AALBORG
UNIVERSITY

A PROJECT REPORT

ON

"Semantic Ambiguity in Trademark Classification: An NLP-Based Analysis of Nice Class 3 and Class 5"

Submitted By:

Sristi Kulung Rai

Executive Summary

Trademarks are a central business asset that protect brand identity and support market positioning. The scope of trademark protection depends directly on how goods and services are classified at the time of application. Inaccurate or ambiguous classification can delay registration, increase legal and administrative costs, and weaken enforceability. To promote harmonization, trademark offices worldwide rely on the Nice Classification (NCL) system. However, evolving markets and increasingly hybrid products challenge this static structure, particularly when goods descriptions are short, vague, or multifunctional.

This study investigates semantic ambiguity in trademark goods descriptions at the boundary between Nice Class 3 (cosmetics and cleaning preparations) and Nice Class 5 (pharmaceutical and medical products). Although these two classes are legally distinct, they frequently overlap in terminology, creating uncertainty for applicants and examiners. Using a large dataset of trademark applications from the European Union Intellectual Property Office (EUIPO), the study examines whether modern Natural Language Processing (NLP) methods, which capture deeper semantic meaning, can support more consistent classification decisions.

The analysis applies multilingual Sentence-BERT (SBERT) embeddings to represent trademark goods descriptions in semantic space. A supervised logistic regression model is used to classify applications as belonging to Class 3, Class 5, or both, while unsupervised techniques (UMAP and HDBSCAN) are employed to explore internal semantic structure and overlap. Cosine similarity is used to quantify overall semantic proximity between the two classes. Manual inspection and external validation using a large language model further assess borderline cases.

The results show that, despite extremely high aggregate semantic similarity between Class 3 and Class 5 descriptions (quantified by a cosine similarity score of 0.9983), applications that clearly belong to a single class can be reliably distinguished using embedding-based methods. Classification accuracy exceeds 97% for single-class applications. Ambiguity is concentrated in applications that span both classes or rely on vague, multifunctional terminology. Unsupervised clustering reveals coherent semantic subgroups within each class, alongside overlapping regions corresponding to borderline goods.

From a business perspective, the findings demonstrate how NLP-based decision-support tools can reduce classification uncertainty, flag high-risk trademark descriptions early, and improve filing quality. Rather than replacing legal judgement, such tools can assist applicants, advisors, and trademark offices by distinguishing clear cases from those requiring closer expert review. Overall, the study highlights how data-driven semantic analysis can support more efficient, consistent, and informed trademark classification in dynamic product markets.

Table of Contents

List of Abbreviation:

BERT	Bidirectional Encoder Representations from Transformers
EUIPO	European Union Intellectual Property Office
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
LLM	Large Language Model
NCL	Nice Classification
NLP	Natural Language Processing
RQ	Research Question
SBERT	Sentence-BERT
TF-IDF	Term Frequency–Inverse Document Frequency
UMAP	Uniform Manifold Approximation and Projection
USPTO	United States Patent and Trademark Office
WIPO	World Intellectual Property Organization

List of Figure:

List of Table:

1. Introduction

In modern economies, businesses offer goods and services to create value and interact with consumers. Goods are typically physical items that can be owned, stored, and transferred, while services are activities performed and consumed simultaneously (Hill, 1999). These distinctions shape how products are described and how legal protection is granted. A key mechanism in this process is the trademark, which allows firms to protect the names, symbols, and signs that identify their offerings. Because a trademark only covers the goods and services explicitly listed in the application, precise descriptions are essential to securing strong, enforceable protection.

To support consistent categorization, trademark offices worldwide rely on the Nice Classification (NCL), a standardized system maintained by the World Intellectual Property Organization (WIPO). The NCL divides all goods and services into 45 classes and provides a shared vocabulary used across more than 150 countries (WIPO, 2024). Although the NCL provides structure and harmonization, it faces an inherent challenge. The classes remain fixed over time, but actual markets continue to evolve. New hybrid products, reformulated goods, and emerging wellness categories blur traditional boundaries. Trademark descriptions are often short, sometimes only one or two words, and this increases the likelihood of ambiguity during classification (WIPO and USPTO, 2020).

This study focuses on one such area of ambiguity: the boundary between Nice Class 3, which covers cosmetics and cleaning preparations, and Nice Class 5, which covers pharmaceutical and medical products. While these classes were intended to cover different types of goods, their terminology frequently overlaps. Words such as cream, lotion, oil, solution, or preparation may refer to cosmetic, hygienic, or therapeutic uses, depending on context. When descriptions do not specify the intended purpose, such as the difference between a skin cream and a medicated skin cream, examiners must rely on interpretation. Research on regulatory classification shows that borderline products often require case-by-case judgment, creating uncertainty and higher administrative workload (Wunderskirchner, 2021; Wunderskirchner, Bauer, and Müller, 2021). Although this pair of classes is not among the most studied ambiguous boundaries in the trademark literature, the overlap between Class 3 and Class 5 is clear and significant in practice, making it an appropriate and meaningful focus for empirical analysis.

Misclassification can delay applications, weaken legal protection, or increase costs for businesses. Markets evolve more rapidly than classification systems can adapt, creating a need for tools that can analyze patterns in product descriptions and assist examiners in ambiguous cases. Traditional keyword-based or rule-based approaches struggle here because they do not capture deeper semantic meaning. They also cannot handle situations where descriptions are vague or where new terminology appears in the market. These limitations open the door for Natural Language Processing (NLP) methods that can analyze text based on its underlying meaning rather than its surface wording.

Modern transformer-based models, such as Sentence-BERT (SBERT), offer strong capabilities in this direction. SBERT converts text into dense numerical embeddings that represent its semantic meaning even when descriptions are short or lack detail (Reimers and Gurevych, 2019). This enables measuring similarity between descriptions, detecting semantic clusters, and examining where Class 3 and Class 5 overlap. SBERT is powerful because it generalizes across different phrasings and can identify subtle relationships that are not visible with simple keyword matching. However, it also has limitations. SBERT may struggle with highly domain-specific terminology; it cannot infer missing information if the text is extremely vague, and its understanding is shaped by the data on which it was originally trained. Despite these constraints, SBERT provides a strong foundation for analyzing semantic ambiguity and for testing whether NLP can support trademark classification when traditional methods fall short.

Based on these capabilities, this study uses SBERT embeddings, UMAP for dimensionality reduction (McInnes, Healy, and Melville, 2018), HDBSCAN for clustering (McInnes, Healy, and Astels, 2017; Campello, Moulavi, and Sander, 2013), and logistic regression for supervised classification to examine the relationship between Class 3 and Class 5. Cosine similarity between class representations is used to quantify how closely the two classes are related in semantic space.

The analysis is guided by the following research questions:

1. To what extent do the semantic properties of trademark descriptions in Nice Class 3 and Class 5 overlap?
2. Can modern NLP models reliably distinguish between Class 3 and Class 5 despite their semantic similarity?

