IJSRA

International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(Research Article)

Check for updates

# What is forensic chemistry research in Brazil made of? A descriptive and topic modeling analysis of CAPES theses and dissertations

Flávio Leite Rodrigues [1, *], Galileu Batista de Sousa [2] and Marcelo de Andrade Lima Maia [1]

[1] *Universidade Federal do Rio Grande do Norte (UFRN), Natal, RN, Brazil.*
[2] *Instituto Federal do Rio Grande do Norte (IFRN), Natal, RN, Brazil.*

International Journal of Science and Research Archive, 2025, 15(02), 695-705

Publication history: Received on 04 April 2025; revised on 11 May 2025; accepted on 13 May 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.15.2.1372

## Abstract

This article investigates the landscape of forensic chemistry research in Brazil by analyzing 118 master's and doctoral theses and dissertations retrieved from the CAPES Thesis and Dissertation Catalog using the keyword “química forense”. Through text mining and Latent Dirichlet Allocation (LDA) topic modeling, we identified thematic patterns and institutional trends in graduate-level research. Descriptive statistics revealed a concentration of academic production in southeastern Brazil and a predominance of master’s theses. The topic modeling analysis, applied to both unigrams and n-grams, uncovered research clusters focused on analytical instrumentation, electrochemical methods, drug detection, post-blast residue analysis, chemometric tools, and forensic education. Notably, topics such as synthetic drugs and pedagogical applications emerged as distinct and underexplored areas. The results offer a structured overview of the academic development of forensic chemistry in Brazil and highlight the utility of topic modeling as a tool for meta-research and policy planning in science education.

**Keywords:** Forensic Chemistry; Topic Modeling; LDA; Text Mining; CAPES; Graduate Education; Brazil; Scientific Mapping; Chemometrics; Electrochemistry

## 1. Introduction

Forensic chemistry is a critical discipline at the intersection of analytical science and criminal justice, encompassing techniques for the detection, identification, and quantification of substances involved in criminal investigations. Its importance lies in the production of scientifically validated evidence that can support legal decisions, contributing to the credibility of forensic science. In Brazil, the institutional and academic development of forensic chemistry has advanced significantly over recent decades, yet there remains a lack of systematic evaluations of how this field is represented in graduate-level research.

Recent advances in forensic chemistry have shown significant developments in analytical techniques and methodologies. However, comprehensive bibliometric assessments, particularly through topic modeling and text mining, are still sparse in this field. Such methods, widely adopted in social sciences and information science, hold the potential to reveal unseen thematic structures in scientific production, thus providing strategic insights for academic planning and policymaking.

To address this gap, this study draws on data extracted from the CAPES Thesis and Dissertation Catalog (Catálogo de Teses e Dissertações da CAPES, available at https://catalogodeteses.capes.gov.br/). Managed by the Coordination for the Improvement of Higher Education Personnel (CAPES), a foundation linked to Brazil’s Ministry of Education, this open-access platform is the central repository of postgraduate academic output in Brazil. It encompasses metadata and,

[*] Corresponding author: Flávio Leite Rodrigues

in many cases, full-text access to theses and dissertations defended in accredited master's and doctoral programs across the country. The platform plays a vital role in supporting transparency, evaluation, and policymaking in Brazilian postgraduate education and is widely used for scientometric studies and research diagnostics.

This study collected data from the Catalog in April 2025 using the keyword "química forense", which returned 161 records. After abstract screening, 118 documents—comprising both master's theses and doctoral dissertations—were retained for analysis.

Descriptive statistical analysis reveals that most of the graduate research output is concentrated in public institutions, with Universidade Federal do Espírito Santo (UFES), Universidade de São Paulo (USP), and Universidade Federal de Minas Gerais (UFMG) leading in volume. The number of submissions has grown consistently over the past decade, peaking in 2023. Master's theses account for 69.5% of the documents, while doctoral dissertations represent 30.5%. There is also a marked geographic imbalance: more than two-thirds of the academic production originates from the Southeast region, whereas the North contributes only marginally (Figure 1).

