Beyond Words: Using Natural Language Processing to Explore the Online Presence of Canine Professionals

SRN: NKXM1

September 2025

STAT0034 MSc Dissertation

Supervisors: Dr. Chak Hei Lo, Dr. Sarah Weidman, Dr. Jana Muschinski

Word count: 15405

Reference style: Nature

This dissertation is submitted in partial fulfilment of the requirements for the Master's degree in [Data Science], UCL.

ABSTRACT

As more people turn to online platforms for canine-related services, distinguishing between training approaches has become increasingly important. This study presents a keyphrase-driven model to evaluate whether professional dog trainers' stated methodologies align with their website language, using a semantic alignment validation approach in which survey-based groupings are reconstructed from extracted linguistic patterns.

Using data from 91 canine professionals, we examined whether their declared training methods were reflected in the language used on their websites. Unlike conventional workflows that derive clusters directly from raw text, this study employed a reversed workflow: survey-informed categories were defined upfront and then traced through website language. Keyphrases were extracted and refined using a customised pipeline built on KeyBERT with extensions for semantic scoring, exclusivity filtering, and gating, ensuring that only distinctive and interpretable terms were retained as features.

Following preprocessing, 88 trainers remained in the final dataset. The resulting NLP pipeline was modular, comprising four independent classifiers for method, role, gender, and aid type. These could be used in single-task mode for targeted evaluations or combined in a stricter independent multi-task setting to enforce consistency across dimensions. In single-task evaluation, strong macro F1 scores were achieved (0.93 for method, 0.94 for role, 0.94 for aid type), while gender proved more variable (0.79-0.99). Independent multi-task evaluation with a hard-match criterion naturally reduced accuracy but demonstrated that complete trainer profiles could still be reconstructed from text. Out of 88 trainers, 74 were fully aligned with their survey profiles, while 14 showed at least one area of misalignment, particularly in the gender and method categories.

These results confirm that linguistic patterns provide reliable signals of professional identity and that systematic misalignment can be detected. By combining reversed workflow design, validated semantic features, and modular evaluation, the study delivers both methodological novelty and operational relevance. The findings also aim to support dog owners by contributing to the development of accessible resources that reflect how various aspects of training philosophy and professional identity are communicated through website content. It further explores the feasibility of language-based inference methods as a technical approach that may support future research, particularly in contexts where the collection of survey-based data involves substantial cost and effort.

The framework introduced in this study provides a transferable design for NLP-based research concerned with evaluating the alignment between stated identity and linguistic expression, particularly in domains where interpretability and trust in textual communication are essential.

Keywords: survey-guided clustering, pre-trained language model, Keyphrase extraction

DECLARATION

I have read and understood the College and Departmental statements and guidelines concerning plagiarism. I declare that:

- This submission is entirely my own original work.
- Wherever published, unpublished, printed, electronic or other information sources have been used as a contribution or component of this work, these are explicitly, clearly and individually acknowledged by appropriate use of quotation marks, citations, references and statements in the text. It is 15405 words in length.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

NLP : Natural Language Processing

ACKNOWLEDGEMENTS

I would like to begin by expressing my deepest gratitude to my academic supervisor, Dr. Chak Hei Lo, whose guidance and insightful advice were invaluable in directing this project. His support was essential in navigating both the technical and conceptual aspects of the research. Our discussions not only motivated me but also helped clarify and strengthen my ideas, which I believe was fundamental to shaping a high-quality project. I am especially thankful for the way he encouraged me to reflect deeply on core concepts and examine methodological choices with precision, which greatly improved the clarity of this work and gave me the confidence to approach complex problems thoughtfully. The opportunity to exchange ideas and gain new perspectives was highly valuable, pushing me to apply my knowledge in innovative ways. His supervision inspired me to expand my analytical thinking and explore unfamiliar techniques, while reinforcing the importance of connecting theory with practice. This made the journey both intellectually stimulating and rewarding.

I am sincerely thankful to my industrial supervisors from Dogs Trust, Dr. Sarah Weidman and Dr. Jana Muschinski, who were incredibly supportive, motivated me throughout, and guided me in navigating a domain different from my background. Their thoughtful discussions and consistently helpful feedback broadened my perspective and strengthened this project. I would also like to thank Chris Newton for his constructive input, which helped refine several aspects of the study and added valuable depth. I feel truly fortunate to have collaborated on this project, gaining new insights and the chance to apply my skills in a different field. This experience has not only enriched my understanding of the welfare sector but has also inspired me to pursue more interdisciplinary work in the future.

This research was made possible through the invaluable support of the LPDP Endowment Fund for Education. The support has provided continuous motivation and a strong sense of responsibility, fostering my academic journey and encouraging me to pursue ambitious research goals while reminding me of the greater purpose of contributing to society. I have always believed in the value of helping others through data, and this project has strengthened my commitment to conducting research that creates meaningful benefits for society. It has inspired me to think beyond individual achievement and to align my work with the vision of nurturing a new generation of scholars who will help shape Indonesia's future through innovation in science and technology.

Finally, I am profoundly grateful to my parents and sisters, whose love, support, reassurance, and motivation have been the foundation of this journey. Their quiet strength and constant encouragement helped me stay resilient through each phase of this work and life. They have also taken care of the other half of my whole life, carrying responsibilities and worries without ever letting me break alone. Instead, they lifted me with calm hands and steady hearts, allowing me to keep moving even when the weight felt too heavy. Their constant presence, patience, and sacrifices have been my deepest source of strength, and I know I could not have come this far without them. I owe much of this achievement to the care and devotion they have given me, which I will always carry with gratitude.

1. INTRODUCTION

Choosing a dog trainer can be a challenging process, especially in a context where the industry remains largely unregulated (Johnson et al. 2023). Many dog owners rely on trainers' websites as their primary source of information when making decisions about who to trust with their pets. These websites are not only marketing tools, but also public expressions of professional values and training philosophy. However, the terminology used to describe methods is often vague, inconsistently applied, or strategically framed. This reflects a broader ambiguity in how humans perceive and interact with dogs, as both valued companions and behavioral projects to be managed (Greenebaum 2010). While trainers may offer different philosophies, all forms of training involve symbolic processes where humans assign meaning to dog behavior and their role in family life. Greenebaum (2010) argues that dog training is not only about shaping canine behavior, but also about educating humans to interpret and navigate relationships with nonhuman animals through culturally embedded expectations. Terms such as *positive*, *balanced*, or *science-based* are used across a wide range of practices, some of which operate on fundamentally different assumptions about canine welfare (Gabrielsen 2017). Without a shared regulatory framework or common linguistic standards, the same term may be used to signal very different practices, intentions, or levels of aversiveness, potentially obscuring important ethical or methodological distinctions. The consequences of this are not trivial, since the choice of trainer influences not only the training process itself but also the welfare of dogs and the trust of owners who seek professional guidance.

This ambiguity introduces a mismatch between what trainers claim and what their language implies. As a result, dog owners may interpret shared terms in highly inconsistent ways, assuming agreement where none exists. The gap between advertised messaging and actual training practices can pose real challenges for both organizations and dog owners. Although organisations such as Dogs Trust aim to support owners in interpreting online information, studies show that public-facing content does not always clearly reflect a trainer's core philosophy (Johnson et al. 2023). Phrases like "positive methods" are often interpreted by dog owners as signaling specific ethical standards, yet in reality, such language spans a wide range of practices. Even within reward-based systems, there is significant variation in how techniques are applied, and vague terminology often hides key differences in trainer behavior and its consequences for dog welfare (Hiby et al. 2004). Owners may therefore be misled into choosing services that do not reflect their expectations, and the resulting mismatch may contribute to training failure, deteriorating welfare, or even the relinquishment of dogs.

In this context, efforts to understand whether a trainer's stated language aligns with their actual methods have become increasingly important. Recent studies have raised concerns that inconsistently defined terminology in dog training discourse can mislead dog owners and hinder informed decision-making. For example, descriptors such as "intrusive" or "minimally aversive" are frequently under-defined and open to interpretation, limiting public understanding (Fernandez 2024). This linguistic ambiguity poses a challenge for dog owners

seeking ethical and transparent training options, especially in the absence of formal regulation in the UK. Dog welfare charities such as Dogs Trust play an important role in supporting owners by helping them interpret online materials and identify the language typically associated with different training approaches.

Research Objectives

To address the difficulty that owners face in navigating the unregulated dog training sector in the UK, this study sets out two main objectives:

1. Are there patterns in language use between different types of professionals that can be identified using an NLP-based pipeline and model? Is this model robust enough (evaluated through five-fold cross-validation) to generalise under real-world data challenges, such as imbalanced labels and underrepresented trainer groups? Such findings could be used to create informative resources that help owners interpret key phrases and their associations when navigating an unregulated industry.