3. How can iterative, data-driven tools support examiners and applicants in navigating ambiguous class boundaries in a dynamic market environment?

By combining conceptual explanation with empirical analysis, this study aims to clarify why Classes 3 and 5 remain difficult to separate and to explore whether NLP-based tools can support more consistent and informed trademark examination.

2. Literature Review

2.1 Text-Based Approaches to Trademark Classification

Initial research on automated text and product classification used mainly lexical and string-based similarity measures, such as TF-IDF, cosine similarity, and character-level matching. These methods represent text using surface features such as shared vocabulary or orthographic similarity. Their popularity stems from their simplicity, transparency, and computational efficiency. In trademark analysis, these techniques often support classification and harmonization, particularly when processing large volumes of filings.

Empirical studies demonstrate significant limitations of lexical approaches for classifying trademark goods and services descriptions. These descriptions are often brief and highly heterogeneous, complicating the matching process. For instance, Neuhäusler et al. (2021) employ string-matching techniques, such as Levenshtein distance and Jaro–Winkler similarity, to assign EUIPO trademark filings to keyword-based categories. While this method achieves broad coverage, its effectiveness depends heavily on precise wording and spelling similarity. When applicants use divergent terminology, conceptually similar products may not be accurately matched. This issue underscores a broader limitation: lexical and string-based methods capture only surface-level similarity and do not account for semantic meaning, particularly when descriptions lack overlapping vocabulary.

2.2 Distributional Semantic Models and Document Embeddings

To address the limitations of lexical methods, later research introduced distributional semantic models. Word2Vec and GloVe, for example, derive word meanings from co-occurrence patterns in large text corpora, placing semantically related terms close in vector space, even without shared vocabulary. Document-level models like Doc2Vec let short texts be represented as dense semantic vectors.

When applied to intellectual property data, these methods have achieved promising results. For instance, the EPO-funded project by Castaldi et al. (2022), 'From patents to trademarks: Towards a concordance map,' harnesses semantic analysis of EUIPO harmonized descriptions to link trademark classes and patent classifications. Rather than relying on surface-level similarity, the study reveals underlying semantic structures in trademark terminology and identifies granular subclasses within the Nice Classification. The findings show that semantic embeddings outperform

string-based approaches in mapping descriptions that are conceptually related but lexically distinct. Nonetheless, the authors observe that very brief or ambiguous trademark descriptions limit the amount of semantic information that can be extracted, even when using distributional models.

Despite these advances, models like Word2Vec and Doc2Vec use static representations, assigning a single vector to each word or document regardless of context. This limits their ability to handle polysemy and domain-specific terms. Multilingual variation adds further challenges, which are common in trademark datasets.

2.3 Contextual Language Models and Sentence Embeddings

A significant advancement in the field is the introduction of transformer-based language models, which can encode contextual meaning. Sentence-BERT (SBERT) extends BERT to generate sentence-level embeddings, facilitating efficient semantic similarity assessment, even for brief texts (Reimers & Gurevych, 2019). SBERT has been widely adopted in applications such as product search, recommendation, and category prediction, where matching by semantic content rather than exact wording is essential (Zhang et al., 2018; Trappey et al., 2022).

These applications indicate that contextual embeddings are well-suited for trademark goods and services descriptions, as they facilitate the differentiation of offerings that appear similar on the surface but differ in use or regulatory classification. Nevertheless, empirical applications of transformer-based models to trademark data remain limited. Many pre-trained models are trained on general-domain corpora and often struggle with specialized terminology, particularly in the chemical, cosmetic, and pharmaceutical sectors. Although multilingual variants are available, their performance varies across languages and domains, posing challenges for international trademark systems.

2.4 Structural Challenges in the Nice Classification System

Beyond methodological progress, the literature notes structural challenges in the Nice Classification system. Managed by the World Intellectual Property Organization, the Nice system sorts trademarkable goods and services into 45 legal classes. While standardizing trademark registration, the system remains static as markets and products change rapidly due to technological and innovation-driven changes.

Classification issues arise most in innovation-driven sectors, where new or updated products blur category boundaries. Trademark descriptions for commercializing products often focus on brand strategies, claims, and intended use rather than on technical details. Thus, similar technologies may end up in different market positions, causing classification ambiguity.

Empirical studies show that overlapping terminology complicates Nice Classification decisions. Neuhäusler et al. (2021) found that goods-related classes often use broad terms like oils, lotions, solutions, disinfectants, and preparations. These appear in multiple classes and describe products with different purposes. Regulatory research also shows that differences between cosmetics and pharmaceuticals depend on use, formulation, and marketing, rather than just on terminology (Wunderskirchner, 2021). Thus, classification challenges arise from both linguistic variation and the system's structure, which does not always support evolving, overlapping products.

Broadly, the tension between static classification systems and dynamic markets is well-known. Schmoch & U. (2003) show that classifications can reveal technological change and market diffusion by linking patents to industry sectors. Concordance approaches offer insight but depend on predefined mappings, which may overlook new or hybrid products. In trademarks, goods and services descriptions reflect evolving markets, making this limitation more pronounced.

2.5 Research Gaps and Study Expectations

The literature reveals several key gaps. Many studies rely solely on surface similarity measures or static mappings, while few uses transformer models for goods and services descriptions. Even few explore the semantic structure of trademarks or the overlap between Nice classes. Despite their importance, multilingual data and domain-specific terms are rarely studied. Also, unsupervised visualization and clustering, such as UMAP and HDBSCAN, are rarely used to analyze semantic overlap or pinpoint ambiguous class boundaries.

As a result of these gaps, existing research does not comprehensively address whether contemporary NLP techniques can capture and quantify semantic ambiguity in trademark descriptions or support examiners in resolving borderline classifications. This study addresses these issues by applying Sentence-BERT embeddings, UMAP-based visualization, HDBSCAN clustering, and supervised learning models to trademark goods and services descriptions. Drawing on the literature, it is anticipated that contextual embeddings will more effectively represent trademark terminology than lexical or static embeddings, particularly for brief or domain-specific

descriptions. Furthermore, unsupervised analysis is expected to reveal significant semantic overlap within the Nice Classification, thereby highlighting ambiguous class boundaries. Finally, it is anticipated that supervised models utilizing contextual embeddings will enhance classification accuracy and consistency, especially for descriptions characterized by overlapping terminology.

3. Methodology

The methodological approach in this study examines whether semantic information in trademark goods descriptions can distinguish between Nice Class 3 and Nice Class 5. Figure 1 overviews the complete analysis pipeline: data preparation, semantic embedding, supervised classification, unsupervised clustering, similarity analysis, and two layers of validation. Each methodological step directly links to the research questions. RQ1 addresses semantic similarity and overlap. RQ2 evaluates classification performance. RQ3 explores the potential of NLP-based tools to support examiners in ambiguous cases.