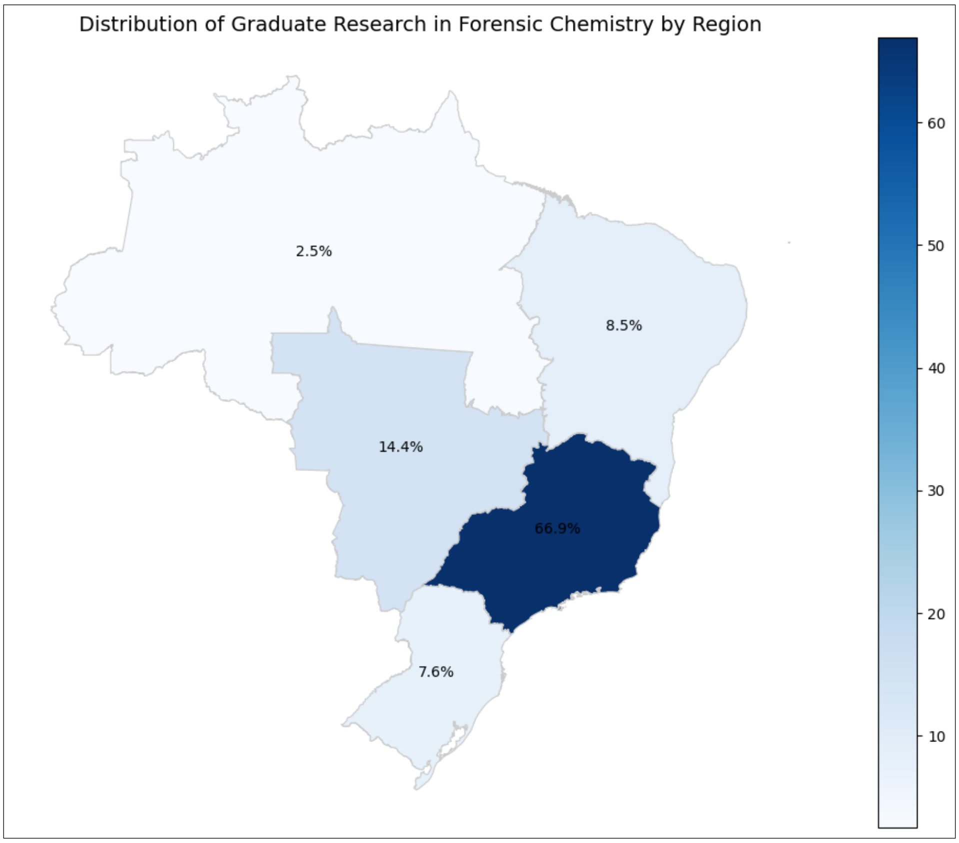
Distribution of Graduate Research in Forensic Chemistry by Region
2.5%
8.5%
14.4%
66.9%
7.6%
60
50
40
30
20
10

**Figure 1** Distribution of graduate research in forensic chemistry by Brazilian region. Source: Map geometry adapted from Click That 'Hood (Code for America), and data from the CAPES Theses and Dissertations Catalog (2025)

To explore the thematic structure of this body of literature, we applied Latent Dirichlet Allocation (LDA), a probabilistic topic modeling method well suited for text mining of academic corpora [1]. Prior to modeling, all abstracts were pre-processed, resulting in a word count distribution centered between 130 and 150 words per document (Figure 2), a length considered appropriate for topic modeling stability.

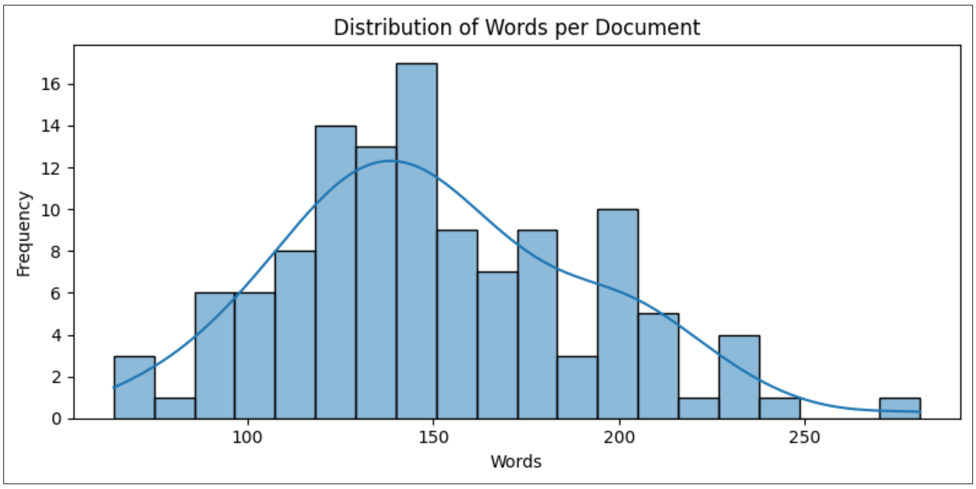

**Figure 2** Histogram and density curve showing the distribution of word counts per thesis/dissertation abstract included in the dataset. The most frequent word count is centered around 140–160 words, with a slightly right-skewed distribution. Source: Elaborated by the authors from data collected on CAPES Theses and Dissertations Catalog

While LDA and related techniques have been successfully applied in organizational and political sciences [2,3, 4], their use in forensic science remains emergent. By applying these tools to the Brazilian corpus of forensic chemistry research, this study aims not only to identify dominant thematic trends but also to contribute methodologically to the integration of text mining into forensic scholarship.

## 2. Materials and Methods

### 2.1. Data Collection

The data analyzed in this study were retrieved from the CAPES Thesis and Dissertation Catalog (https://catalogodeteses.capes.gov.br/), the official repository of graduate academic output in Brazil. A search was conducted in April 2025 using the keyword "química forense" (forensic chemistry), which returned 161 entries spanning from 2014 to 2024. Each record includes metadata such as title, abstract, author, year of defense, type of degree, and affiliated institution.

After manually reviewing the abstracts to ensure thematic alignment, 118 documents were retained. These comprised both master's theses and doctoral dissertations specifically focused on forensic chemistry topics. The filtering process excluded works in adjacent fields (e.g., general toxicology or environmental chemistry) unless explicitly connected to forensic applications. The resulting dataset was compiled into a comma-separated values (CSV) file, which served as the basis for all subsequent analyses.

