IAB Taxonomy Classification Using Graph-Based Retrieval-Augmented Generation (RAG)

Niketh Bayya Mahesh , Divya Prashanth Paraman
Affiliation {nbayyam, dparaman}@iu.edu

December 20, 2024

Abstract

In this work, we explore the use of a hierarchical knowledge graph-based Retrieval-Augmented Generation (RAG) model to classify webpage content into the IAB Content Taxonomy. By leveraging Neo4j for graph representation, our approach integrates contextually constrained synonyms, multilingual embeddings, and tier-specific semantic structures to enhance retrieval quality. We incorporate LLM-driven summarization and a weighting-based node refinement strategy to address ambiguity, reduce misclassification, and improve cluster loyalty. Experimental results show notable gains in classification accuracy, particularly when combining original and summarized inputs. We discuss observed biases, the impact of pruning low-weight nodes, and provide insights into future directions such as fine-tuning embedding models, incorporating more contextual triplets, and expanding evaluation across tiers and languages.

1 Introduction

Accurate classification of webpage content into the IAB Content Taxonomy is essential for applications such as targeted advertising, content filtering, and analytics. The taxonomy's hierarchical structure, however, presents challenges in modeling and classification, particularly when semantic overlaps occur between closely related categories. For example, distinguishing categories like "Education" and "Family and Relationships" often leads to misclassifications if contextual nuances are not properly captured.

Retrieval-Augmented Generation (RAG) offers a powerful solution by integrating pre-trained language models with external knowledge retrieval. In this study, we address the core challenges of semantic ambiguity, feature overlap, and scalability by employing Neo4j, a graph database optimized for hierarchical structures. Our hypothesis is that a graph-based retrieval framework, enriched with context-constrained synonyms and weighting-based node refinement, can significantly enhance classification accuracy while maintaining scalability across multilingual and complex datasets.

To achieve this, we constructed a hierarchical knowledge graph of the IAB taxonomy, integrated LLM-driven contextual synonyms, applied a weighting approach to favor cluster loyalty, and explored summarization to refine semantic retrieval. This research examines the limitations of traditional classification methods and highlights the potential of graph-based RAG systems in tackling intricate, large-scale classification tasks.

2 Previous Work

Traditional approaches to webpage classification often relied on rule-based systems or conventional machine learning techniques (e.g., Support Vector Machines, Decision Trees). While these methods performed reasonably well on small-scale, structured datasets, they frequently struggled with:

1. Capturing semantic relationships in hierarchical data.
2. Scaling effectively to large and linguistically diverse datasets.
3. Handling feature overlap and contextual ambiguity, especially when closely related categories share similar terminology.

In recent years, graph-based databases such as Neo4j have emerged as robust alternatives for modeling complex relationships, demonstrating their utility in recommendation systems, social network analysis, and other domains. However, the application of graph-centric approaches to taxonomy-based classification, particularly when combined with Retrieval-Augmented Generation (RAG) techniques, remains underexplored.

Beyond academic exploration, industry practitioners have adopted the IAB Content Taxonomy for contextual targeting and brand safety. According to the IAB Tech Lab's official documentation [1], the Content Taxonomy provides a standardized framework for categorizing content, enabling more consistent and transparent alignment between publishers and advertisers. While tools and services (e.g., Integral Ad Science, DoubleVerify) leverage the taxonomy for brand suitability and contextual targeting, much of this integration is proprietary and lacks detailed accounts of underlying graph-based methodologies.

Some exploratory work exists in applying machine learning methods directly to the IAB taxonomy. For instance, Lefranc and Risson [2] discussed using ML techniques to classify webpages into IAB categories. Additionally, research by Al-Olimat et al. [3] showcased a graph-based approach for context-aware advertisement recommendations using hierarchical taxonomies, although not specifically the IAB taxonomy. These initiatives highlight the potential of combining structured taxonomies with semantic retrieval methods, but a clear gap remains in fully integrating graph-based retrieval with RAG models for improved IAB taxonomy classification.