2. To what extent does misalignment occur at the level of individual trainers, as assessed by comparing model-predicted results based on website content with the self-reported profiles collected in the survey?

Approach to Addressing the Objectives

The two objectives outlined above were addressed through a combination of structured data collection, semantic feature engineering, and supervised model evaluation. Rather than treating prediction as the end goal, the focus was on developing interpretable tools to test whether website language provides a reliable signal of training orientation. The key design decisions and computational steps used to achieve each objective are summarised below.

1. This study uses a reversed NLP workflow and treats the classifier as a form of sanity check. These two design choices shape the overall methodology. Instead of grouping website text and inferring categories from it, the analysis starts with the categories already declared by trainers in the survey. These include training method, tool use, and professional role. The goal is to examine whether the language on each trainer's website aligns with what they say about themselves. By beginning with real, self-reported data, the study avoids guessing and keeps the analysis grounded. The classifier is not used just to predict outcomes, but to test whether language patterns consistently match declared identities.

   The pipeline is built on a modular system connected to a database. Keyphrases are extracted using KeyBERT and then scored across three dimensions: semantic similarity to the website, alignment with method-specific prompts, and exclusivity between trainer groups, calculated with Jensen-Shannon divergence. A gating function filters out vague or general terms so that only strong, method-specific signals contribute to the feature space. These gated scores are combined into interpretable vectors and passed into a Random Forest classifier.

To ensure the model is robust, we apply five-fold cross-validation and test different thresholds, class weights, and light-touch oversampling strategies. This setup acts as a simulation of real-world challenges, where imbalanced and sparse label distributions often make training and evaluation more difficult. It ensures that robustness is assessed in a setting that reflects practical constraints.

2. The second objective was addressed by evaluating how well website language aligns with trainers' declared identities. After training the classifier, we applied it to predict profiles based solely on website content. For trainers who had submitted survey responses, these predictions were compared against self-reported categories to detect potential misalignment. This procedure serves as a structured mechanism for identifying cases where public messaging may obscure or misrepresent actual practice. To support flexible evaluation, both single-task and independent multi-task classification models were trained independently for each label group: method, role, gender, and aid type. This modular design allows Dogs Trust to evaluate specific dimensions one by one or combine them into more complex profiles. Inspired by stepwise selection strategies, we also tested combinations of labels in a independent multitask setting. These combined models were evaluated using a strict hard-match criterion, where predictions were only considered correct if all labels were correct at once. This approach offers a more conservative alignment check, tightening the criteria for consistency and helping surface subtler forms of mismatch.

Significance

This study is particularly relevant for both Dogs Trust and dog owners. For Dogs Trust, the findings offer a practical way to assess whether a trainer's public messaging reflects welfare-focused practices, supporting their mission to promote humane standards in an unregulated industry. For dog owners, clearer understanding of trainer orientation helps reduce the risk of choosing inappropriate services, which can lead to lasting consequences for animal wellbeing. Since behaviour-related issues are a major factor in dogs being surrendered to shelters, identifying misleading or inconsistent language plays a key role in supporting better outcomes.

Beyond the dog welfare domain, this project also presents a sustainable and modular NLP framework that can be adapted to other sectors where trust and credibility rely on public-facing communication. The pipeline's structure separates keyphrase scoring, semantic alignment, and classification, allowing for flexible reuse and task-specific adaptation while maintaining interpretability. This makes it suitable for addressing practical challenges in other domains such as education, healthcare, or behavioural services, where clarity and accountability are essential.

Challenges

At the same time, several challenges accompany this research. The dataset represents real-world constraints, with a relatively small number of trainers and substantial class imbalance across methods and tools. This makes prediction inherently more difficult and

increases the risk of overfitting. For instance, some training aids such as shock collars or noise devices are reported by very few trainers, while positive aids like treats and clickers are common. Similarly, while reward-based methods dominate, other orientations such as aversive or balanced approaches are sparsely represented. These imbalances make it difficult to build models that are both fair and robust, as minority categories may be under-predicted.

Website texts also present inherent noise. Many contain marketing content, repeated slogans, or generic claims designed for search engine optimisation rather than for transparent communication of training philosophy. This ambiguity blurs the signal and makes it challenging for automated models to identify reliable linguistic markers. Moreover, terms with similar surface forms may have different meanings depending on context. For example, “positive” can be used to describe both positive reinforcement and positive punishment, despite their opposite implications for welfare. Without careful semantic scoring, such terms risk being misclassified, reducing both accuracy and interpretability.

A further challenge lies in the reliance on survey data as the reference point for “ground truth.” While survey responses provide the most direct self-declaration available, they remain self-reported and may be subject to bias, exaggeration, or selective omission. Trainers may under-report their use of aversive tools, or may describe their philosophy in aspirational rather than operational terms. This means that even the anchor labels used to guide the NLP pipeline cannot be assumed to be perfectly accurate. The misalignment detection component of this study is therefore particularly important, since it recognises the potential for declared and actual practices to diverge.

Finally, the small scale and domain-specific nature of the dataset raises broader questions about generalisability. Unlike large-scale NLP applications that benefit from millions of documents, this study is constrained by the real-world availability of canine professionals’ websites. The resulting models are therefore highly interpretable and tailored to the domain, but may not generalise to other sectors or even to significantly larger samples without careful retraining and validation. These challenges highlight the need for transparent, interpretable pipelines that can withstand noisy, imbalanced, and small-scale data while still producing actionable insights.

Positioning within the Literature

Previous work shows that trainer language reflects training philosophy: force-free or science-based trainers emphasise empathy, welfare, and scientific grounding, while traditional or blended approaches highlight obedience, authority, or practical results (Johnson et al. 2023). Although textual content can signal orientation, systematically capturing these cues is difficult, especially with small and noisy datasets like trainer websites.

Many existing approaches rely on unsupervised workflows, where website content is clustered without labels and only later compared to known categories. While flexible, this approach risks generating patterns that do not clearly relate to real-world attributes, making the interpretation difficult. It also increases the likelihood that resulting clusters reflect stylistic or surface-level features rather than meaningful distinctions in philosophy or practice. By

contrast, this study introduces a survey-guided NLP pipeline that reverses the conventional direction of analysis. Instead of clustering website texts and inferring categories afterwards, the analysis begins with structured groupings based on trainers' declared roles, methods, tools, and gender, and uses these as the foundation for feature extraction and classification. This reverse workflow enables more transparent and grounded evaluation of language patterns, by anchoring analysis in how trainers identify themselves rather than relying on assumption-driven clustering. It also ensures that the features used in classification, e.g., gated semantic-prompt-exclusive scores, are directly linked to meaningful categories, rather than to stylistic patterns with limited interpretability.

Summary

Together, these elements introduce the project's aims, challenges, and significance. By framing the research within canine welfare and industry regulation, identifying key linguistic and methodological issues, and outlining the approach and its limitations, the introduction builds a clear rationale. It positions the work as both technically innovative and practically relevant, offering an interpretable and deployable framework for evaluating alignment between language and practice in dog training.

2. LITERATURE REVIEW

Natural Language Processing (NLP) has become increasingly central in handling unstructured data across low-resource domains. In this project, we use pretrained language embeddings within a semantic profiling pipeline that leverages structured survey responses as semantic anchors. These responses guide how we interpret, extract, and model linguistic patterns from canine trainer websites.

This chapter explains the reasoning behind each core component of our pipeline: (1) embedding-based representation, (2) Sentence-Level Embeddings with Sentence-BERT, (3) keyphrase extraction and filtering, (4) semantic similarity scoring, and (5) multi-label classification using interpretable models. The literature is presented with direct links to our implementation decisions, with emphasis on practical alignment rather than theoretical breadth.

2.1. Embedding-Based Representation

Early NLP systems commonly represented text using surface-level statistical features such as word frequencies or TF-IDF scores, which rely on simple co-occurrence patterns (Zhang et al. 2024). While effective in early applications, these methods fail to capture semantic nuance, especially in specialised domains where language use is highly context-dependent and heterogeneous.

To address these limitations, we employ dense vector representations generated by `all-mpnet-base-v2` from `Hugging Face`, a transformer-based model from the Sentence-BERT family (Reimers et al. 2019). Instead of counting terms, these models encode sentences into fixed-length vectors in a high-dimensional semantic space, where syntactic and contextual relationships are captured through self-attention mechanisms (Vaswani et al. 2017).

