informatics

MDPI

Article

Exploring Scientific Literature Using Topic Modeling: A Practical Framework for Discovery and Classification

Amir Alipour Yengejeh [1], Larry Tang [1,2,*], Candice M. Bridge [2,3] and Chandra Kundu [1]

[1] Department of Statistics and Data Science, University of Central Florida, Orlando, FL 32816, USA; amir.alipouryengejeh@ucf.edu (A.A.Y.); chandra.kundu@ucf.edu (C.K.)
[2] National Center for Forensic Science, University of Central Florida, P.O. Box 162367, Orlando, FL 32816, USA; cbridge@ucf.edu
[3] Department of Chemistry, University of Central Florida, Orlando, FL 32816, USA
* Correspondence: liansheng.tang@ucf.edu

Abstract

The increasing volume and diversity of scientific publications poses challenges for scalable and interpretable topic discovery and automated document categorization. This study proposes an integrated framework that combines probabilistic topic modeling with supervised classification to support large-scale scientific literature analysis. Using 3689 abstracts from the Journal of Forensic Sciences (2009–2022), Latent Dirichlet Allocation (LDA) is applied to uncover latent thematic structures, assess topic diagnosticity across forensic disciplines, and analyze temporal research trends. Bayesian model selection with repeated resampling identifies a stable topic resolution, with the number of topics T lying in the range 83–88, yielding semantically coherent and discipline-aligned topics. The resulting document–topic representations are then used for supervised abstract classification. Across multiple models and resampling scenarios, the strongest and most stable performance is achieved under a Grouped Category configuration. In particular, XGBoost attains an Accuracy of 0.754 and a Macro-averaged F1 score of 0.737 at $T = 88$, with comparable results at neighboring topic counts, indicating robustness to topic granularity. Overall, the proposed framework provides a reproducible, interpretable, and computationally efficient pipeline for literature organization, trend analysis, and metadata enhancement in scientific domains.

Keywords: topic modeling; latent dirichlet allocation; scientific literature analysis; text classification; supervised learning; temporal trend analysis; metadata enrichment; forensic science

Academic Editor: Olga Kurasova

Received: 3 December 2025
Revised: 11 January 2026
Accepted: 26 January 2026
Published: 30 January 2026

1. Introduction

The rapid growth of scientific literature across disciplines has introduced substantial challenges for knowledge discovery, thematic organization, and large-scale content analysis. Manual categorization and keyword-based indexing approaches struggle to scale effectively and often fail to capture latent semantic structure within large document collections. Consequently, automated topic modeling methods have become essential tools for exploring, summarizing, and organizing scientific texts.

While topic modeling has been widely applied in scientific literature analysis, many existing studies focus on unsupervised topic discovery or downstream classification tasks in isolation. Such fragmented approaches limit reproducibility and hinder systematic evaluation of how inferred topics support interpretability, temporal analysis, and predictive

performance. Moreover, the absence of a unified analytical framework often makes it difficult to assess topic quality, robustness, and practical utility within applied settings.

This study addresses these limitations by presenting a structured and reproducible topic modeling framework for scientific literature analysis. The proposed approach integrates unsupervised topic discovery, topic evaluation and diagnostic analysis, temporal trend assessment, and supervised document classification within a single end-to-end pipeline. Topic-based representations derived from Latent Dirichlet Allocation (LDA) are used not only to uncover latent thematic structure but also to support interpretable and effective classification of documents.

The framework is demonstrated using a large corpus of forensic science abstracts, a domain characterized by heterogeneous topics, evolving research priorities, and imbalanced category distributions. This application highlights the practical relevance of combining topic modeling with systematic evaluation and supervised validation in real-world scientific literature.

The main contributions of this work are summarized as follows:

- A structured topic modeling framework that integrates topic discovery, evaluation, temporal analysis, and supervised classification.
- A systematic approach for assessing topic interpretability, diagnostic relevance, and robustness across topic granularities.
- An empirical evaluation of topic-based representations for document classification under class imbalance and varying label granularity.
- A domain-specific case study demonstrating the applicability of the framework to large-scale forensic science literature.

The remainder of this paper is organized as follows. Section 2 reviews the background and related works on topic modeling and scientific text analysis. Section 3 describes the proposed analytical framework and methodological components. Section 4 presents empirical results from topic modeling, trend analysis, and classification experiments. Section 5 discusses implications, limitations, and directions for future work. Finally, Section 6 concludes the paper.

2. Background

Computational approaches for large-scale scientific literature analysis have evolved substantially over the past two decades, with probabilistic topic modeling playing a central role in uncovering latent thematic structure in text corpora. A foundational contribution in this area is Latent Dirichlet Allocation (LDA) [1], which formulates topic discovery as a probabilistic generative process based on word co-occurrence patterns. By representing documents as mixtures of latent topics, LDA provides an interpretable and flexible framework for thematic analysis in large document collections.

Early studies demonstrated the effectiveness of LDA for analyzing scientific literature. Griffiths and Steyvers [2] showed that topics inferred from academic articles align closely with journal-defined disciplines and can be used to examine the temporal evolution of research themes. These findings established LDA as a practical tool for mapping the structure and dynamics of scientific fields. From an implementation perspective, Ponweiser [3] further supported the adoption of LDA by providing a detailed guide for applying topic modeling in the R environment using the `topicmodels` package [4], emphasizing reproducibility and transparency.

Building on these foundations, numerous studies have applied LDA to large-scale scientific corpora to analyze research trends, journal similarity, and disciplinary evolution. For example, Gatti et al. [5] modeled more than 80,000 abstracts in operations research and management science to reveal shifting thematic structures and journal relationships.

In transportation research, Sun and Yin [6] analyzed over 17,000 articles to identify dominant research themes and regional patterns. Similarly, Xiong et al. [7] examined manufacturing research literature spanning multiple decades and reported increasing interdisciplinarity across technical and managerial domains.

More recent work highlights the scalability of LDA for mapping complex and evolving scientific landscapes. Yu and Xiang [8] analyzed over 170,000 publications to identify major subfields and temporal trends in artificial intelligence research, while Zhang et al. [9] applied LDA to millions of PubMed-indexed articles to uncover latent disease–gene associations. Topic modeling has also been widely used to support structured and systematic literature reviews. For instance, Madzík and Falát [10] analyzed more than 35,000 documents to study thematic evolution in the Analytic Hierarchy Process literature, and related approaches have integrated topic modeling with review methodologies to enable scalable and reproducible literature synthesis [11,12].

Beyond traditional literature review settings, topic modeling has been combined with complementary analytical techniques to examine research evolution and discourse in adjacent contexts. Examples include integrating LDA with bibliometric mapping and keyword co-occurrence analysis [13], combining topic modeling with sentiment analysis to study public discourse [14], and applying topic modeling to quantify thematic diversity within specialized scientific domains such as dental research [15] and blockchain-enabled food supply chains [16].

In parallel, several studies have investigated how topic representations can be incorporated into broader analytical pipelines. Debortoli et al. [17] showed that topic proportions inferred from LDA can serve as informative features for downstream machine-learning models, bridging unsupervised topic discovery and predictive modeling. However, such approaches are often limited to specific applications and are not formulated as unified, end-to-end frameworks for large-scale scientific literature classification or metadata enhancement.

At the methodological level, multiple surveys have summarized advances in topic modeling. Kherwa and Bansal [18] reviewed probabilistic topic modeling methods, inference strategies, evaluation metrics, and application domains, providing a comprehensive overview of classical approaches. More recent extensions include dynamic topic models designed to capture temporal evolution in scientific discourse, with Guillén-Pacho et al. [19] demonstrating interpretable labeling of evolving research themes.

In recent years, neural and embedding-based topic models have gained increasing attention due to their representational flexibility and scalability. A comprehensive survey by Wu et al. [20] reviews neural topic modeling methods, applications, and challenges. Representative approaches include BERTopic [21], which combines transformer-based embeddings with clustering, as well as its application to scientific literature screening tasks [22]. Large language models (LLMs) have also been explored for topic generation and evaluation through prompt-based frameworks such as TopicGPT [23] and related methods for short-text modeling and topic interpretation [24–27]. Despite their promise, these approaches typically operate outside probabilistic frameworks and do not naturally yield structured topic representations suitable for downstream supervised learning.

Overall, despite substantial progress in topic modeling and its applications to scientific literature, most existing work either emphasizes unsupervised topic discovery or treats topic modeling and classification as separate tasks. Relatively few studies provide a unified, reproducible framework that integrates interpretable topic modeling with systematic evaluation, temporal analysis, and supervised classification. To address this gap, the present study proposes an end-to-end framework that bridges unsupervised topic discovery with supervised learning, enabling both exploratory analysis and structured document classification using interpretable topic-based representations.

3. Methodology

Building on the background and related work discussed in Section 2, this section presents a modular and generalizable framework for large-scale scientific literature analysis that integrates probabilistic topic modeling with supervised classification. The proposed framework is designed to support both exploratory thematic discovery and structured document categorization, while maintaining interpretability, reproducibility, and scalability.

As illustrated in Figure 1, the analytical workflow proceeds in a structured, step-by-step manner and consists of five interconnected components: (1) text preprocessing, (2) topic modeling, (3) topic diagnostic, (4) trend analysis, and (5) supervised classification with performance assessment. Each component addresses a distinct stage of the analysis while contributing to a coherent end-to-end pipeline.

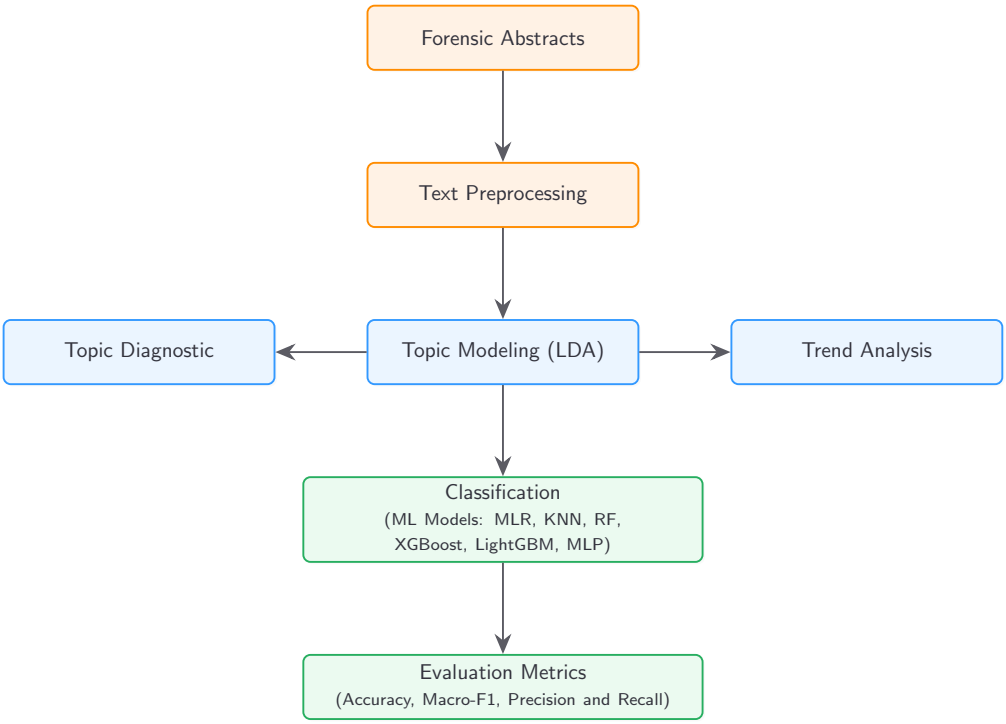

Figure 1. Schematic overview of the proposed analytical workflow. The pipeline includes text preprocessing, topic modeling via Latent Dirichlet Allocation (LDA), parallel topic evaluation and trend analysis, supervised classification using multiple machine-learning models, and performance assessment using standard evaluation metrics. Colors are used for visual clarity only and do not encode additional information.

3.1. Text Preprocessing

The initial phase of the analysis focuses on preparing forensic science abstracts for topic modeling and downstream classification. All documents were processed using a standardized text normalization pipeline designed to improve semantic consistency while reducing noise. This pipeline removes non-ASCII characters, punctuation, numeric tokens, stop words, and frequently occurring boilerplate terms common in scientific writing (e.g., "study", "method", "results"). Lemmatization is applied to reduce inflectional variation, and excess whitespace is normalized. After text cleaning, 3668 abstracts retained non-empty content.

To ensure that retained documents contained sufficient semantic information for modeling, abstracts with fewer than 60 words after preprocessing were excluded based on an empirical assessment of document length distributions. This step removed 59 abstracts (1.61%), resulting in a final corpus of 3609 abstracts used in all subsequent analyses.

Following cleaning and length filtering, the abstracts were tokenized into bigrams rather than unigrams in order to capture meaningful domain-specific phrases and improve semantic coherence. This representation is particularly appropriate for forensic literature, where key concepts are frequently expressed as multi-word terms (e.g., "crime scene", "cause of death", "mass spectrometry"). To further reduce noise, common demographic, administrative, and methodological expressions—including age descriptors and repeated statistical phrases—were explicitly filtered at both the token and bigram levels.

An exploratory overview of the resulting bigram vocabulary is presented in Figure 2 using a TF–IDF–weighted word cloud. Compared to frequency-based visualizations, TF–IDF weighting downweights ubiquitous phrases while emphasizing discriminative terms that characterize distinct forensic subdomains. Prominent themes include forensic chemistry (e.g., "mass-spectrometry", "gas-chromatography"), death investigation (e.g., "cause-death", "postmortem-interval", "manner-death"), biological evidence analysis (e.g., "dna-profile", "latent-fingerprint", "skeletal-remain"), and applied forensic practice (e.g., "crime-scene", "medical-examiner"). This visualization provides an interpretable summary of the dominant thematic structure present in the corpus after preprocessing.