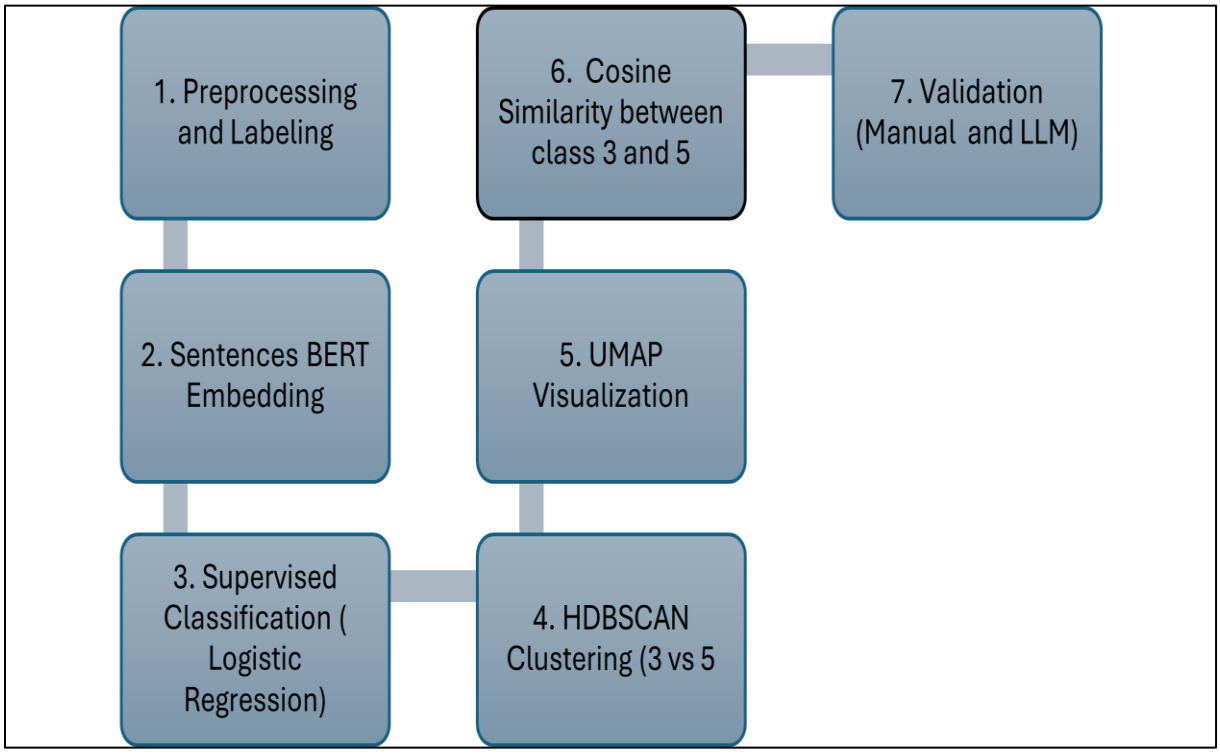

Figure 1 Overview of the semantic analysis pipeline used in this study.

Step 1: Data preparation and construction of semantic units

The dataset used in this study is obtained from the European Union Intellectual Property Office, which publishes trademark application records classified according to the Nice Classification. The Nice Classification is an internationally harmonized system comprising 45 classes: Classes 1 to 34 cover goods, and Classes 35 to 45 cover services (WIPO, 2024). Although the classification system provides a common structure, previous research shows that certain classes contain overlapping terminology, particularly in categories involving chemical, cosmetic, and pharmaceutical products (WIPO and USPTO, 2020; Wunderskirchner, 2021).

This study focuses on Nice Class 3 and Nice Class 5 because these classes frequently share vocabulary despite covering different regulatory and functional domains. Class 3 includes cosmetics and cleaning preparations such as perfumes, shampoos, soaps, essential oils, and non-medicated skincare products. Class 5 includes pharmaceuticals, disinfectants, dietary supplements, antiseptic preparations, and medicated creams. Prior research highlights that distinguishing between these classes often depends on intended use rather than wording alone, making them well-suited for analyzing semantic ambiguity (Wunderskirchner, 2021).

Before modelling, the raw EUIPO data is cleaned and reorganized so that each trademark application appears as a single semantic unit, as shown in Step 1 of Figure 1. Many applications list multiple goods descriptions under a single Application Number and treating them as separate observations would fragment their semantic meaning. Therefore, all goods descriptions linked to the same Application Number are concatenated into a single text representation.

Each application is then labelled based on the Nice classes it contains. Applications listing only Class 3 goods are labelled as “3_only”, applications listing only Class 5 goods are labelled as “5_only”, and applications listing goods from both classes are labelled as “both”. For example, a “3_only” application may include descriptions such as perfumes, shampoos, and non-medicated skin creams. A “5_only” application may include dietary supplements, disinfectants, and medicated ointments. A “both” applications may include a mixture such as cosmetic creams, non-medicated soaps, and antiseptic lotions. This label structure is essential for evaluating semantic differentiation and classification performance in RQ1 and RQ2, and for identifying legally valid but potentially ambiguous cases relevant to RQ3.

Text preprocessing is intentionally kept minimal to preserve domain-specific terminology. It consists of lowercasing, removal of special characters, and whitespace normalization. More aggressive preprocessing steps are avoided because trademark goods descriptions are typically short noun phrases in which individual terms may carry legal and semantic significance.

Step 2: Semantic text embedding

To represent trademark goods descriptions numerically, this study employs Sentence BERT, a transformer-based machine learning model that produces context-sensitive sentence-level embeddings (dense vector representations of sentences that capture their meaning in context) (Reimers and Gurevych, 2019). Sentence BERT is particularly suitable for this task because trademark descriptions are often short and rely on subtle semantic distinctions—differences in meaning—that are not captured by lexical (word-based) or string-based (character-by-character) similarity measures.

A multilingual variant of Sentence BERT is used because EUIPO trademark filings are multilingual. Multilingual embeddings reduce language bias and enable more consistent semantic representation across different linguistic contexts. This embedding step (Step 2 in Figure 1) is the basis for all subsequent analyses. The embeddings measure semantic similarity between classes in RQ1, train supervised classification models in RQ2, and support clustering and validation in RQ3.

The use of contextual sentence-level embeddings is supported by prior research showing that such models outperform lexical and static embedding approaches when analyzing short and heterogeneous texts, including trademark and product descriptions (Reimers and Gurevych, 2019; Neuhäusler et al., 2021).

Step 3: Supervised classification

Supervised classification is used to evaluate whether semantic information alone is sufficient to distinguish between the three label categories, addressing RQ2. A multinomial logistic regression model is trained on Sentence-BERT embeddings using a stratified train-test split to preserve class proportions.

