*Journal of the Brazilian Computer Society, 2026, 32:1*doi: 10.5753/jbcs.2026.6635

AI-Driven Hierarchical Taxonomy Generation from Emergency Call Transcripts

Juan Gabriel Flores Sanchez [Computer Science Research & Development Laboratory (LIDI), Universidad del Azuay, Cuenca - Ecuador | juanfloressanchez@es.uazuay.edu.ec]
Marcos Orellana [Computer Science Research & Development Laboratory (LIDI), Universidad del Azuay, Cuenca - Ecuador | marore@uazuay.edu.ec]
Patricio Santiago García-Montero [Computer Science Research & Development Laboratory (LIDI), Universidad del Azuay, Cuenca - Ecuador | santyg20@est.uazuay.edu.ec]
Jorge Luis Zambrano-Martinez [Computer Science Research & Development Laboratory (LIDI), Universidad del Azuay, Cuenca - Ecuador | jorge.zambrano@uazuay.edu.ec]

Computer Science Research & Development Laboratory (LIDI), Universidad del Azuay, Cuenca - Ecuador

Received: 01 October 2025 • **Accepted:** 21 December 2025 • **Published:** 25 March 2026

Abstract This article presents a case study on hierarchical topic modeling for emergency call transcripts from Ecuador's ECU 911 service. We introduce a hybrid methodology that first generates a taxonomy from unlabeled data using *BERTopic* and agglomerative clustering, and then employs embedding-based similarity for multi-label classification. By leveraging multilingual embeddings (*LaBSE*) and clustering algorithms (*UMAP & HDBSCAN*), we identified 23 coherent topics, demonstrating a practical balance between accuracy and operational applicability. The key result is a significant reduction in Hamming Loss and an F1-score of 0.4951, achieved without the need for pre-labeled data. This underscores the method's primary practical significance: offering a scalable, automated solution for emergency management centers to rapidly categorize complex incidents, thereby enhancing situational awareness and resource allocation. The integration of *LLaMA 3* for automated label generation further optimized semantic interpretation, highlighting the potential of language models in critical, resource-constrained domains.

Keywords: Hierarchical Text Classification, Emergency Call Analysis, BERTopic, Large Language Models, Natural Language Processing, Multilingual NLP, Emergency Communication Systems

1 Introduction

Recent advances in Artificial Intelligence (AI) have revolutionized large-scale data management, particularly in the field of automated text classification. By leveraging machine and deep learning techniques, AI enables accurate analysis and hierarchical categorization of unstructured text, uncovering complex linguistic patterns critical for time-sensitive domains such as emergency services [Kowsari *et al*., 2019]. Hierarchical classification further enhances information accessibility, addressing the inefficiencies of manual methods—such as prolonged processing times and human error—while optimizing decision-making in high-volume environments such as public safety [Pacheco *et al*., 2023; Palanivinayagam *et al*., 2023].

The Integrated Security Service (ECU 911) in Ecuador faces considerable operational challenges due to the absence of an automated system for classifying emergency call transcripts. Presently, dependence on manual annotation is not only labor-intensive and susceptible to human error but also impairs swift response coordination with entities such as the National Police. The unstructured nature of these transcripts, which are characterized by colloquial language and multi-thematic content, further complicates analysis. Although Large Language Models (LLM), including Bidirectional Encoder Representations from Transformers (BERT), present a promising approach through bidirectional contextual understanding that could facilitate the processing of these communications without necessitating extensive labeled datasets, their implementation encounters practical limitations. Obstacles such as the demand for substantial computational resources, elevated training costs, and the need for multimodal alert analysis may limit their feasibility in resource-limited settings, as seen in ECU 911 [Yao *et al*., 2024; Li *et al*., 2023].

The primary issue examined in this study concerns the lack of an automated and scalable system for the hierarchical classification of multi-thematic emergency calls within resource-constrained environments, distinguished by the absence of a predefined taxonomy and manually annotated data. This operational gap at ECU 911 results in three distinct technical challenges: (i) the incapacity of unsupervised techniques such as Latent Dirichlet Allocation (LDA) to attain adequate semantic coherence in colloquial, low-density contexts text; (ii) the impracticality of supervised models requiring extensive annotated datasets; and (iii) the lack of interpretability in existing topic modeling outputs for critical decision-making. To address this disparity, this study delineates the subsequent quantifiable objectives: (1) to autonomously produce a semantically coherent hierarchical taxonomy from an unlabeled Spanish emergency corpus transcripts; (2) to accomplish this, employing a topic diversity score and coherence scores that demonstrate meaningful thematic consistency separation; and (3) to implement a multi-label classification system on this generated taxonomy that maintains a reasonable F1-score,

thereby providing a functional and scalable alternative to manual processing.

Consequently, this methodology effectively mitigates the limitations associated with traditional methods, such as LDA, when applied to contexts with low semantic density. Given the absence of predefined categories in the ECU 911 transcripts, this method facilitates automated hierarchical classification, thereby optimizing the processing of emergency calls without requiring extensive volumes of labeled data. The aim of this research extends beyond merely enhancing the efficiency of managing critical information; it also seeks to contribute to the advancement of scalable solutions within the realm of Natural Language Processing (NLP) as applied to public services.

Although the task is delineated as a hierarchical multi-label classification challenge, our methodology diverges from conventional supervised techniques. Rather than employing a predetermined taxonomy, we initially identify a topic hierarchy directly from the corpus through unsupervised topic modeling (BERTopic) and agglomerative clustering. This inferred taxonomy is subsequently utilized to categorize new texts based on semantic similarity, thereby effectively integrating topic modeling with hierarchical classification.

The structure of this article is organized as follows: Section 2 presents the scientific works relevant to this research. Section 3 demonstrates the proposed methodology. Section 4 evaluates the findings of the study; finally, Section 5 presents the conclusions of the research.

2 Related Works

Recent studies have highlighted the increasing use of LLMs for hierarchical text classification, particularly in the context of topic identification. However, unsupervised methods face limitations due to their dependence on unlabeled data, struggling with colloquial and context-dependent language common in command and control communications. Consequently, supervised or semi-supervised approaches dominate the literature, despite their demand for extensive labeled data and computational resources [Li *et al*., 2022; Topal *et al*., 2021].

The analyzed literature presents a diverse array of methodologies for text classification within emergency contexts, with a notable predominance of supervised models. Wang *et al*. [2023] and Yuan and Wang [2022] demonstrated the efficacy of BERT, enhanced through specialized loss functions (Cross-Entropy Weighted Focal (CEWF)) and hybrid architectures such as Recurrent Convolutional Neural Network (RCNN), in addressing imbalances in air accident and traffic data. Conversely, traditional methodologies, such as decision trees and Support Vector Machines (SVM) [Andirov *et al*., 2023], have exhibited shortcomings in managing complex categories, thereby underscoring the need for more advanced models capable of overcoming the limitations associated with labeled data.

