https://www.periodicos.unis.edu.br/mythos/

CYBERSECURITY DISCOURSE IN ONLINE COMMUNITIES: A THEMATIC ANALYSIS OF REDDIT AND DISCORD

DISCURSO SOBRE CIBERSEGURANÇA NAS COMUNIDADES ONLINE: UMA ANÁLISE TEMÁTICA NO REDDIT E DISCORD

Erick De Oliveira[1*], Rodrigo Franklin Frogeri[2], Ana Amélia Furtado de Oliveira[3]

[1] *B.S. in Information Systems (UNIS-MG). University Center of Southern Minas - UNISMG, Varginha, Brazil.* erick.oliveira1@alunos.unis.edu.br

[2] *PhD in Information Systems and Knowledge Management (FUMEC). University Center of Southern Minas - UNISMG, Varginha, Brazil.* rodrigo.frogeri@professor.unis.edu.br

[3] *PhD in Linguistic Studies (UNESP). University Center of Southern Minas - UNISMG, Varginha, Brazil.* ana.furtado@professor.unis.edu.br

Editorial Details
Double-blind review system
Memory of a Scientific Event
XI Simpósio Mineiro de Gestão, Educação, Comunicação e Tecnologia da Informação – SIMGETI (2025).

Article History:
Received: December 18, 2025
Revised: January 18, 2026
Accepted: March 27, 2026
Published online: July 08, 2026

Editor-in-Chief:
Rodrigo Franklin Frogeri
Guest Editors:
Pedro dos Santos Portugal Júnior
Fabrício Pelloso Piurcosky
Technical Editor:
Eufrásia de Souza Melo

Funding:
This study received no external funding.

How to cite this article:
De Oliveira, E.; Frogeri, R. F.; Oliveira, A. A. F. (2026). Cybersecurity Discourse in Online Communities: A Thematic Analysis of Reddit and Discord. **Mythos**, v. 18, 1, 442-467. https://doi.org/10.36674/mythos.v18i1.1048

***Corresponding Author:**
Erick De Oliveira
erick.oliveira1@alunos.unis.edu.br

Abstract

In the current cybersecurity landscape, online discussion groups serve as essential collaborative spaces for knowledge exchange among professionals, researchers, and enthusiasts. Cooperation among participants and the sharing of experiences are critical for identifying and understanding cyber threats, enabling the rapid and effective adoption of preventive and corrective measures. Accordingly, this study aims to analyze the main themes discussed in online groups within specialized virtual communities, such as Reddit and Discord. The central research question guiding the analysis was: What are the main themes discussed in online groups? The data analyzed consist of posts, interactions, and debates extracted from online communities dedicated to cybersecurity, covering the period from 2022 to 2025 and totaling 2,108,803 raw records. To organize and process the data, IRaMuTeQ (Interface of R for Multidimensional Analysis of Texts and Questionnaires) was employed, enabling lexical analysis and multivariate statistical analysis of the corpus. Subsequently, QualiGPT was used to further deepen and interpret the results associated with the identified thematic classes. The findings contribute to the ongoing debate on main cybersecurity themes by offering new insights into the challenges, practices, and impacts present in specialized virtual environments. The analyses indicate that these spaces function as environments for collaborative learning and professional development. Discussions ranged from digital defense practices to career pathways, highlighting the close integration of theory and practice within the cybersecurity field.

Keywords: Lexicographic analysis, Focus group, Iramuteq, Information security, Knowledge sharing, QualiGPT.

Resumo

No atual cenário da segurança cibernética, os grupos de discussão online servem como espaços colaborativos essenciais para a troca de conhecimentos entre profissionais, pesquisadores e entusiastas. A cooperação entre os participantes e o compartilhamento de experiências são fundamentais para identificar e compreender as ameaças cibernéticas, permitindo a adoção rápida e eficaz de medidas preventivas e corretivas. Assim, este estudo teve como objetivo analisar os principais temas discutidos em grupos online dentro de comunidades virtuais especializadas, como o Reddit e o Discord. A questão central de pesquisa que orientou a análise foi: Quais são os principais temas discutidos nos grupos online? Os dados analisados consistem em postagens, interações e debates extraídos de comunidades online dedicadas à segurança cibernética, cobrindo o período de 2022 a 2025 e totalizando 2.108.803 registros. Para organizar e processar os dados, foi utilizado o software IRaMuTeQ (Interface R para Análise Multidimensional de Textos e Questionários), permitindo a análise lexical e a análise estatística multivariada do corpus. Posteriormente, a ferramenta QualiGPT foi utilizado para aprofundar e interpretar os resultados associados às classes temáticas identificadas. As descobertas contribuem para o debate em curso sobre os temas centrais em segurança cibernética, oferecendo novos insights sobre os desafios, práticas e influências presentes em ambientes virtuais especializados. As análises indicam que esses espaços funcionam como ambientes de aprendizagem colaborativa e desenvolvimento profissional. As discussões variaram de práticas de defesa digital a planos de carreira, destacando a estreita integração entre teoria e prática no campo da segurança cibernética.

Palavras-chave: Análise lexicográfica, Grupo de discussão, Iramuteq, Segurança da informação, Troca de conhecimento, QualiGPT.

Data available

The data and analytical resources supporting the findings of this study are publicly available in the Zenodo repository:

De Oliveira, E., Frogeri, R. F., & Oliveira, A. A. F. de (2025). Cybersecurity Discourse in Online Communities: A Thematic Analysis of Reddit and Discord (Version 1). Zenodo. https://doi.org/10.5281/zenodo.17980683

The repository includes the preprocessing scripts, analytical procedures, and aggregated datasets necessary to reproduce the study. In accordance with ethical considerations and platform terms of service, raw user-generated content is not fully redistributed; however, the provided materials ensure transparency and reproducibility of the research workflow.

Declaration on the Use of Artificial Intelligence to Prepare the Manuscript

The authors declare that artificial intelligence (AI) tools were used during the preparation of the manuscript submitted to Revista Mythos. The platform used was ChatGPT (GPT-4.5, free version), developed by OpenAI. The tool was employed between December 15 and December 17, 2025, specifically for tasks such as text editing, language refinement, translation between Brazilian Portuguese and American English, and supporting qualitative analysis through the QualiGPT approach (check Appendix 1 for details). The prompt used by the authors was as follows: “Translate the text from Brazilian Portuguese into American English. Then perform proofreading and a readability check of the translated text.”

All authors take full responsibility for the accuracy, integrity, and originality of the content generated or enhanced with the assistance of AI.

1 INTRODUCTION

In the current digital era, cybersecurity has become an essential issue for society and organizations due to the increasing sophistication and frequency of threats in the virtual environment (Furfaro et al., 2017; Ramaseri-Chandra & Pothana, 2024). The contemporary technological landscape is characterized by the widespread adoption of fiber-optic networks, which drive global interconnectivity and significant advances in data transmission speeds (Madani & McGregor, 2024). This movement is further strengthened by the emergence of fourth- and fifth-generation mobile networks (4G and 5G), which not only expand the capacity to integrate billions of devices and support society's growing connectivity demands (Salahdine et al., 2023) but also increase exposure to digital threats and the complexity of cybersecurity.

Cyber incidents can lead to financial losses, legal liabilities, privacy breaches, and reputational damage, while also affecting national security and the continuity of strategic operations (Georgakopoulos & Jayaraman, 2016). In the current cybersecurity landscape, online discussion groups represent essential collaborative spaces for knowledge exchange among professionals, researchers, and enthusiasts in the field (Wu et al., 2021). Cooperation among participants and the sharing of experiences are critical for identifying and understanding cyber threats, enabling the rapid and effective adoption of preventive and corrective measures. In this context, the establishment of mechanisms for data exchange plays a strategic role in strengthening security capabilities and promoting a more resilient and trustworthy cyber ecosystem (Kanca & Sağiroğlu, 2021).

Accordingly, this study aims to analyze the main themes discussed in online groups within specialized virtual communities, such as Reddit and Discord. The central research question guiding this analysis was: What are the main themes discussed in online groups? To achieve the proposed objective, a large volume of textual data was analyzed through the application of different Natural Language Processing (NLP) techniques.

It is expected that the results of this study will provide insights into emerging concerns and discussions related to cybersecurity. The findings may help professionals develop proactive initiatives aligned with current themes discussed into online cybersecurity communities. In the academic context, these results may reveal emerging topics and provide a foundation for further in-depth research.

2 THEORETICAL BACKGROUND

2.1 DINÂMICA EVOLUTIVA DAS AMEAÇAS DIGITAIS

The modern internet, characterized by its speed and efficiency, originated from a project known as ARPANET (Advanced Research Projects Agency Network), developed in 1969 by the Advanced Research Projects Agency of the United States Department of Defense to interconnect research centers and universities (Walden, 2019). In the networks that preceded the internet during the 1970s, security was not a central concern, as these structures operated in restricted, experimental environments under institutional control (Safaei Pour et al., 2023).

ARPANET initially operated using the Network Control Protocol (NCP). However, as the network expanded, the need arose for a more flexible and scalable protocol. As a result, in 1983, TCP/IP (Transmission Control Protocol/Internet Protocol) was introduced, proposed by Robert Kahn and Vinton Cerf (IEEE, 2025). Once incorporated into ARPANET, TCP/IP enabled the interconnection of multiple independent networks.