Formally, for any input sentence s, a transformer-based sentence encoder produces a dense embedding as:

$$\text{Embed}(s) \in \mathbb{R}^d$$

where $\text{Embed}(s)$ is the output vector representation of sentence s, and d is the dimensionality of the embedding space (typically 768 for `all-mpnet-base-v2`). This formulation follows the architecture of Sentence-BERT (Reimers et al. 2019), where a pooling strategy (such as mean pooling) is applied on top of transformer outputs to obtain a fixed-size embedding for each sentence. This embedding captures the semantic content of the input and serves as the basis for all subsequent operations in our pipeline.

Unlike traditional approaches requiring domain-specific training, this method works in low-resource settings without extra supervision. It introduces a flexible architecture where semantic representations are reused across classification, filtering, and alignment. Sentences are encoded into embeddings optimised for semantic similarity, forming the basis for cosine comparison, keyphrase scoring, class-level representation, and prompt-based filtering. As the

dog training domain lacks large-scale corpora, no additional pretraining or fine-tuning is applied. Instead, general-purpose embeddings are adapted through structured postprocessing, combining semantic clustering with class-level alignment to improve keyword relevance. This inference-time adaptation enables nuanced filtering without the cost of retraining or added supervision.

2.2. Survey-Informed Grouping

Topic modeling is a common method for identifying patterns in text, but it often produces abstract groupings that are difficult to interpret or align with real-world categories. In contrast, we organize our data using responses from a structured survey, which asked trainers about their methods, roles, gender identity, and tools used (Johnson et al. 2023). These categories act as ground truth for profiling.

This explicit grouping approach avoids ambiguity and allows us to trace language patterns to concrete behavioral variables. It also facilitates targeted comparisons, such as examining how language differs between those who use positive-only techniques and those who include aversive aids. This method reflects calls for contextualized grouping in behavior-focused domains (Fernandez 2024), where understanding practitioner intent and identity is key.

2.3. Keyphrase Extraction

2.3.1. Comparison to Related Work

This approach resembles the class-specific keyword extraction pipeline proposed by (Meisenbacher et al. 2024), who extended the KeyBERT library by introducing seed keyword guidance and iterative score refinement. However, while their method relies on seed keywords and document-level scoring, our modification introduces group-level representations through class centroids and evaluates keywords holistically based on both alignment and exclusivity.

We independently develop the scoring logic, using KeyBERT only to generate initial candidates. These keywords are then filtered and re-ranked based on their similarity to the document, alignment with prompt definitions, and exclusivity to each class. The system works over clusters defined by survey responses, aiming to identify keyphrases that best represent each group as a whole. This approach, which we call reversed grounding, maps class-level definitions back to representative keywords to align website language with survey concepts.

2.3.2. Adaptive Weighting Based on Prompt Alignment

In keyword extraction, frequency or semantic similarity alone may not suffice, as a keyword can be class-specific without aligning conceptually with its definition. This issue is critical when distinguishing subtle philosophical or methodological differences, such as dog training approaches. To address this, we use a customised gating mechanism that adjusts keyword weights according to their alignment with a reference definition, or “prompt”, for each method.

Our approach is loosely inspired by soft gating strategies found in sparse models, particularly the Mixture of Experts (MoE) architecture (Shazeer et al. 2017), where components are weighted based on input relevance. While MoE models use a softmax function over multiple experts, we adopt a simplified version of this idea: adjusting individual keyword weights based on their conceptual alignment.

Each candidate keyword is first assigned a raw score based on a weighted combination of three components: its semantic similarity to the website content, its exclusivity to a specific method (measured using Jensen-Shannon divergence), and its alignment with a textual definition prompt.

To refine this further, we apply a non-linear gating function to downweight keywords that fall below a certain alignment threshold. This gating function is defined as:

$$\text{Gate}(x) = \epsilon + (1 - \epsilon) \cdot \left(\frac{x - \text{floor}}{1 - \text{floor}} \right)^{\gamma} \tag{2.1}$$

Where:

- x: prompt alignment score (semantic similarity between the keyword and the definition),
- floor: the minimum threshold of acceptable alignment,
- ϵ: a small baseline to preserve minimal contribution,
- γ: a curve-shaping parameter that accentuates the impact of well-aligned keywords.

While not a softmax, this gating formula serves a similar purpose: adjusting contribution strength based on conceptual relevance. It provides an interpretable and flexible mechanism for reweighting keywords according to how well they reflect the class-level prompt.

2.3.3. Discriminative Filtering Using Class Exclusivity

Semantic alignment alone may yield keywords that are relevant but too general. To ensure that selected keywords are discriminative, we incorporate a class exclusivity score using Jensen-Shannon divergence (JSD), inspired by contrastive clustering approaches such as SimCKP (Choi et al. 2023). For each keyword, we compute the JSD between its distribution across classes and a uniform distribution:

$$\text{JSD}(k_i) = \frac{1}{2} \left[\text{KL}(P_{k_i} \parallel M) + \text{KL}(Q \parallel M) \right] \tag{2.2}$$

Here, P_{k_i} denotes the class distribution of keyword k_i, Q is the uniform distribution, and $M = \frac{1}{2}(P_{k_i} + Q)$ is the midpoint between the two. The KL terms represent Kullback-Leibler divergence. A higher $\text{JSD}(k_i)$ indicates that keyword k_i is more specific to a single class, making it more discriminative.

2.4. Random Forest Classification

The final stage of our pipeline classifies individual documents based on filtered and embedded keyphrases. We use a Random Forest classifier due to its effectiveness with high-dimensional, sparse data and its ability to measure feature importance.

Random Forest is an ensemble learning method that constructs multiple decision trees using random subsets of data and features (Breiman 2001). For classification, it aggregates predictions through majority voting:

$$\hat{y} = \text{mode}\left(\{h_t(x)\}_{t=1}^{T}\right) \tag{2.3}$$

where $h_t(x)$ is the prediction from the t^{th} tree, and T is the total number of trees.

For multi-label tasks, a separate sub-forest is trained per target variable (Kocev et al. 2013). The final prediction vector is defined as:

$$\hat{\mathbf{y}} = [\hat{y}^{(1)}, \ldots, \hat{y}^{(m)}], \quad \hat{y}^{(j)} = \text{mode}\left(\{h_t^{(j)}(x)\}_{t=1}^{T_j}\right) \tag{2.4}$$

This formulation treats each output as conditionally independent given the input, allowing scalability and modularity. Random Forests are also suitable when interpretability is important, as predictions can be linked to key features (Hastie et al. 2009; Shyrokykh et al. 2023).

2.5. Managing Class Imbalance and Evaluation

The dataset shows significant class imbalance across methods, roles, and demographics. To address this, we apply selective oversampling during cross-validation for better exposure, while preserving the original distribution in final training using class weighting and calibrated thresholds to ensure fair and interpretable predictions.

This dual strategy of applying oversampling during validation and class weighting during deployment is supported by previous research showing that such combinations enhance recall without introducing structural bias (Mujahid et al. 2024; Karaman et al. 2024; Khvatskii et al. 2025). The class weight w_c for a class c is computed as:

$$w_c = \frac{n}{|C| \cdot n_c}$$

To ensure balanced evaluation, stratified cross-validation is applied, preserving label distributions across folds and reducing variance. As outlined by Moreno-Torres et al. (2012), improper partitioning can introduce bias in imbalanced datasets. To implement stratification, the multi-label annotations are first represented as a binary matrix $Y \in \{0,1\}^{n \times m}$, where $Y_{ij} = 1$ indicates that instance i is assigned to label j.

2.5.1. Evaluation Metrics and Robustness Strategies

To address class imbalance and validate generalisability, we apply stratified k-fold cross-validation, ensuring each fold maintains the same class distribution as the full dataset. This

prevents evaluation bias and is recommended when class distributions are skewed (Moreno-Torres et al. 2012).

Our primary metrics are F1-score and Accuracy. These are defined as follows:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \quad \text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{2.5}$$

These evaluation metrics are widely used in imbalanced classification settings (Goutte et al. 2005; Sokolova et al. 2009).

2.6. Integrated Database-Driven Architecture for NLP Pipelines

The NLP pipeline is integrated into a SQL-backed architecture that supports traceability, reproducibility, and adaptability. Structured and unstructured data are stored in normalized schemas, with stored procedures such as `fn_get_baseline_combination` enabling dynamic retrieval of trainer responses. Keyword extraction remains adaptable to changes in survey inputs or label definitions.

Each keyword is linked to its cluster ID, source document, cosine score, and filtering status, enabling transparent auditing and monitoring of semantic shifts. Logs of keyphrase matches, similarity scores, and predictions support exploratory analysis and model debugging, in line with responsible NLP deployment principles (Ye et al. 2024).