Unsupervised models are emerging as a viable alternative to address this issue. Haj-Yahia *et al*. [2019] and Stammbach and Ash [2021] achieved results comparable to those of supervised models by employing semantic embedding and document clustering methodologies. In contrast, *BERTopic* [Tang *et al*., 2024] has demonstrated its effectiveness in identifying dynamic topics across various contexts, including health applications and governance. Notably, its integration with techniques such as Uniform Manifold Approximation and Projection (UMAP) and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [Liu and Wan, 2024] facilitated the hierarchical organization of substantial volumes of data with thematic coherence (0.7 vs. 0.27 for LDA).

In the context of hierarchical classification, LLMs represent a significant advancement. The algorithms Hierarchy-guided BERT with Global and Local hierarchies (HBGL) [Jiang *et al*., 2022] and Hierarchy-Aware Global Model (HiAGM) [Zhou *et al*., 2020] have optimized the management of intricate taxonomies by utilizing hierarchical structure encoders. Concurrently, TELEClass [Zhang *et al*., 2025] has successfully diminished computational expenses by utilizing LLMs for automatic taxonomy enrichment.

These advancements, assessed via specific metrics such as Hierarchical F1 (HiF) [Gargiulo *et al*., 2019], effectively address the constraints of flat methodologies, providing enhanced accuracy even within low-frequency categories. This capability is critical for emergency systems, where information is subject to rapid evolution.

Current research emphasizes LLM-based hierarchical text classification for topic identification, though unsupervised approaches struggle with unlabeled colloquial communications common in command centers. While literature favors supervised methods, their reliance on costly manual annotations remains a significant problematic. Our solution leverages *BERTopic*, combining topic modeling with agglomerative clustering and *LLaMA*-enhanced semantic analysis to enable precise emergency classification into multi-tiered taxonomies through the contextual interpretation of documents.

To synthesize the literature and provide a critical framework for situating our contribution, **Table 1** presents a comparative analysis of the dominant text classification paradigms applied to emergency and related domains. This synthesis surpasses merely providing a descriptive summary by explicitly contrasting the fundamental mechanisms, prerequisites, and inherent trade-offs of each approach, thereby elucidating the methodological niche occupied by our hybrid methodology.

3 Methodology

To precisely delineate the scope and contributions of this work, we define the core tasks involved. Hierarchical Topic Modeling refers to the unsupervised process of discovering latent topics and organizing them into a semantic hierarchy, encompassing both topic identification and the construction of a taxonomy. This directly enables Hierarchical Taxonomy Generation, a specific task that involves creating a multi-level structure where broad themes encompass specific sub-themes. Unlike Hierarchical Multi-label Classification, a supervised task of assigning documents to a pre-existing hierarchy, our approach generates the taxonomy directly from the data. Consequently, our methodology focuses on Multi-level Topic Identification, a process that extracts topics at varying levels

Table 1. Comparative Analysis of Text Classification Paradigms in Emergency Contexts

Paradigm	Core Mechanism	Data Requirements	Key Strengths	Key Limitations	Representative Works
Supervised	Learns from labeled examples to predict predefined categories.	Large volumes of high-quality, manually annotated data.	High accuracy with clear, predefined categories; effective for well-defined tasks.	Impractical where labeled data is scarce; fails to discover novel or emerging topics.	Wang *et al*. [2023]; Zhou *et al*. [2020]
Unsupervised	Discovers latent topics and structures directly from unlabeled text.	Unlabeled corpus only.	No need for manual labeling; adaptable to novel and evolving data.	Struggles with semantic coherence on colloquial language; topics can be less interpretable.	Haj-Yahia *et al*. [2019]; Stammbach and Ash [2021]
Hybrid (Our Work)	Integrates unsupervised topic discovery with semantic interpretation for classification.	Unlabeled corpus; minimal human validation for labels.	Scalable taxonomy generation without pre-defined schema; balances discovery with interpretability.	Semantic overlap in complex domains can challenge cluster separation; relies on model interpretation.	Proposed Method

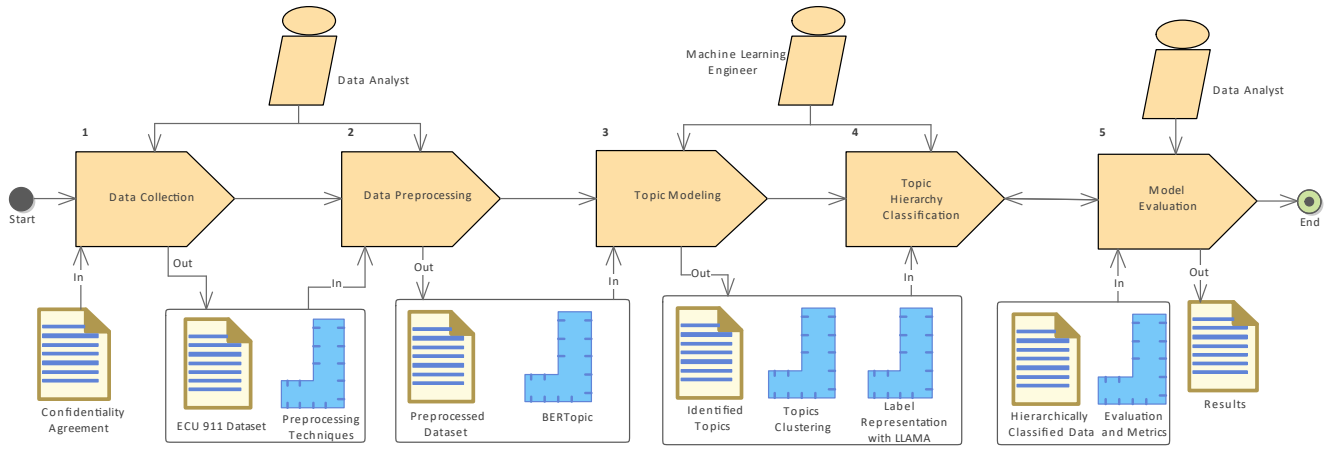


Figure 1. Methodological Workflow Overview.