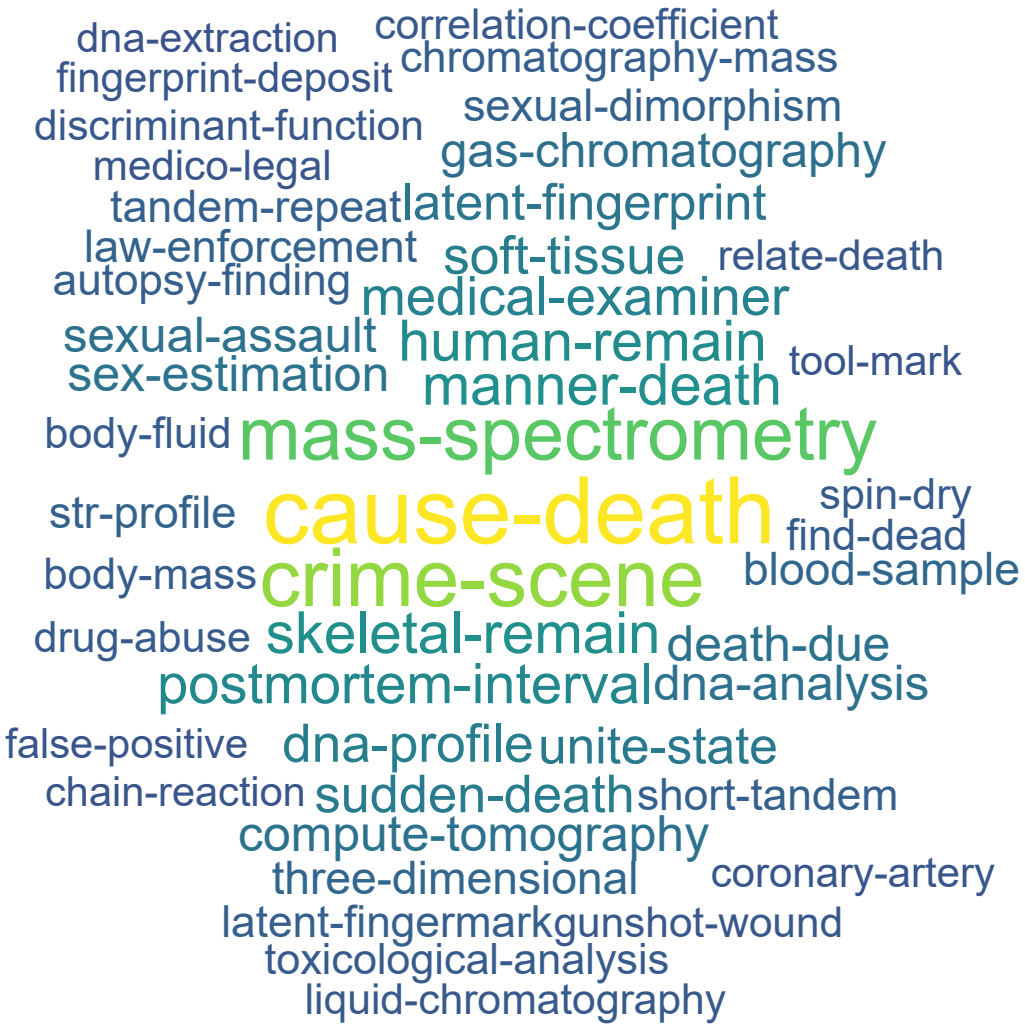

Figure 2. TF–IDF–weighted word cloud of bigram tokens extracted from the preprocessed forensic science abstracts. The visualization highlights discriminative multi-word terms representing major forensic subdomains, including analytical chemistry, death investigation, biological evidence analysis, and crime scene examination. Colors are used for visual emphasis only and do not encode additional information.

Subsequently, a document–term matrix (DTM) was constructed from the filtered bigram tokens, where each entry represents the frequency of a given bigram within a document. To reduce sparsity and improve model stability, rare terms appearing in five or fewer documents and overly common terms appearing in more than 90% of documents were removed. This filtering step reduced the vocabulary size from 227,162 to 4658 bigram terms while retaining all 3609 documents, as no document became empty after vocabulary pruning. The resulting sparse but informative DTM serves as the input to the Latent Dirichlet Allocation (LDA) model.

An overview of the preprocessing workflow used to construct the DTM is illustrated in Figure 3.

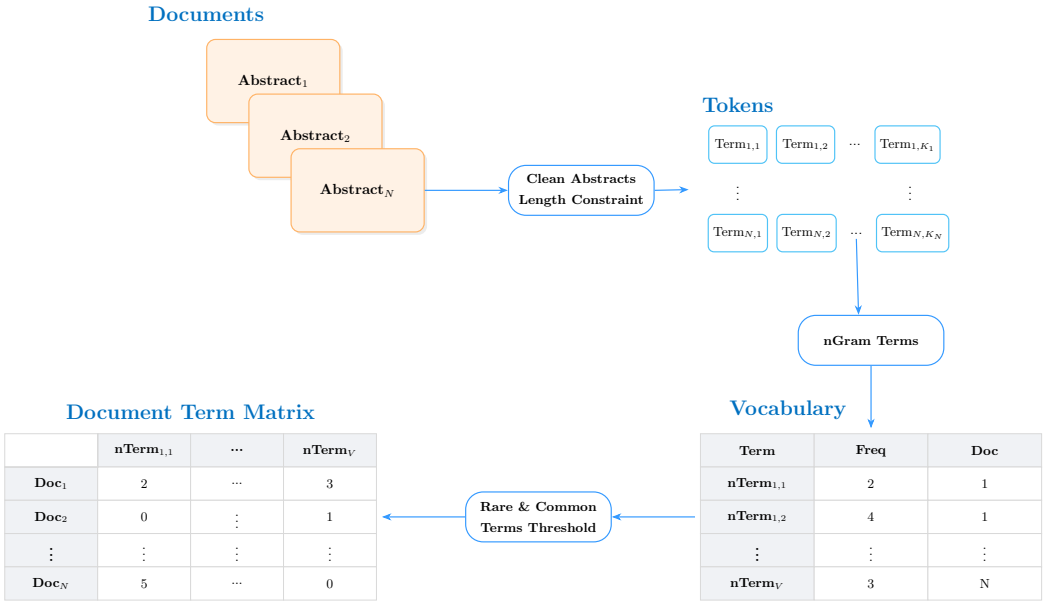

Figure 3. Workflow for Preparing Forensic Science Abstracts for Topic Modeling. The process begins with the collection of documents and proceeds through cleaning the abstracts, applying a length constraint, tokenizing the content into meaningful terms, generating bigrams to enhance contextual understanding, and ultimately forming the Document Term Matrix (DTM). Rare and common terms are then filtered out based on a predefined threshold to refine the vocabulary, which is then utilized in the topic modeling process. Numbers in the matrix are arbitrary examples to demonstrate the concept.

3.2. Topic Modeling via Latent Dirichlet Allocation

Following text preprocessing and construction of the document–term matrix, Latent Dirichlet Allocation (LDA) is applied to uncover latent thematic structure within the forensic science abstracts. LDA models each document as a mixture of latent topics, where each topic is characterized by a probability distribution over words. This formulation enables interpretable, low-dimensional representations of documents while preserving semantic structure.

Conceptually, LDA factorizes the conditional word distribution $P(w \mid d)$ into two latent components: a topic–word distribution $P(w \mid z)$ and a document–topic distribution $P(z \mid d)$, parameterized by $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, respectively. Figure 4 illustrates this decomposition, showing how observed word frequencies are explained through latent topic assignments.

Latent Dirichlet Allocation was introduced by Blei et al. [1] as a probabilistic generative model in which both topic distributions within documents and word distributions within topics are assigned Dirichlet priors. Figure 5 presents the graphical model underlying this generative process.

Formally, LDA assumes the following prior distributions:

$$\theta_d \sim \text{Dirichlet}(\alpha), \tag{1}$$

$$\phi_t \sim \text{Dirichlet}(\beta), \tag{2}$$

where α controls the concentration of topics within documents and β governs the concentration of words within topics. Smaller values of α encourage documents to focus on fewer topics, while smaller values of β produce more sharply defined topic vocabularies.

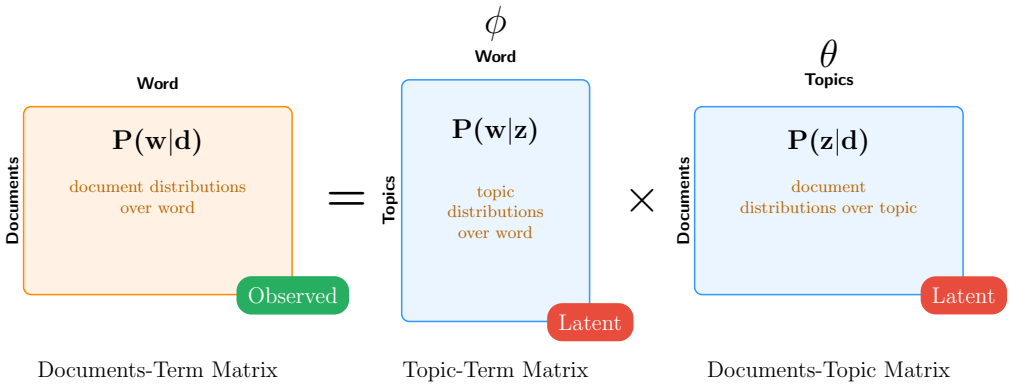

Figure 4. Decomposition of a document–term matrix into latent topic–word ($\boldsymbol{\phi}$) and document–topic ($\boldsymbol{\theta}$) distributions. The observed word distribution $P(w \mid d)$ is factorized through latent topic assignments, providing interpretable representations of documents and topics.

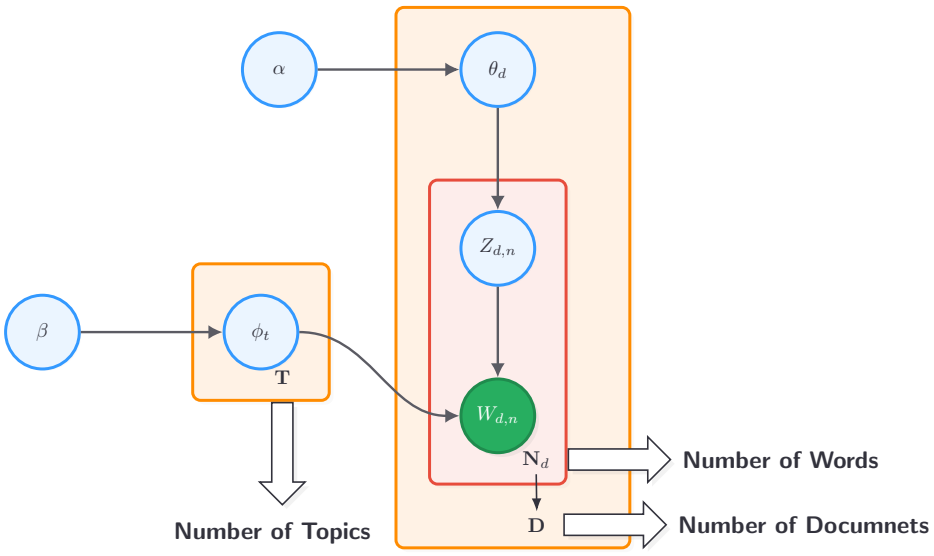

Figure 5. Graphical representation of the Latent Dirichlet Allocation generative process, illustrating the interaction between document–topic distributions ($\boldsymbol{\theta}$), topic–word distributions ($\boldsymbol{\phi}$), and latent topic assignments.

For each word token $w_{d,n}$ in document d, the generative process proceeds as:

$$z_{d,n} \sim \text{Multinomial}(\theta_d), \tag{3}$$

$$w_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}}), \tag{4}$$

where $z_{d,n}$ denotes the latent topic assignment.

Inference in LDA requires estimation of the posterior distribution $P(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{W}, \alpha, \beta)$, which is analytically intractable due to dependencies among latent variables. Approximate inference methods include variational Bayes [1] and collapsed Gibbs sampling [2].

In this study, collapsed Gibbs sampling is employed following Griffiths and Steyvers [2], integrating out $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ and sampling only the latent topic assignments $\mathbf{Z}$. Given posterior samples of topic assignments, the topic–word and document–topic distributions are estimated as:

$$\hat{\phi}_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + V\beta}, \tag{5}$$

$$\hat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_\cdot^{(d)} + T\alpha}, \tag{6}$$

where $n_j^{(w)}$ is the number of times word w is assigned to topic j, $n_j^{(\cdot)}$ is the total number of assignments to topic j, $n_j^{(d)}$ is the number of words in document d assigned to topic j, $n_\cdot^{(d)}$ is the total number of words in document d, V is the vocabulary size, and T denotes the number of topics.

The resulting document–topic matrix $\boldsymbol{\Theta}$ provides an interpretable, low-dimensional representation of each abstract and serves as the foundation for topic evaluation, temporal trend analysis, and downstream supervised classification.

3.3. Identification of Diagnostic Topics

To assess the extent to which individual topics are characteristic of specific forensic science categories, we quantify *topic diagnosticity* using a category-specific prevalence ratio. This measure identifies topics that are not only frequent within a given category but also relatively uncommon in other categories, thereby serving as discriminative thematic markers.

Let $\theta_j^{(c)}$ denote the average proportion of topic j among documents belonging to category c, computed from the document–topic matrix $\boldsymbol{\Theta}$. For each category c, the diagnostic topic d_c is defined as:

$$d_c = \arg\max_{j \in J} \left(\frac{\theta_j^{(c)}}{\sum_{\substack{r \in C \\ r \neq c}} \theta_j^{(r)}} \right), \tag{7}$$

where J denotes the set of all topics and C represents the set of all forensic science categories.

This ratio emphasizes topics that are disproportionately prevalent within a given category relative to all others, thereby highlighting category-specific thematic structure. The approach follows earlier work on topic diagnosticity in scientific literature analysis [3] and supports interpretability by linking latent topics to established disciplinary labels.

3.4. Topic Trend Analysis: Hot and Cold Topics

To examine the temporal evolution of thematic emphasis in forensic science research, we categorize the topics inferred by the LDA model into *hot* and *cold* topics based on longitudinal trends in topic prevalence, following established approaches in topic evolution analysis [2,3]. This analysis is conducted using annual aggregates of topic proportions derived from the document–topic matrix $\boldsymbol{\Theta}$.

