Migration Letters

Volume: 20, No: S3(2023), pp. 332-345

ISSN: 1741-8984 (Print) ISSN: 1741-8992 (Online)
www.migrationletters.com

The Role of Legal and Technical Cybersecurity Systems in Mitigating Hate Speech

Dr. Mohamed F. Ghazwi[1], Dr. Hassan Al-Tarawneh[2], Dr. Mohmmad Husien Almajali[3], Dr. Faisal Tayel Alqudah[4], Dr. Ala' Fahed Al-Adwan[5]

Abstract

The research addresses the problem of the increasing phenomenon of hate speech, especially with the technical revolution and the multiplicity of social media, and the role of cybersecurity in mitigating this phenomenon. The analytical and descriptive approach is adopted to profoundly discuss the legal and technical system of cybersecurity and its role in combating hate speech. The research comprises the legal and technical system for cybersecurity in confronting hate speech and the scope of legal texts facing the problem of hate speech. Altogether, the study incorporates the degree of addressing hate speech by protecting service providers in the field of cybersecurity, whether from a legal or technical perspective, especially in the absence of a clear definition of what is called hate speech in international law. The results find that there are several problems related to identifying hate speech on social networking platforms via the Internet and the role that cybersecurity plays in reducing it through the techniques regulating its work, taking into account the necessity to be consistent with the legislative frameworks regulating this phenomenon at the level of various legislations. The results also show a promising path for advancing hate speech recognition within the realm of cybersecurity. Given the results, the research recommends that future research endeavors should focus on refining algorithmic approaches, addressing ethical considerations, fostering interdisciplinary collaboration, and developing solutions that prioritize user well-being while ensuring the security of digital environments.

Keywords: *cybersecurity law, digital environments, hate speech, mitigation, protection.*

1. Introduction

Over the past few years, plenty of events have taken place making it clear that hate speech in general, and online hate speech in particular, can threaten the livelihood, safety, and even lives of people, especially those belonging to minorities in terms of race, religion, or sexual orientation. Frequent hate speech incidents are a sign of the spread of hate speech at the international level (Brown, 2017).

Hate speech has dramatically escalated, as hatred and intolerance are now worrisome, especially through social media used as platforms to spread hate speech that threatens social peace and the unity of society (Simpson, 2018). This discourse contains a destructive ideology aiming at tearing the fabric of society, and therefore it is necessary to

[1] Associate professor at Faculty of Law - Al-Zaytoonah University of Jordan, Law@zuj.edu.jo
[2] Associate Professor Department of Data sciences and Artificial Intelligence Al-Ahliyya Amman University Amman, Jordan, H.Altarawneh@Ammanu.edu.jo
[3] Associate professor at Faculty of Law - Al-Zaytoonah University of Jordan, Moh.almajali@zuj.edu.jo
[4] Assistant professor at Faculty of Law - Al-Zaytoonah University of Jordan, F.alqudah@zuj.edu.jo
[5] A practicing lawyer at the Jordan Bar Association, Alaa.adwan@yahoo.com

combat hate speech, mitigate its intensity, and confront its impacts without restricting freedom of expression (Alkiviadou, 2018).

Now that violence is on the rise on the ground in the world in general and the Middle East in particular, alongside the emergence of negative results from time to time, this subject matter requires discussing its causes and effects (Spanje & Rekker, 2022). One of the most prominent features of hate speech is some countries' calls and statements of the necessity of dividing society based on color, while others call for the need to return to race and expel their refugees because they are outsiders to their countries. The wave of violence and hatred has globally increased as a result of these practices, and this is noticeable in more than one country, which stimulates the growth of terrorism and the escalation of conflicts. In any slightest statement, a person can smell the language of hatred and detestation of others for racist reasons, which appears on the surface now and then (Al-Nasser, 2022).

Hate speech is a broad term that refers to political speech that incites hostility to others from other religions, colors, and races. Since there is no specific definition in legal studies, this term includes every expression insulting to an ethnic, religious, or national group in any form of racism, xenophobia, enmity between religions, fanaticism, or incitement to violence, hatred, or discrimination. Within this context, hate speech is like a wake-up call - the louder it sounds, the greater the risk of genocide. It precedes and reinforces violence (Chen, 2022).

Individuals' opinions differ on hate speech and its relationship to the right to freedom of opinion and expression, as Article 19 of the International Covenant on Civil and Political Rights states "The right to freedom of expression entails special duties and responsibilities and is therefore subject to certain restrictions". Meanwhile, another side believes that freedom of expression is a basic human right. Accordingly, this situation has created a case of controversy that calls for study; especially at this time when several platforms have appeared in social media to produce and spread hate speech.

2. Research Problem

Hate speech incites violence and intolerance, as the devastating impact of hate is nothing new. However, its size and influence have been recently amplified by new communication technologies, especially online social media platforms. Hate speech - including online - has become the most common method of spreading divisive speech on a global scale, threatening peace around the world. The impact of hate speech, especially online social media platforms, extends across many areas, from protecting human rights and preventing atrocities to maintaining peace, achieving gender equality, and supporting children and youth (Guney, Davies, & Lee (2022).

Since combating hatred, discrimination, racism, and inequality is at the heart of the principles of human peace, there is a need to confront hate speech at every turn. Importantly, this mission is emphasized in international human rights frameworks and global efforts to achieve sustainable development goals. The wide dissemination of hate speech through online social media platforms confirms that there are several problems arising related to identifying the concept of hate speech and how to deal with it to reach an appropriate technical mechanism to combat it. The hat discourse must be combated appropriately through the preparation of strong technical mechanisms that are capable and consistent with the legislative frameworks regulating this phenomenon at the level of various legislations.