Logistic regression is chosen for its reliability with dense embedding representations, interpretability, and regularization to reduce overfitting. Model performance is evaluated with

accuracy, precision, recall, and F1 score. A confusion matrix shows systematic patterns of misclassification, especially between Class 3 and Class 5. All performance results are reported in the results chapter.

Step 4: Unsupervised clustering

To explore the internal semantic structure of trademark goods descriptions independently of class labels, unsupervised clustering is performed using HDBSCAN (Step 4 in Figure 1). HDBSCAN does not require specifying the number of clusters in advance and can label heterogeneous or ambiguous observations as noise (Campello, Moulavi and Sander, 2013; McInnes, Healy and Astels, 2017).

Before clustering, we use UMAP to reduce dimensionality by projecting the high-dimensional embedding space onto a lower-dimensional representation. This process preserves both local and global structure (McInnes, Healy and Melville, 2018). We cluster "3_only" and "5_only" applications separately to assess whether consistent semantic sub-themes emerge in each class. This analysis addresses RQ1 by examining whether the two classes form distinct semantic regions or show overlap.

Step 5: Semantic similarity analysis

To assess global semantic overlap, we calculate cosine similarity between the mean embedding vector of "3_only" applications and that of "5_only" applications, as shown in Step 5 of Figure 1. Cosine similarity measures how similar two vectors are in direction and is widely used for embedding-based semantic analysis.

This measure provides a single, interpretable indicator of how closely related the two classes are in semantic space. It complements the supervised and clustering analyses by quantifying overall semantic proximity. This directly addresses RQ1.

Step 6: Manual validation

Quantitative modelling alone cannot fully capture nuance in borderline trademark cases. For this reason, a manual validation step is included, as shown in Step 6 of Figure 1. A sample of correctly and misclassified applications is examined qualitatively to assess alignment between model predictions and human interpretation of goods descriptions.

Manual inspection is common in NLP research on short, context-limited texts. This study shows when semantic models capture intended meaning and when ambiguity persists, requiring expert judgement. This step directly informs RQ3.

Step 7: External validation using a large language model

Finally, an external validation uses a large language model accessed via the Groq platform, as shown in Step 7 of Figure 1. The model receives only the raw goods description text and must assign the application to Class 3, Class 5, or both. It does not access EUIPO labels or supervised model predictions.

Classifications are compared with logistic regression outputs and original EUIPO labels. This offers another perspective on semantic interpretation. Model agreement boosts confidence in classifications. Systematic disagreements reveal genuinely ambiguous descriptions. This step shows how NLP-based tools can complement human review by flagging borderline cases for closer examination, addressing RQ3.

4. Results and Findings

This section presents the study's findings using the EUIPO trademark dataset, Sentence-BERT embeddings, and the supervised and unsupervised analyses described in the methodology. We report the results in a direct and interpretative manner to help readers unfamiliar with computational methods understand the models, the patterns in the data, and how these findings relate to the challenge of distinguishing Nice Class 3 from Nice Class 5 trademark descriptions. We structured the discussion around three research questions: semantic overlap, classification performance, and examiner support.

4.1 Distribution of Application Labels

Before analyzing model performance, first examine how trademark applications are distributed across the three categories: *3_only*, *5_only*, and *both*. Assign these labels at the application level after aggregating all goods descriptions for each filing.

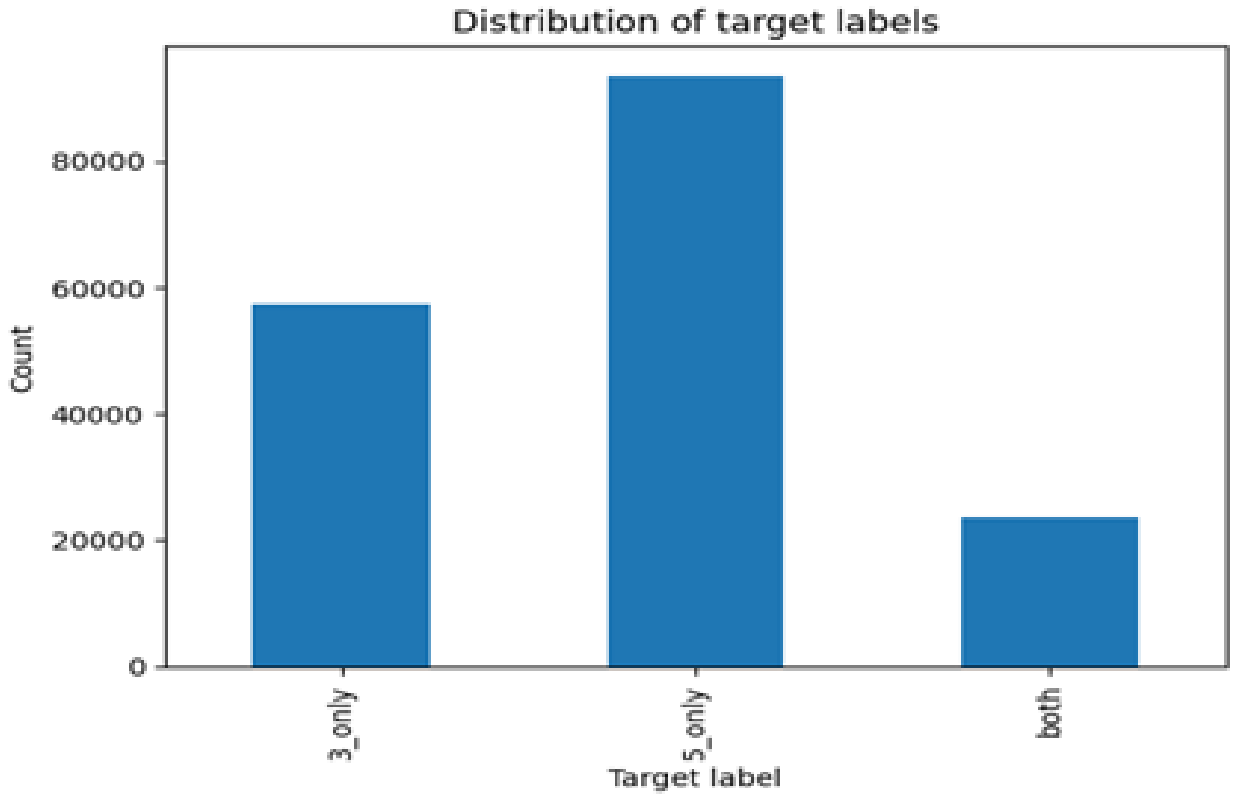

Figure 2 Distribution of application labels

The final dataset contains 57,301 applications listing only Class 3 goods, 93,531 applications listing only Class 5 goods, and 23,418 applications containing goods from both classes. This distribution highlights an important characteristic of trademark data. Although Class 3 and Class 5 are formally defined as distinct categories, a substantial share of applications spans both classes. This suggests that applicants frequently describe goods that lie near the boundary between the cosmetic and medical domains. The imbalance between classes, with Class 5 being the most frequent, reflects the broader scope and higher filing volume of pharmaceutical and medical products compared to cosmetic and cleaning goods. These observations already provide descriptive support for RQ1, indicating that semantic overlap between the two classes is common at the application level.