The continuous growth of the internet led to the emergence of new risks and vulnerabilities. A milestone in this context was the Morris Worm in 1988, widely recognized as the first malware of major impact, infecting approximately 10% of the systems connected to the network (FBI, 2018).

Early security initiatives included the implementation of basic firewalls, the creation of incident response teams (CERTs), the use of local passwords, and cryptography, the latter being largely restricted to military or academic environments (Daya, 2013). At the same time, the first commercial antivirus solutions emerged

(Islam & Xiangdong, 2025). Nevertheless, security awareness among ordinary users remained limited, highlighting the need for greater digital education and more robust protection practices.

In the early 1990s, Tim Berners-Lee, then a researcher at the European Organization for Nuclear Research (CERN), developed the World Wide Web (WWW). By integrating hypertext concepts with TCP/IP and DNS (Domain Name System) protocols, he established a standardized way to access and share information over the internet (CERN, 2019).

Companies soon began offering online services, giving rise to electronic commerce, with pioneers such as Amazon and eBay marking a new era of interactivity and digital business (Ntumba et al., 2023).

The spread of the internet among non-specialist users led to the emergence of early macro viruses, such as Concept (1995) and XM/Laroux (1996). These demonstrated how Word documents and Excel spreadsheets could be used to propagate malware through macros, exposing their vulnerabilities (TrendMicro, 2015). During the same period, the first social engineering attacks appeared, driven by the development of software such as AOHell, which automated password-stealing attacks and gave rise to the term *phishing* (Rekouche, 2011).

To protect online transactions, the internet adopted the HTTPS (HyperText Transfer Protocol Secure) protocol, based on SSL (Secure Sockets Layer) (Tanenbaum & Feamster, 2021), while firewalls evolved and became increasingly sophisticated. With the emergence of Web 2.0 and the "data explosion" of the first decade of the 2000s, the internet assumed a highly interactive character. This transformation reshaped how individuals and organizations relate to one another, fostering social collaboration and collective intelligence, and offering new opportunities to engage users more effectively (Murugesan, 2007). Social networks, blogs, wikis, and cloud services emerged as central platforms for communication, collaboration, and learning. Systems such as Facebook, YouTube, and Twitter enabled users to share information, create collective content, and actively participate in digital communities (Hsu et al., 2014).

At that time, although security measures focused on protecting networks, transactions, and malware, they did not keep pace with the rapid growth in users and network traffic. In the first half of 2004 alone, more than 1,200 software vulnerabilities were identified, nearly 5,000 viruses and worms were detected, and botnet incidents increased from 2,000 to more than 30,000 cases. This significantly amplified the impact of threats such as spyware, spam, and phishing (Wired, 2004).

The increasing volume of data also facilitated Distributed Denial of Service (DDoS) attacks, such as the one against Yahoo!, which rendered its system inoperable for two hours, causing advertising losses and underscoring the need for more sophisticated security strategies and robust information protection policies (Specht & Lee, 2003)

With the spread of mobile internet and the consolidation of cloud computing, connectivity became fully portable, enabled by smartphones and high-speed networks such as 4G mobile technology (Shuai Zhang et al., 2010). At the same time, the expansion of cloud computing, materialized by platforms such as AWS, Microsoft Azure, and Google Cloud,- allowed users and organizations to store and process data remotely, redefining the digital infrastructure paradigm (Kaushik et al., 2021).

However, the exponential growth of mobile devices increased the incidence of attacks targeting applications, including the spread of malware specifically designed for Android and iOS platforms (Yang et al., 2018), which exploit vulnerabilities and carry out malicious actions (Seo et al., 2014)

Massive data breaches affected social networks and large corporations such as Yahoo! and LinkedIn (Wired, 2016), leading to the emergence of data protection regulations such as the GDPR (General Data Protection Regulation) (Lambrinoudakis, 2018) in Europe and the LGPD (General Data Protection Law) in Brazil (Piurcosky et al., 2019).

Driven by the rise of the Internet of Things (IoT), connectivity expanded beyond computers and mobile devices to include a wide range of everyday and industrial devices, such as security cameras, vehicles, and industrial sensors (Georgakopoulos & Jayaraman, 2016). Fifth-generation mobile technology (5G), by

significantly increasing both transmission speed and connection density, enabled the development of a highly complex and interconnected digital (Ma et al., 2019).

In cybersecurity, new challenges emerged with attacks targeting IoT devices, as exemplified by the Mirai Botnet (Margolis et al., 2017), which exposed vulnerabilities resulting from improper configurations. Digital security became a strategic priority, as governments and organizations began to recognize cyberattacks as national threats. Consequently, protection ceased to be an exclusively corporate concern and became a matter of national sovereignty (Möllers, 2021).

Artificial Intelligence (AI) has been applied both defensively, through anomaly detection, authentication, and access control, and offensively, enabling automated phishing and spoofing, assisted malware generation, and AI-driven DDoS attacks (Gao, 2024).

According to recent studies (e.g., Tzavara & Vassiliadis, 2024), the expansion of the internet and the evolution of cybersecurity have occurred in an interdependent manner. As connectivity, network complexity, and online interactivity increased, they also drove the development of increasingly sophisticated defense mechanisms based on prevention, governance, and emerging technologies (Tzavara & Vassiliadis, 2024).

3 METHODOLOGY

3.1 Research Design

To achieve the objective of this study, a descriptive, qualitative, and observational research design was adopted, focusing on the indirect analysis of textual content spontaneously produced in online environments. This approach is appropriate for examining naturally occurring discourse in digital communities, as it allows for the investigation of patterns, meanings, and thematic structures without interfering in the interactions among participants. Accordingly, no form of mediation, suggestion, or interaction was carried out by the researchers, ensuring that the data reflect authentic community dynamics.

The methodological framework combines lexicographic statistical analysis with inductive thematic interpretation. Specifically, the study integrates computational textual analysis using IRaMuTeQ with AI-assisted qualitative coding through QualiGPT (Zhang et al., 2023), enabling both structural segmentation of the corpus and interpretive exploration of emergent themes.

3.2 Data Sources and Community Selection

Data collection focused on the Reddit and Discord platforms, selected due to their prominence as digitally engaged environments hosting active cybersecurity communities. These platforms function as spaces where practitioners, learners, and enthusiasts exchange technical knowledge, discuss emerging threats, and share professional experiences.

The temporal scope of the study covers the period from 2022 to 2025, allowing for the capture of recent developments in cybersecurity discourse. The initial selection of communities was informed by recommendations from reference sources in the cybersecurity domain, including medium.com, which highlights the TryHackMe Community on Discord, and guardz.com, which identifies the r/netsec subreddit as a relevant discussion space. These communities were adopted as the core of the dataset, and additional publicly accessible groups were subsequently incorporated to expand the corpus and ensure broader representativeness of cybersecurity-related discussions.

Only open and publicly accessible communities were included in the analysis. Closed or private groups were excluded due to access restrictions and ethical considerations, particularly the need to ensure transparency and compliance with data collection standards. A complete list of analyzed communities is provided in the research repository.

3.3 Data Collection Procedures

Data collection was conducted using platform-specific tools that allow systematic extraction of textual content. For Reddit, scripts were developed in Python using the PRAW (Python Reddit API Wrapper, version 7.8.1) library, executed in the Visual Studio Code environment (version 1.103.0). These scripts enabled the retrieval of posts and comments from r/netsec and other selected subreddits, with the data organized in comma-separated values (.csv) format.

For Discord, data were collected using the DiscordChatExporter tool (version 2.46.0), which allowed the extraction of complete message histories from the TryHackMe Community server and other selected public servers. The exported data were stored as plain text (.txt) files.

The resulting dataset was entirely in English, and all analyses were conducted in this language to preserve technical terminology and semantic precision. In total, the raw dataset comprised 2,108,803 textual records, forming a substantial empirical basis for subsequent analysis.

3.4 Data Preprocessing and Corpus Construction

The collected data were subjected to a preprocessing and organization pipeline implemented in Python. This process involved the removal of blank lines, URLs, emojis, and other non-relevant textual elements, as well as the normalization of text to ensure consistency in analysis.

To ensure domain specificity, a filtering procedure was applied based on cybersecurity-related terminology derived from widely recognized frameworks, including MITRE ATT&CK, OWASP, NIST, and SANS (NIST, 2013). These sources provide standardized vocabularies encompassing key concepts, techniques, vulnerabilities, tools, and practices used by cybersecurity professionals. The filtering process retained only textual segments containing at least one relevant term, covering categories such as authentication and access control, malware, social engineering, vulnerabilities, defensive and offensive techniques, network and cloud security, tools and platforms, team practices (Blue/Red/Purple), certifications, and contemporary developments in the field.

This procedure enabled the construction of a domain-focused corpus that captures technically relevant discourse. However, it is acknowledged that dictionary-based filtering may introduce selection bias by privileging established terminology and potentially excluding emerging or non-standard expressions. Therefore, the results should be interpreted as reflecting structured patterns within a filtered cybersecurity corpus.

After preprocessing and filtering, the final corpus consisted of a 4.54 MB textual file containing 831,838 words and 22,404 distinct lexical forms. The formatting of the corpus followed the guidelines provided in the IRaMuTeQ documentation, ensuring compatibility with the software's analytical requirements (IRaMuTeQ, n.d.).