By combining Python-based semantic processing with SQL-based data management, the system achieves a hybrid structure that is technically well-maintained and clearly auditable, in line with best practices in explainable NLP (Oniani et al. 2023).

2.7. Summary

This literature review has explained how each aspect of our profiling pipeline is supported by existing research and aligned with the goals of behaviorally informed text classification. Instead of treating NLP as a black box, we build the pipeline around interpretable components anchored in real-world categories. By combining pretrained semantic representations, structured survey design, and transparent classification logic, our system serves as a practical model for profiling in domains where labeled data is limited and stakeholder accountability is critical. Each design decision reflects the dual needs of performance and transparency. As such, the system is both research-informed and deployment-ready.

3. METHODOLOGY

This project followed a structured pipeline to process, analyze, and interpret text-based data from canine professionals. The methodology is divided into three main stages: Survey data preparation and preprocessing, keyword extraction and semantic feature construction, and feature-based modelling. Each stage was designed to maintain alignment between survey responses and the textual representation found on trainer websites.

3.1. System Environment and Dependencies

All processes were implemented in Python 3.10 within a Conda-managed environment. SQL Server 2019 was used as the database system, with scripts developed for data access, transformation, and modelling. Core packages included:

- **Data Handling:** `pandas`, `numpy`, `sqlalchemy`, `pyodbc`
- **NLP and Embedding:** `sentence-transformers`, `keybert`
- **Machine Learning:** `scikit-learn`, `torch`, `torchvision`
- **Environment:** SQL Server Management Studio v18

All code was modularised for reusability. Each stage of the workflow was designed to be independently testable and reproducible.

3.2. Stage One: Data Preparation and Cleaning

This stage involved assembling and preparing two integrated datasets: structured survey responses from UK-based dog trainers and unstructured textual content obtained from their professional websites. Both sources were stored and managed within Microsoft SQL Server to facilitate standardized access, preprocessing, and downstream analysis.

On the survey side, individual trainer responses were first consolidated into a master dataset. Each record captured multiple categorical selections across dimensions such as training method, tools used, professional role, and gender. These fields were normalized into relational tables through a structured data modeling process. Referential joins were established between each label set and the main trainer profile, enabling modular querying and group-level aggregation. Once normalized, exploratory data analysis (EDA) was performed to assess label distributions, identify class imbalances, and uncover interpretable subgroups. These subgroups later served as semantic anchors and baseline clusters for downstream tasks such as keyphrase extraction and classification. Based on this analysis, combinations such as method–gender–role were selected to represent meaningful trainer categories.

On the web-scraped side, the pipeline began by aggregating all relevant webpages linked to each trainer. These were concatenated into a unified document per trainer, establishing a

one-to-one mapping between survey response and online text. To reduce noise and irrelevant structural content, a dictionary of filtered words was developed and continuously refined. This dictionary captured elements commonly found across websites such as headers, footers, contact information, placeholders (e.g., “[REDACTED]”), and navigation elements. Preprocessing routines, implemented using SQL stored procedures, were then applied to clean the web text. This included HTML tag removal, whitespace normalization, and duplicate section elimination. Cleaned texts were saved into a designated SQL table (`t_web_cleaned`) as the final processed corpus.

The final step in this stage involved defining baseline configurations for analysis. Using the survey-derived label combinations, a set of stored procedures was written in SQL Server to automate feature extraction and cluster assignment for each configuration. Each procedure was executed and evaluated independently, producing a set of structured outputs to be used in Stage Two. After cleaning, 88 out of the original 91 profiles were retained. Three profiles were excluded due to missing or unusable website content, and were treated as null values for the rest of the pipeline.

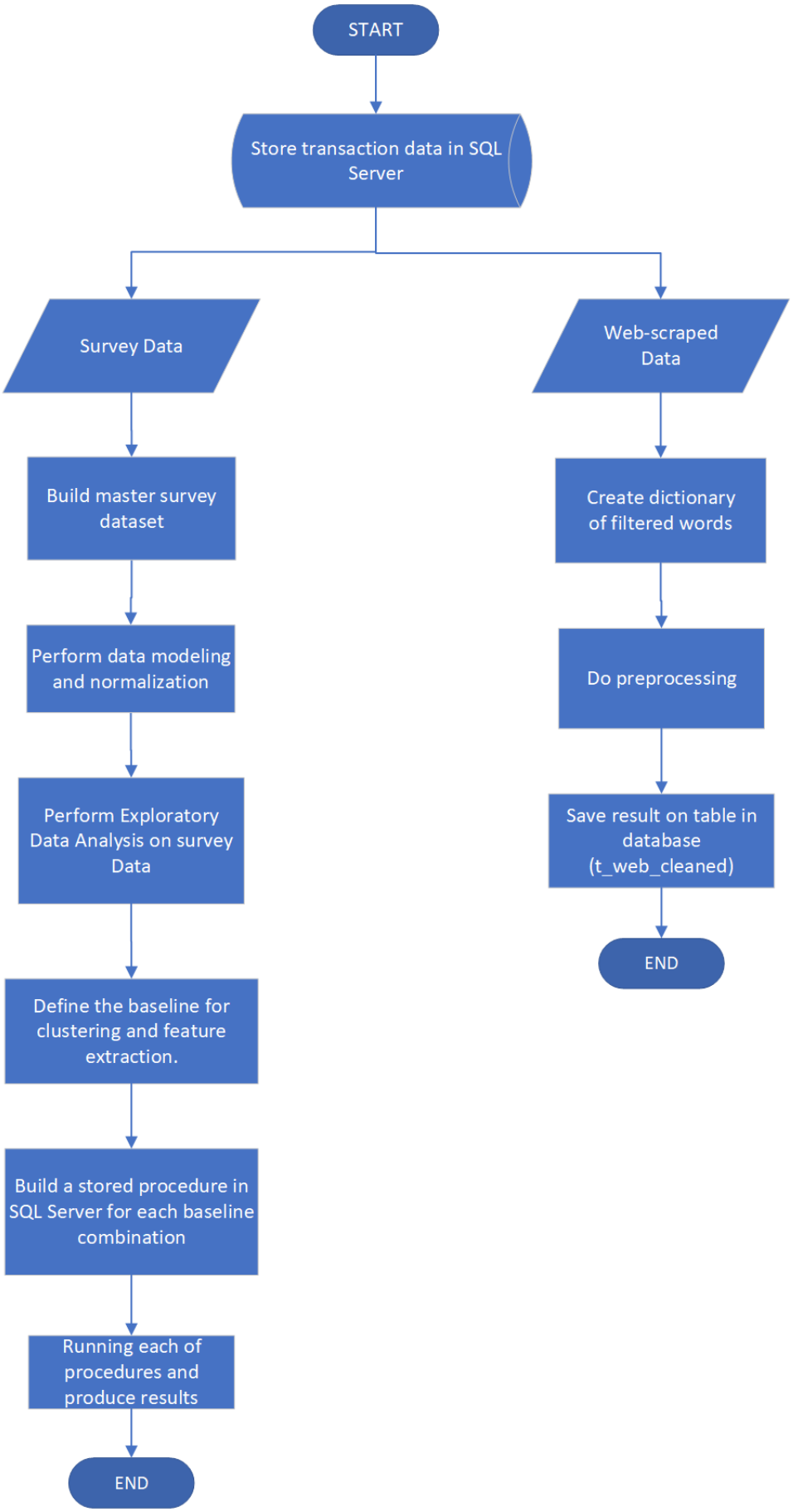

Figure 3.1: Stage 1: Text Cleaning and Survey Structuring

3.2.1. Survey Data Preparation

Survey Data Relation Modelling

To support the training and prediction pipeline, the backend was built around a structured relational database using SQL Server 2019. The data came from two main sources: survey responses and website content. Since these sources differ in structure and format, we designed the database to keep everything normalized, traceable, and easy to query.