Our work builds on this foundation by integrating Neo4j with state-of-the-art language models and contextual augmentation. This novel combination enables dynamic taxonomy adjustments, semantic retrieval, and context-aware classification. In doing so, we address critical gaps in existing methodologies, offering a scalable and more accurate approach to IAB taxonomy-based content classification.

3 Methodology

3.1 Datasets and Data Modeling

1. **Data Collection**: To construct a representative data set aligned with the IAB Content Taxonomy, we began by selecting 8 Tier-1 categories, each covering distinct domains (e.g., "Automotive," "Business and Finance," "Education," "Entertainment," "Family & Relationships," "Hobbies & Interests," "Medical Health," and "Personal Finance"). For each category of Tier 1, we identified 10 subcategories (Tier-2) to ensure a more granular coverage of the content.

   Initially, synonyms for these keywords were generated using conventional tools like NLTK's WordNet. However, these synonyms often lacked the nuanced context required for an accurate

classification, especially in closely related categories where generic or irrelevant synonyms introduced ambiguity.

To address this, we employed Large Language Models (LLMs) to generate context-constrained synonyms for each keyword. The LLM prompt included information about the hierarchical levels, guiding the model to produce synonyms that were semantically aligned with the specific tier context rather than generic equivalents. This approach ensured that the expanded keyword sets more accurately reflected the subtle distinctions between categories and tiers, thereby improving the quality of subsequent retrieval and classification steps.

Using the refined set of keywords and their synonyms, we queried search engines to retrieve relevant URLs. For each keyword, top URLs were collected, resulting in a total dataset of approximately 800 webpages, of which 150 were non-English (e.g., German, French, Chinese), ensuring multilingual coverage for further testing of the RAG framework.

2. **Data Modeling**: With the curated dataset and enriched keyword sets in hand, we proceeded to model the IAB taxonomy using a graph-based representation in Neo4j. The hierarchical structure of the taxonomy naturally lent itself to a graph format, where nodes and relationships could explicitly capture parent-child categories and semantic linkages. We represented Tier-1, Tier-2, and Tier-3 categories as distinct node types. Each Tier-1 node connected to its Tier-2 children through an IS_PARENT relationship, and similarly, each Tier-2 node connected to its Tier-3 categories using the same relationship type. This structure established a clear hierarchical graph that mirrored the taxonomy's conceptual design. Keywords and their contextually derived synonyms were incorporated as nodes within the graph, linked to their respective Tier-3 categories. A SIMILAR_TO relationship connected each keyword node to its synonym nodes, forming semantic clusters around each category. This allowed for easier retrieval of category-relevant terms and provided a richer semantic context that could be leveraged by RAG-based retrieval methods. By representing categories, keywords, and synonyms as nodes, and using relationships such as IS_PARENT (for hierarchy) and SIMILAR_TO (for semantic similarity), the resulting knowledge graph captured the multi-level structure of the IAB taxonomy. It also embedded the contextual nuances introduced by the LLM-generated synonyms, thereby forming a robust semantic backbone for subsequent classification tasks.

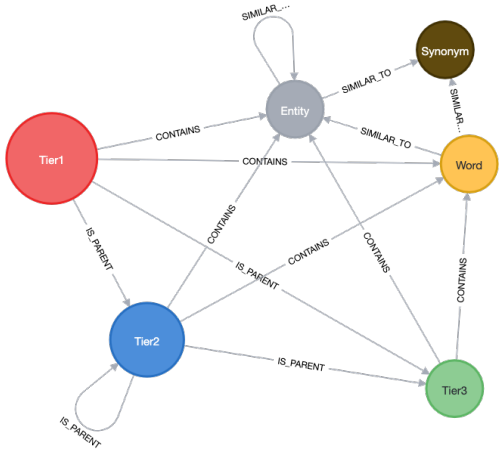


Figure 1: Data Model

3.2 RAG Framework

1. **Base RAG Framework**:
   Our initial methodology combined three core steps to classify webpages into the IAB taxonomy:

   (a) Embedding Generation with Multiple Models: We started by encoding the raw webpage text using multiple multilingual embedding models (e.g., xlm-roberta-base, mBERT, mDeBERTa-v3, LaBSE, DistilmBERT, mT5, XLM). Each model produced an embedding vector that captured semantic nuances of the webpage content.