### 2.2. Descriptive Analysis

Descriptive statistics were applied to characterize the dataset in terms of:

- Institutional affiliation: number of documents per university.
- Geographic region: distribution across Brazil's five official regions.
- Degree type: master's vs. doctoral programs.
- Year of defense: annual trends in submission.

Visualization of results was performed using Matplotlib and Seaborn, generating bar plots and pie charts. Figure 2 presents the distribution of word counts across abstracts, confirming text adequacy for modeling.

### 2.3. Text Preprocessing

To optimize LDA modeling performance, hyperparameter tuning (e.g., alpha and beta parameters) was carried out using a grid search method. The model's semantic coherence was measured using c_v scores, ensuring topic interpretability and coherence. Portuguese-specific preprocessing, including stopword removal and lemmatization using spaCy, was essential to accurately reflect the linguistic nuances of the dataset. Text preprocessing followed best practices for social science text analysis [5]

- Lowercasing
- Removal of punctuation, numbers, and special characters
- Tokenization
- Portuguese stopword removal (NLTK)
- Lemmatization (spaCy's Portuguese model)

After preprocessing, abstracts had a mean length of approximately 145 words, with distribution shown in Figure 2. This word count falls within the optimal range for topic modeling performance [5].

### 2.4. Topic Modeling

To uncover latent thematic patterns in the corpus, we applied Latent Dirichlet Allocation (LDA), a generative probabilistic model commonly used in natural language processing to discover abstract topics from a collection of documents. LDA assumes that each document is a mixture of topics, and each topic is a distribution over words. This approach allows researchers to identify hidden structures in large, unstructured text datasets.

All modeling procedures were implemented using the Gensim library [6] in Python, executed within Google Colab notebooks. Two modeling strategies were employed:

- Unigram model: based on individual words, commonly used in standard topic modeling workflows.
- N-gram model: based on sequences of two or more contiguous words (bigrams and trigrams), which help capture domain-specific expressions such as mass spectrometry, gunshot residue, or raman spectroscopy.

N-grams are particularly useful in forensic science texts, where technical terminology often appears in multiword units. To assess the lexical structure of the corpus, we extracted the most frequent filtered bigrams and trigrams (Figure 3), which revealed terms associated with analytical instrumentation, chemical substances, and forensic techniques.

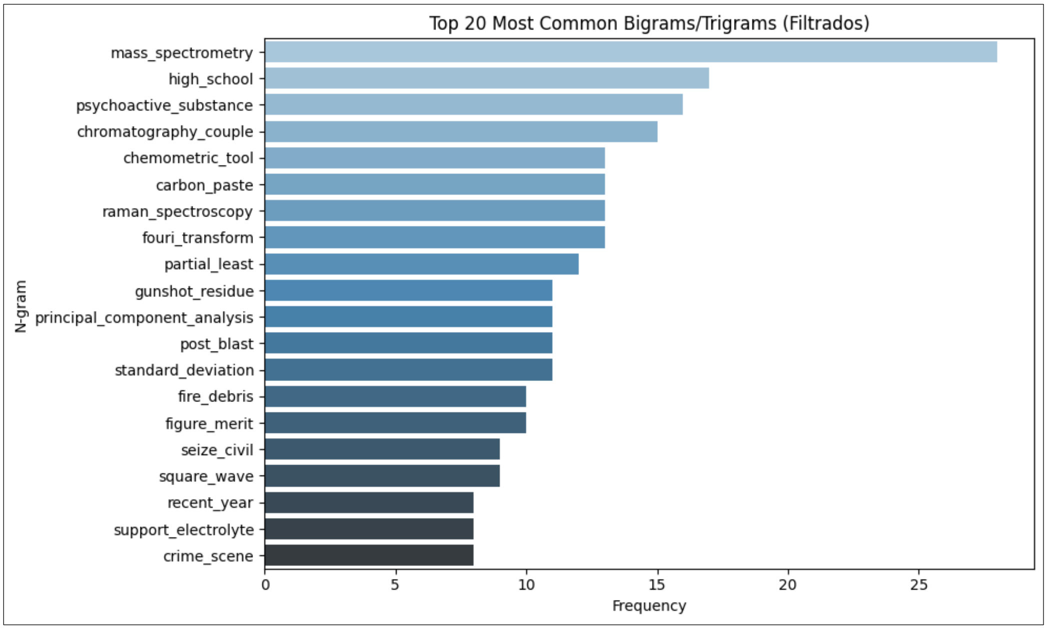

**Figure 3** Top 20 most common filtered bigrams and trigrams. The results highlight frequent use of expressions such as mass spectrometry, psychoactive substance, chromatography couple, and fire debris, reflecting the analytical and investigative focus of the forensic chemistry literature. Source: Elaborated by the authors from data collected on CAPES Theses and Dissertations Catalog

The number of topics and hyperparameter tuning were defined in an iterative process based on semantic coherence, exclusivity, and expert validation, as will be detailed in the following section.

## 3. Results and Discussion

### 3.1. Topic Modeling Based on Unigrams

The selection of the number of topics (K) for the LDA model was guided by coherence metrics and topic sparsity [7]. As illustrated in Figure 4, coherence values increased from K=8 to K=13, but higher values also generated several sparse or semantically redundant topics. After balancing these criteria, a six-topic model (K=6) was selected as optimal for interpretability and thematic compactness.