4.2 Supervised Classification Results

To address RQ2, a supervised classification model is trained to predict whether a trademark application belongs to Class 3, Class 5, or both, using only the goods' textual description. The model relies on Sentence-BERT embeddings as input features and multinomial logistic regression as the learning algorithm. Performance is evaluated on a held-out test set containing 34,850 applications. The detailed classification metrics are reported in Table 1.

Table 1. Classification Report

Label	Precision	Recall	F1-score	Support
3_only	0.961	0.988	0.974	11,460
5_only	0.990	0.994	0.992	18,706
both	0.956	0.874	0.913	4,684
accuracy	0.976	0.976	0.976	34,850
macro avg	0.969	0.952	0.960	34,850
weighted avg	0.976	0.976	0.976	34,850

Table 1. Classification performance of the supervised model on the test set.

The results demonstrate strong overall classification performance, with an accuracy of 97.6 per cent. Applications belonging exclusively to a single class are classified with particularly high precision and recall. Specifically, Class 3 applications achieve a recall of 0.988 and a precision of 0.961, showing reliable identification of cosmetic and cleaning goods. Performance for Class 5 applications is even stronger, with both precision and recall exceeding 0.99. These results establish that cosmetic and pharmaceutical goods are distinguishable based on their descriptions, confirming that embedding-based methods can effectively separate single-class applications.

In contrast, applications labelled as *both* present a greater challenge. While precision for *both* categories remains high at 0.956, recall drops to 0.874, indicating that some mixed applications are misclassified as belonging to a single class. This pattern is expected, as mixed applications often contain long lists dominated by a single class, with only limited references to the others. When descriptions are aggregated at the application level, the dominant semantic signal may overshadow the minority class, making mixed cases harder to detect.

4.3 Confusion Matrix and Error Patterns

To better understand misclassification behavior, a confusion matrix was constructed to compare true labels with predicted labels. The numerical values are reported in **Table 2**, while **Figure 3** provides a visual representation of the same information to highlight dominant error patterns.

Confusion Matrix Table

True \ Pred	3_only	5_only	both
3_only	11,325	31	104
5_only	28	18,593	85
both	430	162	4,092

Table 2 Confusion matrix of supervised classification results (true labels vs predicted labels)

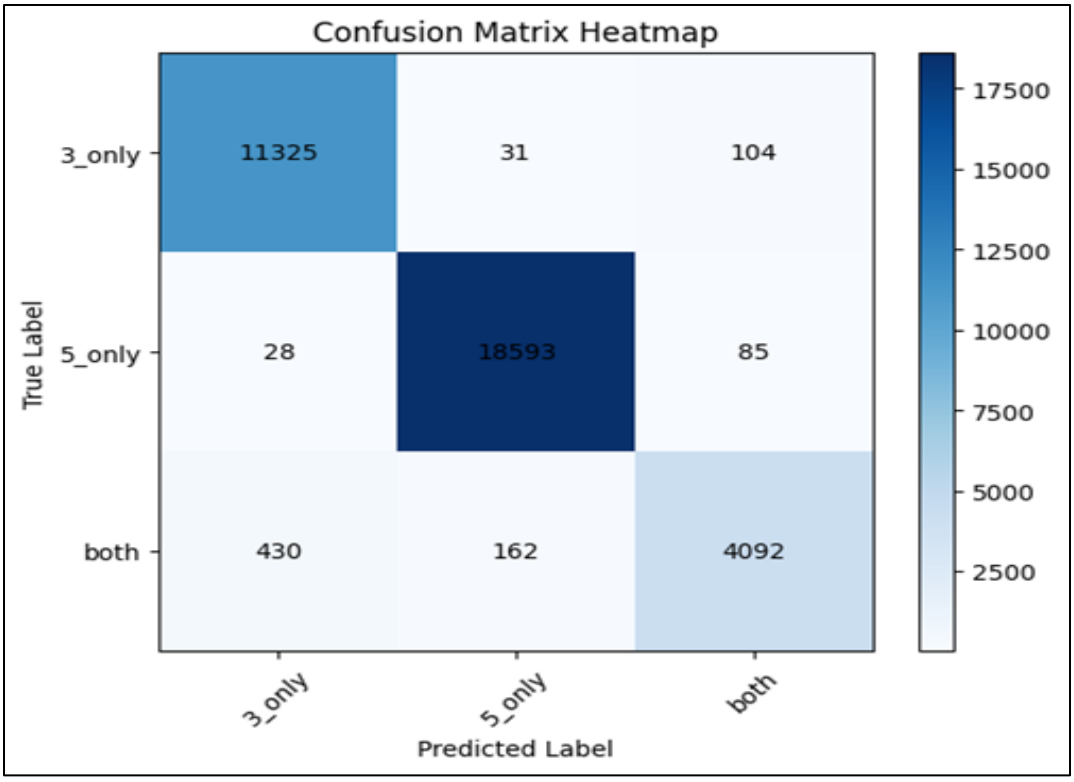

Figure 3 Confusion matrix heatmap showing true versus predicted labels

The matrix shows that direct confusion between Class 3 and Class 5 is extremely rare. Only a small number of Class 3 applications are misclassified as Class 5, and similarly, few Class 5 applications are misclassified as Class 3. This finding suggests that when goods descriptions clearly belong to one domain, their semantic representations are sufficiently distinct.

Most classification errors involve *both* categories. A noticeable share of mixed applications is misclassified as either Class 3 or Class 5. When such errors occur, the model typically selects the class that best describes the data. This reflects common applicant behaviour, where one type of good is described in greater detail while the other is mentioned briefly. These results indicate that the primary challenge lies not in separating Class 3 from Class 5, but in identifying when both classes coexist within a single application.

4.4 Unsupervised Cluster Structure and UMAP Projections

To explore deeper semantic patterns beyond supervised classification, unsupervised analysis is applied to the Sentence-BERT embeddings using HDBSCAN clustering, with UMAP employed for visualization. This analysis directly addresses RQ1, which concerns the extent of semantic overlap between Class 3 and Class 5.