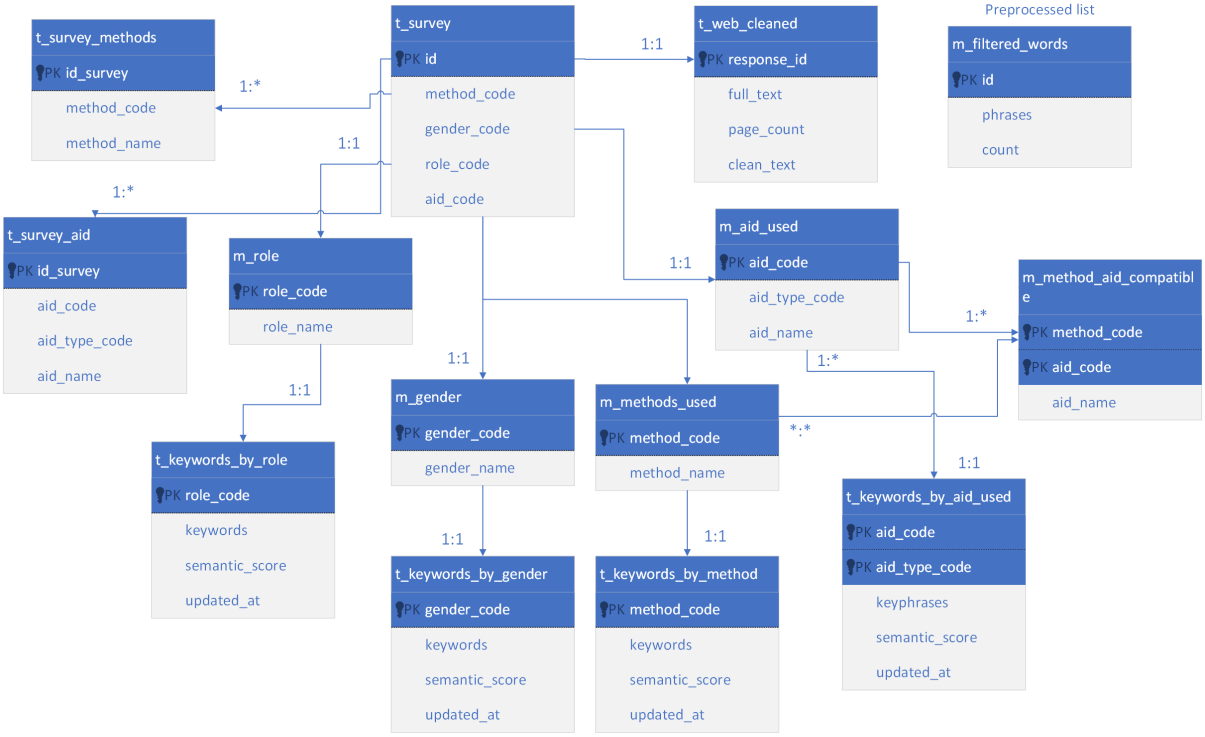

Figure 3.2: Data Modelling (Entity Relationship Diagram)

We followed a master-detail schema, using a clear naming convention:

- Tables starting with `m_` store master data (reference lists)
- Tables starting with `t_` store transactional data (survey input, processed content, extracted keywords)

This setup helped us keep the data clean, reduce duplication, and make the structure easy to extend over time.

Master Tables (`m_`)

These tables define the fixed categories or lookup values used throughout the system:

- `m_role`
  Contains the list of professional roles (e.g., instructor, behaviourist). Each trainer can select one or more roles. This table is referenced in both survey input and keyword mapping.

- `m_gender`
  Stores gender categories from the demographic section of the survey. Used for grouping and semantic comparison.

- `m_methods_used`
  Lists available training methods. Referenced by survey inputs and used in modeling trainer preferences.

- `m_aid_used`
  Defines tools or aids used during training (e.g., clickers, harnesses). Trainers can select multiple aids in the survey.

- `m_method_aid_compatible`
  Maps which aids are commonly used with which methods. Derived from co-occurrence patterns in the survey data and used as a logic filter during prediction. Validated using stored procedures in SQL Server.

- `m_filtered_words`
  A manually curated list of boilerplate or low-value phrases commonly found across trainer websites. This is used to clean scraped content before semantic processing.

Transactional and Processed Tables (`t_`)

These tables hold the actual data collected and processed throughout the project:

- `t_survey`
  The core table for survey responses. Each row represents one trainer and links to selected roles, methods, aids, and demographic attributes.

- `t_survey_methods`, `t_survey_aid`
  These junction tables handle the many-to-one structure of survey selections. Each row represents a single method or aid selected by a trainer.

- `t_web_cleaned`
  Stores website content per trainer. Includes full-page text and a cleaned version ready for NLP processing. Linked to the trainer via a shared ID.

- `t_keywords_by_role`, `t_keywords_by_method`, `t_keywords_by_gender`, `t_keywords_by_aid_used`
  These hold the extracted keywords for each grouping category, along with their semantic similarity scores. Used for building feature vectors for classification.

Table Relationships and Design Logic

The entity relationship diagram reflects these key patterns:

- `t_survey` acts as the central table, linking to all master tables using foreign keys.

- Multi-select fields (methods, aids) are normalized via junction tables to avoid column duplication.
- Website content from `t_web_cleaned` is directly tied to each trainer through their survey ID.
- Keyword tables store semantically ranked terms per group, which are later matched against individual trainer content.
- Compatibility between methods and aids is enforced through `m_method_aid_compatible`, with validation logic handled through SQL stored procedures.

All interactions between Python and SQL Server were managed using `sqlalchemy` and `pyodbc`, allowing smooth integration across the data pipeline. This schema supports consistent querying, scalable input updates, and alignment between text-derived and structured features.

Preprocessing the Webscraped Data

The webscraped data provided by Dogs Trust included multiple individual web pages for each trainer, delivered in a Parquet file. Preprocessing in this study began from the point where these pages were merged into a single text document. Each merged document was then linked to its corresponding survey response using stored procedures in SQL Server, ensuring a clear one-to-one match between the website text and the trainer's profile data.

Cleaning procedures included:

- Removing HTML tags, navigation elements, headers, and placeholder markers
- Filtering UK phone numbers, emails, URLs, and formatting symbols
- Filtering out generic or repetitive marketing phrases using a custom dictionary of filtered terms defined in the `m_filtered_words` table

Survey fields that allowed multiple selections (such as training methods, roles, and aids) were normalized into long-form relational tables. For instance, if a trainer selected three different methods, these were recorded as three separate rows linked to the same trainer ID.

The database followed a master–detail structure:

- Master data such as roles, methods, aids, and gender were stored in dedicated `m_` tables (e.g., `m_role`, `m_methods_used`, `m_aid_used`, `m_gender`)
- Survey selections and text results were stored in transactional `t_` tables (e.g., `t_survey`, `t_survey_methods`, `t_survey_aid`)

Cleaned website texts were stored in `t_web_cleaned`, which includes the full combined text and a cleaned version used for semantic analysis. Each entry in `t_web_cleaned` is linked to the corresponding trainer in `t_survey` via a shared ID, maintaining traceability throughout the NLP pipeline.

Additional quality checks were implemented to detect empty fields, duplicate rows, and mismatches between scraped website content and survey entries. Summary views were

created to support exploratory analysis and to verify the distribution and completeness of data across demographic and methodological groups.