The suitability of cybersecurity to combat it must also be studied, especially in light of the ironic situation that social media platforms are private property managed by the private sector and the main arena in which most people today exercise their right to freedom of expression, which creates additional problems as well, particularly since

companies are not obligated to comply with the international law of human rights. Therefore, it is not obligated to follow the rules of this law, which seeks to strike a balance between protecting people from facing the consequences of hate speech and protecting their right to freedom of expression. Accordingly, the research problem is reflected in identifying hate speech and its legal deterrence in international conventions, the position of Jordanian legislation on hate speech and its deterrence, the degree to which the Jordanian cybersecurity law addresses the phenomenon of hate speech, and the technical aspects of cybersecurity in addressing hate speech.

3. Research Objectives

Given the research problem, the research objective is as follows:

- Addresses the problem of the increasing phenomenon of hate speech, especially with the technical revolution and the multiplicity of social media and the role of cybersecurity in mitigating this phenomenon.

4. Research Significance

The present research is significant due to the great significance of the research problem raised as states signatories to the International Covenant on Civil and Political Rights (ICCPR) are obligated to criminalize hate speech that is based on incitement to discrimination or violence against vulnerable groups or their members. This research is also important as it creates a technical environment that would combat this phenomenon, such as cybersecurity and its technologies to combat this phenomenon, especially in cyberspace. Now that cyberspace has many distinctive features that make it difficult to identify online hate speech due to its cross-border nature, this leads to the emergence of more problems that require the ability to deal with online hate speech in a way that achieves a balance between people's rights to security and their right to freedom expression and privacy.

5. Method

The analytical and descriptive approach is adopted to discuss the legal and technical system of cybersecurity and its role in combating hate speech.

6. Conceptual Framework

The nature of the research requires structuring the conceptual framework to be in two sections. Section one discusses the nature of the cybersecurity legal systems in mitigating hate speech. Section two gives an insight into the cybersecurity technical systems in mitigating hate speech.

6.1 Cybersecurity Legal Systems in Mitigating Hate Speech

The legal framing or approach of hate speech faces a very complex problem because describing this speech as a crime requires the availability of the elements and conditions that the law considers to be a reason for conviction and the possibility of punishment or prohibition, along with assessing the material and moral damages incurred by individuals. Accordingly, agreement on a unified international law on the prohibition of hate speech is difficult, as the conditions for the legal prohibition of hate speech are included in the International Covenant on Civil and Political (ICCPR) and the American Convention on Human Rights (ACHR). However, the European Convention on Human Rights (ECHR) and the African Charter on Human and Peoples' Rights (ACHPR) allow this, as they do

not explicitly require governments to prohibit hate speech due to the lack of a clear definition of what is called hate speech in international law (Abu Yousef, 2022).

The law derives its binding force from the unity of the standards used in defining the concept of hate speech, its sources, and its general characteristics, despite the international consensus that the prohibition of hate speech is represented by the legal control of the approved characteristics of this speech until it becomes a criminal act that must be punished within the various international legislation. Accordingly, this section is divided into two parts as follows:

6.1.1 Hate speech and legal prohibition in international conventions and Jordanian legislation.

In this section, clarification of hate speech and the nature of the prohibition against it in some international conventions and Jordanian legislation are provided in two subdivisions: the legal prohibition of hate speech in international conventions and the legal prohibition of hate speech in Jordanian legislation.

a. Legal prohibition of hate speech in international conventions

Hate speech tackles controversial issues such as human dignity and security, equality between individuals, and freedom of expression. Now that hate speech is not explicitly declared in many international documents and treaties on human rights, the prohibition of hate speech is addressed in some international human rights conventions, namely: the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights.

First: Hate speech in the Universal Declaration of Human Rights

Article (18) stipulates “Everyone has the right to freedom of thought, conscience and religion; this right includes freedom to change his religion or belief, and freedom, either alone or in community with others and in public or private, to manifest his religion or belief in teaching, practice, worship and observance” (UDHR, 1948). Thus, this right is irreducible from its moral and legal value, and this is confirmed by Article 19 of the same declaration.

Though the Universal Declaration of Human Rights does not explicitly address the issue of incitement to or advocacy of hatred, it is evident through the interpretation of several provisions in the Universal Declaration of Human Rights that it allows states to intervene to prohibit hate speech or speech that is considered provocative or inciting hatred. The permission of states to intervene to ban hate speech is understood from Article (1) of the Universal Declaration of Human Rights, which stipulates “Everyone is entitled to all the rights and freedoms outlined in this Declaration, without distinction of any kind, such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status” (UDHR, 1948).

In the same context, Article (7) of the Universal Declaration of Human Rights, which provides more explicitly for protection against discrimination and incitement to discrimination, stipulates “All are equal before the law and are entitled without any discrimination to equal protection of the law. All are entitled to equal protection against any discrimination in violation of this Declaration and against any incitement to such discrimination” (UDHR, 1948). Another confirmation is also shown in Article 29 of the same declaration, stipulating “In the exercise of his rights and freedoms, everyone shall be subject only to such limitations as are determined by law solely to secure due recognition and respect for the rights and freedoms of others and of meeting the just requirements of morality, public order and the general welfare in a democratic society” (UDHR, 1948), which refers to the duties that each person bears towards the group, and is interpreted as a matter of concern for the due recognition and respect for the rights and freedoms of others (Salmani, 2022).

