related to some groups as hate speech while it is not the case. The AI might also generate stereotype if ever the data contain recipient stereotype and make otherwise unfair assumptions. Learning from data that is gathered from online sources can also worsen the situation as it inflates the possibility of overly favouring extreme views. Thus, overgeneralization can lead to the blocking of the legitimate content or not recognizing the implicit hatred, which is undesirable. The reinforcement learning can propose loops of behaving badly if there is no careful supervision, while the lack of posing contexts can make people fail to grasp jokes, sarcasm, or irony and miss phenomena like subtle racism or misogyny. Moreover, underrepresentation of minorities and other oppressed groups in the training data can lead to the AI's inability to recognize hate speech directed at the dis- criminated minorities, thus reinforcing existing prejudice and not providing protection of the endangered individuals."

7. **What do you see as the main limitations of current AI technology in terms of understanding the complexities and nuances of language?**

"In this relation, the current state of AI technology has some serious constraint in comprehending the shades and greys of language. It is also important to mention that AI in most cases has difficulties with understanding of the specifics of the conversation flow, such as irony, tone of voice, hidden meanings, etc., and could misunderstand the so-called ambiguous or polysemic words. Reading sarcasm and irony is quite complex, this is mostly because they require the use of intonation and body language both of which are not translate in the text. Besides, it is also noted that AI cannot properly perceiving secondary forms and cues which might convey poisonous messages with- out the use of the offending words themselves. Another factor that makes it difficult is cultural and linguistic," the models trained to recognize specific cultures or languages have difficulties with the expressions of others. The fast-changing context also becomes a problem as new slang, or expressions may appear and may not be considered by the model. Besides, common sense and intuitive reasoning related to context or intentions are also missing in AI, which sometimes misjudges the cues given. Prejudice can also be emanated in case bias is introduced to the training data which in turn will affect the fairness of the model and dependability of the results in different setting or population. All these challenges mean that it is very complex to design AI systems that can effectively understand the complexity and richness of natural language."

8. **How do you envision the future use of linguistic analysis and AI in combating online hate speech?**

"In the future, the combine technological advancement in linguistic analysis and AI in fighting hate speech have the following potential. AI models will integrate contextual awareness that will help them understand jokes, anger, and hidden/prejudice bias. Thanks to advancements in Natural Language Processing (NLP), it should be easier to filter out the material that poses a risk to the child now that slang and cultural references are constantly changing. Multimedia analysis will combine text analysis with images and videos, which I also find usable although they do add more complexity to the approaches. The future progress of machine learning is ensuring that the existing models are not only less prejudice but also can readily incorporate the different linguistic and cultural domains. Moreover, monitoring and/or learning mechanisms put into practice will also help assess new trends and threats and continuously learn from them, thus being sensitive to the changing environment. Integration of the AI systems with moderation by human beings will enhance the tolerance of hate speech detection and freedom of speech. Altogether, these developments will help in elaborating a more sophisticated and preventive approach to the problem of hate speech in the Internet."

9. **How can the need to effectively detect online hate speech be balanced with concerns about privacy and free speech?**

"Some of the most important strategies of addressing the paradox between protecting the society from online hate speech and at the same time not infringing people's privacy and freedom of speech include assigning concrete and precise meanings to hate speech prevents the serious category from including harmless content that would be borderline on free speech violation or entirely infringe on the constitution. Editorial privacy can be preserved by methods of anonymization and by employing applications that analyze the content of messages securely. Clearness in the detection practices and standards and monitoring by the third party also ensures the work is done in right manner. It is effective to integrate AI with human moderation because there are al- ways some moments that are better to figure out in context, and we do not want to over-censor content. Users should be able to appeal moderation decisions and correct errors for the sake of free speech to prevail but at the same time contacts will need to stay accurate. Designing AI system in an ethic way with an emphasis on fairness and transparency and including diversities helps in minimizing bias. Moreover, teaching people how to behave appropriately and safe on the internet decreases the amount of hatred and forbidden information as well as decreases the necessity to apply aggressive approaches. These approaches are, together, commensurate to a harmonious solution that would protect both privacy and address the issue of hatred speeches."