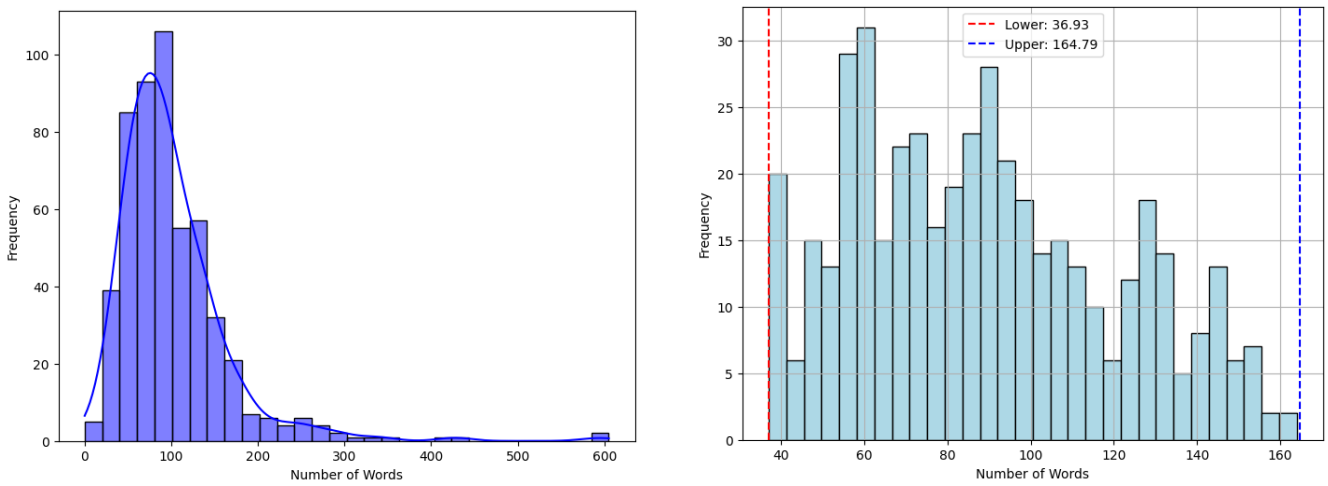
of abstraction to capture both general themes and specific subtopics within a unified analytical framework.

The development of our hierarchical classification methodology follows a structured workflow, as illustrated in **Figure 1**. The process comprises five principal phases: i) Data Collection, ii) Data Preprocessing, iii) Topic Modeling, iv) Topic Hierarchy Classification, and v) Model Evaluation. The core analytical work is executed in two fundamental stages: initially, we utilize BERTopic and hierarchical clustering to identify and structure a taxonomy of topics; subsequently, we employ semantic embeddings to assign texts to one or more nodes within this hierarchy.

The methodological workflow was formally delineated using a Systems Process Engineering Metamodel 2.0 (SPEM 2.0) diagram, an Object Management Group standard for modeling software development processes, which provides a structured visual and conceptual framework for representing method content and processes. Furthermore, to enhance reproducibility, the optimal hyperparameters for UMAP, HDBSCAN, and BERTopic are consolidated. Regarding the LLaMA 3.1-8B-Instruct labeling process, a qualitative analysis is provided, which includes a step-by-step demonstration using a synthetic emergency call transcript. This example explicitly contrasts the AI-generated labels with the broader, human-interpretable thematic categories within the generated hierarchy, illustrating the model's capacity to produce contextually precise and semantically coherent labels that align with expert intuition for emergency scenarios.

3.1 Data Collection

Our study utilizes a corpus of emergency call transcripts provided by ECU 911 to Computer Science Research and Development Laboratory (LIDI) at the Universidad del Azuay [Orellana *et al*., 2024], focusing exclusively on Citizen Security incidents due to their informational richness. The dataset comprises 529 call records, containing dialogues between operators and whistleblowers (reporters). However, our analysis focuses on whistleblower contributions because their dialogues contain idioms or colloquial language with an unstructured or informal style, compared to the operators' structured responses, which offer limited analytical value. Following formal ethics approval, access to sensitive emergency-call data was granted only after a rigorous anonymization protocol was applied; informed consent was obtained where necessary, with a waiver granted by the ethics board for archival data. The selected calls represent diverse citizen security emergencies, providing essential linguistic patterns for technological development.



(a) Text length distribution with Kernel Density Estimation (KDE) curve

(b) Text length distribution (68% percentile-filtered)

Figure 2. Statistical Distribution of Text Length

Although the dataset comprising 529 transcripts is modest in size for hierarchical modeling, it proved adequate to validate the core methodology as a proof-of-concept. Nonetheless, the scope of this investigation was initially designed as a proof-of-concept to assess the proposed hybrid methodology within realistic yet constrained data conditions. While the limited sample size may restrict the generalizability across wider emergency domains, our comprehensive preprocessing procedures and utilization of multilingual embeddings (LaBSE) effectively mitigated overfitting and enabled meaningful topic identification.

Preliminary Exploratory Data Analysis (EDA) revealed critical textual characteristics that informed our preprocessing approach, with text lengths averaging 100.86 words (SD = 63.93) across a 0–605-word range. This analysis, which included Gaussian approximations of distributions, identified outliers and biases that required mitigation to optimize language model performance. **Figure 2** illustrates the distribution of text lengths, highlighting variations that inform the development of preprocessing strategies. Such empirical grounding minimizes artifacts and strengthens pipeline robustness by aligning transformations with observed data properties.

The dataset's composition necessitated specialized handling of colloquial language patterns characteristic of emergency reporting. Statistical characterization of lexical features and discourse structures enabled targeted preprocessing to preserve semantic content while standardizing inputs. This approach ensures that the refined data effectively supports downstream modeling tasks, particularly for hierarchical classification systems that require consistent linguistic representations of emergency scenarios.

To determine a representative range and avoid biases caused by excessively long, short, or nonexistent texts, the empirical rule of the normal distribution was applied, which states that 68% of the data fall within one standard deviation of the mean. In this case, 68% of the texts were identified as being between 37 and 164 words, with 444 of the 529 texts initially used for preprocessing.

3.2 Data Preprocessing

While contextual embeddings encompass extensive linguistic information, our preprocessing pipeline specifically tackles challenges intrinsic to the domain of emergency call transcripts. The rationale for utilizing Named Entity Recognition (NER) and Part-of-Speech (POS) tagging in conjunction with embedding-based techniques is threefold: (i) Entity Preservation: NER detects vital emergency-specific entities (such as locations, individuals, and organizations) requiring explicit retention for effective operational response; (ii) Structural Normalization: POS tagging facilitates the identification and standardization of colloquial grammatical structures prevalent in spontaneous emergency speech; (iii) Noise Reduction: Domain-specific stopwords and polite expressions, although meaningful within semantic contexts of embeddings, tend to introduce noise in topic modeling tailored to our particular application.

The preprocessing pipeline adheres to a sequential algorithm optimized for emergency call transcripts:

1. Text normalization: convert to lowercase, remove numbers and punctuation, and reduce whitespace.
2. Linguistic Analysis: Apply spaCy's for NER and POS tagging.
3. Entity Preservation: Extract and store named entities for subsequent analysis while ensuring referential integrity.
4. Syntactic Filtering: Remove non-essential words based on POS tags, such as conjunctions, prepositions, and auxiliary verbs.
5. Lexical Cleaning: Remove standard stopwords (NLTK Spanish list) and domain-specific frequent terms.
6. Length Normalization: Filter texts outside the 37-164 word range based on empirical distribution analysis.
7. Deduplication: Eliminate redundant terms while maintaining the sequence and semantic coherence of the text.
8. Frequency-Based Filtering: Remove excessively frequent courteous expressions and formalities identified via corpus frequency analysis.