Let $\theta_{d,k}$ denote the proportion of topic k in document d. For each topic k and calendar year t, the annual mean topic proportion is computed as:

$$\bar{\theta}_k(t) = \frac{1}{|D_t|} \sum_{d \in D_t} \theta_{d,k}, \tag{8}$$

where D_t denotes the set of documents published in year t, and $|D_t|$ is the number of such documents. The resulting time series $\{\bar{\theta}_k(t)\}$ summarizes how the prominence of topic k evolves over time.

3.4.1. Hot Topics

A topic k is classified as *hot* if its annual mean proportion exhibits a statistically significant increasing trend over time. Specifically, for each topic, a linear regression model is fitted with $\bar{\theta}_k(t)$ as the response variable and time t as the predictor. A topic is labeled hot if the estimated slope coefficient is positive and statistically significant:

$$\frac{d}{dt}\bar{\theta}_k(t) > 0. \tag{9}$$

Hot topics therefore represent research themes that are gaining prominence in the forensic science literature.

3.4.2. Cold Topics

Conversely, a topic k is classified as *cold* if its annual mean proportion shows a statistically significant decreasing trend over time. Using the same regression framework, a topic is labeled cold if the estimated slope coefficient is negative and statistically significant:

$$\frac{d}{dt}\bar{\theta}_k(t) < 0. \tag{10}$$

Cold topics correspond to areas whose relative emphasis in the literature has diminished over the study period.

The hot–cold topic classification provides a quantitative summary of shifting research priorities within forensic science. Hot topics highlight emerging or expanding areas of inquiry, while cold topics indicate themes that have received decreasing attention over time. Together, these trends offer insight into the dynamic structure of the field and complement the static topic diagnosticity analysis by revealing how disciplinary focus evolves longitudinally.

3.5. Post-Processing: Classification Algorithms

Following the estimation of document–topic proportions via LDA, the resulting document–topic matrix is denoted by $\boldsymbol{\Theta} \in \mathbb{R}^{D \times T}$, where each row $\boldsymbol{\theta}_d$ represents the topic distribution for document d, and T denotes the number of topics. These topic-based representations provide a low-dimensional and interpretable feature space that is well suited for downstream supervised learning tasks.

To assess the predictive utility of the inferred topics, a suite of supervised classification models is employed within a structured modeling and evaluation workflow. This workflow follows a systematic data mining discipline commonly used in applied machine learning, encompassing feature construction, model training, cross-validated tuning, class imbalance handling, and performance evaluation. Within this framework, topic proportions serve as inputs to multiple classifiers, enabling a comparative assessment of how effectively topic-based representations support document categorization.

This subsection outlines the classification setup, including the selected supervised models and dataset configurations, and situates them within the broader evaluation pipeline. Formal definitions of evaluation metrics and imbalance-handling strategies are provided in Section 3.5.3.

3.5.1. Classification Models

Using the document–topic matrix $\boldsymbol{\Theta}$ as input features, a diverse set of supervised learning algorithms is evaluated to assess the effectiveness of topic-based representations for document classification. The selected models span multiple learning paradigms, including instance-based, linear, tree-based, boosting-based, and neural approaches. This methodological diversity enables a robust comparison of classification performance, mitigates algorithm-specific bias, and aligns with established best practices in applied text classification and data mining workflows [28]. All models are trained, tuned, and evaluated using consistent data partitions, cross-validation procedures, and evaluation metrics to ensure fair comparison.

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm that assigns a class label to a document based on the majority class among its k nearest neighbors in the topic space [29]. Given a document represented by its topic proportion vector $\boldsymbol{\theta}$, proximity is measured using the squared Euclidean distance:

$$d(\boldsymbol{\theta}, \boldsymbol{\theta}_d) = \|\boldsymbol{\theta} - \boldsymbol{\theta}_d\|_2^2 = \sum_{j=1}^{T} (\theta_j - \theta_{d,j})^2. \tag{11}$$

The predicted class label is obtained via majority voting:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \sum_{\boldsymbol{\theta}_d \in N(\boldsymbol{\theta})} \mathbb{I}(y_d = y), \tag{12}$$

where $N(\boldsymbol{\theta})$ denotes the set of k nearest neighbors. KNN serves as a nonparametric baseline and is particularly useful for assessing local structure and neighborhood consistency in the topic space.

Multinomial Logistic Regression (MLR)

Multinomial Logistic Regression (MLR) extends binary logistic regression to multiclass classification using a softmax link function [29]. For a document d, the probability of belonging to class c is modeled as:

$$P(y_d = c \mid \boldsymbol{\theta}_d) = \frac{\exp(\mu_{dc})}{\sum_{r=1}^{C} \exp(\mu_{dr})}, \quad \mu_{dc} = \gamma_c + \sum_{j=1}^{T} \beta_{jc} \theta_{d,j}, \tag{13}$$

where $\boldsymbol{\beta}_c \in \mathbb{R}^T$ and $\gamma_c \in \mathbb{R}$ are class-specific coefficients and intercepts. Owing to its linear structure and interpretability, MLR provides a transparent benchmark for evaluating the discriminative power of topic proportions and is widely used in topic-based text classification and document regression tasks [17,30,31].

Random Forest (RF)

Random Forests, originally proposed by Breiman [32], construct an ensemble of decision trees via bootstrap aggregation and random feature selection, and aggregate their predictions to improve generalization performance. By averaging across trees, RF reduces variance and enhances robustness, making it well suited for high-dimensional and potentially correlated feature spaces such as document–topic matrices. RF has demonstrated strong and stable performance across a wide range of text classification applications [28,31].

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a scalable gradient boosting framework that sequentially fits decision trees to correct the residuals of previous models [33]. The prediction for document d is given by:

$$\hat{y}_d = F_K(\boldsymbol{\theta}_d) = f_0(\boldsymbol{\theta}_d) + \sum_{k=1}^{K} f_k(\boldsymbol{\theta}_d), \tag{14}$$

where each $f_k \in \mathcal{F}$ denotes a regression tree. XGBoost incorporates regularization, shrinkage, and column subsampling, enabling robust performance under class imbalance and complex decision boundaries. Its effectiveness for text classification tasks has been demonstrated in multiple applied studies [34].

Light Gradient Boosting Machine (LightGBM)

LightGBM is a gradient boosting framework that improves computational efficiency by employing histogram-based learning and leaf-wise tree growth [35]. Compared to traditional boosting methods, LightGBM scales efficiently to large datasets while maintaining strong predictive accuracy. Its suitability for high-dimensional text representations and imbalanced classification scenarios has been demonstrated in document classification studies [36].

Multilayer Perceptron (MLP)

To capture potential nonlinear relationships among topic proportions, a Multilayer Perceptron (MLP) classifier is also considered. The MLP models complex interactions through stacked fully connected layers with nonlinear activation functions [37]. Neural classifiers have gained increasing attention in topic-based and embedding-based text modeling due to their flexibility and representational capacity [20].

3.5.2. Handling Class Imbalance via SMOTE

Class imbalance is a common and critical challenge in scientific literature classification, particularly in domain-specific corpora such as forensic science abstracts, where some categories are substantially underrepresented. Skewed class distributions can bias supervised learning algorithms toward majority classes, leading to inflated accuracy estimates while masking poor performance on minority categories.

Within the proposed CRISP-DM-style framework, the Cross-Industry Standard Process for Data Mining (CRISP-DM) provides a structured, iterative workflow for data-driven analysis, encompassing data preparation, modeling, evaluation, and interpretation. Here, the term "CRISP-DM-style" is used to indicate a disciplined and modular organization of the supervised learning pipeline rather than adherence to a formal industrial deployment process. classification pipeline, class imbalance handling is applied only to the training set of each fold during cross-validation. To mitigate imbalance-related bias, we employ the Synthetic Minority Over-sampling Technique (SMOTE) [38], which generates synthetic training samples for minority classes by interpolating between nearest-neighbor observations in

feature space. Recent large-scale benchmarking studies have shown that SMOTE-based oversampling substantially improves macro-F1 and balanced accuracy in text classification tasks, particularly when classifiers are sensitive to skewed label distributions [39]. These findings underscore the importance of systematic imbalance handling rather than reliance on raw accuracy alone.

In this study, consistent with best practices in text classification, we examine the impact of imbalance handling under three distinct resampling scenarios:

- Scenario 1: No Resampling. The original imbalanced dataset is used without modification, serving as a baseline for comparison.
- Scenario 2: Global SMOTE. All minority classes are oversampled to match the size of the majority class, yielding a fully balanced training set.
- Scenario 3: Class-wise SMOTE. Oversampling is applied selectively on a per-class basis, allowing minority classes to be increased to predefined thresholds while reducing the risk of overfitting and synthetic redundancy.

This structured design allows the effects of imbalance handling to be isolated and evaluated systematically across classification models and dataset configurations in the subsequent results section.

3.5.3. Classification Metrics

To provide a comprehensive and unbiased evaluation of classification performance—particularly under class imbalance—multiple complementary metrics are reported, including Accuracy, Precision, Recall, and Macro-averaged F1-score. Prior studies in text classification have emphasized that reliance on accuracy alone can be misleading in imbalanced settings, as it may obscure poor performance on minority classes [28,31,39]. Accordingly, all models are evaluated using the same set of metrics across resampling scenarios to ensure fair, interpretable, and reproducible comparison.

Accuracy

Accuracy measures the overall proportion of correctly classified documents:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{15}$$

While accuracy provides a global summary of classifier performance, it does not reflect class-specific errors and may overestimate performance when majority classes dominate the dataset [40].

Precision

Precision quantifies the reliability of positive predictions by measuring the proportion of correctly predicted positive instances:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{16}$$

High precision indicates a low false-positive rate and is particularly important when incorrect assignment of documents to scientific categories may lead to misleading downstream interpretation or analysis.

Recall

Recall (also referred to as sensitivity) measures a classifier's ability to correctly identify all relevant instances:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{17}$$

In imbalanced scientific corpora, recall is essential for assessing whether underrepresented categories are adequately captured by the model [41].

Macro-Averaged F1-Score

The F1-score represents the harmonic mean of precision and recall:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{18}$$

We report the macro-averaged F1-score, which computes the F1-score independently for each class and then averages across classes, assigning equal weight to all categories regardless of class size. Macro-F1 is widely recommended for imbalanced text classification tasks, as it provides a balanced assessment of performance across both majority and minority classes [39,41].

3.6. Materials and Methods

All analyses were conducted using a combination of `R` (version 4.3.2) and `Python` (version 3.10) statistical computing environments. Text preprocessing, Latent Dirichlet Allocation (LDA) modeling, topic diagnosticity analysis, and temporal trend analysis were primarily implemented in `R` using established packages for text mining and probabilistic modeling. Supervised classification, resampling strategies (including SMOTE), cross-validation, and performance evaluation were implemented in `Python` using widely adopted machine-learning libraries. This hybrid implementation leverages the strengths of both ecosystems, enabling reproducible data processing, interpretable topic modeling, and scalable supervised learning within a unified analytical workflow.

3.7. Dataset Description

This study analyzes a curated corpus of 3689 abstracts published in the Journal of Forensic Sciences between 2009 and 2022. The dataset was constructed to support the methodological framework described in Section 3 and was selected due to the journal's consistent use of discipline-specific categorical labels and its broad coverage of forensic science subfields. In addition to abstracts, the dataset includes associated metadata such as article titles, author information, publication year, keywords, and journal-assigned categories.

The journal-defined categories span a wide range of forensic science domains, including Anthropology, Odontology, Questioned Documents, Jurisprudence, Digital and Multimedia Sciences, Toxicology, Engineering and Applied Sciences, Pathology/Biology, Criminalistics, and Psychiatry and Behavioral Science. These categorical labels provide an external reference framework for evaluating the coherence, diagnosticity, and downstream classification utility of topics inferred via Latent Dirichlet Allocation (LDA).

Figure 6 shows pronounced imbalance in the distribution of abstracts across forensic science categories. A small number of disciplines account for a large proportion of the corpus, while several categories are sparsely represented. This imbalance directly motivates the class rebalancing strategies, dataset configurations, and evaluation metrics introduced in the subsequent supervised classification and SMOTE analyses.

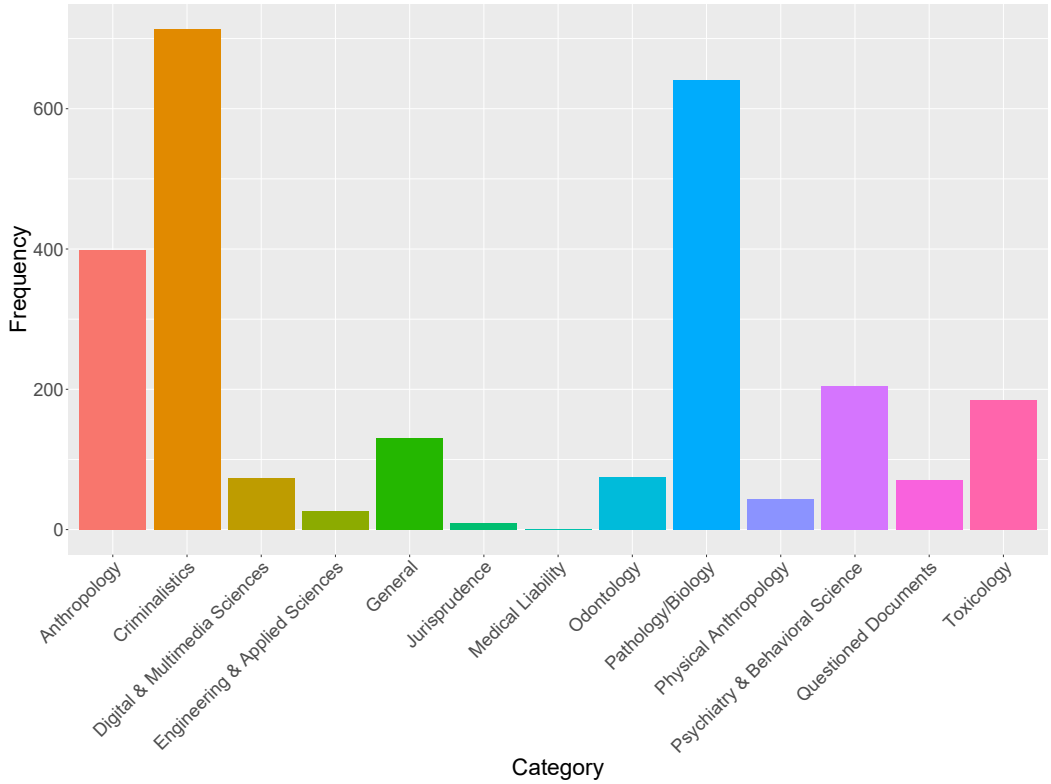

Figure 6. Distribution of abstracts across forensic science categories (2009–2022). The bar chart illustrates substantial variation in category frequencies, with Criminalistics, Pathology/Biology, and Anthropology exhibiting the highest representation.