   (b) Neo4j-Based Retrieval: The generated embeddings were then used to query a Neo4j-based knowledge graph, which contained a hierarchical representation of the IAB taxonomy along with context-constrained synonyms. By performing a vector similarity search, we retrieved the top candidate categories or tiers for each embedding model.

   (c) Simple Voting Mechanism: Each model contributed a single vote for the category it deemed most relevant. We aggregated votes across all models and selected the category with the highest vote count as the final classification. While this approach provided a foundational solution, it encountered limitations, particularly due to model biases and ambiguities in closely related categories.

   During preliminary evaluations (see accuracy table in later sections), we observed that certain models, notably RoBERTa and mDeBERTa-v3, introduced biases that affected the classification accuracy. By excluding these models from the ensemble, we obtained improved performance, reflecting the importance of careful model selection.

   Improvements to the Base Framework: Building upon the initial framework, we implemented several enhancements:

   (a) Model Exclusion: We removed the biased models (RoBERTa and mDeBERTa-v3) from the ensemble, retaining those that demonstrated more consistent and unbiased behavior.

   (b) Incorporation of Summarization: To reduce noise and enhance context, we introduced an LLM-based summarization step. Summarizing the webpage text before embedding enabled the models to focus on core thematic elements, improving classification precision.

   (c) Weighted Aggregation of Original and Summarized Inputs: We experimented with various weight ratios (x/y) to combine embeddings from both original and summarized text, systematically adjusting the influence of each. This approach allowed for a more balanced decision, leveraging the complementary strengths of raw and distilled representations.

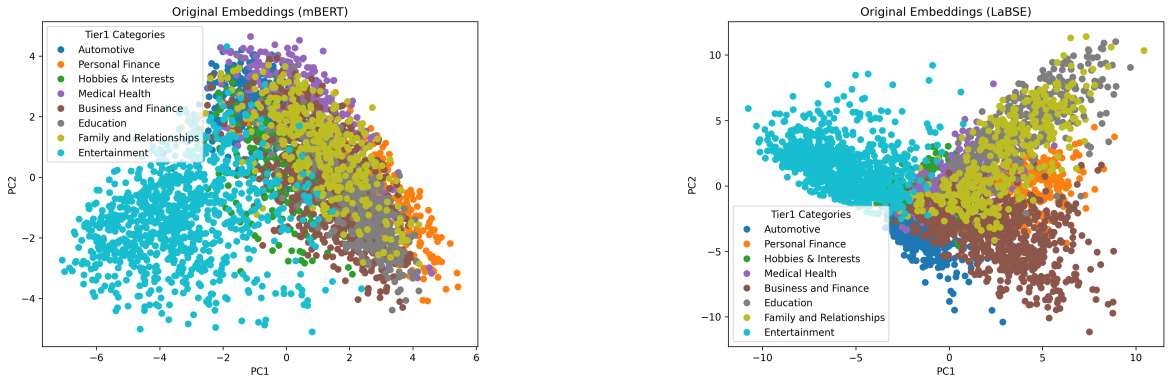
   (d) Refined Model Set (Only mBERT/LaBSE with Summarized Text): Through iterative testing, we found that using only summarized inputs and restricting the ensemble to the two most effective models (mBERT and LaBSE) yielded the highest accuracy results. Specifically, this configuration achieved approximately 70% Tier-1 accuracy and 43% Tier-2 accuracy, outperforming all previous versions. Detailed accuracy metrics can be found in the subsequent tables.

