**REPORT 13**

**Detection of Hate Speech against LGBTQIA+:**

An Approach Using Speculative Design

**[ANONYMIZED REPORT – NO IDENTIFYING INFORMATION]**


## Abstract

With the increasing use of the Internet and social media platforms, new challenges arise, including the proliferation of online hate speech. This study aims to analyze hate speech in [omitted for blind review] against the LGBTQIA+ community through Artificial Intelligence, exploring detection and classification methods, as well as identifying limitations and challenges. Additionally, it investigates the profile of people who propagate hate speech, seeking to understand the characteristics of these personas. The study employs a speculative design approach to map the current state and predict possible futures, proposing innovative solutions for creating a safer and more inclusive online environment.


## 1. Introduction

There has been an increase in the use of the Internet, as well as social media platforms [Gandhi et al.]. They provide a space for individuals to communicate easily with others, freely sharing their ideas and thoughts. However, this freedom of expression on the Internet brings many challenges that need to be addressed. [Istaiteh et al. 2020] states that one of the most important challenges is dealing with the continuous flow of hate speech incidents reported online worldwide. This research proposes to carry out a comprehensive analysis on the topic of hate speech in [omitted for blind review] against the LGBTQIA+ community. From the perspective of Artificial Intelligence, it intends to understand the main approaches used for its detection and classification, identify limitations, challenges, and data availability. In addition, it seeks to understand the profile of people who reproduce hate speech through comparisons.

[Canbay 2024] states that the personality profile of people who reproduce hate speech does not receive enough attention; recognizing the personality of a person who is practicing cyberbullying or offending someone on social media helps to distinguish that person from potential suspects. In addition, [Alharthi 2021] states that recognizing the characteristics of a possible target of online hate speech is important, as it can help predict potential targets and protect them.

This work intends to carry out a speculative design process on this theme, understanding the current moment, raising possibilities and how the theme connects to everyday life, as well as mapping trends regarding the theme. Based on the identified trends, the second stage of the speculative process intends to understand how they shape the future through a reflection exercise. Finally, speculative design leads to reflecting on the desirable future, where a futuristic solution is conceived that reconfigures the scenario of the theme.

## 2. Theoretical Foundation

### 2.1. What Is "Hate Speech"?

According to Singh [Singh et al. 2023], hate speech can be defined as comments directed at race, religion, gender, age, or any comment with the intent to incite hatred or enmity among people. Below, different types of hate speech are conceptualized.

#### 2.1.1. Targeted Hate Speech

Hate speech is a complex phenomenon, intrinsically associated with relationships between groups and dependent on nuances of language, which makes its identification a challenging task. Thus, [Jahan and Oussalah 2023] divides the concept of hate speech into categories of this discourse, specifying "sexism/gender discrimination," "racism," "religious intolerance," and "other concepts." Figure 1 illustrates the categorization diagram of hate speech concepts presented in [Jahan and Oussalah 2023].

### 2.2. Why Research Hate Speech Specifically against the LGBTQIA+ Community?

From the articles found in this review, a scarcity of material directly related to the detection of hate speech against the LGBTQIA+ community was noted. The search results revealed that most hate speech detection methods focus on other minority groups or are generalist hate speech detectors. This fact highlights the need for more studies that address the experiences and challenges faced by the LGBTQIA+ community in the digital world, as generalist detection methods may not consider specific indicators of prejudice against this group.

### 2.3. Concepts for Detection and Classification of Hate Speech

#### 2.3.1. Natural Language Processing (NLP)

According to Sachi et al. [Sachi et al. 2023], Natural Language Processing (NLP) can be defined as a field of computer science and artificial intelligence that focuses on the interaction between humans and machines through the use of natural language. It is a subarea of computational linguistics and allows computers to understand, analyze, and produce human language.

#### 2.3.2. Machine Learning (ML)

Machine Learning is understood by [Singh et al. 2023] as a computational model that can be trained on data to detect and classify hate speech.

## 3. Speculative Design Methodology

### 3.1. Mapping the Current State

The use of artificial intelligence is increasingly present as part of solutions. Many solutions are enhanced through natural language processing, enabling the training of models for hate speech detection. Figure 2 presents a mental map of the

components involved in the theme, highlighting the types of personas involved, actions of nature, and technologies.

**Personas Involved**

- **Victims of hate speech:** People who are targets of hate speech. In this work, the focus will be people belonging to the LGBTQIA+ community.
- **Producers of hate speech:** People who produce, reproduce, and disseminate hate speech.
- **Government/Society:** Entity responsible for promoting and institutionalizing the regulation of social media use.
- **Private Companies:** Companies responsible for social media platforms, in which users interact with each other and are also responsible for ensuring that each country's regulations are complied with within their scope.