3.5 Lexicographic Analysis Using IRaMuTeQ

The processed corpus was analyzed using IRaMuTeQ (version 0.8 alpha 7), an open-source software developed in the R programming language and widely used in academic research for statistical analysis of textual data (Sousa, 2021). The analytical procedure included three complementary techniques: Word Cloud generation, Similarity Analysis, and Reinert's Descending Hierarchical Classification (DHC) (Reinert, 1990).

Word Cloud analysis provided an initial overview of the most frequent terms in the corpus, enabling identification of dominant lexical elements. Similarity Analysis examined co-occurrence relationships among words, revealing semantic networks and central concepts structuring the discourse. The core analytical method, DHC, partitioned the corpus into distinct classes based on vocabulary distribution, using chi-square statistics to identify significant associations between words and textual segments (Reinert, 1990).

The combined application of these techniques is justified by their complementarity: while frequency analysis offers a descriptive overview, similarity analysis captures relational structures, and DHC enables systematic

categorization of the corpus into coherent thematic classes. This integrated approach has been widely adopted in studies employing IRaMuTeQ for large-scale textual analysis (Martins et al., 2022).

3.6 Qualitative Thematic Analysis Using QualiGPT

To deepen the interpretation of the themes identified through lexicographic analysis, an inductive thematic analysis was conducted using the QualiGPT application (Zhang et al., 2023). QualiGPT is a tool based on GPT language models designed to facilitate qualitative data analysis by automating coding processes and generating thematic insights from textual data. QualiGPT is available at https://chatgpt.com/g/g-HtBvI9uXe-qualigpt.

The qualitative analysis was performed on textual subsets corresponding to the four classes generated by IRaMuTeQ. These subsets, totaling approximately 3.8 MB and 32,336 posts, were extracted from the corpus and analyzed separately to preserve class-specific contextual integrity. The analytical procedure followed an inductive approach, in which themes were derived directly from the data without predefined categories.

The process was guided by structured prompts that defined the analytical objective (understanding user perceptions), the nature of the data (spontaneous online discourse), and the expected output (thematic summaries, codes, clusters, and interpretive descriptions). Based on these instructions, QualiGPT generated thematic representations for each class, which were subsequently reviewed and consolidated by the researchers.

Only macro-level themes were retained in the final analytical framework, as these represent the most recurrent and structurally significant patterns in the corpus. This hybrid approach combines the statistical rigor of lexicographic analysis with the interpretive flexibility of AI-assisted qualitative coding.

To ensure transparency and reproducibility, all prompts, model configurations, and analytical parameters are documented in Appendix A. Furthermore, the outputs generated by QualiGPT were systematically reviewed by the authors, and thematic categories were consolidated based on recurrence, coherence, and alignment with the underlying lexicographic structure. We believe that this hybrid approach leverages the interpretive capabilities of large language models while maintaining methodological grounding in data-driven segmentation.

3.7 Ethical Considerations

The study adheres to ethical standards for research involving publicly available online data. All analyzed content was collected from open-access communities, and no interaction with users was performed. The research design ensures that the data reflect naturally occurring discourse without researcher intervention.

User privacy was preserved through the exclusion of personally identifiable information, and no attempts were made to identify or profile individual participants. Closed or restricted communities were excluded to ensure compliance with ethical principles and platform policies. Furthermore, any data sharing associated with this study will be limited to aggregated outputs and analytical artifacts, ensuring alignment with platform terms of service and privacy considerations.

4 RESULTS AND DISCUSSIONS

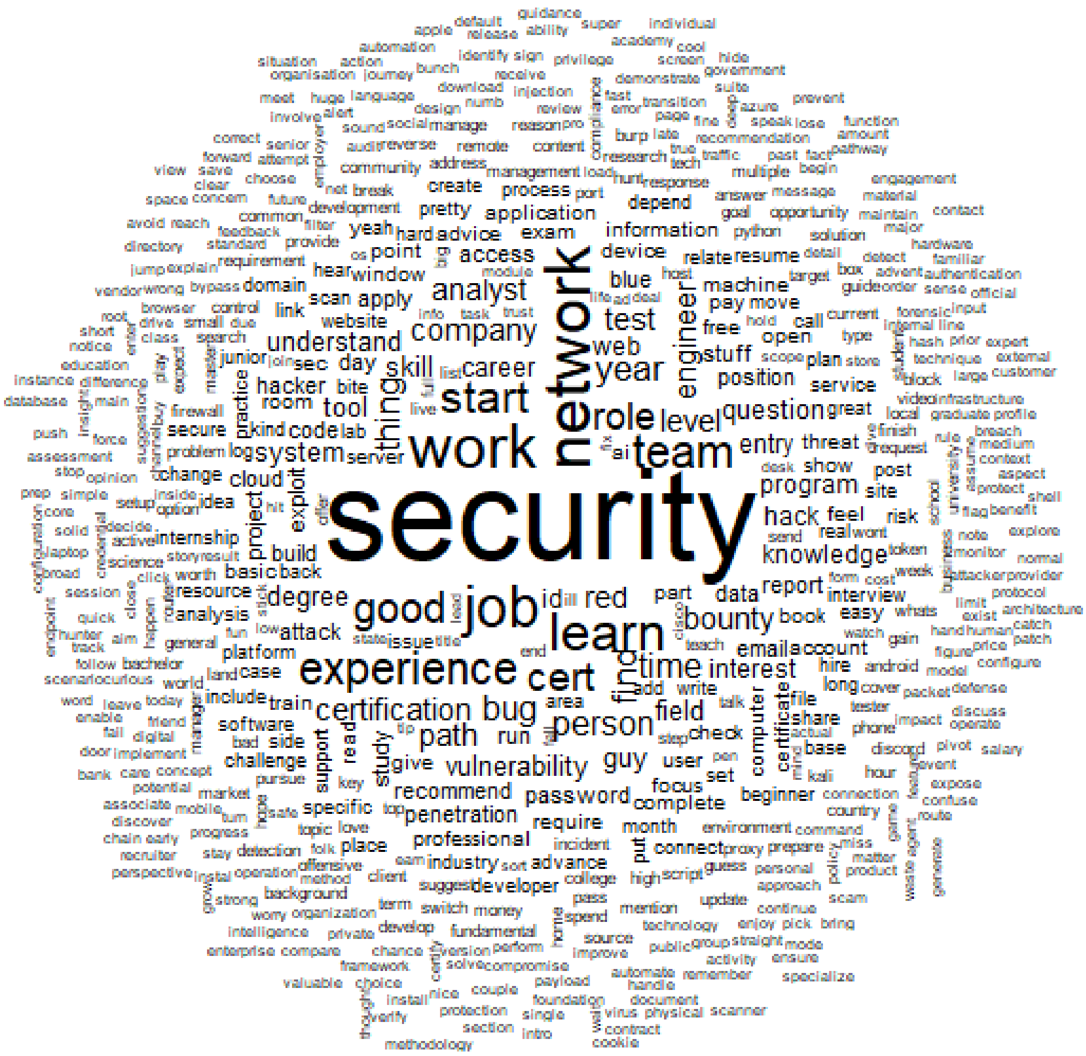
To understand the structure and predominant themes of the corpus, three complementary techniques from the IRaMuTeQ software were employed: Word Cloud, Similarity Analysis, and Reinert's Descending Hierarchical Classification (DHC) (Reinert, 1990). In addition, Inductive Thematic Analysis was conducted using the QualiGPT (H. Zhang et al., 2023) application to deepen the interpretation of the results obtained from the DHC. This combined methodological approach enabled a more detailed, systematic, and organized exploration of the textual content, as presented in the following sections.

4.1 Word Cloud Analysis

The visualization of the most common terms in the corpus was performed through a word cloud, illustrated in Figure 1. This technique allows for the graphical identification of the most frequent terms, highlighting the concepts that predominate in the analyzed texts.

Figure 1

Word Cloud



Source: Prepared by the authors using IRaMuTeQ software (2025).

The word cloud analysis reveals the predominance of terms such as *security*, *work*, *network*, *job*, and *learn*, indicating that interactions within digital communities are strongly oriented toward technical and

professional aspects of cybersecurity. The term *security*, with 7,381 occurrences, highlights the centrality of discussions focused on system protection, risk identification, and vulnerability mitigation. The term *network*, with 3,238 occurrences, underscores the relevance of building connections and collaboration among community members. The terms *work* (3,098) and *job* (2,835) reflect participants' interest in labor market dynamics and professional opportunities in technology. Finally, *learn*, with 2,549 occurrences, reinforces the educational character of these spaces, which are oriented toward technical knowledge acquisition, experience sharing, and continuous skill development.

The presence of terms such as *team*, *experience*, *cert*, and *start* highlights different dimensions of interaction within cybersecurity digital communities. The term *team* emphasizes the importance of collaborative work and knowledge exchange among professionals and learners. *Experience* indicates that users value practical accounts and the sharing of professional experiences that support the development of technical competencies. The term *cert* refers to professional certifications, demonstrating that these credentials are perceived as important milestones for qualification and career recognition. Finally, *start* reflects the interest of many participants in beginning their careers in cybersecurity.