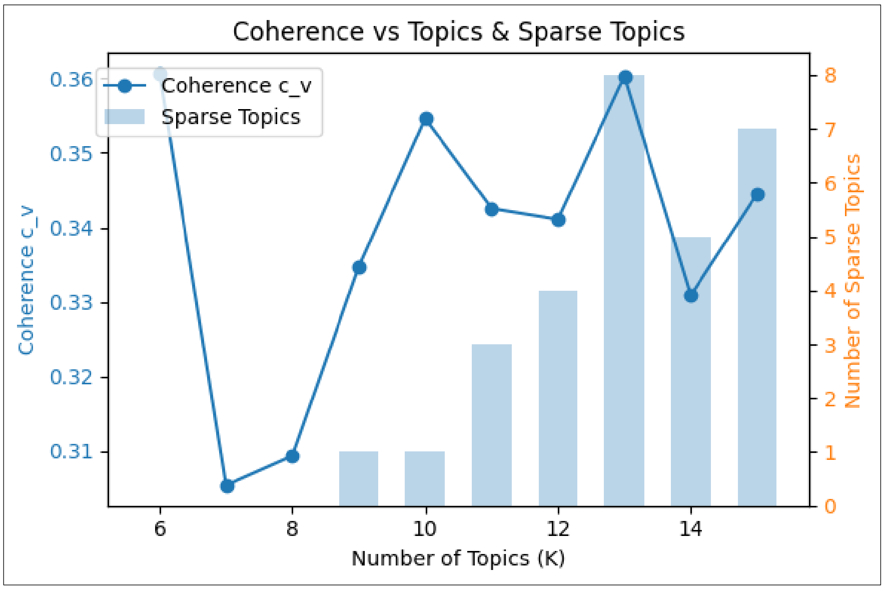

**Figure 4** Coherence score (c_v) and number of sparse topics for values of K ranging from 6 to 15. K=6 offered the best compromise between semantic coherence and model parsimony. Source: Elaborated by the authors from data collected on CAPES Theses and Dissertations Catalog

The distribution of dominant topics among the 118 documents is shown in Figure 5, with Topics 1 and 3 appearing most frequently.

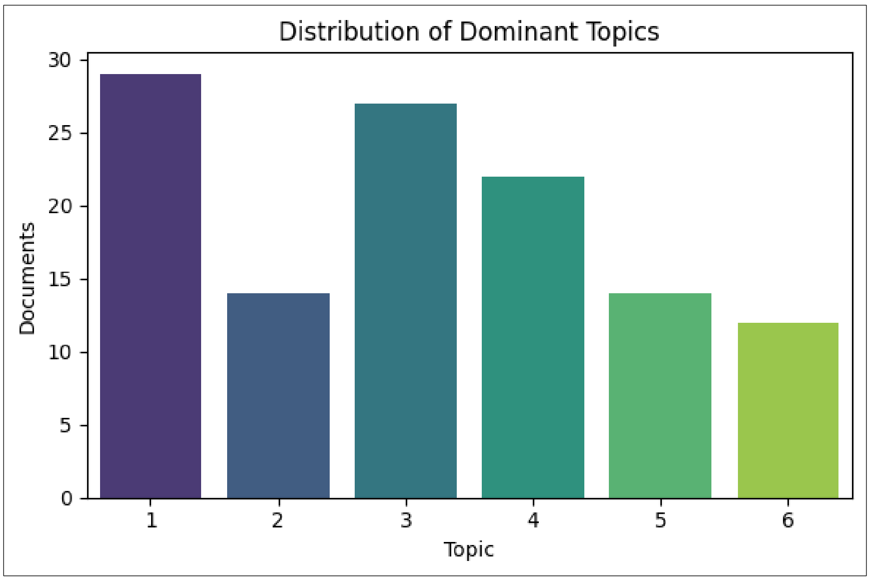

**Figure 5** Frequency of dominant topics across documents. Topics 1 and 3 together represent nearly half of the corpus. Source: Elaborated by the authors from data collected on CAPES Theses and Dissertations Catalog

Figures 6 and 7 present the visualization of each topic. Word clouds (Figure 6) highlight high-probability terms, the size of each word reflects its relevance and frequency within the topic. Common terms include "sample", "method", "analysis", and "drug", with topic-specific variations highlighting analytical techniques, educational aspects, and chemical substances [8]. while semantic graphs (Figure 7) map co-occurrence structures among top terms. Each node represents a keyword, and edges indicate co-occurrence within documents. These networks provide a structural view of the most interconnected terms per topic, revealing underlying thematic relationships in forensic chemistry research.