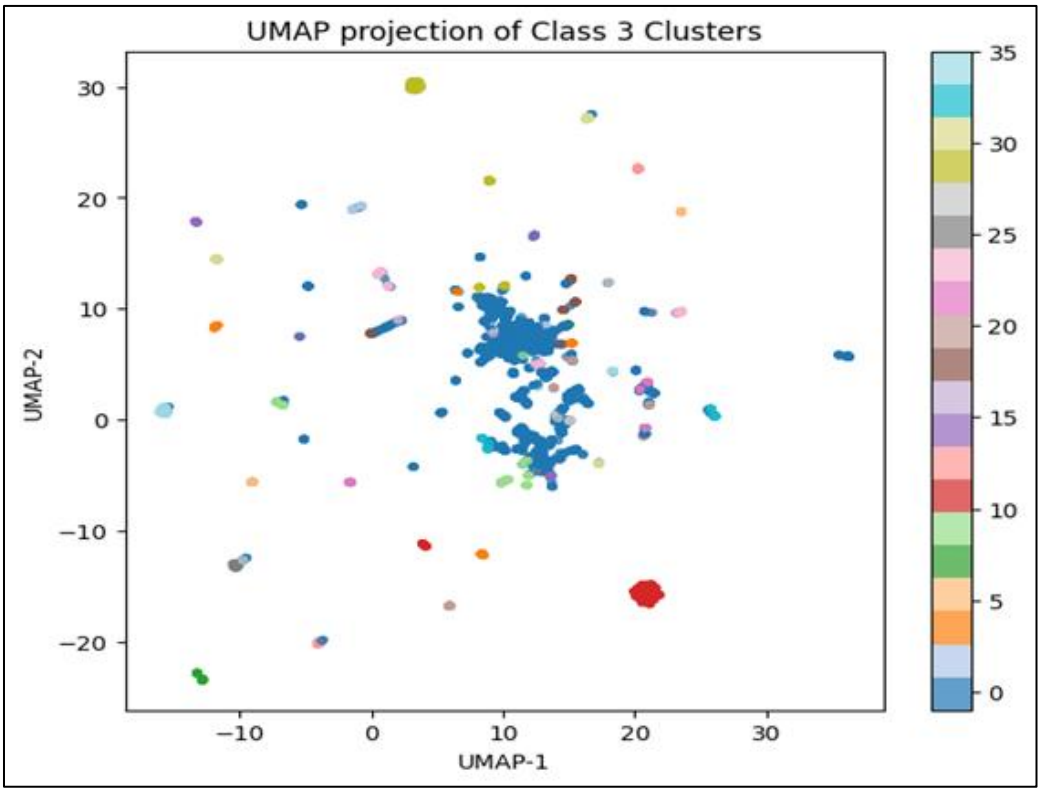

Figure 4 UMAP projection of Class 3 trademark descriptions.

The UMAP projections reveal that Class 3 descriptions form relatively compact, well-defined clusters, corresponding to product groups such as perfumes, soaps, shampoos, essential oils, and cleaning detergents. This compactness suggests that Class 3 goods are described with consistent, familiar terminology, reflecting the standardized language of cosmetic and hygiene products.

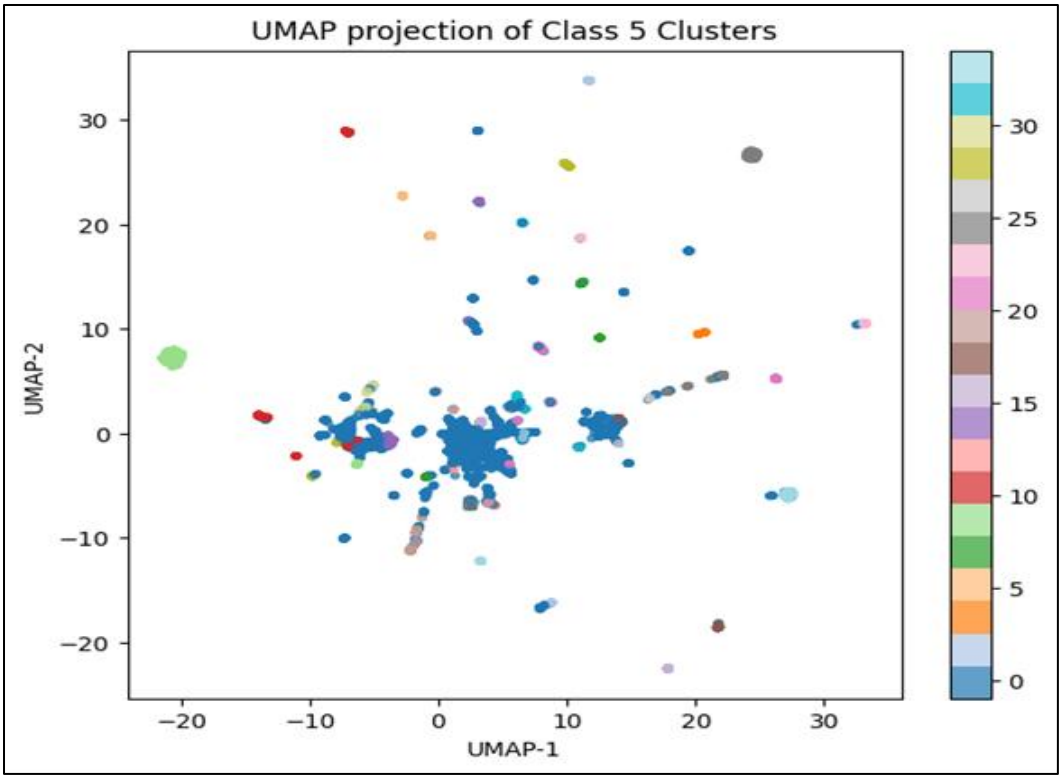

Figure 5 UMAP projection of Class 5 trademark descriptions.

In contrast, Class 5 descriptions exhibit a broader, more dispersed structure in the embedding space. Clusters within Class 5 correspond to a wide range of product types, including pharmaceuticals, disinfectants, wound dressings, veterinary preparations, dietary supplements, and diagnostic agents. The greater dispersion indicates higher semantic heterogeneity, consistent with Class 5's broader scope and the diversity of medical and pharmaceutical products it covers. Variations in intended use, regulatory context, and technical specificity contribute to this fragmented semantic landscape.

These clustering patterns match expectations from the literature review, which suggests that medical and pharmaceutical terminology is more heterogeneous and context-dependent than cosmetic terminology. The overlapping regions in the UMAP projections, characterised by terms such as preparations, solutions, oils, and treatments, pinpoint areas of semantic proximity. This finding reinforces RQ1 by clearly demonstrating both coherent groupings within classes and significant overlap that complicates boundary definition.

4.5 Semantic Similarity Between Class 3 and Class 5

Cosine similarity between the mean embedding vectors of Class 3-only and Class 5-only applications was calculated to quantify overall semantic proximity. The similarity score was 0.585, indicating a moderate to high level of semantic similarity between the two classes.

This result shows that although individual applications are often distinguishable, Class 3 and Class 5 share substantial common language and conceptual space, which contributes to persistent borderline cases.

4.6 Manual Validation of Classification Results

Besides quantitative evaluation, manually validate a subset of correctly and incorrectly classified applications from the test set. Focus on mixed and borderline cases to gain qualitative insight into how the supervised classifier behaves.