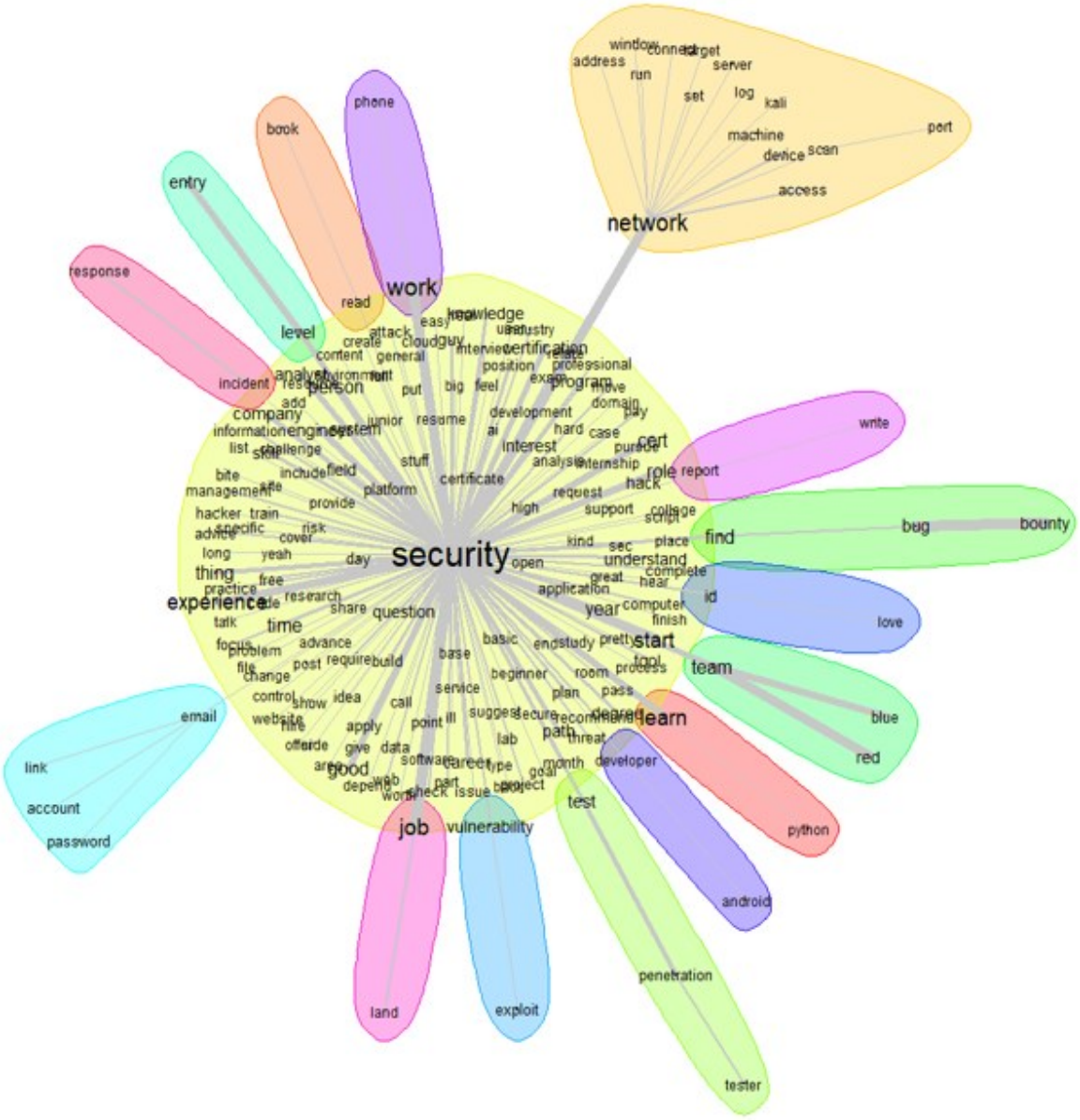
Overall, the recurrent presence of these terms suggests that the analyzed forums and communities function as hybrid environments in which learning, professional practice, and networking are integrated. These spaces play a relevant role in knowledge sharing, vulnerability discussions, and the dissemination of best practices in the cybersecurity field (Wu et al., 2021).

4.2 Similarity Analysis

The similarity analysis, conducted based on the textual corpus and presented in Figure 2, highlights the centrality of the term *security*, which emerges as the semantic core of the network, connecting multiple lexical fields. This central position indicates that the concept of security constitutes the axis around which the main discussion topics are organized. It also reflects the multidisciplinary nature of cybersecurity, which requires the integration of specialized technical knowledge, collaborative educational practices, and strong connections to the labor market (Furnell, 2021)

Figure 2

Similarity Analysis



Source: Prepared by the author using IRaMuTeQ software (2025).

Among the peripheral clusters, the lexical field related to *network* stands out, associated with terms such as *machine*, *port*, *device*, *scan*, and *access*. These terms refer to technical aspects of network infrastructure and system protection, demonstrating an emphasis on practices related to administration and defense of computing environments.

The clusters formed around *work* and *job* suggest discourses focused on professional activity and the practical application of acquired knowledge. In contrast, the cluster combining *learn* and *start* reveals participants' interest in initial learning processes and technical training. The cluster formed by *vulnerability* and *exploit* reflects the practical dimension of cybersecurity within the studied communities, where knowledge exchange extends beyond theoretical concepts to include applied skills such as ethical identification and exploitation of vulnerabilities. These practices strengthen technical learning and preparation for the job market. Additionally, the terms *team* and *find* appear linked to *bug*, *bounty*, and *report*, referring to teamwork and

the culture of vulnerability reward programs (bug bounty) (Walshe & Simpson, 2020), highlighting the cooperative and applied nature of cybersecurity activities.

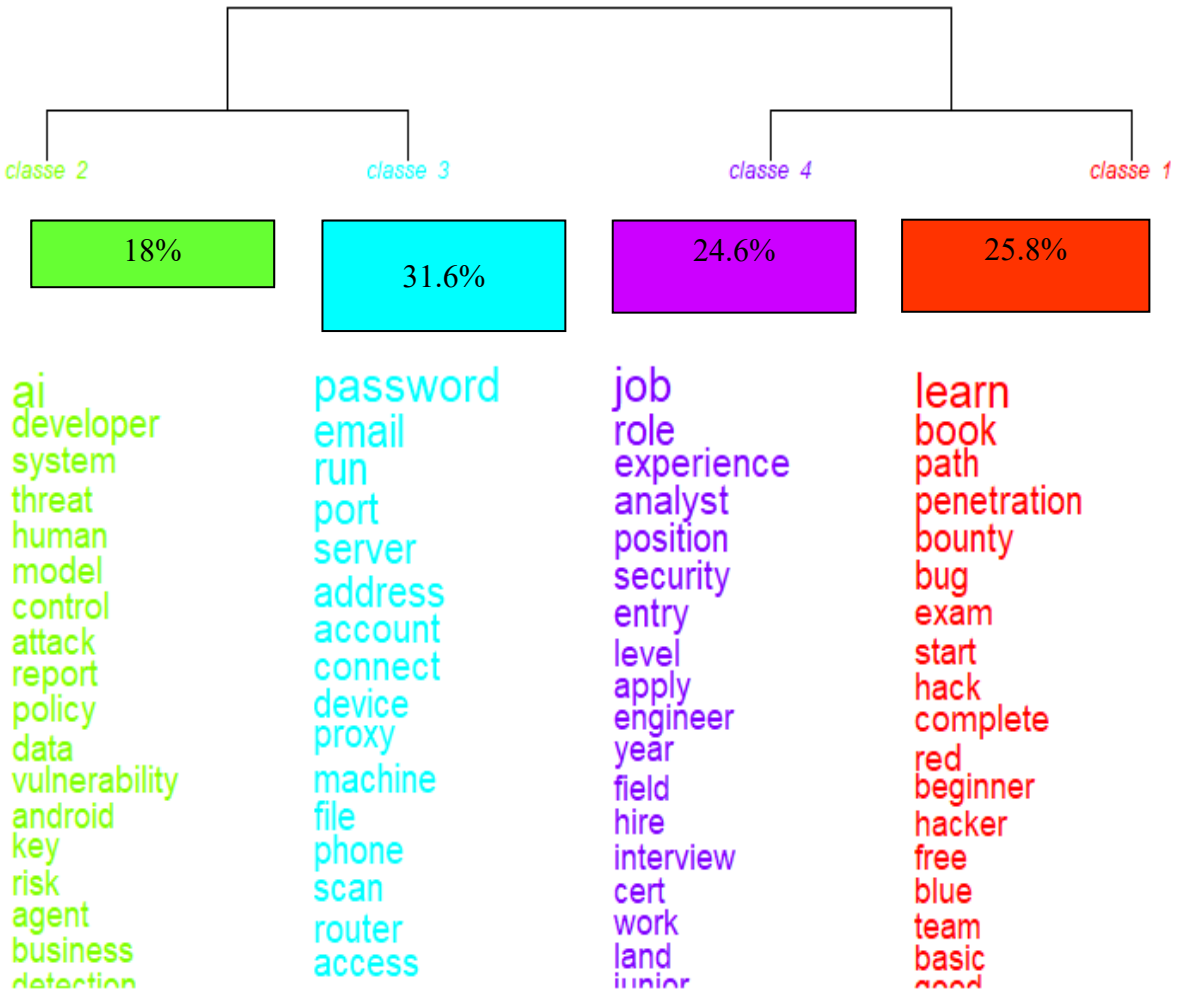
In summary, the analyzed discussions are structured around technical, educational, and professional dimensions, demonstrating that cybersecurity communities operate as hybrid spaces for learning, practice, and professional development. The most recurrent terms, such as *network*, *vulnerability*, *exploit*, *learn*, *start*, *experience*, *work*, *job*, and *company*, reflect this organization. This association reinforces the role of digital communities as spaces for training and knowledge exchange between beginners and experienced professionals (Ho & Gross, 2021).