10. **What kind of oversight or regulation do you believe is necessary to ensure the ethical and responsible use of AI in detecting online hate speech?**

"For responsible and ethics oversight in using AI in detecting on-line hate speech, there- fore, it requires a construct of a multi-layered model of oversight and regulation. Hate speech definitions in terms of regulation procedures should be clear and unambiguous and produced by professionals such as legal advisors and ethicists with the help of community representatives considering balancing of efficiency against freedom of speech. Recommendations on the-transparent approach of AI algorithms and decision- making, which entails the disclosure of detection criteria and decision-making insights by the AI providers. Measures should be taken regarding user data to adhere with data protection policies so that data is stripped and dealt with, discretely. It has been suggested that human decisions should be incorporated into the moderation process, as such decisions tend to be context sensitive, there should also be simple mechanisms for the user to appeal against moderation decisions. There should be regulations on ethical designs of AI system, removing prejudices and including multiculturalism in AI. “Inherent checks on AI systems require frequent impact assessments; the degree and effects of AI systems’ influence must be measured in relation with its influence over privacy, free speech, and potential biases.” Incorporation of other stakeholders such as civil society, and technology gurus make regulations comprehensive and acceptable. Altogether, these measures lead to the effective and non-ethical use of AI in the fight against hatred speech on social media.”

Interview 3

1. **How do you define online hate speech?**

“Show of contempt where its not needed, using bad language, threatening people and showing ill will.”

2. **What are some of the linguistic features and patterns that characterize on- line hateful content?**

“Bad words are often used, the tone is threatening and there is an expression of contempt in the language.”

3. **In what ways can linguistic analysis techniques be adapted to better identify different types of online hate speech?**

“There is already work being done on automatically analysing online text to find the writer’s intent and if the text is negative or positive, even the writer’s emotional state can be guessed.”

4. **How do you see the role of language and linguistics in improving the accuracy of AI models for identifying online hate speech?**

“AI Models are generally trained on the data that is from the same domain, where the model is to be used. Which in this case is hate speech and the writer might or might not have good grammar. That being said the models need linguistics and language to assess the semantics (real meaning) in the text, to correctly interpret ambiguous texts and in case of voice-based data understand different dialects.”

5. **What are the most significant linguistic challenges you have encountered while working with AI to detect and address online hate speech?**

“The pre-processing of the data, so that it is fit to be fed into the algorithm for training and then to improve the accuracy of the model.”

6. **In what ways do you think AI could potentially replicate or perpetuate harmful language patterns or biases if not properly trained or developed?**

“AI will replicate, whatever it is trained on, if AI is trained on data that has bad language, it will reproduce bad language, so care should be taken when training an AI model.”

7. **In your opinion, what are the limitations of current AI technology when it comes to understanding the complexities of language and detecting nuances in meaning?**

“Although NLP (Natural Language Programming) has advanced a lot after Chatgpt, there are still some things that AI cannot do. It misses the subtext when the subject is not from the western culture. It falls short when dealing with eastern languages.”

8. **How do you see linguistic analysis and AI being used in the future to address the problem of online hate speech?**

“Linguistic analysis is used to train bots to automatically remove comments that violate the community guidelines, this can be improved by improving the ability of AI understand the context and cultural subtexts.”

9. **How do you balance the need for effective detection of online hate speech with concerns around privacy and free speech?**

“If AI is used to control online hate speech to need to disable the comment section or similar measure aren’t necessary because AI, if it is trained well has the capability to detect hate speech pretty well.”

10. **In your opinion, what kind of oversight or regulation is necessary to ensure that AI is being used ethically and responsibly in detecting online hate speech?**

“The training data fed to the AI model needs to be diverse enough to include cultural aspects other that the west, it should be diverse and inclusive in its nature and the model should be able to understand the semantics in text, the model should go through a test for biasness before it is available for public use.”

Interview 4

1. **How do you define online hate speech?**

“Definition: Online hate speech refers to any content posted, shared, or communicated through digital means (such as social media, forums, websites, etc.) that expresses or promotes hatred, discrimination, hostility, or violence towards individuals or groups based on attributes such as race, ethnicity, religion, gender, sexual orientation, disability, or other identifiable characteristics. It can include offensive language, derogatory remarks, threats, or any content that seeks to dehumanize or degrade individuals or groups.”