This structured approach guarantees the preservation of semantic information critical to emergencies while enhancing

coherence for topic modeling and improving computational efficiency.

Key techniques such as Named NER and POS tagging have been employed to extract structured information, with the `es_core_news_lg` *spaCy* model facilitating the processing of Spanish text. NER identifies contextual entities such as individuals and locations, while POS tagging categorizes grammatical elements, collectively contributing to the refinement of data quality. Furthermore, text normalization techniques—including the conversion to lowercase, the removal of numbers and punctuation, and the reduction of whitespace—standardize the corpus, ensuring uniformity. These preprocessing steps are crucial for enhancing downstream NLP tasks, as they eliminate irrelevant variations and preserve linguistically meaningful content, thereby facilitating robust semantic analysis and text mining applications.

The text elaborates on advanced filtering mechanisms designed to enhance the semantic purity of the corpus. The removal of stopwords, facilitated by Natural Language Toolkit (NLTK)'s predefined list, eliminates lexically insignificant terms. Moreover, the elimination of short words serves to discard residual noise. A deduplication process further refines the dataset by removing redundant terms without disrupting the text sequence. Additionally, frequency analysis identifies and filters overly recurrent polite expressions and formalities, which, although not classified as stopwords, contribute to noise due to their high frequency of occurrence. This multilayered approach ensures that only words with substantial semantic value are retained, thereby optimizing the dataset for accurate content analysis. By systematically applying these techniques, the preprocessing workflow significantly enhances the quality of the corpus, facilitating more precise and efficient linguistic modeling while maintaining a focus on analytically pertinent elements.

3.3 Topic Modeling

This research utilizes the BERTopic framework as its core topic modeling architecture, employing its standard five-stage pipeline that integrates transformer embeddings with unsupervised clustering [Egger and Yu, 2022]. Nonetheless, our contribution considerably surpasses a typical implementation. While adopting the multilingual LaBSE embedding model, our methodology is specifically tailored to address the linguistic nuances inherent in Spanish emergency discourse. Additionally, we deploy a bespoke hierarchical clustering strategy designed to elucidate the distinctive semantic structure characteristic of emergency calls. Most significantly, we introduce an innovative LLM-driven semantic interpretation module aimed at producing contextually precise topic labels. This constitutes a considerable advancement over BERTopic's default keyword-based labeling system, especially within the scope of domain-specific applications in emergency management.

Dimensionality reduction via UMAP (with a cosine metric and 15 components) preserved semantic relationships, while HDBSCAN clustering (with a minimum cluster size of 8) effectively grouped density-variant patterns and managed outliers [McInnes *et al*., 2018; Malzer and Baum, 2020].

This hybrid approach, enhanced by LLM-based interpretation, achieved contextually accurate clustering while maintaining computational efficiency through the use of the Sentence Transformer library, particularly for the Spanish-language corpus.

Parameter optimization through iterative testing yielded optimal configurations: UMAP used a neighborhood size of 4 and a minimum distance of 0.005, while HDBSCAN applied the Euclidean metric with leaf cluster selection. The resulting 23 clusters demonstrated effective data organization, striking a balance between specificity and broader groupings. Visualization confirmed spatial coherence with distinct cluster boundaries and controlled outlier treatment. This configuration captured the dataset's intrinsic variability while preventing over-fragmentation, as evidenced by semantically meaningful patterns suitable for emergency call analysis. To ensure transparency and reproducibility, the optimized hyperparameters for the core components pipeline are consolidated in **Table 2**.

BERTopic enhanced topic quality through *CountVectorizer* generated frequency matrices and c-TF-IDF optimized coherence, supplemented by Key-*BERTInspired*'s semantic refinement. For automated labeling, *LLaMA 3.1-8B-Instruct* was quantized (4-bit *nf4* with bfloat16) to reduce memory demands. The dual-prompt system (`system_prompt` for role definition and `main_prompt_template` for standardized inputs) generated precise labels when combined with controlled parameters (`temperature = 0.1`, `max tokens = 15`). This integrated pipeline maintained analytical rigor while producing interpretable outputs for emergency response applications.

3.4 Topic Hierarchy Classification

The hierarchical clustering of topics identified via *BERTopic* was performed using agglomerative clustering. This bottom-up hierarchical technique iteratively merges the most similar data points until either a single cluster is formed or a predefined stopping criterion is reached. This approach consolidates topics into coherent groups, enabling a structured hierarchical organization and facilitating the assignment of representative cluster labels.

A linkage matrix was constructed using complete linkage with cosine distance, followed by dendrogram generation to visualize the clustering process. A cut-off distance was then applied to segment topics into final clusters, calculating for each link how different the distance at which it merges is compared to the average of the distances below it. The results were presented alongside their generated labels for clear thematic organization.

As illustrated in **Figure 3**, the dendrogram visualization enables the adjustment of cluster granularity by modifying the cut-off distance, which is essential for identifying hierarchical structures that optimize data categorization into broad themes and subcategories. The agglomerative clustering process employed complete linkage and cosine similarity to generate a linkage matrix (`linkage_matrix`), efficiently capturing inter-topic relationships.

We implemented a pipeline leveraging *LLaMA 3.1-8B-Instruct* for automated cluster labeling. A dictionary was designed to store cluster-assigned labels, with prompts engi-

Table 2. Optimal Hyperparameter Configuration for the BERTopic Pipeline

Component	Hyperparameter	Value	Description / Rationale
UMAP	`n_neighbors`	4	Balances local and global structure preservation in the dimensionality reduction
	`n_components`	15	Number of dimensions for the reduced semantic space
	`min_dist`	0.005	Allows for tight packing of points within clusters while maintaining separation
	`metric`	cosine	Preserves semantic similarity between text embeddings
HDBSCAN	`min_cluster_size`	8	Defines the minimum number of documents required to form a coherent cluster
	`min_samples`	2	Controls the conservatism in declaring core points, set low for fine-grained clustering
	`metric`	euclidean	Distance metric used in the reduced UMAP space for density-based clustering
	`cluster_selection_method`	leaf	Prefers smaller, more refined clusters over large conglomerates
	`prediction_data`	True	Save supplementary information to predict the association of new data.
BERTopic	`embedding_model`	LaBSE	Multilingual sentence transformer model for generating semantic embeddings
	`vectorizer_model`	CountVectorizer	Generates the bag-of-words matrix for class-based TF-IDF computation
	`ctfidf_model`	–	Standard class-based TF-IDF weighting was applied without modifications
LLaMA 3.1-8B	`task`	text-generation	Specific task assigned to the text generator
	`temperature`	0.1	Ensures deterministic and focused label generation with minimal randomness
	`max_new_tokens`	15	Constrains output labels to be concise and directly relevant to the topic content
	`repetition_penalty`	1.1	Impose penalties on repetitions to prevent the model from producing redundant content.