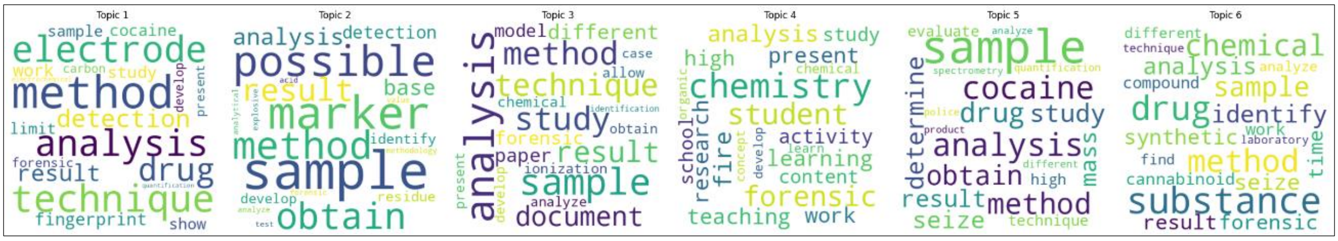

**Figure 6** Word clouds representing the six main topics identified through Latent Dirichlet Allocation (LDA) modeling applied to unigrams from 118 abstracts on forensic chemistry. Source: Elaborated by the authors from data collected on CAPES Theses and Dissertations Catalog

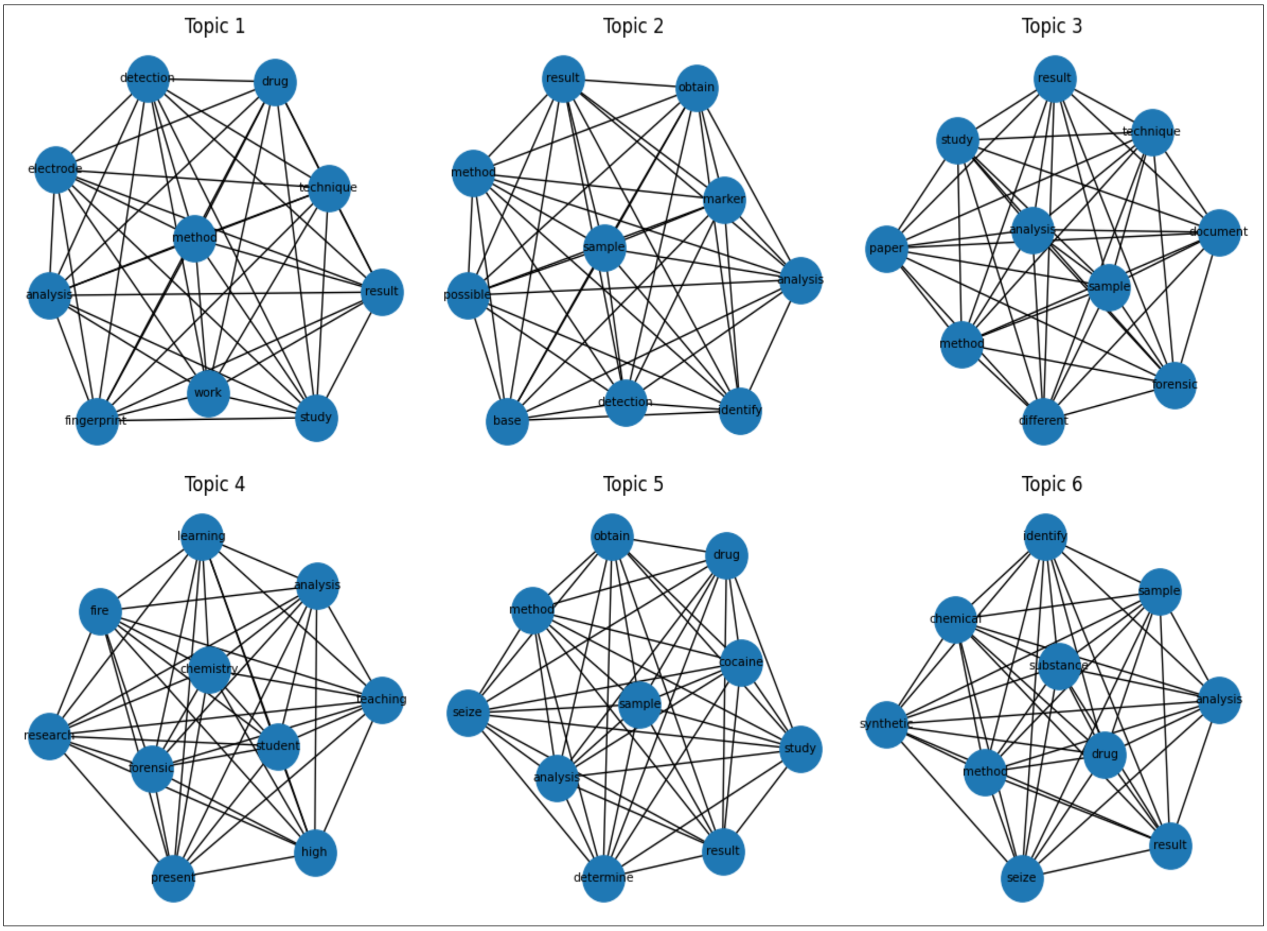

**Figure 7** Visualization of co-occurrence networks for the six topics extracted through Latent Dirichlet Allocation (LDA) using unigrams. Source: Elaborated by the authors from data collected on CAPES Theses and Dissertations Catalog

### 3.2. A detailed interpretation of each topic is presented below

*3.2.1. Topic 1 – Electrochemical Detection and Illicit Drugs*

This topic reflects studies involving electrochemical techniques for the detection of illicit substances. High-probability terms include electrode, detection, fingerprint, and cocaine. FREX and Lift terms suggest specificity to techniques such as voltammetry, use of potentiostat, and applications to explosives or designer drugs.