2. **What are some of the linguistic features and patterns that characterize online hateful content?**

a. “Name-calling and Insults: It often includes hurtful words and insults aimed at attacking someone’s race, religion, gender, or other personal traits.

b. Threats and Aggression: Hateful content may contain threats of violence or aggressive language directed towards individuals or groups.

c. False Generalizations: broad statements to unfairly portray entire groups in a negative way.

d. Dehumanization: It may treat people as less than human, denying them respect or empathy.

e. Emotional and Provocative Language: It uses strong emotions and provocative language to stir up anger or hatred.

f. Misinformation and Conspiracies: Sometimes, it spreads false information or conspiracy theories to undermine certain groups.

g. Echo Chambers: It thrives in online communities where hateful views are shared and reinforced.

h. Pseudoscience or False Intellectualism: It may try to justify discrimination with fake science or flawed logic.”

3. **In what ways can linguistic analysis techniques be adapted to better identify different types of online hate speech?**

a. “Identifying Patterns: By looking for repeated words or phrases used to intimidate or offend others. Understanding Context: Considering where and when certain language is used can reveal if it’s meant to harass.

b. Recognizing Intent: By assessing the purpose behind the language, such as whether it’s meant to hurt or threaten.

c. Examining Tone: Looking at how language is structured and if it shows hostility or aggression.

d. Using Technology: Applying tools like AI to analyze large amounts of text and detect harassment patterns.”

4. **How do you see the role of language and linguistics in improving the accuracy of AI models for identifying online hate speech?**

a. “Understanding Context: Linguistics helps AI understand the context in which words and phrases are used. It can distinguish between harmless discussions and harmful language.

b. Identifying Patterns: Linguistic analysis helps AI detect patterns of hate speech, such as specific words or combinations commonly used to attack or intimidate others.

c. Recognizing Cultural Nuances: Language varies across cultures and communities. Linguistics helps AI account for these differences, ensuring more accurate identification of hate speech that might otherwise be missed.

d. Developing Effective Filters: By studying linguistic patterns, AI can be trained to create better filters that catch offensive content while allowing for legitimate expression.

e. Improving Response Strategies: Linguistic insights enable AI to develop more effective strategies for responding to hate speech, such as suggesting interventions or alerting moderators.”

5. **What are the most significant linguistic challenges you have encountered while working with AI to detect and address online hate speech?**

“The main challenges include understanding the context of language, recognizing sarcasm or slang, and keeping up with how language evolves online.”

6. **In what ways do you think AI could potentially replicate or perpetuate harmful language patterns or biases if not properly trained or developed?**

“AI has the potential to replicate or perpetuate harmful language patterns or biases in several ways if not properly trained or developed:

a. Bias in Training Data: If AI models are trained on biased or unbalanced datasets, they may learn and perpetuate existing prejudices or stereotypes present in the data.

b. Language Generation: AI models that generate text or speech can unintentionally produce offensive or biased language based on patterns in the training data, especially if it lacks diversity or includes discriminatory content.

c. Misinterpretation of Context: AI may misinterpret the context of language, leading to inappropriate responses or the promotion of harmful content without recognizing its negative impact.

d. Amplifying Extremist Views: In social media and online forums, AI algorithms can amplify extremist or hateful content by prioritizing engagement metrics (likes, shares) without considering the harm it may cause.

e. Lack of Oversight: Without proper oversight and monitoring, AI systems may continue to propagate harmful language patterns or biases unchecked, exacerbating societal divisions and discrimination.

f. Echo Chambers: AI algorithms that personalize content based on user preferences may inadvertently reinforce users' existing biases by showing them more of the same type of content, including harmful language.

To mitigate these risks, it's crucial to train AI models on diverse and representative datasets, continuously monitor their outputs for bias and harmful content, and implement robust ethical guidelines and policies in AI development and deployment."