Second: Hate Speech in the International Covenant on Civil and Political Rights of 1966

The International Covenant is of great significance in addressing hate speech, although it does not explicitly use the term hate speech. In the International Covenant on Civil and Political Rights, the person's rights to freedom of thought, conscience, and religion." are reaffirmed in the first paragraph of Article 18 stipulating "Everyone shall have the right to freedom of thought, conscience and religion. This right shall include freedom to have or to adopt a religion or belief of his choice, and freedom, either individually or in community with others and in public or private, to manifest his religion or belief in worship, observance, practice, and teaching" (ICCPR, 1966) and the second paragraph of the same article stipulating "No one shall be subject to coercion which would impair his freedom to have or to adopt a religion or belief of his choice" (ICCPR, 1966).

However, the third paragraph of Article 19 of the International Covenant, stipulating "The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary: (a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order, or public health or morals" (ICCPR, 1966) places a restriction on freedom of expression by imposing respect for the rights or reputations of others, and therefore any hate speech is prohibited (Salmani, 2022).

b. Legal prohibition of hate speech in Jordanian legislation

The penalties set by the Jordanian legislator to combat all forms of hate speech vary imposing many penalties that are stated in much legislation in Jordan as a punishment for actions that the legislator considered a form of hate speech. These penalties are mentioned in the articles (130, 150, 287, and 467 Bis) of the Penal Code (1960), as follows:

Article (130) stipulates "Any person, who at time of war or when the start of such war is anticipated, started in the kingdom a propaganda to weaken the national feeling or stir racist or sectarian differences, shall be punished by temporary imprisonment with hard labor".

Article (150) stipulates "Any writing or speech aims at or results in stirring sectarian or racial prejudices or the incitement of conflict between different sects or the nation's elements, such act shall be punished by imprisonment for no less than six months and no more than three years and a fine not to exceed five hundred dinars (JD 500)".

Article (287) stipulates "Whoever kidnaps or hides a child who is under seven years old or replaced him/her with another one or falsely attributes him/her to a woman, he/she shall be punished by imprisonment from three months to three years. The penalty shall not be less than six months if the aim or the result of the crime is the falsification or alteration of the information related to the child's status or the registration of false personal status information at the official registrars".

Article (467) stipulates "Whoever commits the following shall be punished by a fine not to exceed five dinars (JD5): With no necessity makes noise or clatter in a way disturbing the calm of inhabitants. With no necessity to throw stones, other solid objects, or dirt at cars, buildings, fences, gardens, and pools of others or release a harmful animal, or an insane, under his/her guardianship. Orders his/her dog to attack or follow pedestrians, or not holding it from such actions; even if no harm or damage is made".

Article (41) of the Military Punishment Law stipulates "The following acts committed during armed conflicts are considered war crimes: killing with intent, torture or inhumane treatment, including life science experiments, deliberate inflicting severe pain, serious harm to physical or mental safety or public health, compelling prisoners of war or protected civilians to serve in the armed forces of the enemy country, taking hostages,

unlawful detention of civilians protected under the Fourth Geneva Convention of 1949, unlawful and arbitrary military necessity, destroying or seizing property without justification, attacks directed against the civilian population or individuals, and an indiscriminate attack committed against the civilian population or civilian property in the knowledge that such attack will cause severe loss of life, injury to civilians, or damage to civilian property.

Article (20/L) of the Audiovisual Law of (2015) stipulates "The licensing agreement between the Commission and the licensee shall be regulated, after the approval of the Council of Ministers to grant the broadcasting license, provided that it includes in particular the terms, conditions, and matters shown below, in addition to any other conditions provided for in this law and the regulations issued pursuant thereto:

L- The licensee's commitment to the following:

1- Respect for human dignity, personal privacy, freedoms and rights of others, and pluralism of expression.

2- Refrain from broadcasting anything that offends public decency, incites hatred, terrorism, or violence, stirs up strife and religious, sectarian, or ethnic strife, harms the economy and the national currency, or disturbs national and social security.

3- Not broadcasting false materials that harm the Kingdom's relations with other countries.

4- Not to broadcast media or advertising materials that promote sorcery, misleading, blackmail, and deceiving the consumer.

Article (38) of the Press and Publications Law of (2012) stipulates "Reproduced or quoted press material shall be treated as authored or original material" and article (46) stipulates "A. If the responsible chief editor of the press publication violates any provision of Paragraphs A and B of Article 27 of this law, the lawsuit shall be filed against him by the aggrieved party or the director. If a foreign publication violates the provisions of Paragraph C of Article 27 of this law, the lawsuit against it shall be filed by the director".

More importantly, Cybercrime Law No. (17) of 2023 recently issued in its entirety, and in particular Article (17) thereof, being the closest to the subject of our study, stipulating "Whoever intentionally uses an information network, information technology, information system, website, or social media platform to spread what is likely to stir up racist or sedition, targets social peace, incites hatred, calls for or justifies violence, or insults religions, shall be punished by imprisonment from one year to three years or a fine of no less than (5000) five thousand Dinars and no more than (20000) twenty thousand Dinars, or both penalties". The reader of the texts of the above-mentioned articles in the Jordanian laws finds that they include the prohibition of hate speech and that harsh penalties have been imposed on this speech, without prejudice to the basic right of a person to express their opinion.