### 3.3. Focus: sensor development, electroanalytical detection, drug quantification

*3.3.1. Topic 2 – Chemical Markers and Residue Analysis*

This topic includes terms like marker, residue, obtain, and identify. FREX terms (e.g., hashish, distinction, isomer) indicate interest in substance profiling. The Score terms (blast, explosive, post) suggest forensic applications in post-blast analysis or residue detection.

### 3.4. Focus: marker identification, explosive residues, structural analysis

*3.4.1. Topic 3 – Analytical Validation and Document Forensics*

With central terms like technique, sample, document, and result, this topic includes general analytical method development and possibly document authentication. FREX terms include voltammetry, luminescent, and software. Score terms like ionization and counterfeit support applications in falsification detection.

### 3.5. Focus: analytical protocols, authentication, document analysis

*3.5.1. Topic 4 – Teaching and Learning in Forensic Chemistry*

This topic is distinct in incorporating pedagogical themes: student, teaching, learning, school, research. FREX and Lift terms refer to curriculum, training, and responsibility, indicating integration of forensic chemistry in educational contexts, likely at secondary or undergraduate levels.

### 3.6. Focus: forensic chemistry education, didactic strategies, outreach

*3.6.1. Topic 5 – Cocaine Analysis and Quantification*

This topic emphasizes cocaine, determine, purity, and seize. FREX terms such as film, latent, and journal suggest methodological innovation, while Lift terms reference pulse and thermogravimetric techniques.

### 3.7. Focus: cocaine profiling, purity quantification, forensic method application

*3.7.1. Topic 6 – Synthetic Drugs and Psychoactive Substances*

Includes substance, synthetic, cannabinoid, chemical, and psychoactive. The presence of identify, seize, and laboratory reinforces the topic's focus on emerging synthetic drugs and new psychoactive substances (NPS).

### 3.8. Focus: synthetic drug detection, psychoactive profiling, laboratory-based screening

Figures 9 and 10. Word clouds and semantic networks for each topic. These visualizations aid in topic interpretation by representing term prominence and intra-topic connectivity.

In summary, the LDA model using unigrams revealed six coherent themes within Brazilian graduate research on forensic chemistry. The topics span from sensor development and analytical chemistry to drug analysis and educational outreach, suggesting a methodologically diverse and socially relevant research landscape.

### 3.2 Topic Modeling Based on N-grams

To enrich the semantic representation of the corpus, a second LDA model was constructed using bigrams and trigrams, allowing for the identification of meaningful multiword expressions such as gunshot residue, square wave, or chromatography couple. The inclusion of n-grams enhances topic specificity by preserving domain-relevant collocations often lost in unigram-based models.

The optimal number of topics was determined by analyzing coherence scores and topic sparsity (Figure 8). The model with K=7 was selected due to its high semantic coherence and manageable number of sparse topics [9]. Figure 8 provides visual representations of the topics. Word clouds emphasize the most probable n-grams.

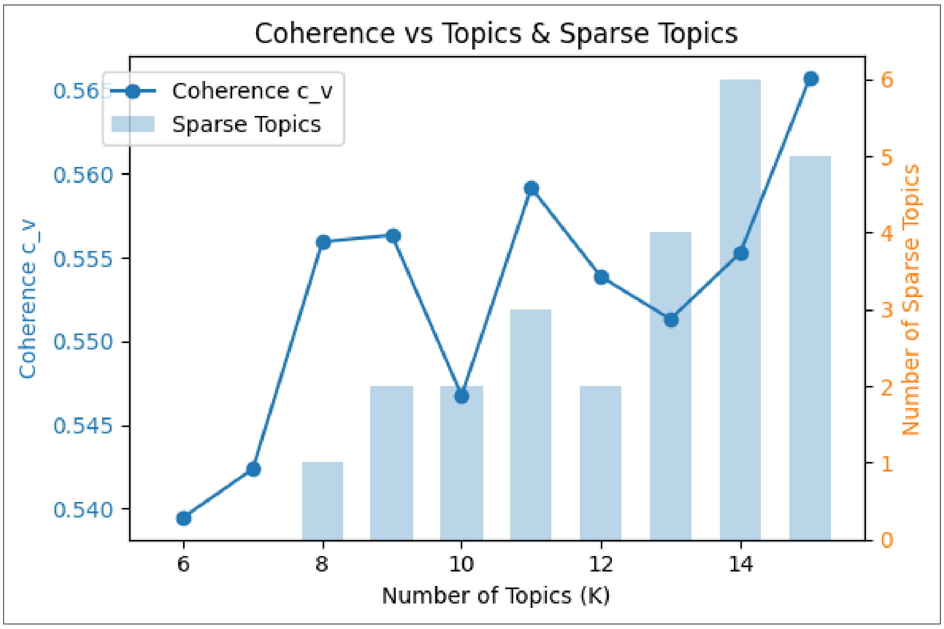

**Figure 8** Topic Coherence and Sparsity across Different Numbers of Topics (Bigrams). Source: Elaborated by the authors from data collected on CAPES Theses and Dissertations Catalog