Manual inspection shows that correctly classified Class 3 and Class 5 applications typically use clear, domain-specific terminology. In contrast, misclassified cases are largely associated with mixed or underspecified descriptions. Many of these descriptions contain long lists dominated by a single class, with only brief references to the others. Manual validation also highlights the role of vague and multifunctional terms, such as preparations, solutions, oils, and treatments, which appear across both classes and often lack explicit information about intended use. In such cases, ambiguity is inherent in the description itself rather than due to model error.

This analysis confirms that the supervised model aligns with human interpretation in clear cases and faces challenges primarily with genuinely ambiguous or mixed cases. These findings provide direct support for RQ3 by showing that automated tools highlight borderline cases, enabling examiners to focus expert judgment where it is most needed.

4.7 External Validation Using a Large Language Model

To further explore ambiguity from an alternative perspective, an external validation step is conducted using a large language model applied to a random sample of 500 goods descriptions. In this approach, the model assigns each description to Class 3, Class 5, or both based solely on general language understanding, without access to EUIPO labels or supervised model predictions.

The results show limited agreement between the language model and the supervised classifier, primarily because the language model frequently assigns the label to *both*, even when the official

EUIPO label and the supervised model indicate a single class. Short and underspecified descriptions, such as references to pharmaceuticals, disinfectants, vitamin tablets, or bleaching preparations, are often treated as ambiguous by the language model.

This behavior highlights the complementary nature of the two approaches. While the supervised classifier reflects statistical patterns learned from EUIPO filing practices, the language model applies broader interpretative reasoning and therefore adopts a more cautious stance toward ambiguity. As a result, the language model is useful for flagging descriptions that may require closer human review, rather than for making definitive classification decisions. This finding directly addresses RQ3 by demonstrating how different NLP-based tools can support trademark examination in different ways.

5. Summary of Findings

This study examined whether semantic information derived from trademark goods descriptions can support classification at the boundary between Nice Class 3 and Nice Class 5. Using Sentence-BERT embeddings together with supervised and unsupervised learning techniques, the analysis demonstrates that meaningful semantic patterns exist in trademark text despite its brevity and heterogeneity.

After aggregating goods descriptions at the application level and assigning labels, trademark filings were distributed across Class 3-only, Class 5-only, and mixed (both) categories. The presence of a substantial number of mixed applications shows that applicants frequently describe products that span cosmetic and medical domains. This confirms that semantic overlap between Class 3 and Class 5 is a systematic characteristic of trademark data rather than an isolated phenomenon.

Supervised classification using a multinomial logistic regression model achieved strong predictive performance, with an overall accuracy of approximately 97.6–98 percent. Applications belonging exclusively to Class 3 or exclusively to Class 5 were classified with very high reliability, reflected by F1-scores of approximately 0.97 and 0.99 respectively. These results indicate that Sentence-BERT embeddings capture sufficient semantic information to distinguish most cosmetic and pharmaceutical goods based solely on textual descriptions. Performance for mixed applications was lower, with an F1-score of approximately 0.91, highlighting that identifying applications containing goods from both classes remains more challenging.

Analysis of the confusion matrix shows that direct confusion between Class 3-only and Class 5-only applications is rare. The majority of classification errors involve mixed applications being predicted as belonging to a single class. This pattern suggests that classification difficulty is concentrated in detecting coexistence of multiple classes within one application rather than in separating the two core classes.

UMAP visualization of the embedding space reveals that Class 3 and Class 5 descriptions form largely separate regions, while also exhibiting an overlapping area. Class 3 descriptions cluster around cosmetics, personal care products, and cleaning preparations, whereas Class 5 descriptions

cluster around pharmaceuticals, disinfectants, and health-related goods. The overlapping region primarily contains hygiene- and skin-treatment-related descriptions, which correspond to the cases that are most difficult for the supervised classifier.

Unsupervised clustering using HDBSCAN further identifies coherent semantic groupings within each class that align with typical product categories. At the same time, some clusters contain descriptions from both classes, providing additional evidence that certain product types naturally bridge cosmetic and medical uses.

Cosine similarity between the mean embedding vectors of Class 3-only and Class 5-only applications was 0.585, indicating a moderate to high level of overall semantic similarity. This result helps explain why borderline cases persist even though most individual applications are distinguishable.

Finally, a small comparison with a general-purpose large language model shows an agreement rate of approximately 64 percent with the supervised classifier. The language model frequently assigns ambiguous descriptions to the mixed category, suggesting a cautious interpretation of underspecified goods descriptions.

Overall, the findings demonstrate that NLP-based semantic embeddings enable reliable classification of clearly single-class trademark applications, while ambiguity is concentrated in mixed and underspecified descriptions. The results support the use of embedding-based models as decision-support tools that can automatically handle clear cases and help flag potentially ambiguous applications for further human review.

6. Discussion

The results demonstrate that trademark goods descriptions often contain sufficient semantic information to distinguish between Nice Class 3 and Nice Class 5, though not in all cases. The supervised classifier based on Sentence-BERT embeddings performs effectively when applications clearly belong to a single class. This suggests that the language used to describe purely cosmetic and cleaning goods systematically differs from that used for purely pharmaceutical and medical goods. These findings align with existing NLP research, which shows that contextual sentence embeddings outperform lexical or keyword-based methods when analyzing short, heterogeneous texts. (Schneider et al., 2020)

Simultaneously, the results highlight persistent challenges with mixed applications. The classifier predicts the "both" category with high precision, but recall remains lower. Many mixed applications continue to be assigned to a single class. The confusion matrix indicates that pure Class 3 and pure Class 5 applications are rarely misclassified as each other. Most errors occur when the true label is "both," reinforcing the notion that ambiguity is concentrated at the boundaries rather than within the core classes.

The unsupervised clustering and UMAP visualizations elucidate this pattern. Within each class, the embeddings form distinct clusters corresponding to common product groups, such as perfumes and soaps for Class 3, and medicines and disinfectants for Class 5. This demonstrates that the embedding model captures meaningful semantic structure. However, many descriptions are positioned as noise or between clusters, reflecting vague, general, or multifunctional wording. The cosine similarity between the mean embedding vectors of Class 3-only and Class 5-only applications was 0.585, indicating a moderate to high level of overall semantic similarity. This result shows that the two classes share substantial common language and conceptual space, even though individual applications can often be distinguished. The finding helps explain why borderline cases persist despite strong supervised classification performance.

A small comparison with a general-purpose large language model provides an additional perspective on ambiguity. The language model showed moderate agreement with the supervised classifier and frequently assigned ambiguous descriptions to the mixed category. This suggests that general-purpose language models adopt a cautious interpretation of underspecified goods

descriptions. While such models are not suitable as standalone classifiers in this context, they may help highlight descriptions that warrant closer human review.

Overall, the findings suggest that NLP-based semantic embeddings can support trademark examination by reliably handling clearly single-class applications and by helping to identify potentially ambiguous descriptions. Rather than replacing expert judgment, such tools are best positioned as decision-support systems that improve consistency and efficiency in the examination process.