4.3 Reinert's Descending Hierarchical Classification (DHC) and QualiGPT Analysis

The DHC identified four distinct clusters of messages within the corpus, defined by lexical proximity among terms and representing the main thematic axes addressed by participants in cybersecurity digital communities. Figure 3 presents the dendrogram resulting from this segmentation, highlighting the predominant concepts and discussions in the analyzed groups.

Figure 3

Dendrogram of Reinert's Descending Hierarchical Classification (DHC)



Source: Prepared by the author using IRaMuTeQ software (2025).

The four classes identified from the corpus are presented below, considering the dendrogram structure shown in Figure 3, their respective analyses, and the interpretive deepening provided by the QualiGPT tool.

4.3.1 Class 1 – Training and Learning Practice

This class, representing 25.8% of the analyzed corpus, includes terms such as *learn*, *book*, *path*, *penetration*, *bounty*, *exam*, *start*, *beginner*, and *basic*. The lexical set indicates an association with the educational and practical dimension of cybersecurity learning, with emphasis on initial technical training and hands-on experimentation.

The term *learn*, as the central core of the class, highlights participants' focus on knowledge acquisition, a characteristic typical of collaborative learning communities (Fisk et al., 2023). Terms such as *book*, *path*, and *exam* suggest the presence of educational resources, learning paths, and certification assessments. Meanwhile, terms such as *penetration*, *bug*, *hack*, and *bounty* demonstrate the practical orientation of learning, connected to simulation activities and ethical system intrusion practices. Table 1 summarizes the main DHC findings for Class 1.

Table 1

Summary of Research Findings – Class 1 (IRaMuTeQ)

Class	Central Theme	Emerging Approaches	Relevance to the Field
Class 1 – Training and Learning Practice (25.8%)	Initial technical and hands-on learning in cybersecurity	Use of simulations (ethical penetration testing, bug bounty), learning through books and certifications, learning communities	Foundational for the development of qualified professionals, especially at entry and early-career levels

Source: Developed by the authors using IRaMuTeQ.

The QualiGPT analysis (Table 2) further deepens the interpretation of Class 1 by identifying dominant themes related to learning resources, certifications, bug bounty practices, hands-on tools, pedagogical dilemmas between theory and practice, community support, and emotional aspects such as motivation and frustration.

Table 2

Summary of Themes Identified in Class 1 – QualiGPT

Theme	Frequency	Brief Description	Associated Excerpts
Learning resources and book recommendations	High	Discussions about study materials, publishers, bundles, and specific recommendations for subfields such as networking, malware, and web hacking	"practical malware analysis… malware analyst's cookbook"
Careers and certifications	High	Questions and guidance on career paths and certifications (Offensive Security Certified Professional, Certified Ethical Hacker, Cisco Certified Network Associate, eLearnSecurity Junior Penetration Tester), including comparisons between academic degrees and professional certifications	"the OSCP is… widely recognized"
Bug bounty and pentesting practices, expectations, and legality	High	Debates on income expectations, legal risks, proper scope definition, and practical preparation for pentesting and bug bounty activities	"is bug bounty legal… when does bug bounty become illegal"
Practical tools and hands-on learning	High	Emphasis on training platforms and tools commonly used in day-to-day technical learning, such as TryHackMe, Burp Suite, Network Mapper, and Kali Linux	"PortSwigger Academy… essential training"

ISSN: 1994-0098 | e-ISSN: 3086-0210

Theme	Frequency	Brief Description	Associated Excerpts
Pedagogical dilemmas: books vs. practice—what to prioritize	Medium	Reflections on balancing extensive reading with continuous hands-on practice in labs, highlighting the complementarity between the two	"you can read hundreds of books and not have any practical skills"
Community, mentoring, and networks	Medium	Recognition of the importance of groups, forums, Discord servers, and mentors for technical and emotional support	"community mentors… help out"
Emotional and motivational aspects	Medium/Low	Feelings of frustration, insecurity, and encouragement, commonly experienced during the early stages of the learning journey	"don't get discouraged", "I felt lost"

Source: Developed by the authors using QualiGPT.

Overall, Class 1 reveals themes focused on cybersecurity education, with strong interest in technical books, certifications, and career paths. It highlights concerns about pentesting and bug bounty practices, particularly regarding legality and adequate preparation, while emphasizing the value of hands-on learning through specialized platforms and tools.

4.3.2 Class 2 – Artificial Intelligence and Protection

Class 2, corresponding to 18% of the corpus, highlights discussions centered on the relationship between artificial intelligence and system protection. The most frequent terms (e.g., *ai*, *developer*, *system*, *threat*, *human*, *model*, *control*, *attack*, and *vulnerability*) suggest technical concerns related to AI models and system and data protection.

Terms such as *ai*, *developer*, *system*, and *threat* point to discussions on the development of artificial intelligence models and the application of machine learning in cybersecurity contexts (Zhang et al., 2022). Other relevant terms, including *attack*, *vulnerability*, *risk*, and *detection*, reflect concerns with cyber threats and risk mitigation in both offensive and defensive security contexts (see Table 3).

Table 3

Summary of Research Findings – Class 2 (IRaMuTeQ)

Class	Central Theme	Emerging Approaches	Relevance to the Field
Class 2 – Artificial Intelligence and Protection (18.0%)	Integration of AI and cybersecurity	Application of machine learning in offensive and defensive security; development of secure and controllable models	Response to emerging threats driven by technological evolution; innovation in cyber defense

Source: Developed by the authors using IRaMuTeQ.

QualiGPT analysis reveals a strong operational focus, with discussions on malware (including LLM-enabled threats), incident detection and response, automation, vulnerability management, Red/Blue/Purple Team practices, secure development, compliance, and governance (see Table 4).

Table 4

Summary of Themes Identified in Class 2 – QualiGPT

Theme	Frequency	Brief Description	Representative Example
Threats and malware (including LLM-enabled threats)	High	Technical discussions on new malware families, the use of Large Language Models to generate or enhance threats, evasion techniques, and Command-and-Control (C2) infrastructures	"rapperbot quick hits uses DNS TXT records to hide rotating C2s"
Detection, SOC, and Incident Response (IR)	High	Debates on Security Operations Center processes, containment versus investigation strategies, and critical telemetry for modern detection, including cloud control-plane abuse	"API telemetry for cloud control-plane abuse"
Tools, automation, and vulnerability management	High	Development of pipelines, automation of vulnerability analysis, reduction of false positives, and integration of scanners and application programming interfaces (APIs)	"vulnparse… automate vulnerability management workflows"
Red/Blue/Purple Team methods and Capture The Flag	High	Write-ups and offensive techniques (e.g., chisel, netexec), collaborative team exercises, and discussions on organizing Capture The Flag events	"purple team episode… netexec usage"
Secure development and issues in app platforms	Medium	Topics related to app publishing, policy-based rejections, build integrity, and privacy issues in Google Play and Android app stores	"Google Play ban due to missing privacy policy"
Compliance, governance, and risk	Medium	References to frameworks (e.g., NIST, FedRAMP), audits, compliance mapping, and the relationship between security practices and regulatory requirements	"FedRAMP… audit mapping notes"
Learning resources, careers, and certifications	Medium	Course announcements, mentions of certifications, and targeted guidance on professional pathways in security	"TPM class announcement… OSCP references"
Development and infrastructure topics	Medium/Low	Issues related to Android development, Jetpack Compose, developer tools, remote VS Code usage, and risks associated with attack surfaces in development environments	"VS Code remote dev… attack surface concerns"

Source: Developed by the authors using QualiGPT.

This Class reflects a discussion toward technical deepening and operational maturity in cybersecurity practices.

4.3.3 Class 3 – Infrastructure and Credentials

Class 3, which represents 31.6% of the textual corpus, constitutes the most prominent thematic group. It focuses on technical infrastructure and credential security, as indicated by frequent terms such as *password*, *email*, and *account*, related to authentication and digital identity protection (see Table 5).

Table 5

Summary of Research Findings – Class 3 (IRaMuTeQ)

Class	Central Theme	Emerging Approaches	Relevance to the Field
Class 3 – Infrastructure and Credentials (31.6%)	Security of technical infrastructure and digital identity management	Continuous monitoring, vulnerability scanning, network best practices, and authentication mechanisms	Essential technical foundation for securing corporate and personal digital environments

Source: Developed by the authors using IRaMuTeQ.

The presence of terms like *run*, *file*, and *scan* suggests operational actions and system analysis associated with monitoring and threat detection. Additionally, terms such as *port*, *server*, *router*, *proxy*, *device*, and *machine* emphasize network connectivity and configuration aspects essential to secure environment management.

This class reflects a strongly technical dimension of cybersecurity discussions, centered on infrastructure protection and identity management (Yigit et al., 2024). QualiGPT analysis further reveals a strong focus on bug bounty activities, web vulnerabilities, reconnaissance automation, hands-on tools, certifications, and legal and ethical boundaries (see Table 6).