3.3. Stage Two: Keyword Extraction and Semantic Feature Construction

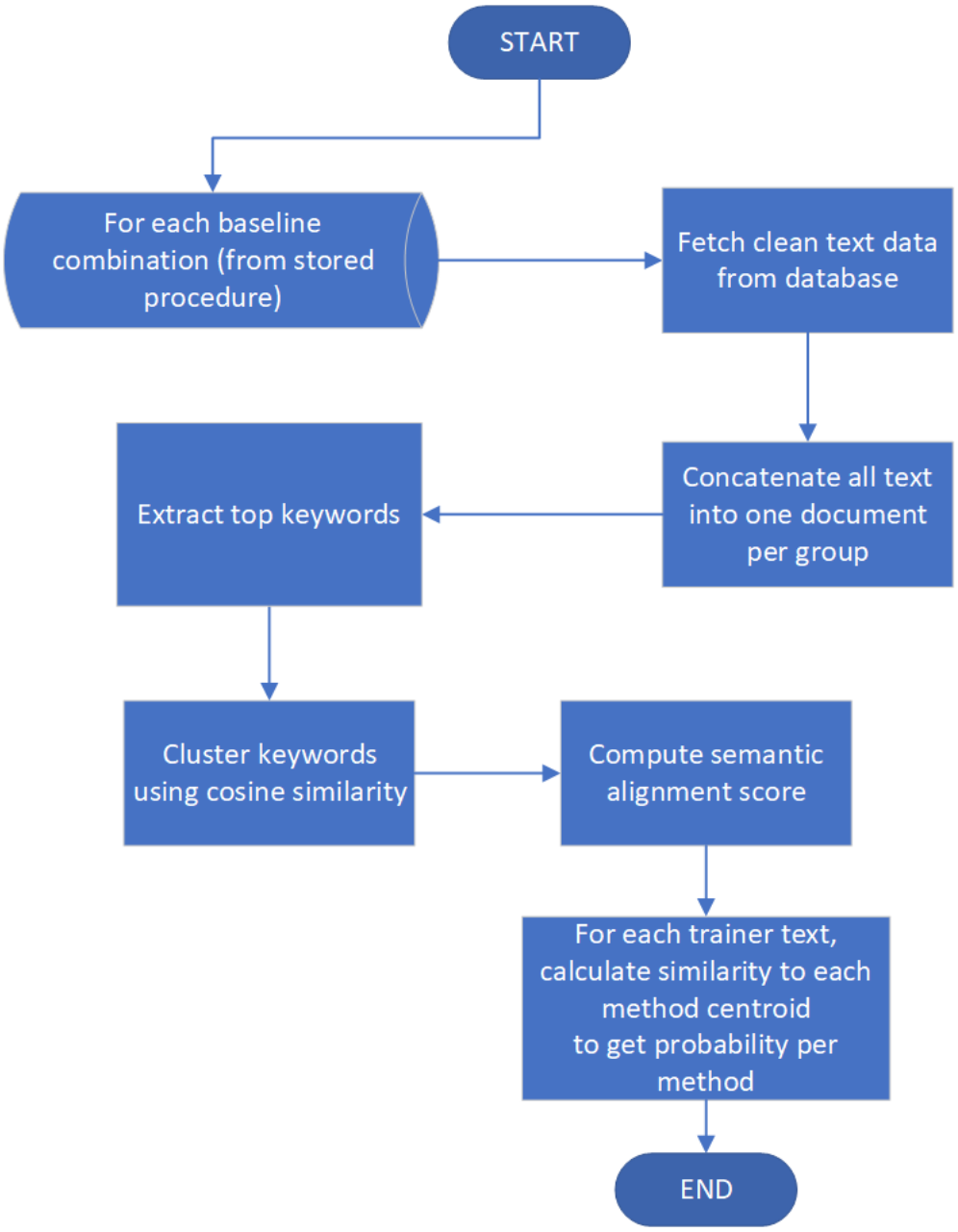

Figure 3.3: Stage 2: Keyword Extraction and Feature Mapping

This stage focused on extracting keyphrases that were semantically discriminative across trainer groups. For each group, the cleaned website content of its members was fetched from the SQL database and concatenated into a single document. This formed the basis for semantic comparison between groups.

A pre-trained sentence embedding model `all-mpnet-base-v2` from `Hugging Face` was applied to generate contextualized representations of these group documents. Using a contextual keyword extraction model, top phrases were identified from each group document. These candidates were filtered through a blacklist of generic or repetitive phrases and further clustered using cosine similarity to reduce redundancy and enhance interpretability. The resulting keyphrases defined each group's semantic profile.

To compute semantic alignment, each trainer's individual website content was compared to every group-level keyword set. Cosine similarity scores were calculated between the trainer's embedding and each group centroid. This produced a probability distribution over all groups for every trainer, indicating their alignment with each declared profile.

In addition to extracting and aligning keywords per task, each keyword was scored using a supervised weighting strategy. Following prior work (Nomoto 2022; Zhang et al. 2024; Kim et al. 2024; Papazis et al. 2025), the raw score was calculated as a weighted combination of three components: semantic similarity, prompt alignment (with task-specific definitions), and class exclusivity measured via Jensen–Shannon divergence (JSD).

Keyword Scoring and Gating

The scoring formula aggregated multiple signals:

- **Semantic score**: cosine similarity between the keyword and the trainer's text.
- **Prompt score**: similarity to the written definition of the class.
- **JSD score**: exclusivity of the keyword across classes.

The raw score was defined as:

$$\text{raw\_score} = 0.45 \times \text{semantic} + 0.45 \times \text{JSD} + 0.10 \times \text{prompt}.$$

A gating function was then applied to adjust this score based on prompt alignment (Shazeer et al. 2017), defined as:

$$\text{gate}(x) = \epsilon + (1 - \epsilon) \left(\frac{x - \text{floor}}{1 - \text{floor}} \right)^{\gamma},$$

where $\text{floor} = 0.40$, $\epsilon = 0.25$, and $\gamma = 1.5$.

The final score became:

$$\text{final\_score} = \text{raw\_score} \times \text{gate}(\text{prompt}).$$

Semantic Feature Representation

Each keyword contributed four individual metrics: semantic similarity, prompt score, gate factor, and final score. For each trainer, three aggregated features were computed per group: mean similarity, maximum similarity, and a binary presence flag. These were concatenated across tasks to produce a comprehensive feature vector representing the trainer's semantic alignment across multiple dimensions.

Pseudo-code Summary

```
for each trainer:
    for each task in [method, role, gender, aid_type]:
        for each keyword in task_keywords:
            semantic = cosine(keyword, trainer_text)
            jsd = exclusivity(keyword, across_classes)
            prompt = cosine(keyword, prompt_definition)
            raw_score = 0.45*semantic + 0.45*jsd + 0.10*prompt
            gate_factor = ((prompt - 0.40)/0.60)^1.5
            gate_factor = max(0.25, min(1.0, 0.25 + 0.75*gate_factor))
            final_score = raw_score * gate_factor
            record [semantic, final_score, gate_factor, prompt]
        compute [mean, max, presence] over group
    concatenate all features → trainer feature vector
```

3.4. Stage Three: Semantic Feature Modelling

Sanity Check via Classification

The final stage focused on transforming the semantic features into a structured format suitable for supervised modelling. Each trainer's profile was represented by a fixed-length vector containing the similarity scores and keyword flags derived earlier.

Since keyword counts varied by group, padding and alignment procedures were used to ensure all vectors had consistent dimensions. Summary statistics were also appended to the vector for broader context.

To organize the data for downstream modelling, all features were stored in a matrix where each row corresponded to a trainer and each column represented a semantic signal. Labels for multi-label classification were drawn from the normalized survey fields.

The concatenated feature vectors were fed into classifiers for each independent task. This step acted as a sanity check: if the scored and gated keywords were valid, they would enable robust supervised classification. Successful performance confirmed that the feature construction pipeline produced representations that were not only linguistically plausible but also empirically discriminative.

The modelling pipeline was built to accept this matrix format while preserving the ability to trace each prediction back to the original trainer ID and source content. Data splits and preprocessing routines were configured to support reproducibility across experiments.