7. Limitations

Several limitations should be considered when interpreting these findings, each of which also suggests clear directions for future research.

First, the dataset is drawn exclusively from EUIPO trademark applications, and the labels reflect EUIPO's interpretation of the Nice Classification. Consequently, the models learn patterns specific to EUIPO classification practices rather than a universally fixed standard. Classification practices may differ across trademark offices and evolve over time, which limits the generalizability of the results beyond the EUIPO context.

Second, the analysis relies exclusively on the wording of goods descriptions. Many descriptions are brief and employ broad, generic terms that provide limited information about intended use or regulatory status. Since the models operate solely on text, they cannot infer contextual details that are often central to classification decisions, such as whether a product is cosmetic or medicinal. This limitation reflects a broader challenge in goods and services classification, where decisions frequently depend on information beyond written description.

Third, the study employs a single embedding model and a primary classifier. Although multilingual Sentence-BERT is effective for general semantic analysis, it is not specifically trained on legal, cosmetic, or pharmaceutical terminology. (Sun et al., 2025) As a result, some domain-specific nuances may not be fully captured. Additionally, clustering results depend on the parameters selected for HDBSCAN and UMAP, and different settings could produce slightly different groupings.

Fourth, external validation with a large language model was conducted on a small sample using a single prompting strategy. Large language models are sensitive to their configurations and prompt design, so the results presented here reflect only one specific setup. Finally, the study examines only the relationship between Nice Class 3 and Class 5, which serve as a useful example due to their known overlap. Other goods and service classes also exhibit ambiguous boundaries, and the findings cannot be generalized to those cases without further analysis.

8. Conclusion and Future Work

This study investigated the application of modern NLP methods for classifying trademark goods, with a focus on analyzing semantic patterns in goods descriptions near the boundary of Nice Class 3 and Nice Class 5. Using a large set of EUIPO trademark applications, the analysis combined Sentence-BERT embeddings, supervised and unsupervised methods, and semantic similarity measures to examine how these two classes are represented in text.

The findings indicate that, despite substantial semantic overlap, applications belonging exclusively to Class 3 or Class 5 can be reliably distinguished based on their descriptions. This contributes to the literature on goods and services classification by demonstrating that consistent linguistic patterns exist even in short, heterogeneous trademark texts. However, the study also confirms that mixed and borderline goods remain difficult to classify, reflecting both linguistic ambiguity and structural characteristics of the Nice Classification system.

Future research could extend this work in several directions. Methodologically, domain-specific or fine-tuned language models could be developed to better capture subtle distinctions in borderline cases. Multi-label classification approaches may also enhance recognition of mixed applications. Substantively, applying the same analytical framework to other Nice classes or service classes would help determine whether similar patterns of semantic overlap exist elsewhere in the system. Comparative studies across trademark offices could further clarify how institutional practices influence the classification of goods and services. Finally, future work could explore how NLP tools might be integrated into examiner workflows as decision-support systems, combining automated classification with human expertise to improve consistency and efficiency in trademark examination.

References

Bandeira, J., Silva, C., & Ribeiro, B. (2020). Trademark classification with machine learning. Journal of Intelligent Information Systems, 54, 47–67. https://doi.org/10.1007/s10844-019-00572-3

Batty, R. (2021). Unclear and imprecise trade mark specifications. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3850429

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). A density-based algorithm for discovering clusters in large spatial databases with noise. In Advances in Knowledge Discovery and Data Mining (pp. 160–172). Springer.

Chiranjeevi, A., & Nunn, N. (2024). Automating Abercrombie: Machine learning approaches to trademark classification. Journal of Empirical Legal Studies, 21(2). https://doi.org/10.1111/jels.12398

EUIPO & EPO. (2020). Intellectual property rights–intensive industries and economic performance. European Union Intellectual Property Office and European Patent Office.

Hill, P. (1999). Tangibles, intangibles and services: A new taxonomy for the classification of output. The Canadian Journal of Economics, 32(2), 426–446. https://doi.org/10.2307/136430

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density-based clustering. Journal of Open Source Software, 2(11), 205. https://doi.org/10.21105/joss.00205

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Neuhaeusler, P., Feidenheimer, A., Frietsch, R., & Kroll, H. (2021). Generating a classification for EUIPO trademark filings: A string matching approach (Report No. 69). Fraunhofer Institute for Systems and Innovation Research ISI.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543).

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3982–3992). https://doi.org/10.18653/v1/D19-1410

Sanghavi, A. (2018). Trademark similarity using semantic and phonetic features. International Journal of Computing, 17(2), 120–130.

Scienovus Intellectual. (2025). Understanding trademark Nice classification: A professional guide. https://scienovus.com/blog/2025/08/15/understanding-trademark-nice-classification-a-professional-guide

Shmatkov, P., Ivanov, I., & Petrov, A. (2023). Trademark text similarity using embedding models. Journal of Intellectual Property Analytics.

Trappey, A. J. C., Ou, J. J., Chuang, K. Y., & Trappey, C. V. (2022). AI-based semantic product classification for e-commerce platforms. Advanced Engineering Informatics, 51, 101601. https://doi.org/10.1016/j.aei.2021.101601

WIPO. (2024a). Madrid System: Filing international trademark applications—Classification of goods and services. World Intellectual Property Organization. https://www.wipo.int/en/web/madrid-system/how_to/file/madrid-system-filing-international-trademark-applications

WIPO. (2024b). Nice Classification—International classification of goods and services. World Intellectual Property Organization. https://www.wipo.int/en/web/classification-nice/index

WIPO & USPTO. (2020). Nice Agreement eleventh edition: General remarks, class headings, and explanatory notes. World Intellectual Property Organization & United States Patent and Trademark Office. https://www.uspto.gov/trademarks/trademark-updates-and-announcements/nice-agreement-eleventh-edition

Wunderskirchner, M. (2021). Borderline products in the EU: Obstacles for manufacturers and importers. International Chemical Regulatory and Law Review, 4(2), 72–74.

Wunderskirchner, M., Bauer, B., & Müller, R. (2021). Borderline classification of cosmetic and medicinal products: A regulatory perspective. Regulatory Toxicology and Pharmacology, 122, 104885. https://doi.org/10.1016/j.yrtph.2021.104885

Zhang, Y., Lu, W., Zhou, R., & He, X. (2018). Deep learning for e-commerce search and recommendation. SIGIR Forum, 52(1), 22–27. https://doi.org/10.1145/3274784.3274788

Zheng, Q. (2025). Intelligent prediction of trademark registration appeal outcomes based on natural language processing and CatBoost algorithm. In Proceedings of the Fifth International Conference on Telecommunications, Optics, and Computer Science (TOCS 2024) (Vol. 13629, p. 136291X). SPIE. https://doi.org/10.1117/12.3067906