6.1.2 The role of the Jordanian cybersecurity law in mitigating hate speech

For regulating cybersecurity in Jordan and its related services, a cybersecurity law has been approved in the Hashemite Kingdom of Jordan under No. (16) Of (2019), in which the National Council for Cybersecurity and the National Center for Cybersecurity have been recently established. A close reading of the articles of the Cybersecurity Law in Jordan demonstrates that they do not clearly and accurately specify the obligations of cybersecurity service providers. Articles of the Cybersecurity Law are limited to stating providers of cybersecurity services in the text of Article (10/a) thereof without referring to the obligations required of them, as the above article stipulates that any person or entity should not provide cybersecurity services except after obtaining the licenses and permits required by law. However, the Cybersecurity Law does not clarify in general the legal

obligations that these individuals and entities bear regarding their provision of these services.

A deep analysis of the aspect of the cybersecurity technical systems in this research and comparing it to what has been stated in the articles of the Jordanian Cybersecurity Law evinces that there is a significant and fundamental shortcoming in cybersecurity technical systems for its providers. Jordanian Cybersecurity Law has also overlooked the penalties imposed on those who violate their obligations while providing those services. Against these findings, an urgent legislative intervention to enact detailed instructions for those services and the responsibilities of their providers is currently required to face the new challenges facing cybersecurity services for their role at all levels, together with mitigating hate speech through technical means, the information network, websites, or any other technical means (Abdulsalam, 2022).

6.2 Cybersecurity Technical Systems in Mitigating Hate Speech

Social media platforms afford users the swift and convenient publication of diverse content spanning various subjects. The ease of disseminating content and the provision of anonymity within these platforms have been identified as factors that might amplify the proliferation of deleterious content. Diverse forms of information can cause harm, whether by design or inadvertently (Giachanou & Rosso, 2020). Such categories encompass misinformation, disinformation, and misinformation. Misinformation, as previously explored (Aswani, Kar, & Ilavarasan, 2019; Kar & Aswani, 2021) entails the circulation of factually inaccurate or fictional content, often without regard for truthful intent. Deliberately fabricated content used to mislead such as disinformation exemplified by fake news is subject to scrutiny (Nasir, Khan, & Varlamis, 2021). Similarly, malformation, including instances like hate speech that aims to incite harm, has been examined within scholarly discourse (Giachanou & Rosso, 2020). This investigation particularly focuses on the task of detecting hate speech.

Prominent social media platforms, including Facebook, Twitter, and YouTube, have asserted their commitment to addressing this issue through policies addressing hate behavior and concerted efforts to combat hate speech (Facebook, 2020). In Q1 2023, hate speech content on Meta's Facebook accounted for a prevalence rate of 0.02 percent. This translates to approximately two instances of hate speech for every 10,000 content views on the platform. Notably, the overall incidence of content categorized as hate speech has exhibited stability since Q1 2022 (See Figure 1).

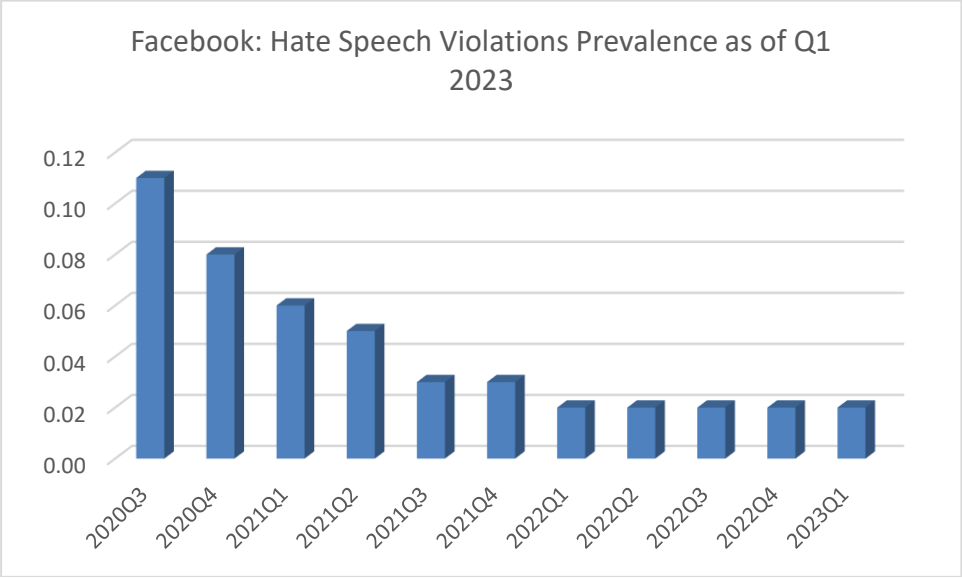

Fig 1: Prevalence of hate speech violations on Facebook worldwide from 3rd quarter 2020 to 1st quarter 2023

Referring to Figure 2, which utilizes data from a report dated March 2022, it can be observed that TikTok addressed 36 percent of the reported cases involving posts promoting anti-Muslim sentiments. Among the 50 flagged posts on TikTok, a total of 12 posts were effectively eliminated, and enforcement measures were applied against 6 user accounts. In contrast, when considering Facebook, out of the 125 posts that were reported, a mere 7 posts underwent any form of action, and no user accounts were deactivated as a result of content reporting. Worth noting is the lack of response from YouTube regarding reported instances of anti-Muslim hate speech.