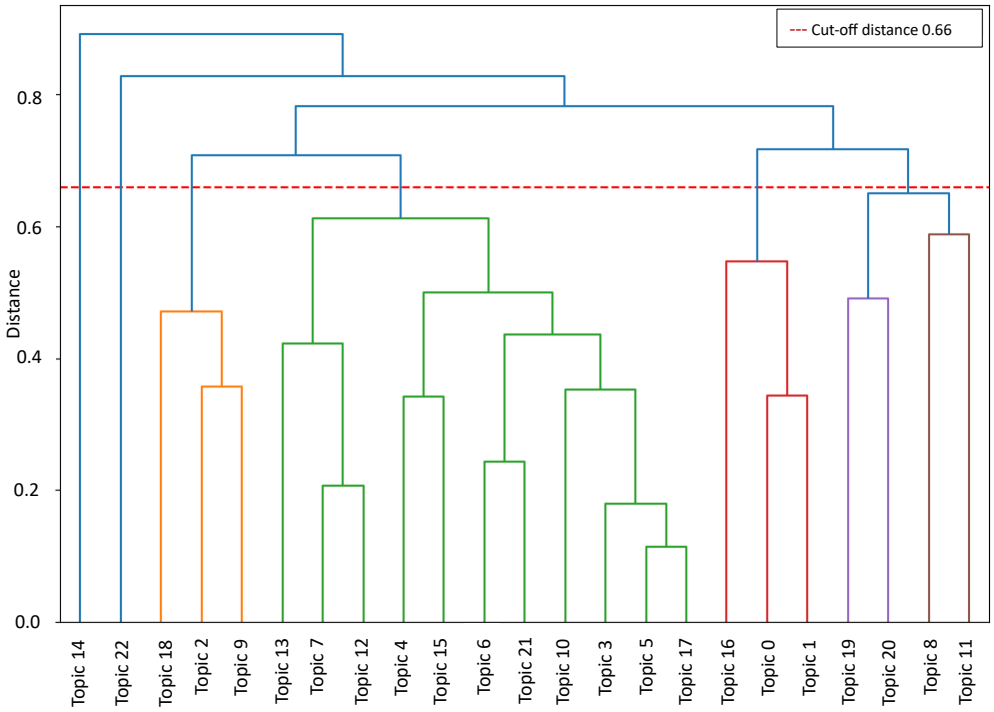


Figure 3. Cluster Dendogram

neered to instruct the model to produce brief, descriptive titles. To ensure consistency, the `transformers.pipeline()` parameters were configured with constraints on output length and specificity. Crucially, all AI-generated labels were subjected to a rigorous human review process to mitigate the risk of erroneous, biased, or potentially harmful categorizations before final approval and deployment.

The three-tier hierarchy was developed through iterative LLM interpretation: labels at Level 3 were initially produced based on the specific contents of clusters; subsequently, Level 2 labels were formulated by identifying semantic similarities among related Level 3 topics through comparative analysis. This bottom-up methodology guaranteed that the taxonomy emerged naturally from the data while preserving semantic coherence throughout all hierarchical levels.

3.4.1 End-to-End Workflow Demonstration

To concretely illustrate the practical application and operational validity of our proposed methodology, this section provides a step-by-step demonstration using a synthetic emergency call transcript. The following example illustrates a single input's progression through the complete analytical pipeline, from raw text to a multi-dimensional hierarchical classification, highlighting the system's ability to capture complex, multi-thematic incidents.

Input Emergency Call Transcript: "Good afternoon, I would like to report that at [anonym] park there is a group of young people using drugs and causing a disturbance. There are also suspicious vehicles parked nearby, and we neighbors are concerned for our safety."

Processing Steps and Outputs:

1. Text Preprocessing:
    - Normalization: Lowercasing and punctuation removal.
    - NER: Identifies and tags [LOCATION: [anonym] park].
    - POS Filtering: Retains nouns, adjectives, and verbs.
    - Final Preprocessed Text: [anonym] park group young people using drugs causing disturbance suspicious vehicles parked nearby neighbors concerned safety.
2. Semantic Embedding Generation: The preprocessed text is encoded by the multilingual Language-agnostic BERT Sentence Embedding (LaBSE) model, producing n-dimensional semantic vector representation ([0.234, -0.567, 0.891, ...]).
3. Topic Modeling and Clustering:
    - Dimensionality Reduction: UMAP projects the embedding into a lower-dimensional space.
    - Density-Based Clustering: HDBSCAN assigns the document to a cluster, identified by its semantic density and characterized by terms related to substance-related public disturbances.
4. Hierarchical Taxonomy Generation:
    - Level 3 (Specific Topic): An LLM synthesizes the top cluster terms ['drug', 'youth', 'park', 'disturbance', 'vehicles'] to generate the human-readable label: "Public Drug Consumption and Disorder Involving Youth."
    - Level 2 (Thematic Group): The LLM contextualizes this topic alongside sibling clusters ('Vandalism', 'Public Intoxication') to generate the broader category: "Substance-Related Public Incidents."
    - Level 1 (Root): The highest-level category is assigned as "Citizen Security."
5. Multi-label Classification: Semantic similarity analysis using Jina embeddings against all topic nodes assigns multiple relevant labels:
    - Primary Label: Public Drug Consumption and Disorder (Confidence: 0.87)
    - Secondary Labels:
      Suspicious Vehicle Presence (Confidence: 0.72)
      Community Security Concerns (Confidence: 0.65)

This example demonstrates the method's capability to accurately classify a single incident across multiple detailed themes within the generated hierarchy, thereby capturing the multifaceted nature of emergency reports.

3.5 Model Evaluation

The evaluation of the proposed methodology was performed in two separate stages: (1) the unsupervised topic modeling and hierarchy generation phase, and (2) the subsequent multi-label hierarchical classification phase. This bifurcated approach guarantees a thorough assessment of both the quality of the discovered taxonomy and its practical applicability for the categorization of new texts.

3.5.1 Evaluation of Topic Modeling and Hierarchical Structure

The coherence scores of the BERTopic model were measured using two key metrics: `C_v` coherence and topic diversity. The `C_v` coherence metric, introduced by Rosner *et al*. [2014], measures the semantic coherence of a set of terms within a topic. Calculated using embeddings generated by the LaBSE model, this analysis measures the interpretability and consistency of keywords within each topic by calculating the cosine similarity between word embeddings. Higher values (closer to 1) indicate that the keywords are semantically consistent and interpretable. Topic diversity evaluates the proportion of unique words among the generated topics by considering the most representative keywords for each. It assesses the extent of overlap between topics, wherein higher diversity signifies a broader coverage of the corpus's thematic content and reduced redundancy.

Additionally, the Silhouette Score was calculated to evaluate the cohesion and separation of the clusters identified by HDBSCAN in the UMAP-reduced space. A higher score (closer to 1) indicates well-separated, dense clusters. A moderate score is expected in this domain due to the inherent semantic overlap between emergency incident types. These metrics collectively validate the internal quality of the unsupervised topic discovery process and the emergent hierarchical organization.