Table 6

Summary of Themes Identified in Class 3 – QualiGPT

Theme	Frequency	Brief Description	Representative Example
Bug bounty, web vulnerabilities, and reconnaissance	Very High	Large volume of practical questions on Cross-Site Scripting (XSS), Cross-Site Request Forgery (CSRF), Server-Side Request Forgery (SSRF), Insecure Direct Object Reference (IDOR), recon automation (e.g., Nuclei, Shodan), proof-of-concept validation, and issues related to legality and scope	"is bug bounty legal?" / "how to confirm XSS/SSRF?"
Learning resources and hands-on platforms (TryHackMe, Hack The Box, Capture The Flag)	High	Recommendations of rooms, courses, books, study roadmaps, and support for beginners seeking simulated environments to build skills	"which TryHackMe room do you recommend to start?"
Tools and hands-on practice (Burp Suite, Network Mapper, Zed Attack Proxy, Nessus, Hashcat)	High	Operational discussions on flags, bypasses, scan optimization, selection of alternative tools, and combined use of scanners	"is RustScan better than Nmap for fast scanning?"
Careers and certifications (OSCP, CEH, PNPT)	Medium	Questions about certification pathways, exam experiences, comparisons between certifications, and strategies for building a professional portfolio	"should I do OSCP or PNPT first?"
Malware analysis and forensics	Medium	Mentions of analysis environments, dynamic and static techniques, sandboxing tools, and legal risks when handling real malware samples	"do I need an air-gapped VM to analyze malware?"

ISSN: 1994-0098 | e-ISSN: 3086-0210

Theme	Frequency	Brief Description	Representative Example
Security Operations Center (SOC), Incident Response, logging, and telemetry	Medium	Questions about logs, Security Information and Event Management (e.g., Splunk), Endpoint Detection and Response–based detection, essential telemetry, and chain of custody	"what logs matter for detecting initial access?"
Ethics, legal risk, and scope management	Medium	Concerns regarding authorization, Vulnerability Disclosure Programs, legal boundaries, and the risk of legal action when testing systems without explicit permission	"can I get sued if I test endpoints outside scope?"
Development, infrastructure, and operations issues (Docker, macOS, networking)	Moderate	Questions about lab setup, virtual machines, VPNs, containers, macOS systems, and tool integrations	"my Docker lab doesn't route traffic—what am I missing?"

Source: Developed by the authors using QualiGPT.

These themes indicate a strong emphasis on offensive security practices combined with concerns about legal scope and operational best practices.

4.3.4 Class 4 – Work and Career

Comprising 24.6% of the corpus, Class 4 reflects discussions related to the labor market and career development in cybersecurity. Frequent terms such as *job*, *role*, *experience*, *analyst*, *position*, *engineer*, *hire*, *interview*, *cert*, and *work* demonstrate participants' interest in professional opportunities, recruitment processes, and career progression (see Table 7).

Table 7

Summary of Research Findings – Class 4 (IRaMuTeQ)

Class	Central Theme	Emerging Approaches	Relevance to the Field
Class 4 – Work and Career (24.6%)	Labor market and professional development in cybersecurity	Interview preparation, certification attainment, and pursuit of specific roles (e.g., analyst, engineer)	Reflects the professionalization of the field and the growing demand for qualified and certified professionals

Source: Developed by the authors using IRaMuTeQ.

The presence of terms like *entry* and *junior* highlights the focus on early-career positions, while *cert* and *interview* indicate discussions around certifications and selection processes. This class emphasizes preparation for employment, practical learning, and career advancement, reinforcing the interdependence between theoretical knowledge and practical application in cybersecurity careers (Bendler & Felderer, 2023).

Table 8

Summary of Themes Identified in Class 4 – QualiGPT

Theme	Frequency	Concise Description	Representative Example
Careers and entry pathways (entry paths)	High	Questions and narratives about possible pathways into the field, including progressive steps (helpdesk → Security Operations Center → pentest) and barriers to junior-level positions.	"cybersecurity isn't entry level" / "helpdesk → sysadmin → cyber"
Certifications and credentials (value and strategy)	High	Discussions about certifications (Security+, Offensive Security Certified Professional, Certified Ethical Hacker, Cisco Certified Network Associate, SANS/Global Information Assurance Certification), their perceived value, regional differences, and impact on recruitment.	"oscp is seen as good by hr" / "Certified Ethical Hacker is valued in India"
Learning resources and hands-on platforms (TryHackMe / Hack The Box / Capture The Flag)	High	Frequent references to practical environments, labs, and training platforms as foundations for technical skill development and evidence of competence.	"tryhackme is a good start... nahamstore room is a good one for bug bounty"
Frustration, anxiety, and impostor syndrome	High	Reports of insecurity, interview anxiety, feelings of inadequacy, and emotional strain during the entry process into the field.	"I feel like a fake cybersecurity professional" / "I can't sleep excited and scared"
Differences between subfields (SOC, pentest, GRC, cloud)	Medium	Comparisons between cybersecurity subareas and questions about required skills and professional profiles, including pentesting, SOC, GRC, and cloud security.	"pentesting is not an entry level job" / "cloud security is in demand"
Interview preparation and portfolio building	Medium	Questions about résumé construction, inclusion of practical evidence (TryHackMe, write-ups), and strategies to stand out in interviews.	"how to put tryhackme on resume?" / "writeups and portfolio building"
Advice and mentoring (networking and contacts)	Medium/Low	Recommendations emphasizing the importance of networking, LinkedIn outreach, and community interactions to facilitate job opportunities.	"it's about who you know... message people on linkedin"

Source: Developed by the authors using QualiGPT.

QualiGPT analysis reveals recurring themes related to entry-level career paths, certification strategies, practical learning platforms, emotional challenges such as anxiety and impostor syndrome, differences between cybersecurity subfields, and the importance of portfolios and professional networking.

4.4 Final Synthesis

The in-depth analysis of the four classes demonstrates that cybersecurity education and practice involve technical, professional, and emotional demands. While some participants seek initial guidance and laboratory-based learning, others engage in advanced discussions on threats, detection, and vulnerabilities.

Questions about certifications, career paths, and interview preparation reveal a demanding and often unclear job market for beginners. Furthermore, recurring feelings of insecurity and frustration highlight that professional development in cybersecurity depends not only on technical skills, but also on community support and continuous guidance.

5 CONCLUSION

The present study aimed to analyze the main themes discussed in online groups within specialized virtual communities, such as Reddit and Discord. Based on the analyses conducted, it was possible to identify the main thematic axes structuring interactions within these environments, revealing that such communities function not only as spaces for technical information exchange but also as formative and professional ecosystems in which learning, collaboration, and career development are mutually reinforcing.

Overall, the results suggest that the analyzed communities are shaped by both educational needs and professional and operational demands. The thematic patterns identified (e.g., through DHC analysis) are consistent with patterns observed in the corpus, such as the rise of artificial intelligence in security, the talent shortage, and the emphasis on basic cyber hygiene (e.g., passwords and authentication). The QualiGPT analysis reinforces these findings by showing that participants recurrently express concerns related to technical training, certification choices, difficulties in entering the field, and the importance of hands-on platforms for skill development. The tool also highlighted emerging themes related to professional trajectories, including anxiety, impostor syndrome, uncertainty about career paths, and the perception of structural barriers for beginners. In addition, it underscored the central role of online communities as spaces for support, networking, learning, and social validation.

The findings indicate that online cybersecurity communities play a significant role in professional training and in the dissemination of technical knowledge. They operate as collaborative environments in which theory and practice are integrated, continuous learning and networking coexist, and professional development is actively encouraged. These spaces can thus be understood as informal learning networks that promote not only technical capacity building but also social and professional engagement, fostering a collaborative and dynamic culture within the field of information security.

From a theoretical perspective, this study advances the understanding of cybersecurity communities by conceptualizing them as hybrid communities of practice, in which knowledge production, professional socialization, and skill development occur simultaneously in decentralized and digitally mediated environments. While prior literature has emphasized formal education and organizational training as primary mechanisms for cybersecurity competence development, the present findings demonstrate that informal, community-driven interactions constitute a parallel and structurally significant layer of knowledge formation. In this sense, the study contributes to theory by extending the concept of communities of practice (e.g., Wenger, 1998) to highly technical, fast-evolving domains, showing how expertise is collectively constructed through continuous interaction, shared problem-solving, and experiential learning in online ecosystems (Fisk et al., 2023).

From a practical perspective, this study contributes to reflections on the future of cybersecurity by demonstrating that competence development does not rely solely on formal education, but also involves online interactions, collaboration, and continuous knowledge exchange. By mapping these dynamics, the study provides insights that may inform educational strategies, organizational policies, and professional practices aimed at strengthening digital security in a proactive and sustainable manner.

Despite its strengths, the methodological approach adopted in this study is subject to several limitations. The reliance on dictionary-based filtering may constrain the scope of analysis by excluding non-standard or emerging forms of cybersecurity discourse. The absence of bot detection, deduplication, and advanced spam filtering introduces potential biases related to user activity and content repetition. The dataset is also limited to English-language content, which may restrict the generalizability of findings to non-Anglophone contexts. Additionally, the use of a large language model for thematic analysis introduces variability related to prompt design and model behavior, which, although mitigated through documentation and human validation, cannot be entirely eliminated.

For future research, it is recommended to expand the corpus of analysis by incorporating multilingual communities and other relevant forums not yet explored in the cybersecurity domain. Additionally, the

combined use of quantitative and qualitative methods is suggested in order to deepen the understanding of how technical knowledge is produced, validated, and disseminated within digital communities.