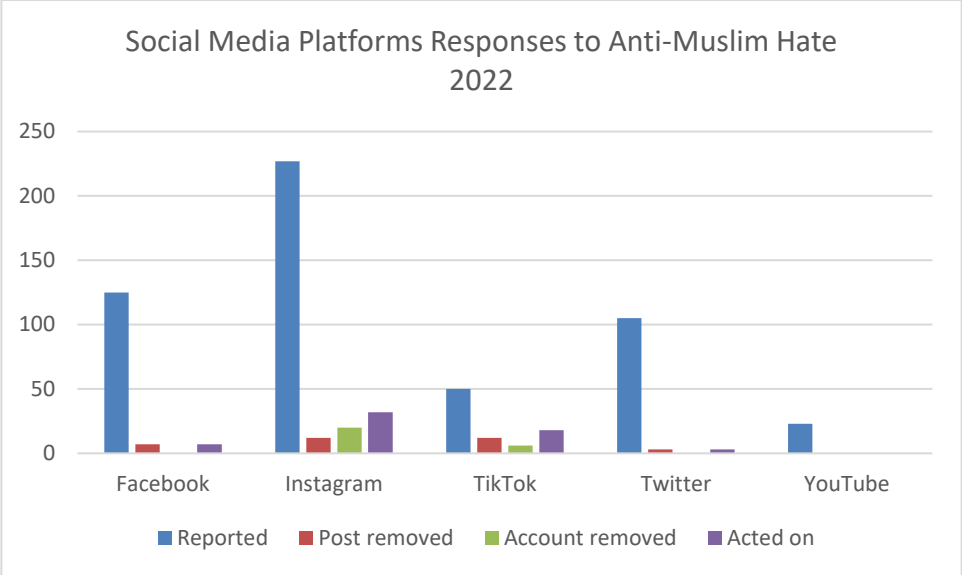

Fig 2: Count of Addressed Anti-Muslim Hate Posts by Leading Global Social Media Platforms as of March 2022

Presently, much of the moderation of such content relies on manual evaluation of potentially problematic content (Waseem & Hovy 2016). Nonetheless, the rapid transmission (sharing) of these messages renders manual content control labor-intensive, time-consuming, costly, and lacking scalability (Cao, Lee, & Hoang, 2020).

The intricate challenge of identifying and mitigating hate speech underscores its explicit intent to foment harm or propagate animosity towards specific groups. This issue is acknowledged as a global concern, affecting various nations and entities. The surge in online social media usage has led to the rapid dissemination of extensive information by millions of users per second, rendering the issue particularly significant. A prevailing understanding posits that individuals tend to express more aggressive speech when they perceive a sense of physical security (Watanabe, Bouazizi, & Ohtsuki, 2018). Additionally, an observable trend exists where hate groups actively seek to enlist individuals for the creation and dissemination of hate speech messages (Del Vigna, Cimino, DellOrletta, Petrocchi, & Tesconi, 2017).

Therefore, the issue of hate speech proliferating effortlessly across online platforms holds significant implications for our society. This is particularly noteworthy due to the potential harm it poses to both individual victims and the broader societal fabric. For instance, the propagation of hate speech can foster animosity among various groups, leading to detrimental consequences (Miškolci, Kováčová, & Rigová, 2020). Specifically, repeated exposure to hate speech may result in desensitization to its violent nature, thereby diminishing victims' perceptions while exacerbating bias against targeted groups (Mathew, Dutt, Goyal, & Mukherjee, 2019).

However, Cybersecurity technology plays a crucial role in combating hate speech by employing algorithms and tools to identify and remove offensive content from online platforms. Research has explored how cybersecurity intersects with hate speech

mitigation, prompting the development of innovative solutions. This convergence addresses both technological and societal challenges, aiming to enhance online safety. Cybersecurity safeguards digital systems against unauthorized access, while hate speech detection targets harmful content. Combining these efforts seeks to create secure online environments and shield users from hate speech's harmful effects. This approach often employs advanced technologies like natural language processing and machine learning to automatically identify and moderate hateful content.

The recognition of hate speech automatically has predominantly been approached as a challenge within the realm of natural language processing. Up to now, researchers have employed automated identification techniques to spot instances of hate speech across various online social platforms. These platforms encompass Facebook and MySpace (Badjatiya, Gupta, Gupta, & Varma, 2017; Maisto, Pelosi, Vietri, & Vitale, 2017). Huang, Inkpen, Zhang, & Van Bruwaene (2018), YouTube, Instagram, and Whisper, in addition to Reddit and Slashdot, along with (Bourgonje, Moreno-Schneider, Srivastava, & Rehm, 2018) utilized machine learning techniques to detect hate speech on social media. Their study demonstrated the efficacy of utilizing a combination of lexical and semantic features for accurate classification. There is also a focus on incorporating contextual information to enhance hate speech detection algorithms. They highlighted the importance of considering user interactions and social context in refining automated detection.