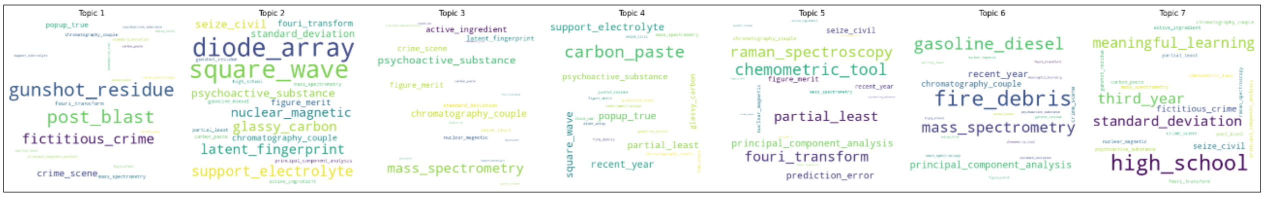

**Figure 9** Visualization of co-occurrence networks for the six topics extracted through Latent Dirichlet Allocation (LDA) using unigrams. Source: Elaborated by the authors from data collected on CAPES Theses and Dissertations Catalog

Below is a thematic interpretation of each topic, using the Highest Probability, FREX, Lift, and Score criteria for validation.

### *3.8.1. Topic 1 – Post-blast Residue and Crime Scene Analysis*

Terms like gunshot residue, post blast, fictitious crime, and crime scene dominate. FREX terms such as fire debris and latent fingerprint reinforce the forensic focus on explosive residues and physical evidence.

## 3.9. Focus: residue analysis, post-blast forensics, crime scene interpretation

### *3.9.1. Topic 2 – Electrochemical Techniques and Instrumentation*

Characterized by expressions like diode array, square wave, and support electrolyte, this topic reflects electroanalytical instrumentation, with applications in substance detection. FREX terms include chemometric tool and high school, suggesting some educational overlap.

### 3.10. Focus: electrochemical sensors, voltammetric methods, technical parameters

#### *3.10.1. Topic 3 – Mass Spectrometry and Psychoactive Substance Detection*

Core terms include mass spectrometry, psychoactive substance, chromatography couple, and active ingredients. This topic likely represents the use of advanced analytical techniques in the detection and classification of drugs [10, 11].

### 3.11. Focus: chemical analysis of drugs, instrumental techniques, forensic toxicology

#### *3.11.1. Topic 4 Voltammetric Platforms and Recent Advances*

With terms like carbon paste, support electrolyte, square wave, and recent years, this topic may reflect recent publications focusing on voltammetric sensor development and methodological refinement.

### 3.12. Focus: voltammetric materials, recent methodological innovations

#### *3.12.1. Topic 5 Chemometric Modeling and Spectroscopic Methods*

Dominated by chemometric tool, Raman spectroscopy, partial least, and principal component analysis, this topic relates to data-driven analytical approaches, particularly using spectroscopy and multivariate statistics.

### 3.13. Focus: chemometric analysis, spectral interpretation, pattern recognition

#### *3.13.1. Topic 6 Fire Debris and Accelerant Detection*

Featuring fire debris, gasoline diesel, mass spectrometry, and crime scene, this topic concerns the chemical investigation of incendiary materials. It overlaps partially with Topic 1 but emphasizes accelerants.

### 3.14. Focus: ignitable liquid residue analysis, fire scene investigation

#### *3.14.1. Topic 7 Forensic Education and Teaching Strategies*

With unique terms such as high school, third year, meaningful learning, and standard deviation, this topic clearly aligns with pedagogical research, suggesting educational interventions or curriculum design.

### 3.15. Focus: forensic chemistry education, science teaching, student engagement

The n-gram-based model confirmed several findings from the unigram analysis—such as the centrality of analytical techniques and drug detection—but provided enhanced resolution by clarifying distinct domains like post-blast analysis, chemometrics, and fire debris. Notably, educational research also emerged as a consistent theme across both models, indicating the strategic expansion of forensic chemistry into teaching environments [12].

The dominance of analytical instrumentation and electrochemical techniques mirrors global trends, where sophisticated chemical analytics remain central to forensic investigations [13]. However, the apparent underrepresentation of integrated forensic intelligence and field-based validation studies in Brazil contrasts with recent international emphases, indicating a potential gap that could be strategically addressed through targeted academic policies.

## 4. Conclusion

This study conducted a comprehensive topic modeling analysis on 118 master's and doctoral theses and dissertations related to forensic chemistry, retrieved from the CAPES Theses and Dissertations Catalog. By applying Latent Dirichlet Allocation (LDA) to both unigrams and n-grams, we identified major research themes and structural patterns across the national academic output. It clearly delineates research strengths and identifies critical gaps, particularly the need for greater emphasis on forensic intelligence integration and educational innovation. Strategic investments in these identified areas could further enhance Brazil's forensic chemistry capabilities, aligning national research efforts more closely with global scientific advancements.

The analysis revealed distinct thematic clusters in both models. In the unigram-based model, six topics emerged, centered on electrochemical detection, chemical markers, analytical validation, educational initiatives, cocaine quantification, and synthetic drug monitoring. The n-gram model yielded seven topics, with finer semantic resolution, capturing domains such as post-blast analysis, voltammetric instrumentation, chemometrics, fire debris, and forensic education.