REFERENCES

Bendler, D., & Felderer, M. (2023). Competency Models for Information Security and Cybersecurity Professionals: Analysis of Existing Work and a New Model. *ACM Trans. Comput. Educ.*, *23*(2), 25:1-25:33. https://doi.org/10.1145/3573205

CERN. (2019). *The birth of the Web*. https://home.cern/science/computing/birth-web

D. Walden. (2019). ARPANET 50th Anniversary Marked at 2019 AAAS Annual Meeting. *IEEE Annals of the History of Computing*, *41*(2), 61–63. https://doi.org/10.1109/MAHC.2019.2913715

Daya, B. (2013). Network security: History, importance, and future. *University of Florida Department of Electrical and Computer Engineering*, *4*. https://askcypert.org/sites/default/files/Network%20Security.pdf

De Wever, B., Schellens, T., Valcke, M., & Van Keer, H. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & Education*, *46*(1), 6–28. https://doi.org/10.1016/j.compedu.2005.04.005

FBI. (2018). *The Morris Worm: 30 Years Since First Major Attack on the Internet*. https://www.fbi.gov/news/stories/morris-worm-30-years-since-first-major-attack-on-internet-110218

Fisk, N., Kelly, N. M., & Liebrock, L. (2023). Cybersecurity communities of practice: Strategies for creating gateways to participation. *Computers & Security*, *132*, 103188. https://doi.org/10.1016/j.cose.2023.103188

Furfaro, A., Argento, L., Parise, A., & Piccolo, A. (2017). Using virtual environments for the assessment of cybersecurity issues in IoT scenarios. *Simulation Modelling Practice and Theory*, *73*, 43–54. https://doi.org/10.1016/j.simpat.2016.09.007

Furnell, S. (2021). The cybersecurity workforce and skills. *Computers & Security*, *100*, 102080. https://doi.org/10.1016/j.cose.2020.102080

Gao, Y. (2024). Cyber Attacks and Defense: AI-Driven Approaches and Techniques. *Academic Journal of Computing & Information Science*, *7*(7). https://doi.org/10.25236/AJCIS.2024.070706

Georgakopoulos, D., & Jayaraman, P. P. (2016). Internet of Things: From internet-scale sensing to smart services. *Computing*, *98*, 1041–1058. https://doi.org/10.1007/s00607-016-0510-0

Ho, S. M., & Gross, M. (2021). Consciousness of cyber defense: A collective activity system for developing organizational cyber awareness. *Computers & Security*, *108*, 102357. https://doi.org/10.1016/j.cose.2021.102357

Hsu, Y.-C., Ching, Y.-H., & Grabowski, B. L. (2014). Web 2.0 Applications and Practices for Learning Through Collaboration. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds), *Handbook of Research on Educational Communications and Technology* (pp. 747–758). Springer. https://doi.org/10.1007/978-1-4614-3185-5_60

IEEE. (2025). *TCP Design Published | IEEE Communications Society*. IEEE COMSOC. https://www.comsoc.org/node/19581

IRaMuTeQ. (n.d.). *Formatage des corpus texte*. http://www.iramuteq.org/documentation/formatage-des-corpus-texte

Islam, M. S., & Xiangdong, L. (2025). *Guardians of the Web: The Evolution and Future of Website Information Security* (No. arXiv:2505.04308). arXiv. https://doi.org/10.48550/arXiv.2505.04308

Kanca, A., & Sağiroğlu, Ş. (2021). *Sharing Cyber Threat Intelligence and Collaboration*. 167–172. https://doi.org/10.1109/ISCTURKEY53027.2021.9654328

Lambrinoudakis, C. (2018). The General Data Protection Regulation (GDPR) Era: Ten Steps for Compliance of Data Processors and Data Controllers. In *Trust, Privacy and Security in Digital Business* (Vol. 11033). Springer International Publishing. https://doi.org/10.1007/978-3-319-98385-1_1

Madani, P., & McGregor, C. (2024). *Cybersecurity Issues in Space Optical Communication Networks and Future of Secure Space Health Systems*. IEEE Aerospace Conference. https://doi.org/10.1109/AERO58975.2024.10521087

Margolis, J., Oh, T. T., Jadhav, S., Kim, Y. H., & Kim, J. N. (2017). An In-Depth Analysis of the Mirai Botnet. *2017 International Conference on Software Security and Assurance (ICSSA)*, 6–12. https://doi.org/10.1109/ICSSA.2017.12

Martins, C. P. dos S., Vasconcelos, A. M. de, & Aguirre, A. de B. (2022). Uso do software Iramuteq na análise da produção científica sobre as compras públicas. *Anais Do XLVI Encontro Da ANPAD*. XLVI Encontro da ANPAD - EnANPAD, Online. https://anpad.com.br/uploads/articles/120/approved/6a13382a520e0420014027350a0b3eb4.pdf

Möllers, N. (2021). Making Digital Territory: Cybersecurity, Techno-nationalism, and the Moral Boundaries of the State. *Science, Technology, & Human Values*, *46*(1), 112–138. https://doi.org/10.1177/0162243920904436

Murugesan, S. (2007). Understanding Web 2.0. *IT Professional*, *9*(4), 34–41. https://doi.org/10.1109/MITP.2007.78

NIST. (2013, November 12). *Cybersecurity Framework*. NIST. https://www.nist.gov/cyberframework

P. Kaushik, A. M. Rao, D. P. Singh, S. Vashisht, & S. Gupta. (2021). Cloud Computing and Comparison based on Service and Performance between Amazon AWS, Microsoft Azure, and Google Cloud. *2021 International Conference on Technological Advancements and Innovations (ICTAI)*, 268–273. https://doi.org/10.1109/ICTAI53825.2021.9673425

Piurcosky, F. P., Costa, M. A., Frogeri, R. F., & Calegario, C. L. L. (2019). A lei geral de proteção de dados pessoais em empresas brasileiras: Uma análise de múltiplos casos. *SUMA DE NEGÓCIOS*, *10*(23), 89–99. http://dx.doi.org/10.14349/sumneg/2019.V10.N23.A2

Ramaseri-Chandra, A. N., & Pothana, P. (2024). Cybersecurity threats in Virtual Reality Environments: A Literature Review. *2024 Cyber Awareness and Research Symposium (CARS)*, 1–7. https://doi.org/10.1109/CARS61786.2024.10778838

Reinert, M. (1990). ALCESTE UNE METHODOLOGIE D'ANALYSE DES DONNEES TEXTUELLES ET UNE APPLICATION: AURELIA DE GERARD DE NERVAL. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *26*(1), 24–54. https://doi.org/10.1177/075910639002600103

Rekouche, K. (2011). *Early Phishing* (No. arXiv:1106.4692). arXiv. https://doi.org/10.48550/arXiv.1106.4692

Safaei Pour, M., Nader, C., Friday, K., & Bou-Harb, E. (2023). A Comprehensive Survey of Recent Internet Measurement Techniques for Cyber Security. *Computers & Security*, *128*, 103123. https://doi.org/10.1016/j.cose.2023.103123

Salahdine, F., Han, T., & Zhang, N. (2023). 5G, 6G, and Beyond: Recent advances and future challenges. *Annals of Telecommunications*, *78*(9), 525–549. https://doi.org/10.1007/s12243-022-00938-3

Seo, S.-H., Gupta, A., Mohamed Sallam, A., Bertino, E., & Yim, K. (2014). Detecting mobile malware threats to homeland security through static analysis. *Journal of Network and Computer Applications*, *38*, 43–53. https://doi.org/10.1016/j.jnca.2013.05.008

Shuai Zhang, Xuebin Chen, Shufen Zhang, & Xiuzhen Huo. (2010). The comparison between cloud computing and grid computing. *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*, *11*, V11-72. https://doi.org/10.1109/ICCASM.2010.5623257

Sousa, Y. S. O. (2021). O Uso do Software Iramuteq: Fundamentos de Lexicometria para Pesquisas Qualitativas. *Estudos e Pesquisas em Psicologia*, *21*(4), 1541–1560. https://doi.org/10.12957/epp.2021.64034

ISSN: 1994-0098 | e-ISSN: 3086-0210

Specht, S., & Lee, R. (2003). *Taxonomies of Distributed Denial of Service Networks, Attacks, Tools, and Countermeasures* (p. 20). Department of Electrical Engineering Princeton Architecture Laboratory for Multimedia and Security. https://www.princeton.edu/~rblee/ELE572Papers/Fall04Readings/DDoSSurveyPaper_20030516_Final.pdf

Tanenbaum, A., & Feamster, N. (2021). *Redes de Computadores*. Bookman.