Vujičić & Mladenović (2023) presented that the emergence of Hate Speech (HS) within the realm of social media (SM) has served as a catalyst for in-depth exploration into effective methodologies for its identification. Notably, this particular study narrowed its focus to the sports domain and the specific linguistic context of Serbian. A comprehensive lexicon of HS terms was meticulously compiled, complemented by a meticulously annotated dataset of comments sourced from sports news sections on digital platforms and YouTube sports channels. Within this framework, dual sets of word embeddings were trained to fuel the mechanisms of Deep Learning (DL) models to prompt an investigation into the feasibility of deploying domain-agnostic attributes for HS detection. A prominent outcome of this study was the illumination of the profound implications of HS on the lives of athletes, compelling an acknowledgment of their vulnerability as a targeted demographic. Looking ahead, the trajectory of subsequent research endeavors is spearheaded by encompassing the refinement of classification outcomes, expansion of existing datasets, and an exploration of alternative model configurations.

Schmidt & Wiegand (2017) explored the application of NLP techniques to identify offensive and hateful language. Their study emphasized the significance of linguistic patterns and semantic context in distinguishing between different forms of harmful content. Moreover, Fortuna, Nunes, & Gomes (2018) employed NLP tools to classify online hate speech in multiple languages. Their research highlighted the challenges of cross-lingual hate speech detection and the need for adaptable models.

Abozinadah, Mbaziira, & Jones (2015) proposed an approach that detecting abusive accounts in Arabic tweets using text classification. It preprocesses the data to uniquely identify words, enhancing indexing efficiency. Through experiments, the Naïve Bayes classifier outperforms other classifiers with 90% accuracy using a common set of features and tweets. Different tweet and feature combinations are tested, with the Naïve Bayes classifier showing the best performance. The study also highlights the effectiveness of normalized tweets and varied features in identifying abusive accounts.

Nobata, Tetreault, Thomas, Mehdad, & Chang (2016) investigated the use of deep learning models for hate speech detection. They demonstrated the effectiveness of convolutional neural networks (CNNs) in capturing complex linguistic features for classification. Chatzakou, Kourtellis, Blackburn, De Cristofaro, Stringhini, & Vakali (2017) employed machine learning techniques to classify offensive content and hate

speech. Their study highlighted the importance of feature selection and model optimization in achieving high detection accuracy. Kwok & Wang (2013) explored the integration of text and visual features for identifying hate speech in images and accompanying text. Their study emphasized the potential of multimodal analysis in capturing subtle forms of hate speech.

Burnap, Williams, Sloan, Rana, Housley, & Edwards (2015) discussed the implementation of real-time monitoring systems to detect and address hate speech incidents as they occur. They highlighted the role of such systems in facilitating timely interventions. [32[introduces "HaterNet," an intelligent system developed in collaboration with the Spanish National Office Against Hate Crimes, designed for detecting and analyzing hate speech on Twitter. HaterNet consists of two main components: a novel text classification model to identify hate speech and a social network analysis module to visualize its evolution. For the text classification module, the authors conducted experiments with 19 strategies, combining different features and classification models. The best-performing model achieves an AUC of 0.828, utilizing word embeddings, emojis, token expressions, and tf-idf enrichment. Additional features like POS tags and suffixes were tested but showed no significant positive impact. HaterNet's classifier employs a dual deep learning model, combining an LSTM and MLP neural network with frequency features, surpassing previous models in performance.

Schmidt, Zollo, Del Vicario, Bessi, Scala, Giesecke, & Quattrociocchi (2020) examined the ethical implications of using automated tools for hate speech detection. They emphasized the importance of addressing algorithmic biases and ensuring transparency in classification decisions. Kshetri (2020) discussed the potential of interdisciplinary collaboration between cybersecurity experts, linguists, and social scientists to develop holistic approaches for hate speech detection and prevention. In conclusion, the literature on cybersecurity technology's role in reducing hate speech underscores the significance of employing advanced techniques like NLP, machine learning, and multimodal analysis. Studies emphasize the need for real-time monitoring, ethical considerations, and collaboration across disciplines to create effective solutions for promoting a safer digital environment.

7. Discussion

The literature survey highlights the growing significance of automated hate speech recognition within the cybersecurity domain, specifically through the lens of natural language processing (NLP). Various studies have showcased the efficacy of utilizing machine learning techniques for the detection and classification of hate speech across diverse online platforms. The incorporation of contextual information, user interactions, and linguistic patterns has been underscored as pivotal in refining hate speech detection algorithms.

The emergence of domain-specific studies, such as (Davidson, Warmsley, Macy, & Weber, 2017) investigation into sports-related hate speech, accentuates the importance of tailoring detection approaches to unique linguistic contexts. This direction opens avenues for exploring domain-agnostic attributes and model configurations, with the ultimate goal of achieving higher accuracy and adaptability. The ethical considerations brought forth by (Mandl, Gainsford, & Longo, 2019) and the emphases on interdisciplinary collaboration by (Pereira-Kohatsu, Quijano-Sánchez, Liberatore, & Camacho-Collados, 2019) demonstrate the multifaceted nature of tackling hate speech. These aspects highlight the need to address algorithmic biases, ensure transparency, and leverage expertise from various fields.

The integration of deep learning models and multimodal analysis, as demonstrated by (Nobata, Tetreault, Thomas, Mehdad, & Chang, 2016) showcases the potential for

leveraging advanced techniques to capture complex linguistic features in both text and visual content. Furthermore, the implementation of real-time monitoring systems, as discussed by Burnap et al., adds a dynamic layer to hate speech prevention by enabling prompt interventions.