3.5.2 Evaluation of Hierarchical Multi-Label Classification

To validate the performance of the final classification assigned to texts within the generated taxonomy, a distinct set of metrics, standard for multi-label classification tasks, was employed. For a comprehensive and impartial assessment, the classification conducted using LaBSE embeddings was corroborated using an alternative, independent embedding model, Jina Embeddings v3.

The sample size for this evaluation was determined using a 95% confidence level and a 5% margin of error, applying the sample size formula for finite populations. The total population consisted of 441 texts. Assuming a maximum variability ratio of $p = 0.5$, which yields the most precision in the calculation, a sample size of 206 observations was obtained. This sample size is statistically representative for this proof-of-concept study. While it provides a reliable estimate of the model's performance on our current dataset, the performance metrics should be interpreted as initial benchmarks. The application of this methodology to larger, multi-domain corpora will be crucial to further establish the robustness and generalizability of the approach.

The evaluation is focused on the following metrics:

- Exact Match (Accuracy): Measures the proportion of instances where the entire set of predicted labels exactly matches the ground-truth.
- Precision, Recall, and F1-Score: These metrics are reported for a more nuanced view of performance. Precision measures the correctness of the predicted labels, Recall measures the ability to find all relevant labels, and the F1-Score provides their harmonic mean.
- Hamming Loss: Measures the fraction of incorrectly predicted labels to the total number of labels. It is a key metric for multi-label scenarios, where a lower value indicates better performance.

This evaluation was conducted at various levels of the hierarchy to assess the model's performance at different levels of granularity. The assessment was performed by comparing the texts with their respective complete routes, which include three hierarchical levels: the general category (Citizen Security) being Level One, Level Two corresponding to the grouping of topics, and Level Three, which represents the specific final label or topic.

4 Results and Discussion

Our contribution resides in the integration of unsupervised topic discovery with hierarchical text classification, facilitating scalable taxonomy development and precise multi-label assignment without dependence on pre-existing categories or extensive labeled datasets.

The analysis contextualizes these results by examining their alignment with prior studies, discussing practical implications, and highlighting model limitations, while also identifying avenues for improvement. By synthesizing these insights, we highlight the methodological contributions and potential applications for hierarchical topic analysis.

Parameter tuning in UMAP (`reduced n_neighbors` and `low min_dist`) and HDBSCAN (`min_cluster_size` and `min_samples`) enabled the identification of coherent thematic clusters in the citizen security domain, achieving a balance between preserving local relationships and reducing noise. The results, with a Silhouette Score of 0.3275 and 23 clusters (8.39% outliers), reflect a moderate separation commensurate with the complexity of the texts. However, overlaps between clusters were observed due to the high semantic similarity and multi-label nature of the domain, suggesting limitations in the algorithms' ability to discriminate closely related topics.

These limitations could be attributed to the absence of embeddings trained explicitly for the citizen security context. It is worth noting that parameter selection was performed through a non-deterministic iterative process, prioritizing configurations that optimized the thematic structure. While the analysis meets the stated objectives, future research could explore alternative modeling techniques or specialized embeddings to improve accuracy in domains with semantic overlap.

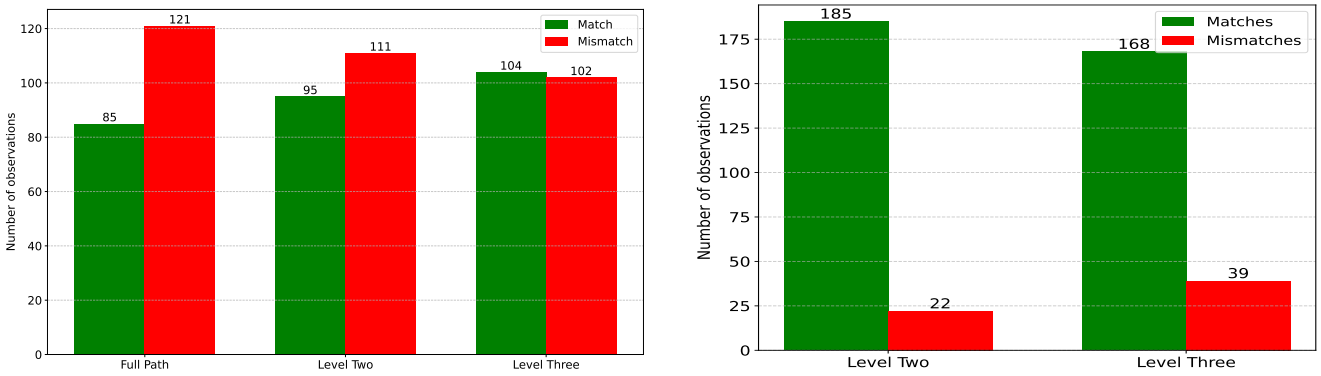
The evaluation of the hierarchical classification model was devised to assess its performance across two distinct operational scenarios, as illustrated in **Figure 4**. The initial scenario captures the model's true multi-label capability, wherein a single document may be associated with multiple relevant nodes within the hierarchy. The secondary scenario, provided for contextual purposes, considers the task as a single-label classification, assigning only the most probable label. This comparative framework is vital for demonstrating the model's proficiency in capturing the thematic complexity inherent in emergency calls, which is not sufficiently represented by traditional single-label metrics.

The *BERTopic* framework generated a hierarchical tree structure of topics, rooted in "CITIZEN SECURITY," through agglomerative clustering of embeddings refined via UMAP and HDBSCAN. While unigrams and bigrams enhanced semantic representation, dependency on keyword extraction necessitated complementary labeling via *LLaMA 3.1-8B-Instruct* to produce cluster-specific descriptors.

This hybrid approach achieved structured topic taxonomy but revealed limitations: the LLM's lack of domain expertise compromised nuanced interpretation of culturally or technically complex themes. Although automated labeling approximated expert judgment through semantic pattern recognition, critical validation remains essential to address gaps in contextual accuracy.

The resultant hierarchy strikes a balance between computational efficiency and interpretability, while underscoring the irreplaceable role of human validation in domain-sensitive applications. These findings advocate for semi-automated systems where LLMs assist, rather than replace, expert curation in specialized domains.

The semantic coherence analysis demonstrated that the topics generated by *BERTopic* exhibited adequate levels of internal coherence, with values ranging from 0.47 to 0.79, indicating effective conceptual alignment between the keywords within each topic. These results confirm the model's ability to identify meaningful patterns in the corpus and differentiate issues effectively. Additionally, a thematic diversity of 0.40 was observed, reflecting significant semantic rela-



(a) Performance Evaluation of *Jina Embeddings* **(b)** Evaluation with the *Jina embeddings* model extended to three possibilities

Figure 4. Comparative Evaluation of Jina Embeddings in Multi-Level Text Validation

tionships between the topics, consistent with the hierarchical approach implemented.