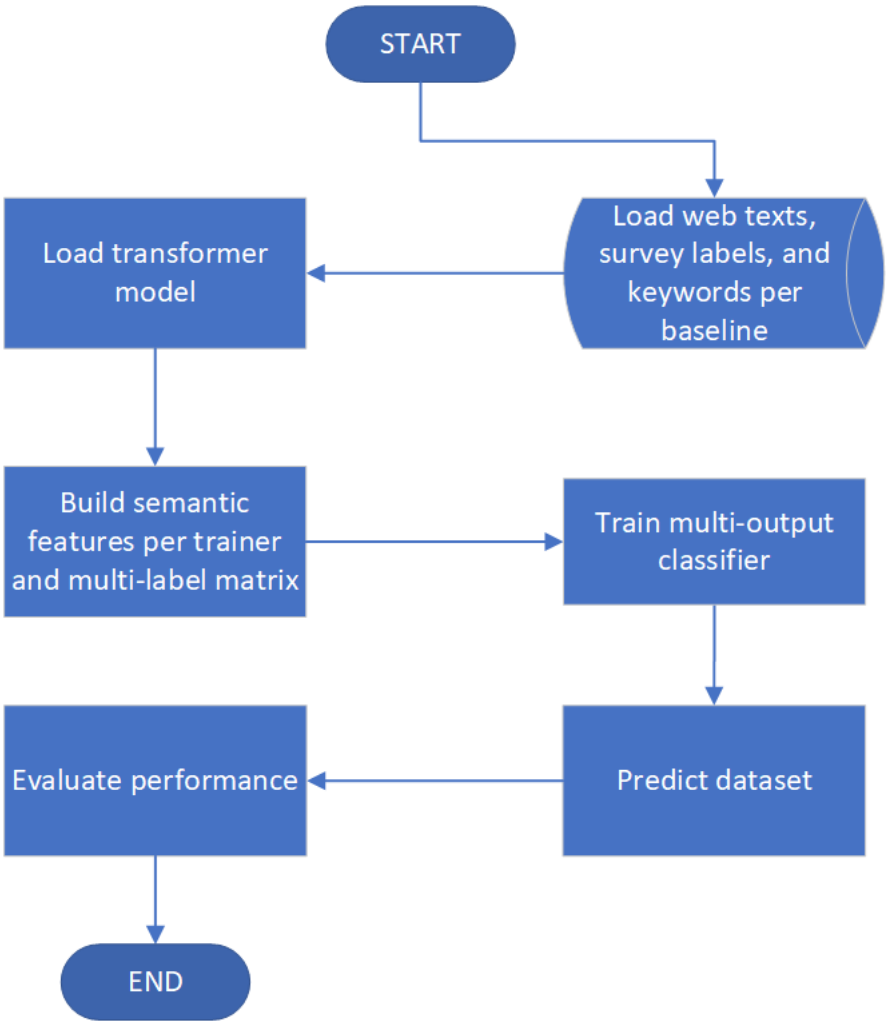

Figure 3.4: Stage 3: Semantic Feature Matrix Construction

3.5. Independent Multitask Modelling

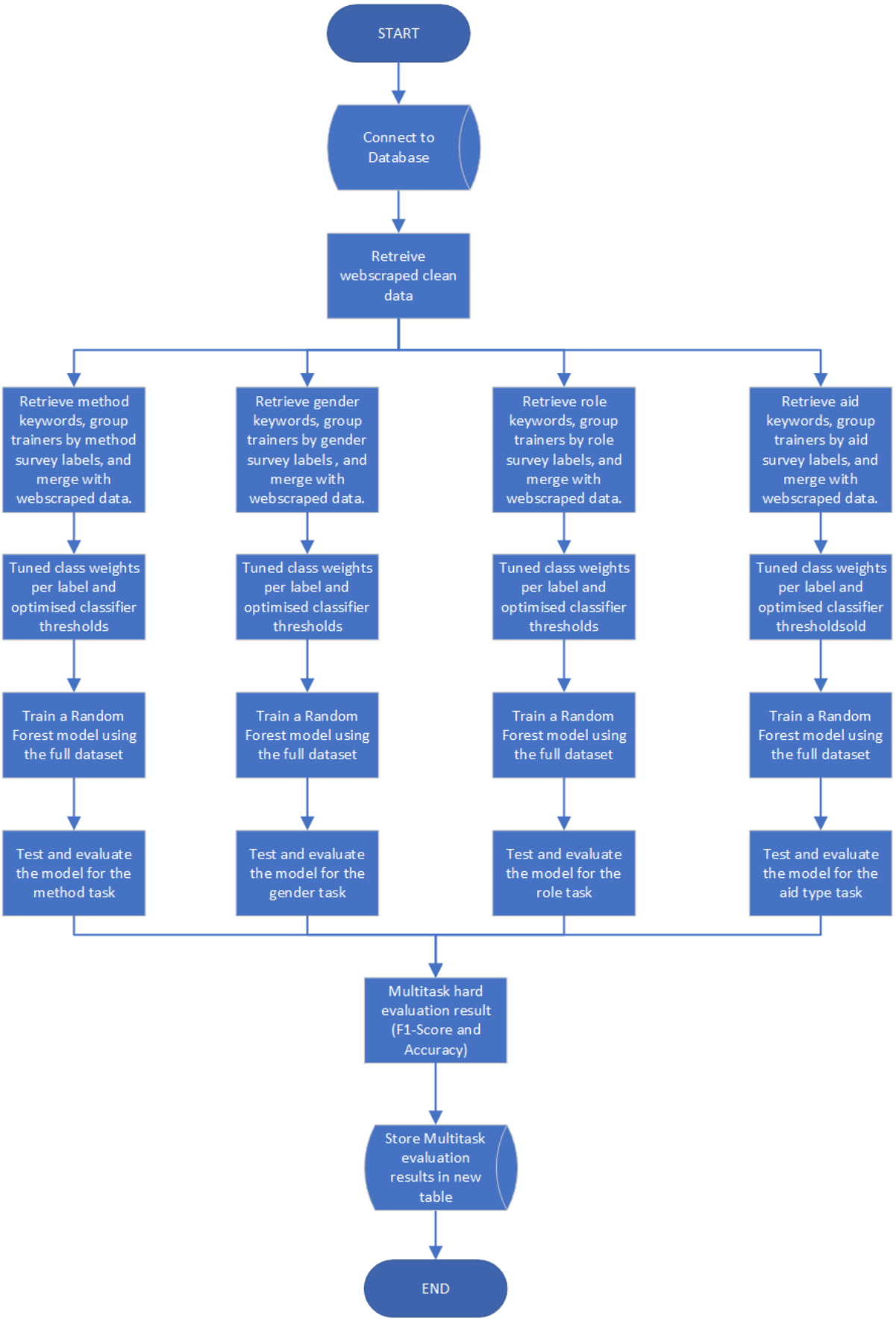

Figure 3.5: Fully Independent Multitask Pipeline with Task-Specific Semantic Representations and Classifiers

Motivation for Independent Parallel Multitask Classification

The decision to adopt an independent multitask classification framework, where each task (method, role, gender, and aid type) is modelled separately, was based on several practical and theoretical considerations.

- **Avoiding label space explosion**: Combining targets like method, gender, and aid type into a single output creates a large and sparse label space (e.g., four methods, three genders, and three aids yield 36 combinations). Many classes would have few samples, harming reliability. Studies show that treating tasks independently improves efficiency and reduces label interference (Ghimire et al. 2024; Xu et al. 2019).
- **Task-specific semantic features**: Keywords in this project were curated per task and aligned to single label groups. Using them across tasks would introduce noise and lower performance.
- **Interpretability and Modularity**: Isolated training enables direct inspection of task-specific keyword effects. This follows recommendations for clearer feature-label analysis (Xu et al. 2019), especially since Random Forests are not optimised for overlapping or loosely related targets like the distinct survey sections used here.

For these reasons, the modelling pipeline adopts four parallel layers, each dedicated to a single task. This setup preserves task-specific features, prevents label interference, and aligns with the conceptual structure of the dataset.

It also improves interpretability. Changes to one task, such as refining aid-related keywords, can be implemented without affecting others. This reflects the real-world independence of attributes like gender and training tools.

Four-Layer Independent Parallel Multitask Classification Training and Evaluation Pipeline

The pipeline begins by loading cleaned website text for each trainer. For every classification task, a task specific set of keywords is retrieved based on survey labels. These keywords are transformed into binary features that record whether each term appears in the website text. Since the keyword sets have been filtered to avoid overlaps, each task operates on a distinct and non interfering feature space. This ensures that the classifier for one task is not influenced by features meant for another.

Once the datasets are prepared, class weights and decision thresholds are tuned for each label within each task. This tuning addresses class imbalance and ensures that predictions remain sensitive to all categories. Each task is then trained using a separate Random Forest classifier on the full dataset. Unlike earlier development stages that relied on cross validation for parameter optimisation, this final training phase uses all available data to capture the most complete set of patterns possible.

Following training, each classifier produces predictions for its respective task. These predictions are evaluated individually using macro averaged F1 score and accuracy, offering detailed insight into performance across all four dimensions of the profile.

In the final step, the outputs of the four classifiers are combined in a multitask evaluation using a hard scoring rule, whereby a prediction is considered correct only **if all four task level predictions are accurate for the same instance**. This strict evaluation helps to identify consistently aligned profiles and determine the most reliable combinations of tasks. The “all-or-nothing” criterion reflects practices in multi-output learning where joint correctness is applied to safeguard the reliability of complete profile predictions (Xu et al. 2019). A comparable concept is employed in composite activity recognition, in which a prediction is accepted only when all constituent activities are classified accurately (Nisar et al. 2023), thereby reinforcing the robustness of the evaluation framework implemented in this study.

Reclassification of Training Aids into Positive, Neutral, and Aversive Categories

Table 3.1: Mapping of specific training aids into grouped aid types used in model training

Aid Code	Aid Name	Mapped Aid Type (Code)
01	Treats	Positive (01)
02	Clicker	Positive (01)
03	Toys	Positive (01)
04	Target stick	Positive (01)
05	Remote treat dispenser	Positive (01)
06	Walking harness	Positive (01)
07	Anti-pulling harness	Neutral (02)
08	Snuffle mat	Positive (01)
09	Crate	Neutral (02)
10	Longline	Neutral (02)
11	Slip lead	Aversive (03)
12	Noise maker	Aversive (03)
13	Water squirter	Aversive (03)
14	Choke collar	Aversive (03)
15	Shock collar	Aversive (03)
16	Citronella collar	Aversive (03)

Exploratory analysis showed that aid type was associated with the trainer’s reported method. Trainers who used aversive aids, such as choke collars, noise makers, and shock collars, tended to report balanced or punishment-based methods, while those using reward-based aids like treats, toys, clickers, and interactive equipment were more commonly linked to positive-only or force-free approaches. This observed relationship provided justification for reclassifying the original sixteen training aids into three semantic categories: positive, neutral, and aversive. Positive aids refer to tools commonly used in reinforcement-based training. Neutral aids, such as crates and longlines, are used to support training logistics and are not inherently reinforcing nor associated with punishment. Aversive aids include tools that apply discomfort or negative stimuli to deter behaviour.