Topic similarity heatmaps indicated that most topics are semantically distinct, with occasional overlaps (e.g., T5–T6 in both models). This supports the diversity of approaches within the field while suggesting potential for integration across subdomains. Furthermore, topic effect plots by academic level show that PhD theses tend to emphasize topics related to fire debris and forensic education, while master's dissertations are more concentrated on instrumental methods and drug analysis. These results suggest differing scopes of complexity and societal engagement between academic levels.

### *Limitations and Future Work*

Despite its insights, the study is limited by its exclusive reliance on abstracts and metadata, which may overlook deeper methodological nuances. Also, the LDA model is sensitive to preprocessing choices, including stopword removal and n-gram generation.

Future studies may expand by

- Including full-text analysis of the theses;
- Combining topic modeling with citation or network analysis;
- Exploring regional or institutional variations more deeply;
- Applying dynamic topic modeling to track temporal evolution.

The domain of forensic analytical chemistry has evolved significantly over recent decades, driven by continuous methodological enhancements and revolutionary technological innovations aimed at improving the precision, reliability, and applicability of forensic evidence analysis. Modern forensic chemists now routinely adopt sophisticated analytical technologies, including portable spectrometric devices and advanced chemometric analyses, facilitating more efficient and accurate in situ forensic investigations.

However, alongside these advancements, contemporary forensic practitioners face intricate operational challenges arising from the intersection of scientific rigor and procedural demands within the criminal justice framework. Issues such as the accuracy and reliability of presumptive testing procedures, potential sample contamination risks inherent to field testing scenarios, and the critical assessment of portable analytical instrumentation performance underscore the complex interplay between scientific capabilities and practical forensic constraints.

Strategic investments in fostering interdisciplinary research initiatives, specifically those integrating forensic intelligence and field-based applications, could substantially enhance the efficacy and impact of forensic chemistry within the national and international criminal justice landscape. Additionally, systematic longitudinal monitoring of evolving research trends and thematic patterns, such as those identified through topic modeling methodologies, promises valuable insights to inform future directions, curriculum development, and policy formulation, thereby elevating Brazil's global standing in both scientific rigor and practical judicial outcomes.

In conclusion, forensic chemistry in Brazil represents a vibrant and progressively expanding field, characterized by robust foundations in analytical science, which has consistently leveraged advanced instrumental and methodological approaches. Notably, recent literature underscores a significant transition from purely laboratory-based analytical chemistry toward more integrative practices that include pedagogical innovation and societal relevance, highlighting the growing interconnection between academia, criminal justice agencies, and broader community contexts.

## Compliance with ethical standards

### *Disclosure of conflict of interest*

No conflict-of-interest to be disclosed.

### *Contributions and Implications*

This research provides a data-driven overview of forensic chemistry research trends in Brazil over the past decade. Key contributions include:

- A structured thematic mapping of graduate research in the field;

- The identification of gaps and strengths, such as underrepresentation of integrated forensic intelligence or field-based validation;
- The demonstration of topic modeling as a viable tool for meta-research and curriculum development in forensic science.

## References

[1] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003; 3:993–1022.

[2] Grimmer J, Stewart BM. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Anal. 2013;21(3):267–97.

[3] Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, Gadarian SK, et al. Structural topic models for open-ended survey responses. Am J Polit Sci. 2014;58(4):1064–82.

[4] Li K, Mai F, Shen R, Yan X. Measuring corporate culture using machine learning. Rev Financ Stud. 2021;34(7):3265–3315.

[5] Banks GC, Woznyj HM, Wesslen RS, Ross RL. A review of best practice recommendations for text analysis in R (and a user-friendly app). J Bus Psychol. 2018;33(4):445–59.

[6] Huang KH. Gensim — Topic Modelling in Python: Medium; 2023 [cited 2025 Apr]. Available from: https://medium.com/@joloiuy/gensim-topic-modelling-in-python-1492a3e9a873

[7] Weston SJ, Shryock I, Light R, Fisher PA. Selecting the number and labels of topics in topic modeling: A tutorial. Adv Methods Pract Psychol Sci. 2023;6(2):25152459231160105.

[8] Veiga M. Forensic analytical chemistry: Connecting science and justice. Braz J Anal Chem. 2022;8(NX2):1–2.

[9] Hannigan TR, Haans RFJ, Vakili K, Tchalian H, Glaser VL, Wang MS, et al. Topic modeling in management research: Rendering new theory from textual data. Acad Manag Ann. 2019;13(2):586–632.

[10] Burks R, Sauzier G, Kammrath BW, Houck MM. Forensic Analytical Chemistry for Minimizing Injustice: Advances and Challenges. 2025 Feb 27 [cited 2025 May 6]; Available from: https://www.annualreviews.org/content/journals/10.1146/annurev-anchem-072624-030546

[11] Yadav VK, Kumar A, Shahid S, Nigam K, Srivastava A. Forensic chemistry. In: Shrivastava P, editor. Textbook of Forensic Science. 1st ed. Singapore: Springer; 2023. p. 661–705.

[12] Illes M, Wilson P, Bruce C. Forensic epistemology: A need for research and pedagogy. Forensic Sci Int Synergy. 2020; 2:51–9.

[13] Dave PY. Review on forensic chemistry. Int J Forensic Sci. 2021;6(2):1–12.