The spatial distribution of the topics reveals conceptual groupings that facilitate their integration at higher levels of the hierarchy. Although some topics exhibit less coherence, attributable to the complexity and interdisciplinarity of the corpus, the overall balance between internal coherence and thematic diversity validates the model's effectiveness for textual analysis and categorization. These findings support the approach's usefulness for processing data with complex thematic characteristics.

To quantitatively demonstrate the enhancement of our hybrid methodology, we conducted a comparison with baseline models such as LDA , which is prevalently employed in unsupervised topic modeling. Within our dataset, LDA yielded a topic coherence (C_v) score of 0.27, as documented in related research [Liu and Wan, 2024] . Conversely, our BERTopic-based methodology achieved coherence scores ranging from 0.47 to 0.79, accompanied by a topic diversity measure of 0.40. This signifies a substantial advancement in semantic interpretability and coverage. For the purpose of error analysis, we scrutinized confusion patterns within the hierarchical taxonomy, recognizing that categories exhibiting significant semantic overlap, such as "Substance-Related Public Incidents" (Level 2) and "Urban Crimes and Violence" (Level 3), were often misclassified owing to their contextual similarities in emergency scenarios. This is evidenced by multi-label assignments, whereby documents received labels from both categories. This confusion arises from the nuanced characteristics of emergency calls, wherein incidents frequently involve multiple interconnected themes, thereby rendering discrete cluster separation challenging. Future efforts to address these overlaps may incorporate domain-specific embeddings or hierarchical constraints.

Figure 4a presents a comparative analysis of classification performance across the three hierarchical levels: Full Path, Level 2, and Level 3. It contrasts the multi-label scenario with a single-label baseline. The results exhibit a consistent and expected trend: performance metrics such as Accuracy, Precision, Recall, and F1-Score are systematically higher under the single-label regime, while the Hamming Loss is lower. This discrepancy is attributable to the inherent simplicity of the single-label task, which necessitates the model to identify only the predominant theme within a document. Conversely, the multi-label task requires a more sophisticated semantic understanding to recognize all relevant themes, thus increasing complexity. The lower *Match* accuracy observed in the multi-label setting is a well-recognized characteristic of such tasks, as it demands correctness of the entire set of predicted labels. The more informative metric, Hamming Loss, which measures the proportion of incorrect labels, indicates acceptable performance given the complexity of the corpus. The observed decline in performance at the *Full Path* level, in both scenarios, can be ascribed to the cumulative difficulty of accurately predicting the entire hierarchical chain of labels for each document; an error at any level invalidates the entire path.

Although the aggregate metrics depicted in **Figure 4a** provide a high-level overview of performance, they do not fully elucidate the model's behavior at individual hierarchy levels. For a more detailed analysis, we examined the per-document label assignments. **Figure 4b** delineates the model's output at Levels 2 and 3, distinguishing between 'Matches', instances in which at least one predicted label was correct, and 'Mismatches', instances in which no predicted label was correct, in the context of the multi-label scenario.

As demonstrated in **Figure 4b**, the model effectively aligned with at least one accurate label for the majority of texts at both Level 2 (broader themes) and Level 3 (more specific topics). The increased match rate at Level 2 is attributable to its reduced granularity and more expansive thematic categories, which are simpler to recognize. The fact that a significant portion of Level 3 topics were also accurately identified, despite their specificity, confirms the model's capacity to capture nuanced semantic details information. This analysis corroborates the findings from **Figure 4a**, demonstrating that although the model may not attain perfect full-path accuracy, it consistently recognizes pertinent thematic content at the suitable level of abstraction, which constitutes a fundamental requirement for effective emergency call response analysis.

The evaluation of the hierarchical classification model at levels two and three reveals a moderate performance, with key metrics such as Accuracy, Precision, Recall, and F1-Score recording approximately 0.47 for level two and 0.49 for level three. The Hamming Loss scores were noted as 0.53 and 0.50, respectively. These results indicate that, although the model does not achieve complete accuracy in assigning paths to texts, it successfully captures a reasonable semantic rela-

Table 3. Label Assessment in Alerting Texts

Alerting text	Label	Level	Label predictions with *jina embeddings*
please, I wanted to say if you can send a patrol to El Ángel park, listen right now a group of like fifteen to twenty people are consuming drugs in the area of the park, but people walk through there and it's really causing discomfort, thank you	Inappropriate behavior problems due to substances	Two	['Inappropriate behavior problems due to substances', 'Urban crimes and violations', 'Social violence and delinquency']
	Problems of drug consumption in public areas	Three	['Problems of drug consumption in public areas', 'Youth altercating under the influence of alcohol','Community delinquency and theft']

tionship between them. This is significant given the complex and subjective nature of the evaluated categories, which include conceptual overlaps. Consequently, its performance is deemed adequate for applications that prioritize semantic understanding over stringent accuracy. Nevertheless, further enhancements to the embeddings and additional tuning are recommended to minimize errors, particularly in classes characterized by fuzzy conceptual boundaries.

Jina Embeddings enables multi-label assignment for a single text, effectively capturing its semantic complexity and richness. For instance, an incident involving drug use in a park may receive contextually valid labels such as "Substance-related misbehavior (Problemas de comportamiento inapropiado por sustancias)", "Urban crimes and violence (Delitos urbanos y violaciones)", and "Social delinquency (Violencia y delincuencia social)" reflecting the event's multifaceted nature. This capability is particularly valuable in domains like citizen security, where incidents often span multiple thematic dimensions simultaneously.

As shown in **Table 3**, the model can propose complementary yet divergent labels, such as "Drug use in public spaces (Problemas de consumo de drogas en áreas públicas)", "Alcohol-fueled altercations among youths (Jóvenes altercando bajo influencia del alcohol)" or "Community theft and crime (Delincuencia y robo en la comunidad)" even if only one aligns with the ground-truth annotation. By identifying interrelated facets of a single event, *Jina Embeddings* enhances semantic analysis, providing nuanced insights for complex scenarios where rigid, single-label classifications fall short.

5 Conclusion

This research successfully developed and implemented a hierarchical classification approach for ECU 911 conversations, demonstrating its capability to address the inherent complexity of multi-label classification tasks. The obtained metrics, including a topic coherence range of 0.47 to 0.79, a topic diversity of 0.40, an F1 score of 0.4951, and a reduction in Hamming Loss with the `jinaai/jina-embeddings-v3` model, confirm the method's effectiveness. The meticulous selection of 444 call texts enabled the identification of 23 topics featuring well-defined hierarchical granularity. This process was significantly supported by comprehensive data preprocessing and embedding implementation, which were instrumental in capturing underlying semantic relationships.