4. Results

This section presents the empirical findings obtained by applying the methodological framework described in Section 3 to the forensic science abstract corpus. The results are organized to mirror the analytical workflow and progress from unsupervised topic modeling to supervised classification.

First, likelihood-based model selection is used to determine an appropriate number of topics for the Latent Dirichlet Allocation (LDA) model. The resulting topic structure is then examined for semantic coherence and interpretability, with many topics exhibiting clear alignment with established forensic science domains.

Next, temporal analyses are conducted to assess changes in topic prevalence over time. These analyses highlight shifts in research emphasis and reveal emerging and declining themes within the forensic science literature.

Finally, the document–topic matrix derived from the selected LDA model is used as input for supervised classification. Multiple machine-learning models are evaluated under different resampling and validation settings, demonstrating the effectiveness and robustness of topic-based representations for abstract classification.

4.1. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is used to uncover the latent thematic structure of forensic science abstracts. The analysis aims to identify a topic configuration that balances statistical fit with interpretability. The inferred topics are examined for semantic coherence and alignment with established forensic science disciplines. In addition, temporal analyses are conducted to assess how topic prevalence evolves over time, providing insight into longitudinal research trends within the field.

4.1.1. Model Selection

Selecting an appropriate number of topics is a critical step in LDA modeling, as it directly influences topic interpretability and the stability of downstream analyses. Model selection is carried out using the Bayesian marginal likelihood criterion proposed by Griffiths and Steyvers [2], which evaluates the quantity $P(W \mid T)$ by integrating over model parameters and selecting the topic number T that maximizes this likelihood.

To account for variability arising from stochastic initialization and sampling in LDA estimation, ten independent resampling runs are conducted on the forensic abstract corpus. In each run, the initial topic count is set to $T = 300$, and the model is estimated using 1000 iterations of collapsed Gibbs sampling. Hyperparameters are held constant across runs, with $\beta = 0.1$ to encourage sparsity in topic–word distributions and $\alpha = 50/T$ to promote balanced topic proportions across documents.

Across the ten resampling runs, Bayesian model selection yields the following optimal topic counts: $\{73, 84, 76, 96, 83, 86, 83, 88, 81, 100\}$. Among these values, $T = 83$ is selected most frequently and exhibits stable log-likelihood behavior across runs. Based on both selection frequency and likelihood stability, $T = 83$ is adopted as the primary topic configuration for subsequent analyses.

To examine the sensitivity of downstream results to topic granularity, all topic counts identified during Bayesian model selection are retained for further analysis, namely $T \in \{73, 76, 81, 83, 84, 86, 88, 96, 100\}$. This set reflects empirically supported topic configurations rather than arbitrarily chosen alternatives. These topic models are subsequently evaluated in supervised classification experiments (see Section 4.2) to assess the robustness of topic-based representations under moderate variation in T.

Figure 7 shows the log-likelihood values obtained across candidate topic counts during Bayesian model selection. A clear maximum is observed at $T = 83$, indicating an optimal trade-off between topic granularity and model fit. Based on this result, the final LDA configuration is fixed at $T = 83$, with hyperparameters set to $\beta = 0.1$ and $\alpha = \frac{50}{T}$.

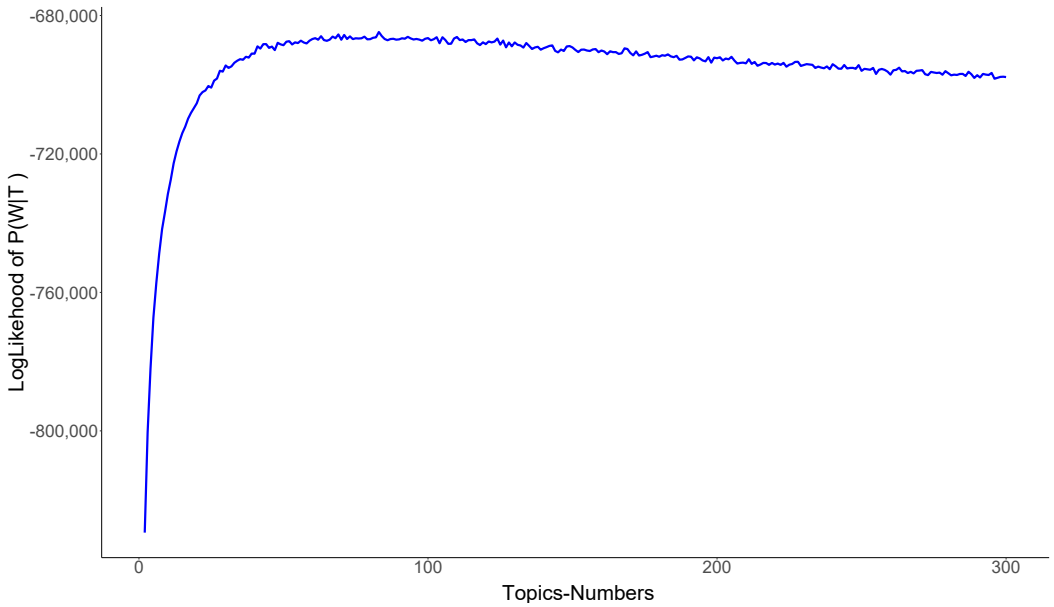

Figure 7. Model selection for LDA. Log-likelihood values across topic counts (T), showing a clear maximum at $T = 83$, which is selected as the primary topic configuration.

Figure 8 presents the most probable terms associated with the first six topics inferred by the selected LDA model. Each panel summarizes the dominant vocabulary defining a topic, enabling qualitative assessment of topic coherence and interpretability. The identified topics align with well-established forensic research areas, including analytical chemistry, forensic genetics, death investigation, behavioral and psychiatric science, spectroscopy, and fingerprint analysis. This correspondence with recognized forensic domains indicates that the LDA model captures meaningful and domain-relevant semantic structure within the corpus.

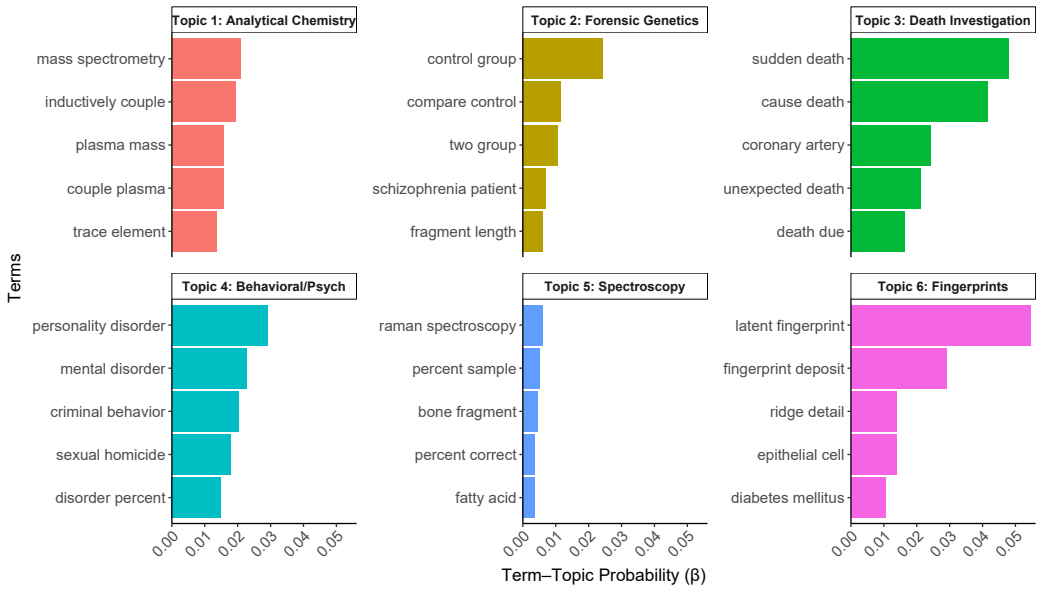

Figure 8. Top-ranked terms for the first six topics inferred by the selected LDA model. Bar lengths represent term–topic probabilities β, illustrating coherent and interpretable thematic structure across major forensic science domains.

4.1.2. Topic Diagnosticity and Forensic Categories

To assess the disciplinary specificity of the inferred topics, we conducted a diagnosticity analysis that combines topic–term representations with the category-specific topic distribution matrix introduced in Section 3.3. Focusing on abstracts published in 2020, this analysis identifies topics that are not only prevalent, but also disproportionately associated with particular forensic science disciplines. By examining topic–category relationships, the analysis evaluates the extent to which individual topics function as distinguishing thematic markers across forensic domains.

Figure 9 displays the diagnostic topic–category matrix for 2020. Distinct concentration patterns are evident, indicating that several topics exhibit strong alignment with specific forensic categories rather than being uniformly distributed across the corpus. These patterns support the interpretability of the topic model and demonstrate its ability to capture discipline-specific thematic structure.

To facilitate semantic interpretation of the diagnostic topics, Figure 10 presents the top-ranked terms associated with a subset of representative topics. These term distributions provide contextual insight into each topic's thematic focus and disciplinary relevance. Together, Figures 9 and 10 offer a complementary perspective on topic diagnosticity, illustrating both cross-disciplinary overlap and domain-specific specialization.

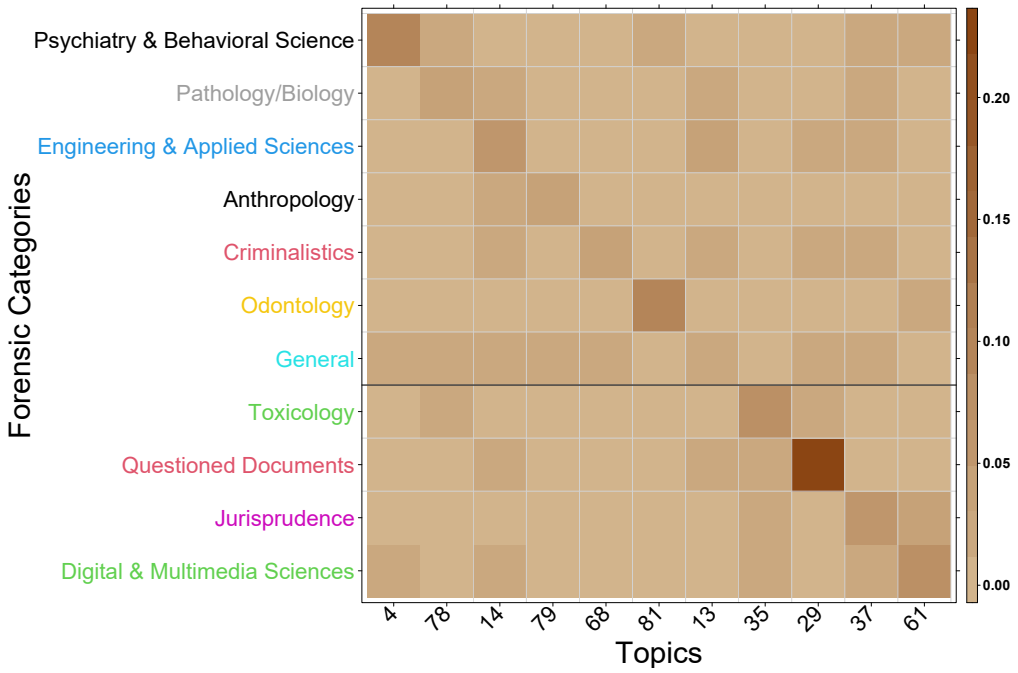

Figure 9. Matrix of topic distributions across forensic categories for 2020 abstracts. Each cell represents the average topic probability within a category, with darker shading indicating stronger diagnostic association. The black horizontal line is used for visual separation only and does not encode additional analytical meaning. Concentrated patterns highlight topics that are characteristic of specific forensic disciplines.

Several illustrative examples highlight the diagnostic relevance of individual topics and their alignment with forensic science disciplines:

- Topic 4 (Psychiatric and Behavioral Analysis) is predominantly associated with *Psychiatry & Behavioral Science*, characterized by terms such as "personality disorder" and "mental disorder," which reflect its focus on psychological assessment and behavioral evaluation.
- Topic 13 (Forensic DNA Analysis Techniques) shows strong associations with *Engineering & Applied Sciences*, *Pathology/Biology*, and *Criminalistics*. Terms such as "sexual assault," "crime laboratory," and "DNA profile" highlight its broad applicability across forensic biology and laboratory practice.
- Topic 14 (Advanced Imaging Techniques) is diagnostic of *Engineering & Applied Sciences*, with prominent terms including "computed tomography," "postmortem scan," and "laser scan," reflecting its use in high-resolution forensic imaging and injury documentation.
- Topic 29 (Questioned Document Analysis) is strongly aligned with the *Questioned Documents* category, featuring domain-specific terminology such as "stamp pad" and "document examination."
- Topic 35 (Biological Fluid Analysis) exhibits high diagnostic value in both *Toxicology* and *Pathology/Biology*, underscoring its relevance to substance detection and postmortem investigations.
- Topic 37 (Crime Scene Analysis) is closely associated with *Jurisprudence* and *Criminalistics*, emphasizing investigative procedures through terms such as "crime scene investigation" and "scene investigator."
- Topic 61 (DNA Analysis) shows prominence in *Digital & Multimedia Sciences* and *Jurisprudence*, reflecting the interdisciplinary role of DNA evidence in legal and digital forensic contexts.