Looking forward, the literature raises several challenges that warrant further exploration. Cross-lingual hate speech detection, for instance, presents a complex puzzle due to language-specific intricacies and cultural variations. Future research could delve into the development of adaptable models capable of transcending language barriers while accounting for these nuances. Moreover, the ever-evolving nature of online platforms demands a continuous adaptation of detection methods to address emerging trends, including hate speech conveyed through images and memes.

Furthermore, it is imperative to evaluate the effectiveness of hate speech detection systems in terms of their impact on user behavior and community dynamics. Research efforts could delve into understanding how the deployment of such systems shapes online interactions and contributes to cultivating a healthier digital ecosystem.

In conclusion, the discussion underscores a dynamic landscape where technological innovations, ethical considerations, and interdisciplinary collaboration converge to combat hate speech within the realm of cybersecurity. Continued research and innovation will play a pivotal role in refining hate speech detection methodologies, mitigating biases, and fostering online spaces that are safer and more inclusive.

8. Future Work

Building on the insights provided by the reviewed literature, there are several avenues for future research in the cybersecurity field pertaining to hate speech recognition:

Refinement of Algorithmic Approaches: Investigate the fusion of lexical, semantic, and contextual features to enhance the accuracy and robustness of hate speech detection algorithms. This includes exploring more sophisticated machine learning models and novel methods within the realm of deep learning.

Cross-Linguistic and Multilingual Analysis: Extend the scope of hate speech detection to encompass a wider range of languages, taking into account the challenges of cross-lingual hate speech detection. This could involve developing adaptable models and linguistic resources for languages with limited data.

Real-Time Monitoring and Intervention: Further develop real-time monitoring systems that can effectively detect and respond to hate speech incidents on various online platforms. Incorporate mechanisms for automatic content moderation and collaboration with platform providers to ensure timely interventions.

Ethical Considerations and Bias Mitigation: Conduct in-depth studies on algorithmic biases in hate speech detection models and propose strategies for minimizing bias. Develop transparent and explainable models to ensure accountability and address potential fairness issues.

Interdisciplinary Collaboration: Promote interdisciplinary collaboration between cybersecurity experts, linguists, social scientists, and legal scholars. This collaboration could lead to the development of holistic approaches that consider both technical and socio-cultural aspects of hate speech.

Impact Assessment and User Well-being: Investigate the broader societal impact of hate speech on individuals and communities. Explore how effective hate speech detection and prevention can contribute to a safer digital environment and mitigate harm.

Dataset Expansion and Standardization: Curate and expand annotated datasets across multiple languages and platforms to create standardized benchmarks for evaluating hate speech detection models. This could facilitate fair comparisons and drive advancements in the field.

Privacy-Preserving Approaches: Explore techniques for hate speech detection that respect user privacy and data protection regulations. Investigate methods for performing analysis while minimizing the exposure of personal information.

9. Conclusion

In a nutshell, it is found that there are several problems related to identifying hate speech on social networking platforms via the Internet and the role that cybersecurity plays in reducing it through the techniques regulating its work, taking into account the necessity to be consistent with the legislative frameworks regulating this phenomenon at the level of various legislations. It is also shown that social media platforms are private property run by the private sector at the same time that they are the main arena in which most people today exercise their right to freedom of expression, creating additional problems. In other words, companies are not obligated to abide by international human rights law, and therefore companies are not obligated to follow the rules of this law that seek a balance between protecting people in the face of the consequences of hate speech and protecting their right to freedom of expression.

Besides, the study indicates that the cybersecurity law in Jordan does not address the legal obligations incurred by individuals and entities when providing these services. The study also indicated that after addressing and comparing the technical aspect with the articles of the Jordanian Cybersecurity Law, there is a significant and fundamental shortcoming in that law relating to service providers. Importantly, the study indicated that the law does not clarify the penalties for those who violate their obligations during the provision of these services, which requires urgent legislative intervention to put in place the detailed regulation of those services and the responsibility of their providers.

More importantly, the study shows that the legislative intervention should be rapid to confront the new challenges facing cybersecurity services and their role at all levels, including the reduction of online hate speech, the information network, any website, or any other technical means. In conclusion, the synthesis of the reviewed literature points to a promising path for advancing hate speech recognition within the realm of cybersecurity. Future research endeavors should focus on refining algorithmic approaches, addressing ethical considerations, fostering interdisciplinary collaboration, and developing solutions that prioritize user well-being while ensuring the security of digital environments.

References

Abdulsalam, S. (2022). Legal Framework for Cybersecurity Services. Hamat Al Haq Publications: Cairo.

Abozinadah, E. A., Mbaziira, A. V., & Jones, J. (2015). Detection of abusive accounts with Arabic tweets. Int. J. Knowl. Eng.-IACSIT, 1(2), 113-119.

Abu Yousef, K. (2022). Legislative confrontation of hate speech. Amman Arab University Journal for Research, 4(1), 1-14.

Al-Nasser, E. (2022). Hate speech, challenges, and ways of confrontation. Istanbul Jordan of Arabic Studies Journal, 1(2), 1-22.

Alkiviadou, N. (2018). The Legal Regulation of Hate Speech: The International and European Frameworks. Croatian Political Science Review, 55(4), 203-229. http://doi.org/10.20901/pm.55.4.08

Aswani R., Kar A. K., & Ilavarasan P. V. (2019). Experience: managing misinformation in social media insights for policymakers from Twitter analytics. Journal of Data and Information Quality, 12(1), 1–18.

Badjatiya P, Gupta S, Gupta M, Varma V. (2017). Deep Learning for Hate Speech Detection in Tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion—WWW '17 Companion. Perth, Australia: ACM Press; 759–60.