**Nature Field**

The following points were identified:

- **Impacts on mental health:** This aspect addresses the impacts on the mental health of people who are victims of hate speech and how this affects their social interactions.
- **Regulation:** Regulation of social media is proposed by the Government persona, which must ensure a healthy virtual environment for all people, in addition to Private Companies needing to ensure its application in their products.
- **Defenses and values:** This aspect addresses the mapping of values defended by people who reproduce hate speech, whether political, economic, or social.

**Technologies**

- **Social Media Platforms:** These are social networks where participants interact with each other. Hate speech is generally produced in this virtual environment.

- **Content Moderation Systems:** Systems that must allow or block content according to its content and the regulation in force at the time.
- **AI Tools for Detection:** This aspect involves a set of techniques and approaches that use Artificial Intelligence to classify and detect hate speech. This component depends on a large amount of data for training.

Twitter's terms of use prohibit the reproduction of offensive content[1]; however, according to the Center for Countering Digital Hate report, 86% of the 300 posts reported for extreme hate speech were still available on the platform one week after the report [Center for Countering Digital Hate 2023]. Despite this fact, Twitter is adopting a new strategy to reduce the circulation of posts that violate its rules. From now on, the platform will display visible labels on tweets that were removed for violating its policies, explaining the reason for removal. Previously, these posts were simply deleted without any public justification[2].

From this information, the platform will have data for learning, including predictive learning. By using the Innovation Map platform[3], it is possible to verify that predictive conflict models are fully functional prototypes ready for testing in an industrially relevant environment. This predictive analysis integrated with geographic data examines past and present communication channels of social media posts to predict areas susceptible to conflict. It identifies ideal locations for intervention and determines the required level of intervention. Thus, it is possible to verify that conflict prediction solutions based on learning data are already functional, with the possibility of adaptation to the context of this work, predicting potential hate reproducers based on their history.

This type of solution meets two Sustainable Development Goals (SDGs)[4]:

- **SDG 16. Peace, Justice and Strong Institutions[5]:** Promote peaceful and inclusive societies for sustainable development, provide access to justice for all, and build effective, accountable, and inclusive institutions at all levels.

- **SDG 17. Partnerships for the Goals**[6]**:** Strengthen the means of implementation and revitalize the Global Partnership for Sustainable Development.

On the same Innovation Map platform, it was possible to verify the existence of an algorithmic bias detection tool[7]. Its importance lies in the fact that with the increasing integration of machine learning into society, it becomes evident that algorithms are not infallible: algorithmic bias has already been identified in several cases. Due to its intrinsic nature, machine learning can perpetuate statistical discrimination.

The most concerning bias is the one that systematically favors privileged groups and disadvantages marginalized groups. There are different examples of this, such as predictive policing systems that generate vicious cycles of discrimination and recruitment processes that exclude candidates from low-income areas or prefer male candidates over women.

The Algorithmic Bias Detection Tool can be applied in research on the analysis and classification of hate speech against the LGBTQIA+ community to ensure that the models used are fair and equitable. By detecting and correcting biases, it is possible to improve precision and effectiveness in identifying hate speech, ensuring that groups such as the LGBTQIA+ community are not discriminated against.

In addition, the application of these techniques can help establish more ethical and transparent practices in the development of new AI technologies. To combat these problems, it is possible to develop systems that inspect algorithmic models and detect bias at different stages of the machine learning process. This type of solution is in a fully demonstrated prototype phase in an operational environment and is fully applicable to the theme of this work.

This solution meets the following Sustainable Development Goals:

- **SDG 05. Gender Equality**[8]
- **SDG 10. Reduced Inequalities**[9]
- **SDG 16. Peace, Justice and Strong Institutions**[10]
- **SDG 17. Partnerships for the Goals**[11]

Given that many bias detection techniques can overlap with ethical challenges in different areas, such as frameworks for good governance, proper data sharing, and model explainability, a comprehensive solution to algorithmic bias must be established both legally and technically to fill the gap and minimize potential conflicts based on bias and prejudice. Otherwise, an unregulated market with access to increasingly powerful predictive tools may gradually and imperceptibly aggravate social inequality, perhaps even leading to a new era of information warfare. Thus, the current state of the theme suggests that there is much ground to be developed in different aspects.

### 3.2. Speculation of Possible Futures

A medium- to long-term future allows for a significant evolution of Artificial Intelligence technologies, including advances in natural language processing, machine learning, and the implementation of more robust policies against hate speech. In addition, medium and long term is sufficient time to observe social and cultural changes that can influence both the prevalence and perception of hate speech against the LGBTQIA+ community.