- Topic 68 (DNA Profiling Techniques) is central to *Criminalistics*, marked by repeated terms such as "DNA profile" and "DNA analysis," which are fundamental to criminal identification.
- Topic 78 (Medicolegal Death Investigation) is diagnostic of both *Pathology/Biology* and *Toxicology*, emphasizing medico-legal evaluation through terms such as "cause of death" and "medical examiner."
- Topic 79 (Anthropological Analysis) is highly specific to *Anthropology*, with terms such as "sex estimation" and "skeletal remains" reflecting its role in human remains identification.
- Topic 81 (Dental Forensics) is clearly diagnostic of *Odontology*, indicated by terms such as "age estimation" and "dental age," which are central to forensic dental profiling.

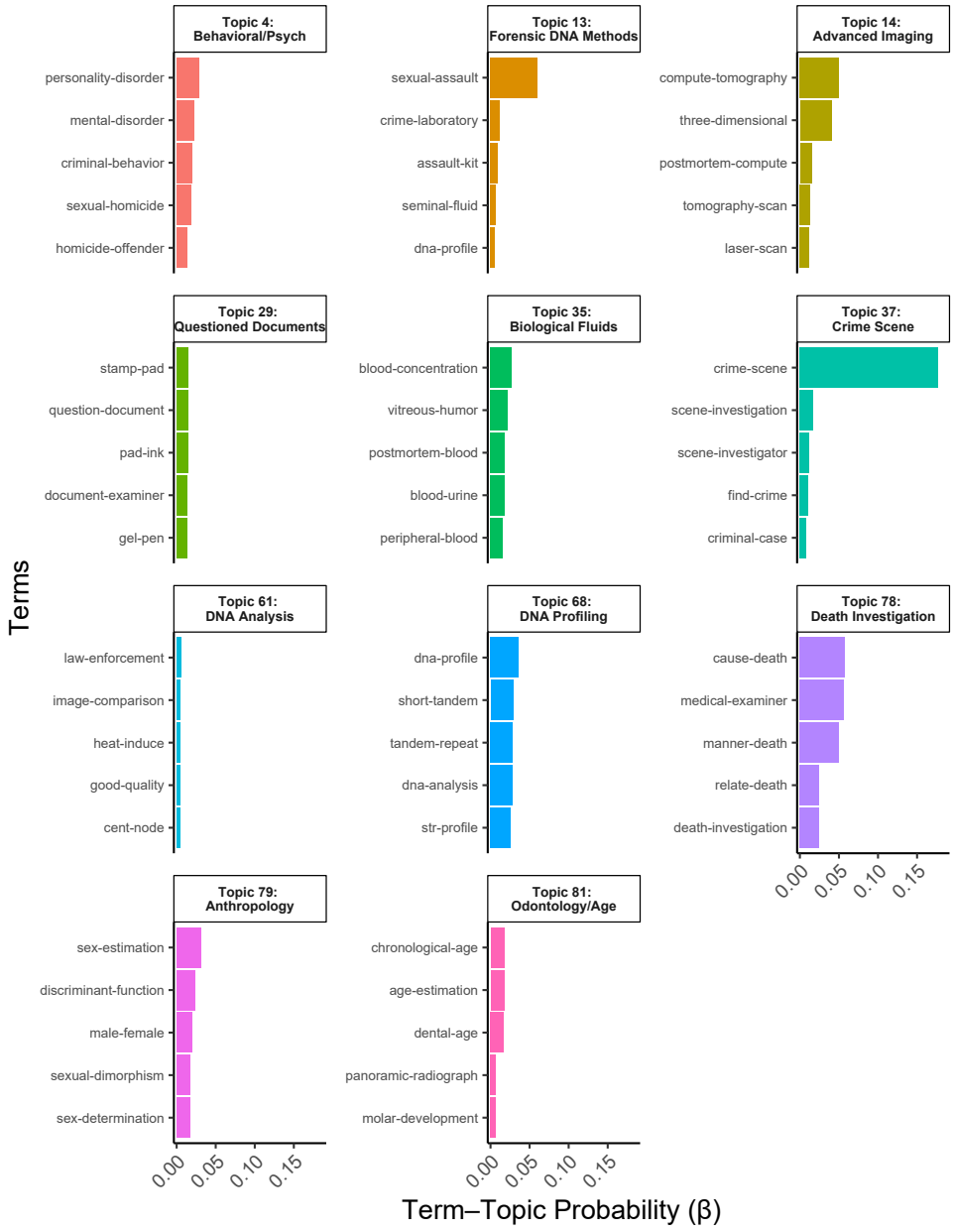

Figure 10. Top-ranked term distributions for selected diagnostic topics inferred from 2020 abstracts. Bar lengths correspond to term–topic probabilities (β), illustrating the semantic content underlying topics with strong category-specific associations.

4.1.3. Hot and Cold Topic Trends in Forensic Science

To examine longitudinal shifts in forensic science research, topic dynamics were analyzed using the estimated document–topic matrix $\boldsymbol{\Theta}$, following the procedure described in Section 3.4. For each topic, annual mean topic proportions ($\bar{\theta}_t$) were computed over the period 2009–2020, providing a normalized measure of topic prevalence through time.

Topics exhibiting sustained increases in mean proportion were classified as *hot* topics, while those showing consistent declines were classified as *cold* topics. This trend-based characterization emphasizes long-term directional changes in research emphasis rather than short-term variability. When interpreted in conjunction with topic–term distributions, the analysis provides insight into evolving thematic priorities within the forensic science literature.

Figure 11 illustrates representative temporal trajectories for selected hot and cold topics. Among the declining themes, Topics 7, 12, 36, and 80 display gradual but consistent reductions in prevalence, indicating a relative decrease in research focus over the study period. In contrast, Topics 13, 26, and 37 demonstrate increasing prominence, suggesting growing scholarly attention and continued methodological development.

To support semantic interpretation of these trends, Figure 12 presents the top five representative terms for each selected hot and cold topic.

Hot Topics

Several topics exhibit clear upward trends, reflecting emerging or expanding areas of forensic research. Topic 13 (*Sexual Assault Forensic Analysis*) shows increasing prevalence and is characterized by terms such as "DNA profiling," "assault kit," and "crime laboratory," indicating sustained methodological and institutional attention to sexual assault casework. Topic 26 (*Digital Forensics and Data Analysis*) captures the growing role of digital evidence and computational analysis in forensic investigations. Topic 37 (*Crime Scene Investigation Techniques*) reflects continued innovation in evidence collection and scene processing, underscoring the field's increasing emphasis on technical rigor and standardization.

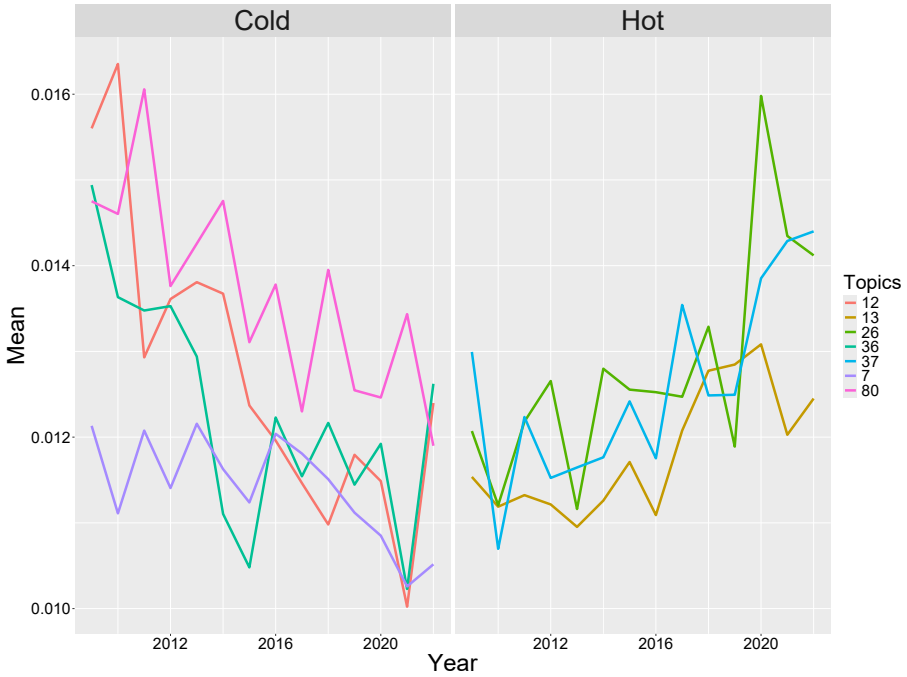

Figure 11. Temporal trends of selected hot and cold topics in forensic science (2009–2020). Each curve represents the annual mean topic proportion ($\bar{\theta}_t$). Topics classified as hot exhibit increasing trends, while cold topics show sustained declines, reflecting shifts in thematic emphasis over time.

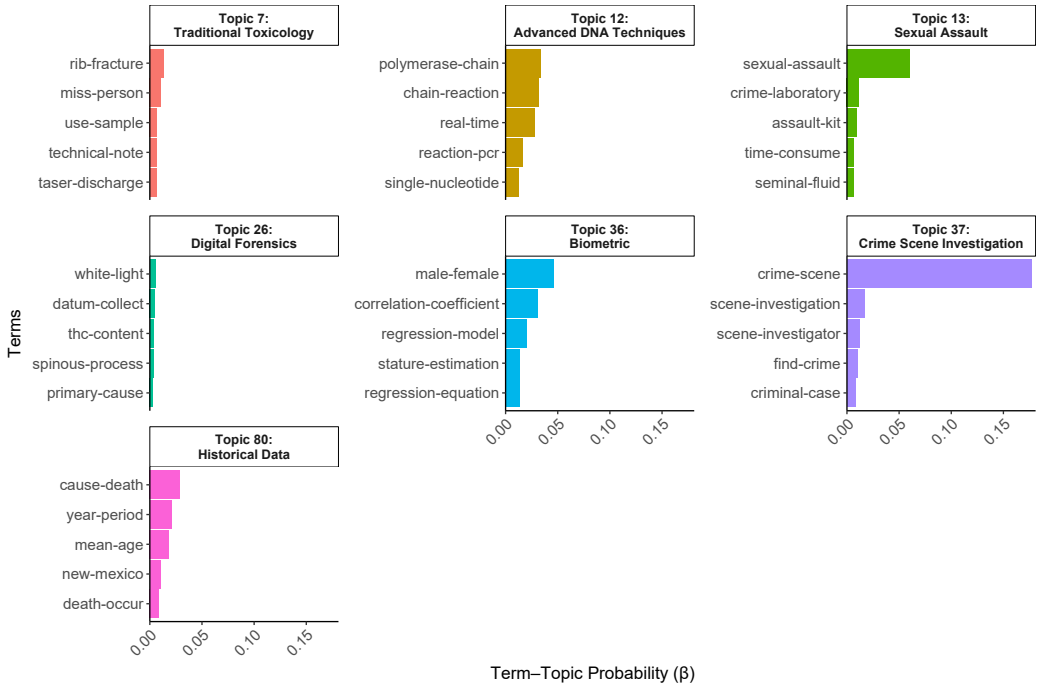

Figure 12. Top-ranked terms for selected hot and cold topics identified through temporal trend analysis. Bar lengths represent term–topic probabilities (β), providing semantic context for topics exhibiting increasing or decreasing prevalence (see Figure 11).

Cold Topics

In contrast, several topics demonstrate declining prevalence over time. Topic 7 (*Traditional Toxicology and Medical Examination*) includes terms such as "rib fracture" and "alcohol intoxication," suggesting a shift away from traditional case descriptions toward more specialized or technology-driven approaches. Topic 12 (*Advanced DNA Analysis Techniques*), historically centered on methods such as PCR, shows a downward trend as these techniques have become routine and less frequently emphasized as standalone research contributions. Topic 36 (*Biometric and Physical Anthropology*) reflects declining emphasis on classical anthropometric approaches, while Topic 80 (*Historical Data and Statistical Reviews*) indicates a gradual move away from retrospective descriptive analyses toward more contemporary, data-driven modeling frameworks.

Overall, the hot–cold topic analysis highlights the evolving research landscape in forensic science. Increasing trends point to areas of active methodological development and technological integration, while declining trends reflect the maturation or consolidation of established research themes. Together, these findings provide a longitudinal perspective on how forensic science priorities have shifted over the past decade.

4.2. Abstract Classification

Building on the unsupervised topic modeling results, this section evaluates the predictive utility of the LDA-derived topic representations for supervised abstract classification. The document–topic matrix $\boldsymbol{\Theta}$ is used as a compact and interpretable feature space to assess whether the inferred topics capture discriminative structure aligned with journal-defined forensic science categories.

To obtain stable and reliable performance estimates, hyperparameter tuning and model selection are conducted using 5-fold cross-validation. For each fold of the cross validation, all classification models are trained using an 80:20 train–test split.

In each fold, class imbalance handling and model fitting are applied exclusively to the training data to prevent information leakage and ensure unbiased evaluation on the test set.

Detailed descriptions of the classification algorithms, resampling strategies, and evaluation metrics are provided in Section 3.5.

As illustrated in Figure 6, the journal-assigned forensic science categories exhibit substantial class imbalance, posing challenges for both predictive accuracy and fair assessment across minority classes. To address this issue during model training, the Synthetic Minority Over-sampling Technique (SMOTE) [38] is incorporated into the supervised learning pipeline. SMOTE is applied exclusively to the training data to improve minority class representation while preserving the integrity of the test set. The original class distributions prior to any resampling are reported in Appendix A for transparency.