   These iterative refinements illustrate how incremental modifications—removing biased models, introducing summarization, adjusting weight ratios, and narrowing the model set—led to a more robust, context-aware, and accurate RAG-based classification system.

Table 1: Accuracy Results for Various Configurations

Approach	Configuration	Tier 1 Accuracy	Tier 2 Accuracy
Base	All 7 Models (Original Text)	47%	26%
Base	Excluding RoBERTa/mDeBERTa (Original Text)	53%	30%
Improved	Only Summarized Text	65%	40%
Improved	Equal Weight (0.5) Original + Summarized	63%	37%
Improved	Only mBERT/LaBSE (Summarized Text)	70%	43%

2. **Addressing Semantic Overlaps Through Weight-Based Refinement** Despite iterative improvements to the RAG framework, misclassifications persisted due to inherent semantic overlaps between closely related categories. As illustrated in the attached embeddings plot (see Figure 2), keywords and synonyms associated with multiple categories often clustered together, making it challenging to distinguish them based on embeddings alone. This overlap caused certain terms—particularly those that appeared contextually relevant to multiple categories—to dilute the semantic clarity of the classification process, ultimately leading to inaccurate assignments.



(a) MBERT Embeddings For Entities

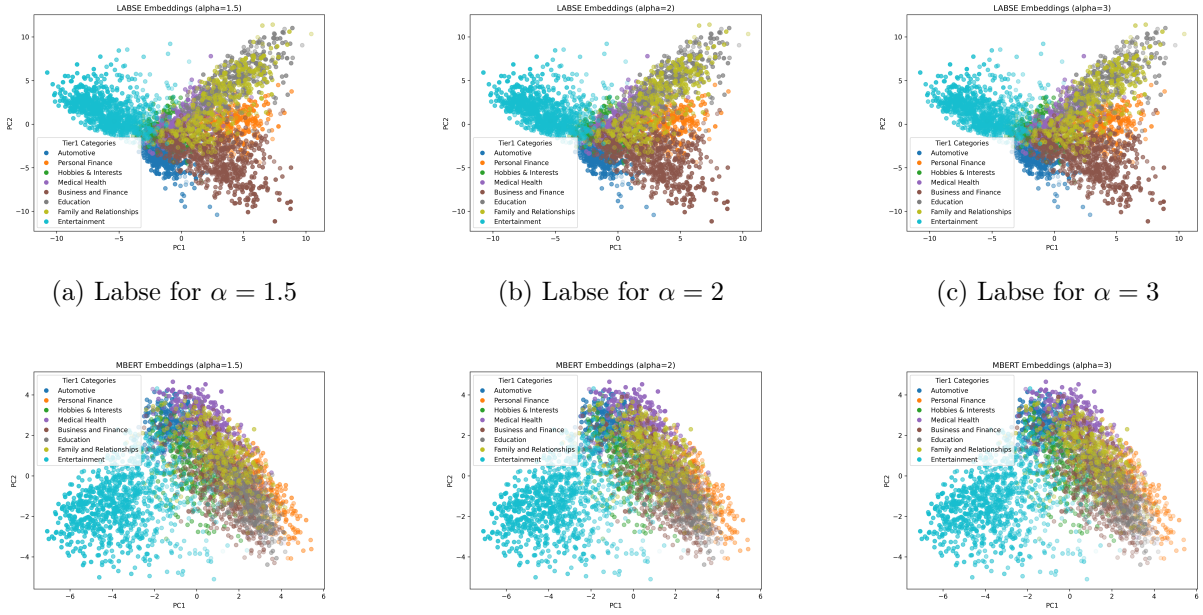
(b) LABSE Embeddings For Entities

Figure 2: Embeddings of Entities

To counter this issue, we introduced a weight-based methodology designed to quantify the "loyalty" of each node (keyword or synonym) to its respective category cluster. The core idea is that terms strongly aligned with a single category should receive a higher weight, while those that appear equally relevant to multiple categories should be penalized. Let S_{intra} denote the intra-cluster similarity of a node (its average similarity to terms within the same category), and let $S_{\text{inter,max}}$ represent the maximum inter-cluster similarity (the highest similarity to any other category's centroid). We define the node weight w as:

$$w = S_{\text{intra}} - \alpha \cdot S_{\text{inter,max}}$$

In this formula, α is a penalty factor controlling the degree to which high inter-cluster similarity reduces a node's weight. By experimenting with different α values (see Figure3), we chose $\alpha = 1.5$ for experimental terms. article graphicx subcaption



(a) Labse for $\alpha = 1.5$ (b) Labse for $\alpha = 2$ (c) Labse for $\alpha = 3$

(d) MBert for $\alpha = 1.5$ (e) MBert for $\alpha = 2$ (f) MBert for $\alpha = 3$

Figure 3: Embedding Plots For Different α Configurations

After computing the weights, we applied min-max normalization to ensure all values were on a consistent scale. If $\min(w)$ and $\max(w)$ represent the minimum and maximum observed weights, respectively, the normalized weight w_{norm} is:

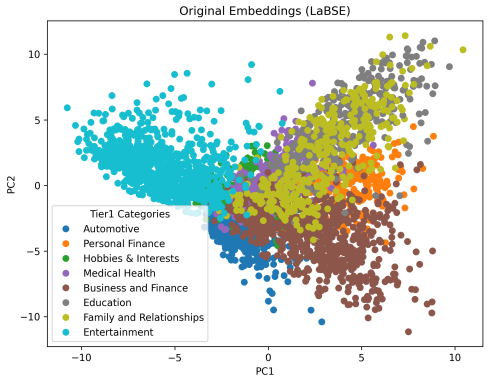
$$w_{\text{norm}} = \frac{w - \min(w)}{\max(w) - \min(w)}$$

Once normalized, we pruned nodes that fell below a chosen threshold ($w_{\text{norm}} < 0.7$). This step aimed to remove ambiguously placed nodes and sharpen the semantic boundaries within the knowledge graph. However, as shown in the attached visualization (Figure4) and subsequent accuracy results, aggressive pruning came at a cost. Specifically, after removing nodes below the 0.7 threshold, **Tier-1 accuracy dropped to 46% and Tier-2 accuracy to 30%,** reflecting a substantial loss of valuable semantic information. article graphicx subcaption

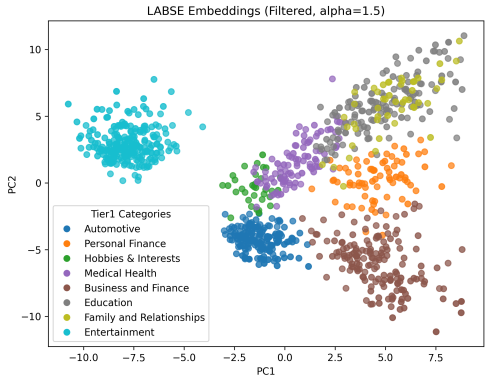
In summary, the weight-based refinement approach offered a structured means to mitigate semantic overlap issues, enabling us to identify and potentially remove nodes that contributed to misclassifications. Although there were trade-offs, this methodology provided critical insights into handling semantic ambiguity within hierarchical taxonomies.