6. **In your opinion, what are the limitations of current AI technology when it comes to understanding the complexities of language and detecting nuances in meaning?**

a. "Contextual Understanding: AI struggles to grasp the nuanced meanings of words and phrases within different contexts. It often lacks the ability to interpret sarcasm, irony, humor, or cultural references accurately.

b. Ambiguity: Language can be ambiguous, and AI finds it challenging to disambiguate words or sentences that have multiple interpretations depending on the context.

c. Figurative Language: AI may struggle with understanding figurative language such as metaphors, idioms, or expressions that convey meaning beyond their literal words.

d. Cultural and Social Context: Language is deeply influenced by cultural and social factors, including regional dialects, slang, and norms. AI may not always recognize these variations and may misinterpret or misclassify language based on its training data.

e. Evolution of Language: Language evolves rapidly, especially online, with new words, phrases, and meanings emerging constantly. AI models trained on static datasets may struggle to keep up with these changes.

f. Subtle Nuances: AI may miss subtle nuances in tone, emotion, or intention conveyed through language, leading to inaccurate interpretations or responses.

g. Bias and Stereotypes: AI can inadvertently perpetuate biases present in its training data, leading to biased language processing and potentially reinforcing stereotypes."

7. **How do you see linguistic analysis and AI being used in the future to address the problem of online hate speech?**

a. "Improved Detection and Monitoring: AI-powered systems can continuously monitor online platforms for hate speech by analyzing linguistic patterns and context. This can help identify and flag harmful content more efficiently than human moderators alone.

b. Contextual Understanding: Advanced AI models can be developed to better understand the context in which language is used, distinguishing between legitimate discourse and harmful speech. This includes recognizing sarcasm, irony, and cultural nuances.

c. Real-Time Intervention: AI algorithms can intervene in real-time by providing warnings, suggesting edits, or temporarily blocking content that violates community guidelines or legal standards.

d. Personalized Feedback: AI can provide personalized feedback to users, educating them on appropriate online behavior and the consequences of hate speech. This proactive approach can help prevent the spread of harmful content.

e. Bias Mitigation: By continually auditing and refining AI models, developers can reduce biases in language processing algorithms, ensuring fairer and more equitable treatment of all users.

f. Collaborative Filtering: AI can facilitate collaborative efforts among platforms, researchers, and policymakers to share insights and best practices in combating hate speech and harassment effectively.

g. Ethical Considerations: Integrating ethical guidelines into AI development ensures that technologies designed to address hate speech uphold principles of fairness, transparency, and respect for freedom of expression."

8. **How do you balance the need for effective detection of online hate speech with concerns around privacy and free speech?**

a. "Clear Guidelines and Policies: Establishing clear guidelines and policies that de- fine what constitutes hate speech and harassment is essential. These guidelines should be informed by legal standards, community values, and human rights principles to ensure they are fair and balanced.

b. Transparency and Accountability: Platforms and AI systems used for detection should operate transparently, clearly explaining how content is evaluated and actions are taken. There should be mechanisms for users to appeal decisions and understand the reasoning behind them.

c. Privacy Protection: AI systems should prioritize privacy by design, minimizing the collection and storage of unnecessary personal data. Content analysis should focus on linguistic patterns and context rather than identifying individuals without clear justification.

d. Contextual Understanding: AI should be trained to understand the context in which speech occurs, distinguishing between legitimate expression, heated debates, and genuinely harmful content. This helps prevent overreach in censorship while effectively targeting harmful behavior.

e. Proportionate Responses: Responses to hate speech and harassment should be proportionate to the severity of the offense. AI systems can be designed to recommend actions such as warnings, content removal, or temporary suspensions based on the nature and frequency of violations.

f. Promotion of Free Speech: Platforms should prioritize promoting a diverse range of perspectives while safeguarding against harmful speech. AI can assist in fostering constructive dialogue by highlighting positive contributions and encouraging respectful interactions.

g. Continuous Evaluation and Improvement: Regular evaluation of AI algorithms and policies is crucial to identify and address biases, unintended consequences, and evolving challenges in detecting online hate speech and harassment.

By integrating these principles into the design and implementation of AI technologies, it is possible to enhance the detection of harmful content while upholding privacy rights and preserving the principles of free speech in online environments. Collabo- ration among stakeholders including tech companies, policymakers, researchers, and civil society organizations is essential to achieving this balance effectively."