The primary contribution of this study resides in the integration of advanced language models and topic modeling methodologies for the analysis of emergency situations in Spanish. This work is informed by an extensive review of the NLP literature. The results underscore the method's capacity to discern intricate thematic patterns while preserving a clear hierarchical structure, a vital characteristic for information management in critical contexts situations. Nevertheless, this study possesses certain limitations, notably the utilization of a modest corpus comprising 529 transcripts and the dependence on general-purpose multilingual embeddings, such as LaBSE and Jina, which may not entirely encapsulate the domain-specific nuances of emergency contexts communications.

Looking ahead, this research establishes a foundational basis for future investigations by delineating explicit pathways for improvement. Future work will concentrate on the integration of the model into real-time emergency response systems and the enhancement of its multilingual capabilities to accommodate various linguistic contexts. Specifically, we intend to incorporate data augmentation techniques, such as paraphrasing and synonym replacement, and expand the corpus with multilingual emergency data from analogous contexts to improve model robustness and cross-lingual applicability, ultimately resulting in more tailored and resilient solutions.

Acknowledgements

This work was supported by the vice rectorate of Research at Universidad del Azuay. Therefore, we thank them for their financial and academic support and the entire Computer Science Research & Development Laboratory (LIDI) staff.

Funding

This research was funded by the vice rectorate of Research at Universidad del Azuay.

Authors' Contributions

Juan Gabriel flores Sanchez contributed to Data Curation, Investigation, Methodology, validation, Writing - original draft; Marcos Orellana, Patricio Santiago Garcia-Montero, & Jorge Luis

Zambrano-Martinez contributed to the conceptualization, Investigation, Methodology, Project administration, Supervision, Validation, and Write - review & editing of this study.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to their sensitive nature and to protect participant confidentiality in accordance with COPE guidelines and research ethics.

Further relevant information

In the manuscript, the authors explicitly state that formal ethics approval was obtained prior to data collection. Access to the sensitive emergency call transcripts from ECU 911 was granted only after a rigorous anonymization protocol was applied to protect participant confidentiality. Furthermore, informed consent was obtained where necessary, with a waiver granted by the ethics board for the use of archival data.

References

Andirov, M., Assan, Z. Z., Nopembri, S., Seilkhan, A., and Myrzakhmetov, D. (2023). Classification of texts on emergency situations in almaty. *Kompleksnoe Ispolzovanie Mineralnogo Syra= Complex use of mineral resources*, 327(4):23–31. DOI: 10.31643/2023/6445.36.

Egger, R. and Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 7:886498. DOI: 10.3389/fsoc.2022.886498.

Gargiulo, F., Silvestri, S., Ciampi, M., and De Pietro, G. (2019). Deep neural network for hierarchical extreme multi-label text classification. *Applied Soft Computing*, 79:125–138. DOI: 10.1016/j.asoc.2019.03.041.

Haj-Yahia, Z., Sieg, A., and Deleris, L. A. (2019). Towards unsupervised text classification leveraging experts and word embeddings. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pages 371–379. DOI: 10.18653/v1/P19-1036.

Jiang, T., Wang, D., Sun, L., Chen, Z., Zhuang, F., and Yang, Q. (2022). Exploiting global and local hierarchies for hierarchical text classification. *arXiv preprint arXiv:2205.02613*. DOI: 10.48550/arXiv.2205.02613.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4):150. DOI: 10.3390/info10040150.

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., and He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2):1–41. DOI: 10.1145/3495162.

Li, Z., Zhu, H., Lu, Z., and Yin, M. (2023). Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*. DOI: 10.18653/v1/2023.emnlp-main.647.

Liu, Y. and Wan, F. (2024). Unveiling temporal and spatial research trends in precision agriculture: A bertopic text mining approach. *Heliyon*. DOI: 10.1016/j.heliyon.2024.e36808.

Malzer, C. and Baum, M. (2020). A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE international conference on multisensor fusion and integration for intelligent systems (MFI)*, pages 223–228. IEEE. DOI: 10.1109/MFI49285.2020.9235263.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*. DOI: 10.48550/arXiv.1802.03426.

Orellana, M., Molina Pinos, P. A., García-Montero, P. S., and Zambrano-Martinez, J. L. (2024). Pre-processing of the text of ecu 911 emergency calls. In *Conference on Information and Communication Technologies of Ecuador*, pages 271–284. Springer. DOI: 10.1007/978-3-031-75431-9_18.

Pacheco, S. A. d. J. S., Romero, F. C., Domínguez, R. G., and Vasconcelos, M. P. (2023). Clasificación jerárquica de texto con machine learning en la industria petrolera. *Innovación y Desarrollo Tecnológico*. Available at:`https://iydt.wordpress.com/wp-content/uploads/2023/11/4_64_clasificacion-jerarquica-de-texto-con-machine-learning-en-la-industria-petrolera_.pdf`.

Palanivinayagam, A., El-Bayeh, C. Z., and Damaševičius, R. (2023). Twenty years of machine-learning-based text classification: A systematic review. *Algorithms*, 16(5):236. DOI: 10.3390/a16050236.

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., and Both, A. (2014). Evaluating topic coherence measures. *arXiv preprint arXiv:1403.6397*. DOI: 10.48550/arXiv.1403.6397.

Stammbach, D. and Ash, E. (2021). Docscan: Unsupervised text classification via learning from neighbors. *arXiv preprint arXiv:2105.04024*. DOI: 10.48550/arXiv.2105.04024.

Tang, Z., Pan, X., and Gu, Z. (2024). Analyzing public demands on china's online government inquiry platform: A bertopic-based topic modeling study. *Plos one*, 19(2):e0296855. DOI: 10.1371/journal.pone.0296855.

Topal, M. O., Bas, A., and van Heerden, I. (2021). Exploring transformers in natural language generation: Gpt, bert, and xlnet. *arXiv preprint arXiv:2102.08036*. DOI: 10.48550/arXiv.2102.08036.

Wang, Z., Wang, L., Huang, C., Sun, S., and Luo, X. (2023). Bert-based chinese text classification for emergency management with a novel loss function. *Applied Intelligence*, 53(9):10417–10428. DOI: 10.1007/s10489-022-03946-x.

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., and Zhang, Y. (2024). A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211. DOI:

10.1016/j.hcc.2024.100211.

Yuan, S. and Wang, Q. (2022). Imbalanced traffic accident text classification based on bert-rcnn. In *Journal of Physics: Conference Series*, number 1 in 2170, page 012003. IOP Publishing. DOI: 10.1088/1742-6596/2170/1/012003.

Zhang, Y., Yang, R., Xu, X., Li, R., Xiao, J., Shen, J., and Han, J. (2025). Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision. In *Proceedings of the ACM on Web Conference 2025*, pages 2032–2042. DOI: 10.1145/3696410.3714940.

Zhou, J., Ma, C., Long, D., Xu, G., Ding, N., Zhang, H., Xie, P., and Liu, G. (2020). Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1106–1117. DOI: 10.18653/v1/2020.acl-main.104.