In addition to imbalance correction, classification performance is evaluated under three complementary dataset configurations designed to assess robustness under varying levels of class sparsity and label granularity:

- **All Categories.** All 11 journal-defined forensic science categories are retained, preserving the original imbalanced label structure. This setting represents the most realistic and challenging classification scenario.
- **Dropped Categories.** Categories with fewer than 100 abstracts—*Digital and Multimedia Science*, *Engineering and Applied Sciences*, *Odontology*, *Questioned Document*, and *Physical Anthropology*—are excluded to reduce extreme sparsity and improve training stability. The resulting distribution is shown in Figure 13.
- **Grouped Categories.** Semantically related categories are merged to form a more compact and balanced label space, reflecting practical forensic classification logic. As illustrated in Figure 14, the grouping scheme is defined as follows:
  - **Criminalistics:** *General*, *Digital and Multimedia Science*, *Jurisprudence*, and *Medical Liability*.
  - **Anthropology:** *Anthropology*, *Odontology*, and *Physical Anthropology*.
  - **Pathology/Biology:** *Toxicology* and *Engineering and Applied Sciences*.
  - **Psychiatry:** *Psychiatry and Behavioral Science* and *Questioned Document*.

Together, these configurations enable a systematic evaluation of classification performance across increasing levels of class balance and label abstraction, providing insight into the stability, robustness, and generalizability of topic-based representations for forensic abstract classification.

Building on this evaluation framework and the performance metrics defined in Section 3.5.3, Figure 15 summarizes the macro-averaged F1 scores obtained from 5-fold cross-validated classification across the three dataset configurations—*All Categories*, *Dropped Categories*, and *Grouped Categories*— as a function of the number of topics T. Each curve corresponds to a classification model trained on LDA-derived document–topic representations, enabling direct comparison of classifier behavior under varying levels of class balance, label granularity, and topic resolution.

The topic counts shown in Figure 15 correspond to candidate values identified during Bayesian model selection (see Section 4.1.1) and are used here to examine how supervised classification performance varies with topic granularity. Evaluating Macro-F1 across this empirically supported range allows assessment of whether classification performance remains stable under moderate changes in T, thereby providing evidence for the robustness of topic-based representations in downstream supervised tasks.

In the *All Categories* configuration, Macro-F1 scores remain low across all topic counts, ranging approximately from 0.25 to 0.40. This pattern reflects the severe class imbalance and label sparsity present in the original dataset, which substantially limits minority-class recognition. Within this challenging setting, ensemble and neural models—most notably XGBoost, LightGBM, and MLP—consistently outperform the instance-based KNN classifier, although absolute performance gains remain constrained by the underlying imbalance.

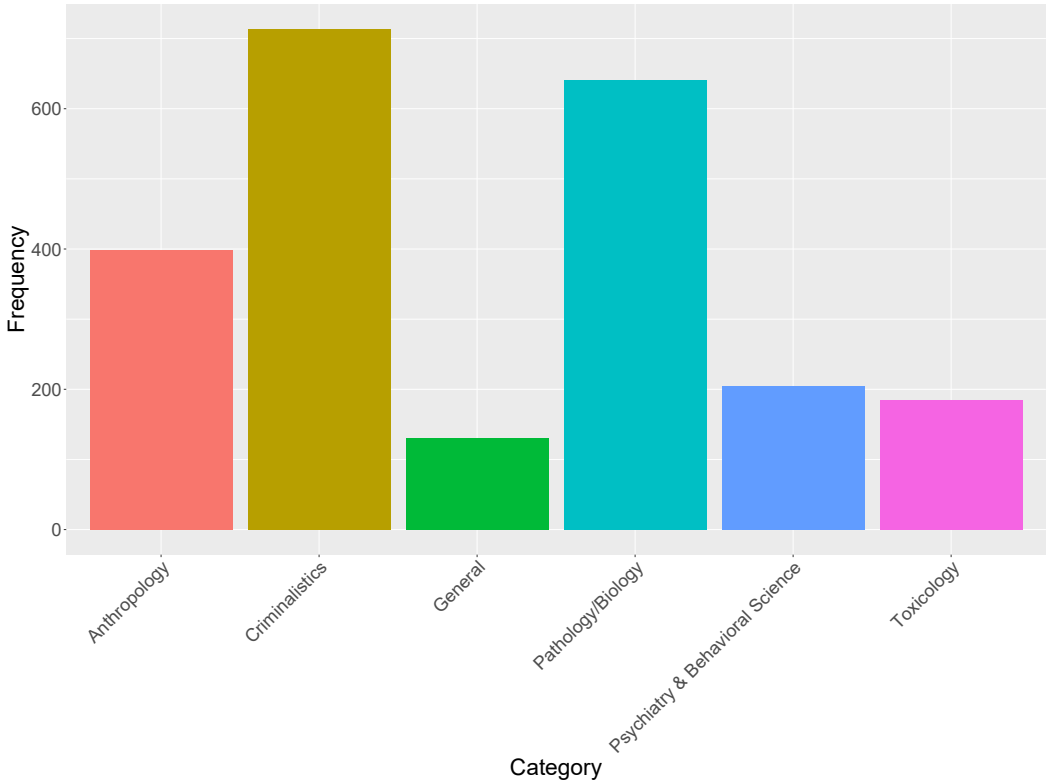

Figure 13. Distribution of forensic science categories after frequency filtering. This bar chart displays the frequencies of forensic science categories retained after excluding those with fewer than 100 abstracts. This filtering step reduces extreme class sparsity and improves the stability and fairness of supervised classification models while preserving the major forensic science domains.

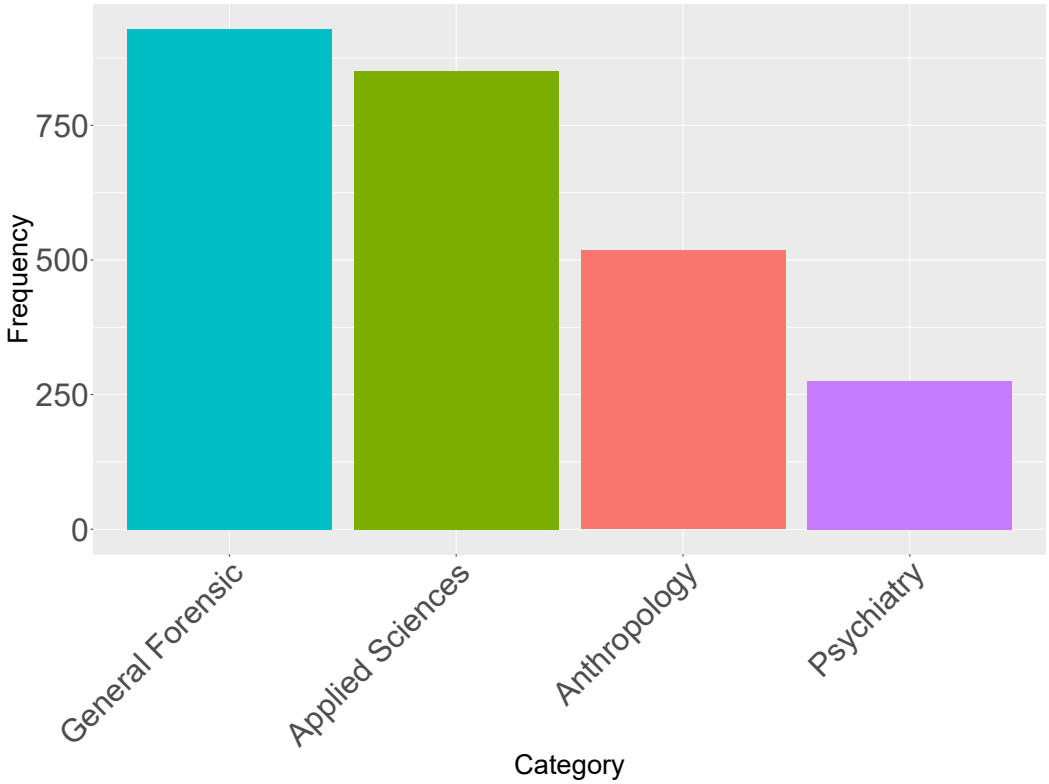

Figure 14. Grouped forensic science categories by thematic and logical similarity. This figure illustrates the frequency of grouped categories, simplifying the classification task by consolidating related disciplines based on conceptual proximity.

Removing extremely sparse categories in the *Dropped Categories* scenario results in a clear and uniform upward shift in Macro-F1 across all classifiers. Scores increase into the 0.45–0.60 range as topic granularity increases, indicating improved recognition of previously underrepresented classes. Performance peaks are observed consistently around $T = 86$ and $T = 88$, suggesting that moderate increases in topic resolution enhance discriminative capacity once extreme sparsity is reduced.

The strongest and most stable performance is achieved in the *Grouped Categories* configuration. In this setting, Macro-F1 values exceed 0.70 for several classifiers, most notably XGBoost, Logistic Regression, LightGBM, and MLP. Across these models, performance remains consistently high within the topic range $T = 86$–88, with only minor fluctuations across neighboring topic counts. The resulting Macro-F1 curves are smooth and closely aligned across classifiers, indicating that label abstraction effectively mitigates class sparsity while preserving the discriminative structure of the topic representations.

Importantly, the topic range yielding the strongest classification performance coincides with the values identified through Bayesian model selection. This alignment provides empirical support for the robustness of the selected LDA representations and confirms their suitability for downstream supervised classification tasks.

To complement the Macro-F1 trends shown in Figure 15, Table 1 reports detailed classification performance for the *Grouped Categories* configuration at two representative topic counts ($T = 86$ and $T = 88$). Results are summarized as mean (standard deviation) over 5-fold cross-validation and include Accuracy, Macro-F1, Precision, and Recall.

Consistent with the patterns observed in Figure 15, the results in Table 1 indicate that classification performance is highly stable between $T = 86$ and $T = 88$. Differences in mean performance across topic counts are small and accompanied by comparable standard deviations, suggesting that supervised outcomes are not sensitive to minor variations in topic granularity within this empirically supported range. This stability provides further evidence for the robustness of the LDA-derived topic representations identified during Bayesian model selection.

At $T = 88$, XGBoost achieves the strongest overall performance among the evaluated classifiers, attaining an Accuracy of 0.754 and a Macro-averaged F1 score of 0.737. This result reflects strong and balanced discrimination across forensic categories, with high Precision indicating relatively conservative decision boundaries and fewer false positives.

To provide additional insight into class-level prediction behavior, confusion matrices for the best-performing boosting models (XGBoost at $T = 88$) are reported in Appendix B. These matrices further illustrate the balanced error structure and confirm that performance gains are not driven by overfitting to a subset of dominant categories.

Taken together, Figure 15 and Table 1 demonstrate that (i) topic representations within the candidate range identified by Bayesian model selection yield stable and reliable classification performance, and (ii) multiple modeling families achieve competitive and consistent results. These findings confirm that the proposed topic-based pipeline provides a robust, interpretable, and largely model-agnostic feature representation for large-scale forensic abstract classification.

Table 1. Classification performance for the *grouped categories* scenario at the best-performing topic numbers ($T = 86$ and $T = 88$). Values are reported as mean (standard deviation) over 5-fold cross-validation. **Bold values** indicate the highest performance observed across all classifiers and topic settings.

Classifier	*T*	Accuracy	Macro-F1	Precision	Recall
LightGBM	86	0.748 (0.009)	0.730 (0.014)	0.743 (0.011)	0.721 (0.017)
LightGBM	88	0.746 (0.013)	0.728 (0.017)	0.743 (0.021)	0.718 (0.016)

Table 1. *Cont.*

Classifier	*T*	Accuracy	Macro-F1	Precision	Recall
Logistic Regression	86	0.751 (0.007)	0.734 (0.009)	0.727 (0.010)	**0.744 (0.010)**
Logistic Regression	88	0.748 (0.010)	0.731 (0.008)	0.725 (0.009)	0.741 (0.006)
MLP	86	0.730 (0.013)	0.710 (0.013)	0.709 (0.013)	0.712 (0.013)
MLP	88	0.734 (0.013)	0.713 (0.012)	0.711 (0.012)	0.716 (0.012)
XGBoost	86	0.745 (0.009)	0.727 (0.010)	0.737 (0.009)	0.720 (0.012)
XGBoost	**88**	**0.754 (0.005)**	**0.737 (0.011)**	**0.743** (0.006)	0.732 (0.015)