9. **In your opinion, what kind of oversight or regulation is necessary to ensure that AI is being used ethically and responsibly in detecting online hate speech?**

a. "Clear Guidelines: Establishing clear guidelines and definitions of what constitutes hate speech and harassment, aligned with human rights principles and legal standards.

b. Transparency Requirements: Requiring transparency from platforms and AI developers about how algorithms are designed, trained, and deployed for content moderation.

c. Privacy Protection: Ensuring that AI systems prioritize user privacy and minimize unnecessary data collection, with strict controls on how personal information is used.

d. Bias Audits: Requiring regular audits to detect and mitigate biases in AI algorithms, ensuring fair treatment across different demographics and viewpoints.

e. Accountability Mechanisms: Implementing mechanisms for users to appeal decisions and understand how content moderation actions are taken, promoting account- ability and fairness.

f. Cross-Platform Collaboration: Encouraging collaboration among platforms, researchers, and regulators to share best practices and address challenges collectively.

g. Continuous Evaluation: Mandating ongoing evaluation of AI systems to adapt to evolving forms of hate speech and harassment, with updates based on new insights and user feedback."

Interview 5

1. **How do you define online hate speech?**

"Online hate speech typically involve language that targets individuals or groups based on attributes like race, religion, gender, or sexual orientation, intending to harm, demean, or incite violence. It's not just about offensive language; it's about words or expressions that have the potential to cause real harm, either by perpetuating stereotypes, encouraging discrimination, or directly threatening someone's safety. Harassment often involves repeated or persistent behavior aimed at intimidating or silencing the target."

2. **What are some of the linguistic features and patterns that characterize online hateful content?**

"Online hateful content often includes linguistic features like derogatory terms, slurs, and offensive stereotypes. Beyond explicit language, it may involve subtle patterns like sarcasm, threats veiled in humor, or dehumanizing metaphors. Repetition of certain negative themes or labels (e.g., associating a group with violence or criminality) is also common. Additionally, the use of hyperbolic language, aggressive tone, and the presence of group-based generalizations or "us vs. them" framing can signal hate speech. The challenge is that these features can vary widely across contexts, making detection complex."

3. **In what ways can linguistic analysis techniques be adapted to better identify different types of online hate speech?**

"Linguistic analysis techniques can be adapted by focusing on context, tone, and intent rather than just keywords. For instance, incorporating natural language processing (NLP) models that understand sarcasm or irony can help in detecting subtle forms of harassment. Using sentiment analysis combined with contextual embeddings, like BERT or GPT, allows models to grasp nuanced language patterns. Additionally, adapting these techniques to recognize evolving slang, coded language, or cultural references ensures better identification of harassment that might otherwise go unnoticed. Regularly updating models with new data is also crucial for keeping up with language changes."

4. **How do you see the role of language and linguistics in improving the accuracy of AI models for identifying online hate speech?**

"Language and linguistics play a crucial role in improving the accuracy of AI models for identifying online hate speech. Understanding linguistic nuances helps models better capture context, intent, and the subtleties of human communication that are often missed by simple keyword-based approaches.

For example, linguistics can help AI differentiate between a sarcastic comment and a genuine threat, or understand when a seemingly neutral phrase carries a harmful connotation in certain contexts. Analyzing syntax, semantics, and pragmatics allows models to grasp these complexities, leading to more accurate detection.

Moreover, linguistic research can guide AI in recognizing evolving language patterns, like new slang or coded language, that hate speech often uses to evade detection. By incorporating insights from sociolinguistics, pragmatics, and discourse analysis, AI models can become more adaptable to different cultural and social contexts, reducing both false positives and false negatives."

5. **What are the most significant linguistic challenges you have encountered while working with AI to detect and address online hate speech?**

"The most significant linguistic challenges include:

Context Understanding: Sarcasm, irony, and nuanced expressions can be difficult for AI to interpret correctly.

Evolving Language: Slang and coded language frequently change, making it hard for models to keep up.

Multilingual Issues: Detecting hate speech in different languages and dialects adds complexity.