4 Discussion

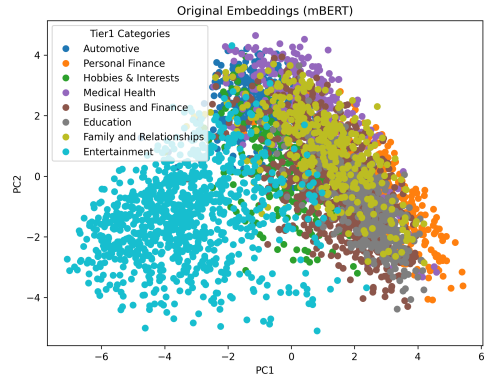
The results of our experiments underscore the potential of graph-based Retrieval-Augmented Generation (RAG) frameworks in navigating the complexities of hierarchical taxonomy classification. By capturing multi-level relationships within a Neo4j knowledge graph and leveraging context-aware embeddings, our approach addresses some of the limitations encountered in traditional classification methods, such as difficulty in distinguishing between closely related



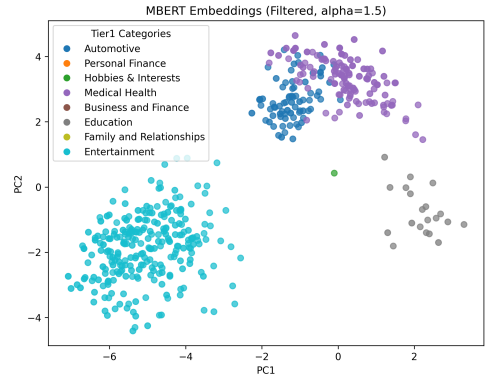
(a) Original Labse Embedding



(b) Labse Pruned



(c) Original MBert Embedding



(d) MBert Pruned

Figure 4: Figure to show information loss due to pruning

categories and handling large, multilingual datasets.

Despite these advances, several challenges remain. The observed misclassifications often stemmed from semantic overlaps between thematically adjacent categories, indicating the need for more robust regularization strategies. Our weight-based refinement and pruning introduced a structured way to mitigate these issues, yet the trade-off between semantic clarity and information loss remains delicate. Similarly, the integration of multilingual data complicated embedding alignment across languages, reinforcing the need for more sophisticated fine-tuning or domain-specific embeddings that can adapt to a wider range of linguistic contexts.

Future works include adding richer contextual triplets—such as entity-property-object relations—and integrating domain-specific corpora into the knowledge graph can further refine category distinctions, thereby reducing ambiguity and enhancing retrieval precision.

5 Conclusion

This research demonstrates the efficacy of integrating Neo4j with a Retrieval-Augmented Generation (RAG) framework to achieve scalable and accurate taxonomy classification. By modeling the IAB taxonomy hierarchically and incorporating LLM-based summarization, we addressed core challenges such as semantic overlaps and multilingual embedding alignment. Our results underscore that graph-driven approaches not only facilitate more nuanced retrieval but also enable flexible, fine-grained refinements through weight-based node management.
Looking ahead, the synergy between graph databases and advanced language models offers immense potential for further improving both the precision and adaptability of classification pipelines. As we extend the taxonomy to additional tiers and experiment with more sophisticated regularization, the methods proposed here can serve as a foundation for developing highly context-aware systems across a variety of domains and languages.

References

[1] IAB Tech Lab. (2021). IAB Tech Lab Content Taxonomy 2.2

[2] Lefranc, J., Risson, S. (2020). Using Machine Learning for IAB Category Classification.

[3] l-Olimat, H. S., Thirunarayan, K., Shalin, V. L., Sheth, A. (2018). A graph-based approach for context-aware advertisement recommendation using hierarchical content taxonomies.

[4] Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust.* New York, NY: Pantheon Books.

[5] Robinson, I., Webber, J., & Eifrem, E. (2015). *Graph databases: New opportunities for connected data* (2nd ed.). Sebastopol, CA: O'Reilly Media.

[6] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding.* In *Proceedings of NAACL-HLT* (pp. 4171–4186).

[7] Mitrovic, S., & Wray, S. (2021). *Applications of Neo4j in taxonomy classification systems.* Journal of Graph Databases and AI, *10*(3), 215–230.

[8] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). *Roberta: A robustly optimized BERT pretraining approach.* arXiv preprint arXiv:1907.11692.

[9] International Advertising Bureau. (2020). *IAB content taxonomy version 2.2.* Washington, DC: IAB.