Bourgonje P., Moreno-Schneider J., Srivastava A., & Rehm G. (2018). Automatic classification of abusive language and personal attacks in various forms of online communication. In: Rehm G, Declerck T, editors. Language technologies for the challenges of the digital age, vol. 10713. Cham: Springer International Publishing; 2018. p. 180–91.

Brown, A. (2017). What is hate speech? Part 2: Family Resemblances. Law and Philosophy 36(1), 561–613. http://doi.org/10.1007/s10982-017-9300-x

Burnap, P., Williams, M. L., Sloan, L., Rana, O., Housley, W., & Edwards, A. (2015). Tweeting the terror: Modelling the social media reaction to the Woolwich terrorist attack. Social Network Analysis and Mining, 5(1), 29-50.

Cao R., Lee R. K. W., & Hoang T. A. (2020). Deephate: Hate speech detection via multi-faceted text representations [Conference session]. In: 12th ACM Conference on Web Science, WebSci '20. ACM (p. 1120)

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean Birds: Detecting Aggression and Bullying on Twitter. In Proceedings of the 2017 ACM on Web Science Conference.

Chen, G. (2022). How equalitarian regulation of online hate speech turns authoritarian: a Chinese perspective. Journal of Media Law, 14(1), 159–179 https://doi.org/10.1080/17577632.2022.2085013

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In Proceedings of the 11th International AAAI Conference on Web and Social Media.

Del Vigna F., Cimino A., DellOrletta F., Petrocchi M., & Tesconi M. (2017). Hate me, hate me not: Hate speech detection on Facebook [Conference session]. Proceedings of the First Italian Conference on Cybersecurity (ITASEC17) (pp. 86–95).

Fortuna, P., Nunes, S., & Gomes, P. (2018). A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys, 51(4), 1-30.

Giachanou, A & Rosso, P. (2020). The battle against online harmful information: The cases of fake news and hate speech [Conference session]. Proceedings of the 29th ACM International Conference on Information & Knowledge Management. ACM (pp. 3503–3504)

Guney, G., Davies, D., & Lee, P. (2022). Towards Gender Equality in Law: An Analysis of State Failures from a Global Perspective. Palgrave Macmillan: Switzerland. https://doi.org/10.1007/978-3-030-98072-6

Huang, Q., Inkpen, D., Zhang J., & Van Bruwaene, D. (2018). Cyberbullying intervention interface based on convolutional neural networks. In: Proc First Workshop Trolling Aggress Cyberbullying. 2018. p. 42.

Kar A. K & Aswani R. (2021). How to differentiate propagators of information and misinformation–insights from social media analytics based on bio-inspired computing. Journal of Information and Optimization Sciences, 42(6), 1307–1335.

Kshetri, N. (2020). Big Data's Role in the Fight against Pandemics: A Literature Review. Technology in Society, 62, 101374.

Kwok, I., & Wang, Y. (2013). Locate the Hate: Detecting Tweets against Blacks. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics.

Maisto A, Pelosi S, Vietri S, Vitale P. (2017). Mining Offensive Language on Social Media. In: Basili R, Nissim M, Satta G, editors. In: Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017. Academia University Press; 2017. p. 252–6

Mandl, S., Gainsford, A., & Longo, Y. (2019). Hate Speech Detection on Social Media: The Case of the UK. In Proceedings of the European Conference on Social Media.

Mathew B., Dutt R., Goyal P., & Mukherjee A. (2019). Spread of hate speech in online social media [Conference session]. Proceedings of the 10th ACM Conference on Web Science (pp. 173–182).

Miškolci J., Kováčová L., & Rigová E. (2020). Countering hate speech on Facebook: The case of the Roma minority in Slovakia. Social Science Computer Review, 38(2), 128–146.

Nasir J. A., Khan O. S., Varlamis I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. International Journal of Information Management Data Insights, 1(1), 100007.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive Language Detection in Online User Content. In Proceedings of the 25th International Conference on World Wide Web.

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech on Twitter. Sensors, 19(21), 4654

Salmani, H. (2022). Criminalizing the sins of hate in international human rights conventions. Journal of Comparative Legal Studies, 7(1), 1416-1441.

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. Computational Linguistics, 43(4), 837-889.

Schmidt, A., Zollo, F., Del Vicario, M., Bessi, A., Scala, A., Giesecke, J., & Quattrociocchi, W. (2020). Anatomy of news consumption on Facebook. Proceedings of the National Academy of Sciences, 117(6), 2765-2773.

Simpson, R. (2018). 'Won't Somebody Please Think of the Children?' Hate Speech, Harm, and Childhood. Law and Philosophy, 38(2), 79–108. https://doi.org/10.1007/s10982-018-9339-3

Spanje, J & Rekker, R. (2022). Hate Speech Prosecution of Politicians and its Effect on Support for the Legal System and Democracy. British Journal of Political Science, 52(1), 886–907. http://doi.org/10.1017/S000712342000068X

Vujičić, S., Mladenović, M. (2023). An approach to automatic classification of hate speech in sports domain on social media. J Big Data 10, 109 (2023).

Waseem, Z & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on Twitter [Conference session]. Proceedings of the NAACL Student Research Workshop. ACL (pp. 88–93).

Watanabe H., Bouazizi M., & Ohtsuki T. (2018). Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE Access, 6, 13825–13835.