The trends identified in Section 3.1 will have a significant impact on the future of hate speech detection against the LGBTQIA+ community. The main reconfigurations may include:

- **Increased Presence of AI in Solutions:** Artificial Intelligence will continue to become an integral part of technological solutions, increasing the capacity for detection and classification of hate speech due to advances in NLP and machine learning.
- **Regulation and Usage Policies:** Regulation of social media may continue to be insufficient. Despite policies prohibiting offensive content, application of these policies may remain weak, allowing much hate speech to remain available on platforms.
- **AI Technologies and Tools:** AI tools may become more sophisticated and accurate, but without a coordinated

implementation strategy, their effectiveness may be limited. Lack of high-quality and diverse data may restrict model training.

- **Content Moderation:** Content moderation systems will continue to evolve, but without clear regulation and rigorous enforcement, moderation may be inconsistent and ineffective, increasing user distrust.

Figure 3 demonstrates a cause-and-effect scheme in the context of this work's theme, highlighting positive, negative, and neutral effects.

### 3.3. Projection of a Desirable Future

Despite advances in solutions for detecting hate speech against the LGBTQIA+ community, many limitations are still found, such as insufficient data and data annotations for training, in addition to algorithms built under biases that do not reflect reality.

The "Future Scenario Maker" tool[12] was used to generate a possible future for 2034. The listed futures aim to solve current limitations and consider an advance in society.

**Future Scenario: The Virtual Reality Enclave**
A solution to combat hate speech against the LGBTQIA+ community in virtual communications. Posts are monitored in real time, even before being published. Texts, images, and videos are analyzed to ensure a safe online environment for all people. This solution is trained with data annotated by the LGBTQIA+ community itself. The solution is supported by unbiased algorithms and applies to all communication media, being a feature of the devices themselves, not limited to specific social networks. In addition, this solution allows the person to learn the reason why content constitutes hate speech, thus providing literacy for users.

To project a possible and desirable future, a series of changes and adaptations to the current scenario is necessary. As explained in this work, regulation, data diversity for training solutions, and, above all, changes in social behavior are essential.

## 4. Conclusions

This work presented a comprehensive analysis on the detection of hate speech against the LGBTQIA+ community using Artificial Intelligence. The main contributions include identifying the main techniques and tools for detection and classification of hate speech, understanding current challenges and limitations, and proposing innovative solutions through the speculative design methodology.

One of the main contributions was the creation of a desirable future scenario, such as the "Virtual Reality Enclave," which offers a holistic approach to monitoring and educating about hate speech in real time. In addition, the work highlighted the importance of using data annotated by the LGBTQIA+ community itself to ensure AI models are fair and representative.

However, the application of speculative design presents some limitations. Its speculative nature may lead to scenarios that do not consider all real-world variables, resulting in solutions that may be difficult to implement in practice. Elements of the methodology, such as data selection and trend interpretation, could have led to different results if conducted differently. For example, reliance on available data may introduce bias, since not all experiences and challenges of the LGBTQIA+ community are equally represented in existing datasets.

In addition, some important issues may have been left aside, such as intersectionality of discrimination and the complexity of social dynamics on social media platforms. To minimize these problems, the study sought to include a diversity of sources and perspectives, as well as validate identified trends based on multiple scenarios.

Future work may expand this research by exploring the integration of emerging technologies, such as emotion analysis and AI-based content moderation, in greater depth. It may also conduct empirical studies to test the effectiveness of proposed solutions in controlled and real environments. In addition, further investigations could focus on the creation of public policies and regulatory frameworks to ensure ethical and effective implementation of hate speech detection technologies.

In summary, this work provides a solid foundation for future research and developments in hate speech detection against the LGBTQIA+ community, promoting a safer and more inclusive online environment.

**References *(kept verbatim from the original)***


Alharthi, R. (2021). Recognizing hate-prone characteristics of online hate speech targets. page 153–155.

Canbay, P. (2024). Predicting discriminative personality profile of haters from digital texts. *Knowledge-Based Systems*, 287.0:111460.

Center for Countering Digital Hate (2023). *X content moderation report*. Technical report.

Gandhi, A. et al. Hate speech detection: A comprehensive review of recent works. *Expert Systems*, e13562.

Istaiteh, O., Al-Omoush, R., and Tedmori, S. (2020). Racist and sexist hate speech detection: Literature review. pages 95–99.

Jahan, M. S. and Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546.0:126232.

Sachi, S. et al. (2023). Hate speech detection using the gpt-2 and natural language processing. pages 276–280.

Singh, R. K. et al. (2023). NLP based hate speech detection and moderation.