Subtlety and Ambiguity: Indirect or veiled threats and dehumanizing language can be hard to spot.

Cultural Differences: The same language or phrase can have different connotations in different cultural contexts."

6. **In what ways do you think AI could potentially replicate or perpetuate harmful language patterns or biases if not properly trained or developed?**

"AI could replicate or perpetuate harmful language patterns or biases in several ways:

Amplifying Biases: AI can amplify existing biases from training data, reinforcing stereotypes and prejudiced views.

Misinterpreting Context: Without proper context understanding, AI might misinterpret neutral or ambiguous language as harmful, or vice versa.

Reinforcing Negative Patterns: AI might replicate harmful language patterns it has learned from biased or unrepresentative datasets.

Exacerbating Inequities: Models might disproportionately target or misrepresent certain groups if the training data reflects societal inequalities.

Failing to Adapt: AI that doesn't update with evolving language and social norms may perpetuate outdated or offensive language patterns."

7. **In your opinion, what are the limitations of current AI technology when it comes to understanding the complexities of language and detecting nuances in meaning?**

"Current AI technology has several limitations in understanding language complexities and detecting nuances. The most important one is context sensitivity, AI often struggles with nuanced contexts, like sarcasm or irony, which can lead to misinterpretation. Another one is ambiguous language which can be challenging for AI, especially when words or phrases have multiple meanings. Moreover, cultural and regional variations, AI models may not fully grasp regional dialects or cultural references, leading to inaccurate detections. Subtlety is also an issue for AI model. Detecting subtle nuances, such as veiled threats or implicit biases, can be difficult for AI. Lastly, evolving Language, AI models may lag behind current language trends and new slang, impacting their relevance and accuracy."

8. **How do you see linguistic analysis and AI being used in the future to address the problem of online hate speech?**

"In the future, linguistic analysis and AI can be used to address online hate speech and harassment in several ways:

Enhanced Contextual Understanding: Improved models will better grasp context, sarcasm, and subtle nuances, making detection more accurate.

Real-Time Adaptation: AI systems will continuously learn from new data and evolving language patterns to stay relevant and effective.

Cross-Language Capabilities: Advanced models will handle multilingual and multicultural contexts more effectively, identifying hate speech across different languages and dialects.

Bias Mitigation: Incorporating advanced linguistic techniques and diverse training data will help reduce biases and ensure fairer detection.

User Customization: AI tools will allow for more personalized settings, enabling users to tailor content moderation to their specific needs and community standards."

9. **How do you balance the need for effective detection of online hate speech with concerns around privacy and free speech?**

"Establishing transparent policies is crucial in balancing effective detection of online hate speech with concerns about privacy and free speech. Clear guidelines about what constitutes hate speech and how data is handled, ensuring users understand the rules and their rights. Secondly, minimized data collection, which means collecting only the necessary data for detection and avoid retaining unnecessary personal information to protect user privacy. In addition, use anonymization techniques to safeguard user identities during analysis and detection processes. Implement AI models that understand context and intent to differentiate between harmful content and legitimate expression, minimizing overreach. Allow users to set their own content preferences and moderation levels to give them more control over their online experience while still maintaining safety."

10. **In your opinion what kind of oversight or regulation is necessary to ensure that AI is being used ethically and responsibly in detecting online hate speech?**

"Effective oversight and regulation for ethical and responsible AI use in detecting online hate speech should include clear guidelines and standards. Establishing comprehensive guidelines on what constitutes hate speech and how AI should be used to detect it, ensuring consistency and fairness. Then mandating transparency about how AI models are trained, the data used, and how decisions are made, so users and stakeholders can understand and trust the process. Moreover, setting up accountability mechanisms for when AI systems make errors or cause harm, including clear procedures for addressing grievances and correcting mistakes. Ethical review boards should be established to oversee the development and deployment of AI systems, ensuring they adhere to ethical standards and respect human rights. Implementing regulations that protect user privacy and ensure informed consent, making sure that personal data is handled responsibly is also very important in this regard. In addition, creating regulatory frameworks that are adaptable to technological advancements and evolving language patterns, ensuring that regulations remain relevant and effective.”