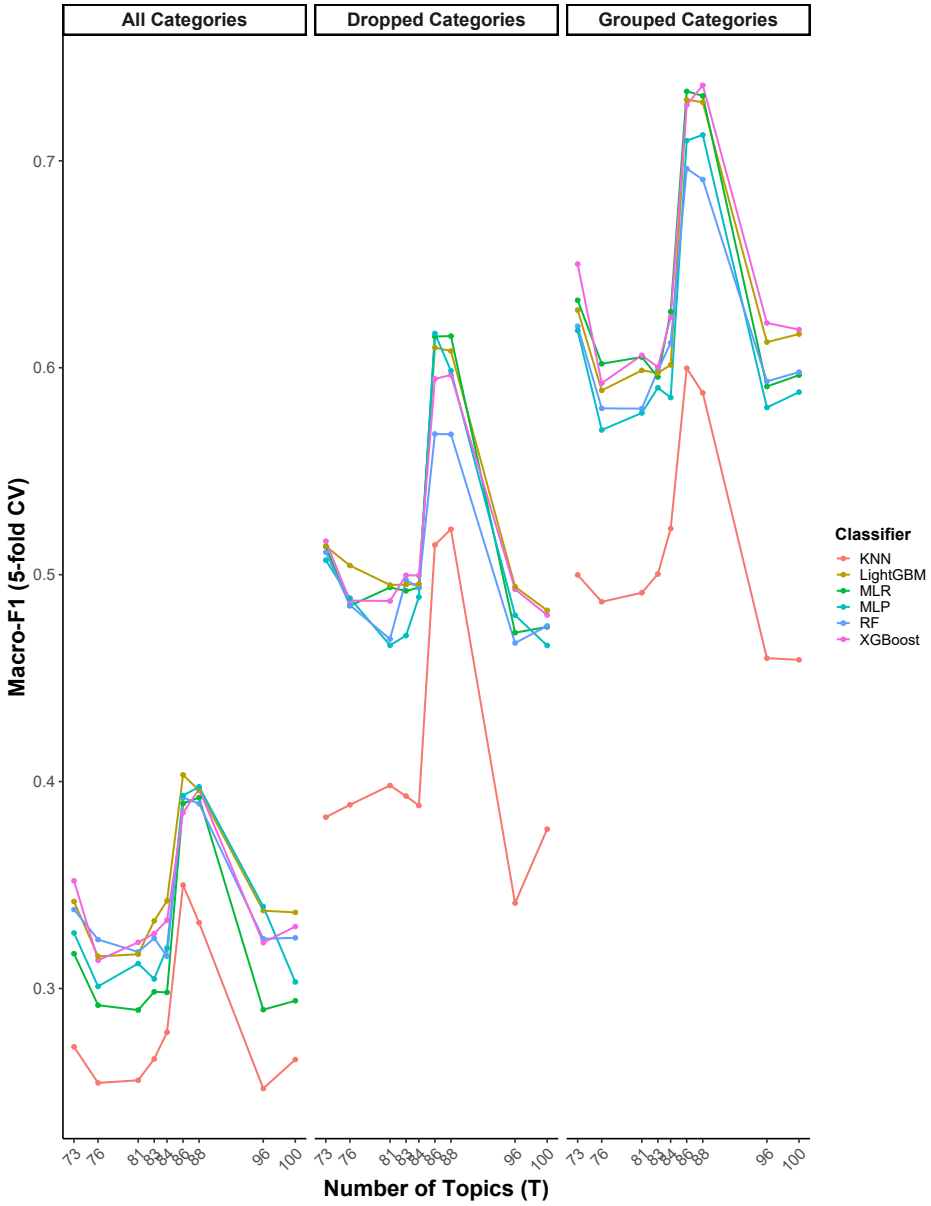

Figure 15. Macro-averaged F1-score (5-fold cross-validation) as a function of the number of topics T across three dataset configurations: *All Categories*, *Dropped Categories*, and *Grouped Categories*. Each curve corresponds to a classification model trained on LDA-derived document–topic representations. Across all scenarios, performance improves with increasing topic granularity and stabilizes within $T = 83$–88, indicating that this interval provides a robust balance between semantic resolution and discriminative capacity. The *Grouped Categories* configuration yields the highest and most stable Macro-F1 scores, reflecting the combined benefits of reduced label sparsity and increased class balance.

5. Discussion

This study presents an integrated and interpretable framework for large-scale scientific literature analysis that combines probabilistic topic modeling with supervised classification. By unifying topic discovery, diagnostic evaluation, temporal trend analysis, and downstream classification within a single analytical pipeline, the proposed approach addresses a key limitation of prior work, where unsupervised topic modeling and predictive modeling are often treated as separate or loosely connected tasks. The results demonstrate that topic-based representations can simultaneously support meaningful exploratory analysis and robust document categorization.

The following discussion interprets the empirical findings in relation to the study objectives, situates the proposed framework within the broader topic modeling and text mining literature, and addresses key methodological considerations. In particular, the discussion highlights the interpretability and stability of the topic representations, examines trade-offs relative to contemporary neural and large language model–based approaches, and explicitly outlines limitations and directions for future research.

5.1. Interpretability and Temporal Dynamics of Topics

The topic modeling results demonstrate that probabilistic topic modeling can recover semantically coherent and domain-relevant themes from forensic science abstracts. The inferred topics align closely with established forensic disciplines, including criminalistics, pathology, toxicology, anthropology, and digital forensics, supporting the interpretability of the learned topic structure. This alignment is further reinforced by the diagnosticity analysis, which shows that several topics are strongly associated with specific journal-defined categories rather than being diffusely distributed across the corpus. Together, these findings indicate that the LDA model captures meaningful latent structure that reflects disciplinary organization within forensic science literature.

Beyond static topic interpretation, the temporal analysis provides additional insight into how forensic science research evolves over time. Topics classified as *hot* exhibit sustained growth in prevalence and correspond to areas of increasing scholarly attention, such as digital forensics, advanced crime scene investigation techniques, and sexual assault analysis. These trends are consistent with broader technological and societal developments, including the expanding role of digital evidence and advances in forensic analytical methods.

In contrast, *cold* topics tend to represent research areas that have reached a stage of methodological maturity, including traditional toxicological analyses and established DNA techniques. Declines in topic prevalence should therefore be interpreted as a transition from active research frontiers to routine forensic practice, rather than a loss of relevance. Taken together, the combined diagnostic and temporal analyses demonstrate that the inferred topics are not only interpretable and discipline-specific, but also capable of capturing longitudinal shifts in research emphasis within the forensic science literature.

5.2. Topic Granularity and Model Robustness

Selecting the number of topics is a longstanding challenge in topic modeling, as it directly affects interpretability, stability, and downstream analytical performance. To address this issue, the proposed framework adopts a Bayesian model selection strategy combined with repeated resampling, allowing stochastic variability in LDA estimation to be explicitly accounted for. Across resampling runs, $T = 83$ emerged as the most frequently selected configuration and exhibited stable likelihood behavior, providing empirical support for this topic resolution.

Rather than viewing this value as a single fixed optimum, the analysis further examines the sensitivity of results to topic granularity by considering a range of empirically supported topic counts identified during resampling. The observed stability of both topic structure and classification performance across neighboring values ($T = 83$–88) indicates that the learned representations are robust to moderate changes in topic resolution. This behavior is particularly important for reproducibility and practical deployment, as it suggests that small variations in model configuration do not substantially alter the substantive conclusions drawn from the analysis.

5.3. Supervised Classification and Topic-Based Representations

The supervised classification results demonstrate that LDA-derived topic proportions constitute an effective and interpretable feature representation for forensic abstract classification. Across all experimental settings, performance improves as class imbalance is reduced and label sparsity is mitigated, with the strongest and most stable results consistently observed in the *Grouped Categories* configuration. In this setting, Macro-averaged F1 scores exceed 0.70 for several classifiers, including Logistic Regression, XGBoost, LightGBM, and MLP, indicating reliable discrimination across forensic science categories.

Within this configuration, performance differences among classifiers are modest and reflect complementary strengths rather than clear dominance by a single model. Logistic Regression achieves strong Macro-F1 and Recall at moderate topic resolutions, suggesting a well-balanced ability to recover instances across categories. This outcome indicates that the topic-based feature space captures substantial linearly separable structure aligned with disciplinary boundaries, reinforcing the interpretability of the learned representations. Because topic proportions are directly interpretable, the linear coefficients of this model offer transparent insight into how specific topics contribute to category distinctions.

Boosting-based models such as XGBoost and LightGBM achieve comparably strong performance and tend to yield higher Precision, reflecting more conservative decision boundaries with fewer false positives. The MLP classifier performs slightly below the linear and boosting-based approaches but remains stable across topic counts, suggesting that nonlinear structure in the topic-feature space can be learned without compromising robustness.

Importantly, classification performance remains stable across topic counts within the empirically supported range identified during Bayesian model selection. This consistency indicates that the effectiveness of topic-based representations does not rely on a single, finely tuned topic configuration. Instead, the results show that the proposed framework produces robust and generalizable features that support downstream supervised learning across a range of reasonable modeling choices.

5.4. Methodological Implications and Relation to Neural Network and LLM-Based Approaches

Recent advances in text mining have been driven by transformer-based models and large language models (LLMs), including BERT and domain-specific variants such as SciBERT, as well as neural topic modeling approaches that integrate contextual embeddings [21,23,42–44]. These methods have demonstrated strong performance in scientific text classification and topic generation, particularly in capturing contextual semantics and producing fluent topic descriptions. At the same time, recent empirical studies have highlighted practical challenges associated with these approaches, including limited transparency, substantial computational cost, sensitivity to prompting or model configuration, reliance on closed-source systems, and constraints imposed by fixed context-length limits [23,45].

Parallel to neural network and LLM-based developments, recent work has also focused on improving the interpretability and coherence of classical topic models through structural,

semantic, and hierarchical extensions. Examples include structural–semantic term-weighting strategies designed to reduce token overlap and enhance topic coherence [46], as well as multi-view and hierarchical topic modeling frameworks that emphasize robustness and interpretability [47]. While these approaches represent important advances in topic model design, they are primarily oriented toward exploratory analysis and are rarely integrated into end-to-end pipelines that jointly support topic discovery, evaluation, and downstream supervised classification.

The framework proposed in this study is best viewed as complementary to both neural network and interpretability-focused topic modeling approaches rather than competitive with them. By integrating probabilistic topic modeling with supervised classification, the proposed pipeline emphasizes interpretability, transparency, and analytical stability properties that are particularly important in forensic science, where analytical outputs may inform legal, investigative, or policy-related decisions. Unlike high-dimensional embedding-based representations or prompt-generated topics, LDA-derived topic proportions provide an explicit and low-dimensional feature space that can be directly inspected, visualized, and linked to well-defined disciplinary concepts.

From a practical standpoint, the proposed approach is computationally efficient and well suited to moderate-sized corpora such as the dataset analyzed in this study. Transformer-based and LLM-driven frameworks typically require large training datasets, specialized hardware (e.g., GPUs), external APIs, and nontrivial computational resources, and may require document truncation or chunking to accommodate context-length constraints [23]. In contrast, the LDA-based pipeline operates on complete documents, requires no external services, and can be trained and evaluated using modest computational resources while still yielding stable and competitive classification performance.

Overall, this work highlights a central methodological trade-off in contemporary text analysis: while neural and LLM-based models offer powerful representation learning and flexible topic generation capabilities, interpretable topic-based frameworks provide advantages in transparency, reproducibility, and resource efficiency. These properties make the proposed approach particularly suitable for forensic science applications and position it as a strong foundation for future hybrid extensions that may incorporate neural or LLM-derived representations where appropriate.

5.5. Limitations and Future Work

Several limitations of the present study should be acknowledged, which also motivate important directions for future research.

First, the empirical analysis is conducted on abstracts from a single journal, the Journal of Forensic Sciences (JFS). This choice was driven by the availability of consistent and well-defined journal-assigned disciplinary categories, which are essential for the downstream supervised classification experiments. Although abstracts from other major forensic venues (e.g., Forensic Science International) were also collected, those journals do not provide standardized category labels, making them unsuitable for the classification and diagnosticity analyses pursued in this study. As a result, the generalizability of the supervised findings is most directly applicable to corpora that include comparable domain metadata. Future work will extend the framework to larger and more diverse scientific collections that provide structured categorical annotations, enabling broader validation across disciplines and publication venues.

Second, while Latent Dirichlet Allocation is a well-established and widely used topic modeling method, it is known to exhibit a precision–recall trade-off in topic discovery. In particular, standard LDA formulations tend to favor higher recall by capturing a wide range of co-occurring terms, which can lead to the inclusion of noisy or weakly informative

topics. Recent work has shown that topic modeling can be viewed through an information retrieval lens, where controlling false positives is critical for improving topic coherence and downstream utility [48]. Although Bayesian model selection and robustness checks mitigate some of these effects in the present study, topic redundancy and noise remain inherent challenges in large-scale unsupervised modeling.

To address this limitation, an important direction for future research is the integration of principled Bayesian variable selection methods into the supervised stage of the pipeline. In particular, recent advances in mixed-type multivariate Bayesian shrinkage priors provide a theoretically grounded mechanism for selecting informative topic features while suppressing redundant or noisy dimensions [49]. Such approaches go beyond standard regularization techniques (e.g., LASSO), which may be unstable or overly sensitive in high-dimensional and correlated topic spaces. Incorporating Bayesian sparse selection is expected to further enhance classification performance and improve interpretability by identifying a compact subset of diagnostically relevant topics.

Finally, although this study does not directly incorporate transformer-based or large language models, the proposed framework is well positioned for hybrid extensions. Topic-based representations can serve as interpretable, low-dimensional complements to contextual embeddings derived from models such as BERT or LLM-based topic discovery methods. Future work will explore combining probabilistic topic modeling with neural or LLM-derived features, balancing the strengths of deep representation learning with the transparency, stability, and resource efficiency of topic-based approaches.

This study shows that combining probabilistic topic modeling with supervised learning yields an interpretable, robust, and computationally efficient framework for scientific literature analysis. The results demonstrate that LDA-derived topic representations capture meaningful disciplinary structure, remain stable across reasonable topic granularities, and support reliable classification performance. Together, these findings position topic-based pipelines as a practical and transparent complement to neural and LLM-based approaches, particularly in domain-sensitive settings such as forensic science.

6. Conclusions

This study proposed an integrated and interpretable framework for large-scale scientific literature analysis that combines probabilistic topic modeling with supervised classification. By unifying topic discovery, diagnostic evaluation, temporal trend analysis, and downstream classification within a single pipeline, the framework bridges the gap between unsupervised topic modeling and predictive modeling. Applied to forensic science abstracts, LDA was shown to recover semantically coherent and diagnostically meaningful topics aligned with established disciplinary categories, while topic-based representations supported robust and stable classification performance across multiple modeling families. These results highlight the practical utility of interpretable topic features for literature organization, trend analysis, and automated document categorization in domains where transparency is essential.

Several limitations suggest directions for future work. First, the empirical evaluation relied on a single journal with structured category metadata, which facilitated supervised analysis but may limit direct generalization to venues lacking such annotations. Second, classical topic models such as LDA are known to exhibit a precision–recall trade-off that can introduce redundant or noisy topics. Future research will explore precision–recall balanced topic modeling and Bayesian variable selection methods with shrinkage priors to further refine topic representations. Finally, while neural network and large language model–based approaches offer powerful representation learning capabilities, their computational cost and limited interpretability pose practical challenges. The proposed framework provides

a transparent and resource-efficient alternative and offers a natural foundation for future hybrid extensions that integrate topic-based representations with neural network or LLM-derived features.

Author Contributions: Conceptualization, A.A.Y. and L.T.; methodology, A.A.Y. and L.T.; formal analysis, A.A.Y. and C.K.; writing—original draft preparation, A.A.Y. and L.T.; writing—review and editing, A.A.Y., L.T., C.K. and C.M.B.; supervision, L.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Class Distributions Before and After SMOTE

This appendix reports the class distributions in the training data before and after Synthetic Minority Over-sampling Technique (SMOTE) application for each dataset configuration used in the supervised classification experiments. As described in Section 3.5.2 SMOTE was applied exclusively to the training portion of each cross-validation fold to mitigate class imbalance while preserving the integrity of the held-out test data.