TrendMicro. (2015). *20 Years of Macro Malware: From Harmless Concept to Targeted Attacks*. https://www.trendmicro.com/vinfo/gb/security/news/cybercrime-and-digital-threats/20-years-of-macro-malware-from-harmless-concept-to-targeted-attacks

Tzavara, V., & Vassiliadis, S. (2024). Tracing the evolution of cyber resilience: A historical and conceptual review. *International Journal of Information Security*, *23*(3), 1695–1719. https://doi.org/10.1007/s10207-023-00811-x

Walshe, T., & Simpson, A. (2020). An Empirical Study of Bug Bounty Programs. *2020 IEEE 2nd International Workshop on Intelligent Bug Fixing (IBF)*, 35–44. https://doi.org/10.1109/IBF50092.2020.9034828

Wenger, E. (1998). Communities of Practice: Learning, Meaning, and Identity. Cambridge University Press. https://doi.org/10.1017/CBO9780511803932

Wired. (2004). Symptoms of Our Time, Part One. *Wired*. https://www.wired.com/2004/09/symptoms-of-our-3/

Wired. (2016). *We run down the worst security breaches of 2016 and reveal how to stay safe in 2017*. https://www.wired.com/story/biggest-hacks-2016/

Wu, M., Aranovich, R., & Filkov, V. (2021). Evolution and differentiation of the cybersecurity communities in three social question and answer sites: A mixed-methods analysis. *PLOS ONE*, *16*(12), e0261954. https://doi.org/10.1371/journal.pone.0261954

Yang, W., Prasad, M. R., & Xie, T. (2018). EnMobile: Entity-based characterization and analysis of mobile malware. *Proceedings of the 40th International Conference on Software Engineering*, 384–394. https://doi.org/10.1145/3180155.3180223

Yigit, Y., Ferrag, M. A., Sarker, I. H., Maglaras, L. A., Chrysoulas, C., Moradpoor, N., & Janicke, H. (2024). *Critical Infrastructure Protection: Generative AI, Challenges, and Opportunities* (No. arXiv:2405.04874). arXiv. https://doi.org/10.48550/arXiv.2405.04874

Z. Ma, M. Xiao, Y. Xiao, Z. Pang, H. V. Poor, & B. Vucetic. (2019). High-Reliability and Low-Latency Wireless Communication for Internet of Things: Challenges, Fundamentals, and Enabling Technologies. *IEEE Internet of Things Journal*, *6*(5), 7946–7970. https://doi.org/10.1109/JIOT.2019.2907245

Zhang, H., Wu, C., Xie, J., Kim, C., & Carroll, J. M. (2023). *QualiGPT: GPT as an easy-to-use tool for qualitative coding*. 1–25. https://doi.org/10.48550/arXiv.2310.07061

Zhang, Z., Ning, H., Shi, F., Farha, F., Xu, Y., Xu, J., Zhang, F., & Choo, K.-K. R. (2022). Artificial intelligence in cyber security: Research advances, challenges, and opportunities. *Artificial Intelligence Review*, *55*(2), 1029–1053. https://doi.org/10.1007/s10462-021-09976-0

ISSN: 1994-0098 | e-ISSN: 3086-0210

Appendix A. QualiGPT Configuration and Prompt Design

A.1 Overview of QualiGPT Usage

Qualitative thematic analysis in this study was conducted using QualiGPT, a large language model (LLM)-based tool designed to support inductive analysis of textual data (Zhang et al., 2023). The role of QualiGPT was strictly interpretive and complementary to the lexicographic analysis performed using IRaMuTeQ. Specifically, the model was applied to textual subsets previously organized into classes through Descending Hierarchical Classification (DHC), ensuring that the structural segmentation of the corpus remained entirely data-driven.

The analytical objective was to identify emergent thematic patterns within each class, based on spontaneous discourse produced in online cybersecurity communities. The approach followed an inductive logic, in which themes were derived directly from the data without predefined coding schemes.

A.2 Model Configuration

The QualiGPT analysis was conducted in **November 2025** using the **ChatGPT Free-tier environment**, which at that time provided access to a GPT-4–class model (commonly referred to as GPT-4o or equivalent variant, depending on system allocation).

The configuration parameters were as follows:

- **Model family:** GPT-4–class model (ChatGPT Free-tier, November 2025);
- **Interface:** ChatGPT web interface (QualiGPT custom GPT);
- **Temperature:** Default system setting (low-to-moderate variability, estimated ~0.2–0.3);
- **Top-p:** Default;
- **Max tokens:** System-managed (no manual override);
- **Frequency penalty:** Default (0.0);
- **Presence penalty:** Default (0.0).

Configuration considerations

As the analysis was conducted using the standard ChatGPT interface, fine-grained parameter control was not available. However, consistency was ensured by:

- maintaining identical prompt structures across all analyses;
- executing analyses within the same temporal window (November 2025);
- avoiding iterative prompt modifications during class-level processing.

Note. This approach reduces variability and enhances procedural consistency, despite inherent limitations of closed-interface LLM environments.

A.3 Input Data and Structure

The input to QualiGPT consisted of textual files corresponding to the four classes generated through IRaMuTeQ:

- Class 1 — Training and Learning Practice;
- Class 2 — AI and Protection;
- Class 3 — Infrastructure and Credentials;
- Class 4 — Work and Career.

These files were previously constructed and thematically organized through Descending Hierarchical Classification and served as the basis for qualitative interpretation.

The dataset used for QualiGPT analysis comprised:

- Total size: approximately 3.8 MB;
- Total textual units: 32,336 posts;
- Timeframe: 2022–2025.

Each class-specific file was uploaded and analyzed independently to preserve thematic coherence and avoid cross-class contamination during interpretation.

All input data were:

- preprocessed using the pipeline described in Section 3 - Methodology;
- anonymized and stripped of identifiers;
- formatted as plain text.

A.4 Prompt Design and Analytical Structure

The QualiGPT analysis was guided by predefined prompts designed to ensure analytical consistency and methodological transparency. The prompts were structured around three key elements:

1. Objective of analysis - to understand users' perceptions expressed in cybersecurity discussions;
2. Nature of the data - spontaneous textual content produced in online communities;
3. Expected output format - comprehensive thematic representations.

The core prompt structure used in the analysis is reproduced below.

Prompt Template

Input (User)
I want to do inductive coding
Output (QualiGPT)
Study background: Goal: Data collection context: Output preference: Data: [Paste your qualitative data here]
Input (User)
Study background: The data analyzed consist of posts, interactions, and debates extracted from online communities dedicated to cybersecurity, covering the period from 2022 to 2025. Goal: to understand users' perceptions; Data collection context: the data collection method- based on spontaneous comments produced within the investigated communities. Output preference: all formats supported by the tool. Data: Class1.txt
Output (QualiGPT)
*As a result, findings were generated in multiple analytical representations, including thematic summaries, codes, clusters, and interpretive descriptions.

Note. Each Class Corpus was attached to the Model. The same procedure was repeated for each Class.
*Text written by the authors.

A.5 Analytical Outputs

Based on the prompt structure, QualiGPT generated outputs in multiple analytical formats, including:

- thematic sumaries;
- coded categories;
- clustered topic groupings;
- interpretive descriptions.

These outputs provided complementary perspectives on the same textual data, enabling a richer understanding of the discourse within each class.

A.6 Analytical Procedure

The analysis followed a standardized procedure for each class:

1. Upload of class-specific textual file;
2. Application of the predefined prompt;
3. Generation of thematic outputs by QualiGPT;
4. Extraction of recurrent themes across outputs.

Each class was analyzed independently, ensuring that thematic interpretation remained grounded in the specific lexical structure identified by IRaMuTeQ.

A.7 Inductive Thematic Consolidation

Following the QualiGPT analysis, the authors conducted a consolidation process to construct the final analytical framework. This process involved:

- reviewing all generated outputs;
- identifying recurrent and overlapping themes;
- grouping related codes into broader categories.

Only macro-level emergent themes were retained, as these represent the most stable and recurrent patterns in the corpus. This decision was made to ensure analytical clarity and avoid over-fragmentation of results.

The final thematic structure reflects an inductive synthesis of QualiGPT outputs, guided by the frequency and consistency of themes across class-specific analyses.

A.8 Reproducibility Considerations

To enhance reproducibility, several measures were implemented:

- use of a standardized prompt across all analyses;
- consistent analytical procedure for each class;
- documentation of model context (ChatGPT Free-tier, November 2025);
- preservation of input files and outputs.

However, it is important to acknowledge that analyses conducted through ChatGPT interfaces are subject to certain constraints, including:

- limited control over model parameters;
- potential variability across sessions;
- dependence on model updates over time.

Despite these limitations, the use of structured prompts and consistent procedures contributes to the reliability of the analytical process.

A.9 Human Validation

All QualiGPT-generated outputs were reviewed by the authors to ensure alignment with the underlying data. This validation process included:

- verification of thematic consistency;
- exclusion of unsupported or weakly grounded interpretations;
- consolidation of overlapping categories.

This step ensures that the final thematic framework reflects both automated analysis and researcher judgment, mitigating risks associated with fully automated interpretation.

A.10 Methodological Positioning and Limitations

QualiGPT is employed in this study as an assistive analytical tool rather than a primary method for knowledge generation. The thematic interpretation remains grounded in the statistically derived structure of the corpus produced by IRaMuTeQ.

Nevertheless, the use of LLM-based tools introduces limitations, including:

- sensitivity to prompt formulation;
- potential influence of model training data;
- lack of full transparency in internal model processes.

Additionally, the use of the ChatGPT Free-tier environment limits parameter control and may affect reproducibility across different system configurations.

These limitations were mitigated through prompt standardization, procedural consistency, and human validation, ensuring that QualiGPT contributes to, rather than determines, the analytical outcomes.

A.11 Data Availability and Replication

To support transparency and replication:

- all prompts used in the analysis are documented in this appendix;
- class-specific input files are preserved in the research repository;
- preprocessing scripts and analytical procedures are available.

In accordance with ethical and platform compliance considerations, raw user-generated content will not be publicly redistributed unless permitted by platform terms of service. Instead, aggregated data and derived analytical outputs will be made available.