Tables A1–A3 summarize the number and percentage of samples per class before and after SMOTE for the *All Categories*, *Dropped Categories*, and *Grouped Categories* scenarios, respectively. These tables illustrate how SMOTE equalizes class representation within each training set, enabling fairer model training and more reliable evaluation of minority-class performance.

Table A1. Class distribution in the training set for the `All Categories` configuration before and after SMOTE. Oversampling was applied only to the training data within each cross-validation fold.

Class	Before SMOTE	%	After SMOTE	%
Anthropology	319	15.6	572	9.1
Criminalistics	572	27.9	572	9.1
Digital & Multimedia Sciences	59	2.9	572	9.1
Engineering & Applied Sciences	21	1.0	572	9.1
General	104	5.1	572	9.1
Odontology	60	2.9	572	9.1
Pathology/Biology	512	25.0	572	9.1
Physical Anthropology	34	1.7	572	9.1
Psychiatry & Behavioral Science	164	8.0	572	9.1
Questioned Documents	57	2.8	572	9.1
Toxicology	147	7.2	572	9.1

Table A2. Class distribution in the training set for the `Dropped Categories` configuration before and after SMOTE. Oversampling was applied only to the training data within each cross-validation fold.

Class	Before SMOTE	%	After SMOTE	%
Anthropology	319	17.5	572	16.7
Criminalistics	572	31.5	572	16.7
General	104	5.7	572	16.7
Pathology/Biology	512	28.2	572	16.7
Psychiatry & Behavioral Science	164	9.0	572	16.7
Toxicology	147	8.1	572	16.7

Table A3. Class distribution in the training set for the `Grouped Categories` configuration before and after SMOTE. Oversampling was applied only to the training data within each cross-validation fold.

Class	Before SMOTE	%	After SMOTE	%
Anthropology	413	20.1	744	25.0
Applied Sciences	680	33.1	744	25.0
General Forensic	744	36.2	744	25.0
Psychiatry	220	10.7	744	25.0

Appendix B. Confusion Matrix Analysis for Grouped Categories

To provide deeper insight into the class-wise behavior of the supervised classifiers, this appendix presents row-normalized confusion matrices for the best-performing models under the *Grouped Categories* configuration. These visualizations complement the aggregate performance metrics reported in Table 1 and the Macro-F1 trends shown in Figure 15 by illustrating the distribution of classification errors across forensic domains.

Figure A1 presents the confusion matrix for XGBoost at $T = 88$, corresponding to the configuration with the highest Macro-F1 for the boosting models. XGBoost exhibits sharper diagonal dominance for *Applied Sciences* and *General Forensic*, indicating more conservative decision boundaries. However, a modest reduction in recall is observed for *Psychiatry*, consistent with the precision–recall trade-off characteristic of boosting-based classifiers.

Overall, the confusion matrices demonstrate that misclassifications are structured and interpretable rather than diffuse or random. The majority of errors occur between closely related forensic categories, supporting the semantic coherence of the LDA-derived topic representations and reinforcing the robustness of the proposed topic-based classification framework.

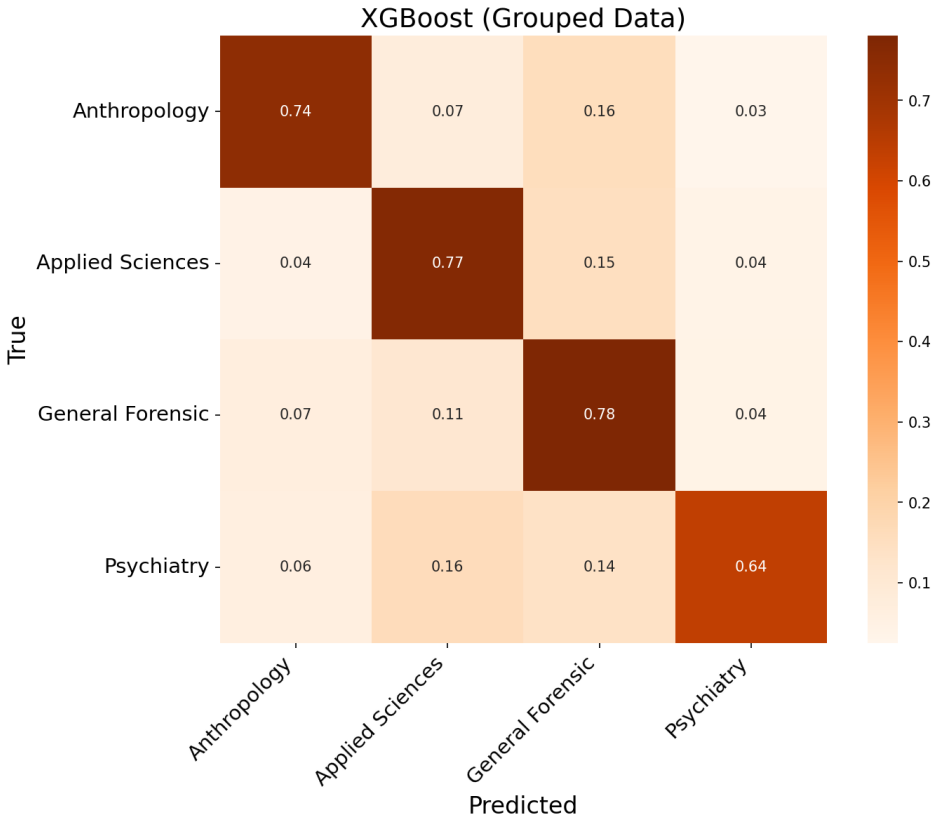

Figure A1. Row-normalized confusion matrix for XGBoost in the *Grouped Categories* configuration at $T = 88$. The model exhibits sharper class separation for dominant categories while showing a modest reduction in recall for Psychiatry, reflecting a precision–recall trade-off typical of boosting-based models.

References

1. Blei, D.M.; Ng, A.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
2. Griffiths, T.L.; Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5228–5235. [CrossRef]
3. Ponweiser, M. *Latent Dirichlet Allocation in R*; Theses/Institute for Statistics and Mathematics; WU Vienna University of Economics and Business: Vienna, Austria, 2012. [CrossRef]
4. Grün, B.; Hornik, K. topicmodels: An R Package for Fitting Topic Models. *J. Stat. Softw.* **2011**, *40*, 1–30 . [CrossRef]
5. Gatti, C.J.; Brooks, J.D.; Nurre, S.G. A historical analysis of the field of OR/MS using topic models. *arXiv* **2015**, arXiv:1510.05154. [CrossRef]
6. Sun, L.; Yin, Y. Discovering themes and trends in transportation research using topic modeling. *Transp. Res. Part C Emerg. Technol.* **2017**, *77*, 49–66. [CrossRef]
7. Xiong, H.; Cheng, Y.; Zhao, W.; Liu, J. Analyzing scientific research topics in manufacturing field using a topic model. *Comput. Ind. Eng.* **2019**, *135*, 333–347. [CrossRef]
8. Yu, D.; Xiang, B. Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling. *Expert Syst. Appl.* **2023**, *225*, 120114. [CrossRef]
9. Zhang, Y.; Shen, F.; Mojarad, M.R.; Li, D.; Liu, S.; Tao, C.; Yu, Y.; Liu, H. Systematic identification of latent disease-gene associations from PubMed articles. *PLoS ONE* **2018**, *18*, e0282763. [CrossRef]
10. Madzík, P.; Falát, L. State-of-the-art on analytic hierarchy process in the last 40 years: Literature review based on Latent Dirichlet Allocation topic modelling. *PLoS ONE* **2022**, *17*, e0268777. [CrossRef]
11. Asmussen, C.B.; Møller, C. Smart literature review: A practical topic modelling approach to exploratory literature review. *J. Big Data* **2019**, *6*, 93. [CrossRef]
12. Sabharwal, R.; Miah, S.J. An intelligent literature review: Adopting inductive approach to define machine learning applications in the clinical domain. *J. Big Data* **2022**, *9*, 53. [CrossRef]
13. Chen, X.; Xie, H.; Tao, X.; Wang, F.L.; Zhang, D.; Dai, H.N. A computational analysis of aspect-based sentiment analysis research through bibliometric mapping and topic modeling. *J. Big Data* **2025**, *12*, 40. [CrossRef]
14. Singh, A.; Glińska-Neweś, A. Modeling the public attitude towards organic foods: A big data and text mining approach. *J. Big Data* **2022**, *9*, 2. [CrossRef] [PubMed]
15. Colangelo, M.T.; Guizzardi, S.; Galli, C. Topic modeling as a tool to identify research diversity: A study across dental disciplines. *Metrics* **2024**, *1*, 3. [CrossRef]
16. Rejeb, A.; Rejeb, K.; Molavi, H.; Keogh, J.G. A Data-Driven Topic Modeling Analysis of Blockchain in Food Supply Chain Traceability. *Information* **2025**, *16*, 1096. [CrossRef]
17. Debortoli, S.; Müller, O.; Junglas, I.; vom Brocke, J. *Text Mining For Information Systems Researchers: An Annotated Topic Modeling Tutorial*; Communications of the Association for Information Systems; Swansea University: Swansea, UK , 2016.
18. Kherwa, P.; Bansal, P. Topic Modeling: A Comprehensive Review. *EAI Endorsed Trans. Scalable Inf. Syst.* **2019**, *20*, e2. [CrossRef]
19. Guillén-Pacho, I.; Badenes-Olmedo, C.; Corcho, O. Dynamic topic modelling for exploring the scientific literature on coronavirus: an unsupervised labelling technique. *Int. J. Data Sci. Anal.* **2025**, *20*, 2551–2581. [CrossRef]
20. Wu, X.; Nguyen, T.; Luu, A.T. A Survey on Neural Topic Models: Methods, Applications, and Challenges. *Artif. Intell. Rev.* **2024**, *57*, 18. [CrossRef]
21. Grootendorst, M. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. *arXiv* **2022**, arXiv:2203.05794.
22. Galli, C.; Cusano, C.; Meleti, M.; Donos, N.; Calciolari, E. Topic Modeling for Faster Literature Screening Using Transformer-Based Embeddings. *Metrics* **2024**, *1*, 2. [CrossRef]
23. Pham, C.M.; Hoyle, A.; Sun, S.; Resnik, P.; Iyyer, M. TopicGPT: A Prompt-Based Topic Modeling Framework. *arXiv* **2024**, arXiv:2311.01449.
24. Doi, T.; Isonuma, M.; Yanaka, H. Topic modeling for short texts with large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), Bangkok, Thailand, 11–16 August 2024; pp. 21–33. [CrossRef]
25. Doi, T.; Isonuma, M.; Yanaka, H. A Comprehensive Evaluation of Large Language Models for Topic Modeling. *arXiv* **2024**, arXiv:2406.00697 [CrossRef]
26. Tan, Z.; D'Souza, J. Toward purpose-oriented topic model evaluation enabled by large language models. *Int. J. Digit. Libr.* **2025**, *26*, 23. [CrossRef]
27. Mu, Y.; Dong, C.; Bontcheva, K.; Song, X. Large language models as alternatives to topic modeling. *arXiv* **2024**, arXiv:2403.16248. [CrossRef]
28. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.E.; Brown, D.E. Text Classification Algorithms: A Survey. *Information* **2019**, *10*, 150. [CrossRef]
29. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009.

30. Blei, D.M. Probabilistic Topic Models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]
31. Shah, K.; Patel, H.; Sanghvi, D.; Shah, M. A comparative analysis of logistic regression, random forest and KNN models for text classification. *Augment. Hum. Res.* **2020**, *5*, 12. [CrossRef]
32. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
33. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
34. Zhang, Q. The text classification of theft crime based on TF-IDF and XGBoost model. In Proceedings of the 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 27–29 June 2020; pp. 1241–1246. [CrossRef]
35. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
36. Lubis, A.R.; Prayudani, S.; Fatmi, Y.; Nugroho, O. Classifying news based on Indonesian news using LightGBM. In Proceedings of the 2022 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM), Surabaya, Indonesia, 22–23 November 2022; pp. 162–166. [CrossRef]
37. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
38. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
39. Taskiran, S.F.; Turkoglu, B.; Kaya, E.; Asuroglu, T. A Comprehensive Evaluation of Oversampling Techniques for Enhancing Text Classification Performance. *Sci. Rep.* **2025**, *15*, 21631. [CrossRef]
40. Powers, D.M.W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 37–63.
41. Sokolova, M.; Lapalme, G. A Systematic Analysis of Performance Measures for Classification Tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
42. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019. [CrossRef]
43. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv* **2019**, arXiv:1903.10676. [CrossRef]
44. Dieng, A.B.; Wang, C.; Gao, J.; Paisley, J. Topic Modeling in Embedding Spaces. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 439–453. [CrossRef]
45. Al Azher, I.; Reddy, V.D.; Akella, A.P.; Alhoori, H. LimTopic: LLM-Based Topic Modeling and Text Summarization for Analyzing Scientific Articles. In Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries (JCDL), Hong Kong, China, 15–19 December 2025; p. 30. [CrossRef]
46. Rodionov, D.; Konnikov, E.; Golikov, G.; Yakob, P. Structural–Semantic Term Weighting for Interpretable Topic Modeling with Higher Coherence and Lower Token Overlap. *Information* **2026**, *17*, 22. [CrossRef]
47. Glunčić, T.; Barić, D.; Glunčić, M. VISTA: A Multi-View, Hierarchical, and Interpretable Framework for Robust Topic Modelling. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 162. [CrossRef]
48. Virtanen, S.; Girolami, M. Precision–Recall Balanced Topic Modelling. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 6750–6759.
49. Wang, S.H.; Bai, R.; Huang, H.H. Two-Step Mixed-Type Multivariate Bayesian Sparse Variable Selection with Shrinkage Priors. *Electron. J. Stat.* **2025**, *19*, 397–457. [CrossRef]