This grouping was also necessary to address modelling constraints. Several of the original sixteen aids were rarely used, resulting in highly sparse and imbalanced labels. These distributions were unsuitable for supervised learning and posed challenges to model stability. By grouping the aids into higher-level categories informed by both functional

characteristics and exploratory findings, the modelling process became more tractable. The revised categories improved label density and interpretability, enabling the use of aid type as a predictive target in the method classification task.

3.6. Robustness Evaluation of Four-Layer Independent Parallel Multitask Classification Using 5-Fold Cross-Validation

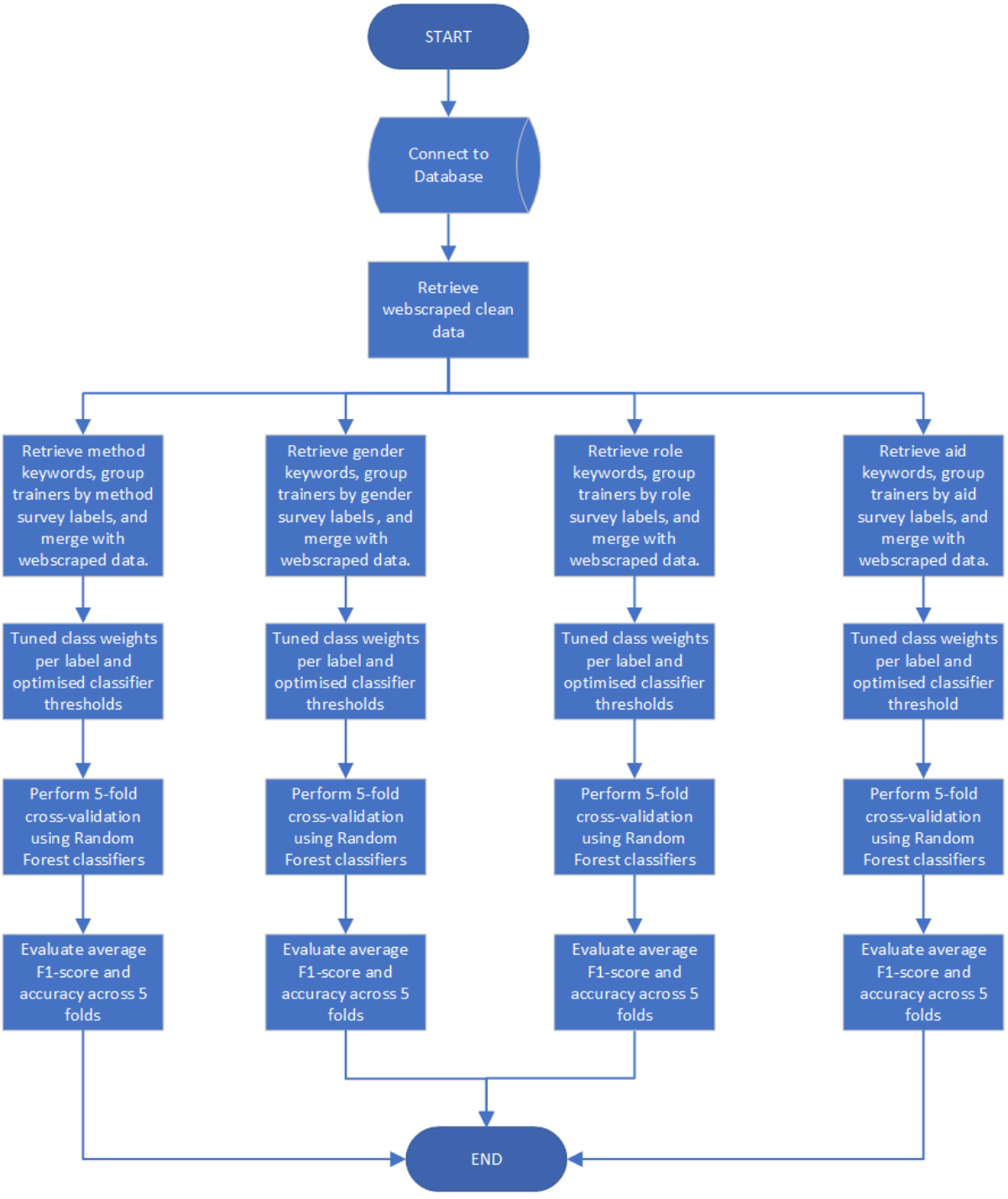

Figure 3.6: Robustness modelling test

This evaluation pipeline assesses performance separately for each classification task: training method, gender, professional role, and training aids. It begins by retrieving pre-processed website text and task-specific keywords derived from survey responses, filtered to ensure exclusivity. These keywords are matched against trainer websites, forming input features for classification.

To address class imbalance, label-specific class weights and classification thresholds are optimized, especially important in multilabel settings such as method or aid type, where label frequencies differ.

Each model is evaluated using five-fold cross-validation. The dataset is split into five parts, training on four and validating on one in rotation. This yields stable and generalisable performance estimates while reducing overfitting risk. Metrics include macro-averaged F1-score and accuracy across folds, offering a balanced view under label imbalance. These results serve as a benchmark for comparing keyword strategies and classifier settings before final deployment.

3.6.1. Controlling Overfitting in Cross-Validation

Initial experiments with grid search and out-of-bag tuning improved cross-validation results but led to poor generalisation, especially on small datasets. The resulting models became overly complex and fitted too closely to training data. This behaviour aligns with prior findings. Han et al. (2021) showed that tuning random forests on small datasets may reduce generalisation and even lower AUC. Probst et al. (2019) noted that tuning node size, sample size, or splitting rules can easily overfit if not controlled.

To address this, we used conservative hyperparameters as structural regularisation. Trees were limited to depth 5, required at least 10 samples to split, and a minimum of 6 per leaf. These choices reduce noisy splits and improve stability. The setting `min_samples_leaf = 6` is supported by Demidova et al. (2019) for improving generalisation.

The number of estimators was fixed at 20 to prioritise speed and avoid excessive variance. Although larger ensembles may improve performance, Breiman (2001) showed that generalisation error stabilises as more trees are added.

Table 3.2: Regularisation settings used in Random Forest classifiers, transposed by task

Parameter	Method	Role	Gender	Aid Type
`max_depth`	5	5	5	5
`min_samples_split`	10	10	10	10
`min_samples_leaf`	6	6	6	6
`max_features`	`sqrt`	`sqrt`	`sqrt`	`sqrt`
`n_estimators`	20	20	20	20

These settings collectively serve as structural regularisation, limiting model complexity and encouraging generalisable patterns across different folds during cross-validation.

3.7. Generating Predicted Profiles for Misalignment Evaluation

Building on the results from Objective 1, where semantic features were validated through k-fold cross-validation, the next phase involved training a final model using the complete dataset. This stage applies the previously validated features to examine how language used on trainer websites corresponds with self-declared identities.

The model was trained on all available data using the same semantic representation, which includes similarity, exclusivity, and prompt alignment scores. This step follows the structured flow of the study. After confirming that these features are meaningful and robust during cross-validation, they are applied to generate predicted profiles for each trainer.

Independent Random Forest classifiers were trained for each task: training method, professional role, gender identity, and use of aids. The class weights and regularisation settings were adopted or slightly adjusted from the previous evaluation to ensure consistency and performance.

The predicted profiles derived from website language were then compared with the corresponding survey responses. These survey labels reflect each trainer's self-declared identity and serve as a reference point for analysing potential mismatches. This phase enables the identification of cases where public-facing communication may diverge from reported practices, offering insight into issues of transparency and consistency in the dog training sector.

3.8. Stepwise Combination Task Selection

In addition to evaluating each task independently and in all possible combinations, a stepwise procedure was used to assess the incremental effect of adding tasks. This reflects real-world deployment, where organisations may prioritise certain predictions before progressing to full profile reconstruction. For example, a regulator might begin with identifying training aids, then extend to role, gender, and method.

The stepwise procedure was implemented as follows:

1. **Identify the strongest baseline task.** Aid type had the most stable performance, with a macro F1-score of 0.9841 and match accuracy of 97.7 percent. It was selected as the starting point.

2. **Sequentially add tasks based on predictive reliability.** Role was added next due to its strong standalone performance and distinct linguistic features. The Aid Type + Role model was evaluated using the same metrics.

3. **Introduce more ambiguous tasks.** Gender was added third, since demographic signals are typically weaker. This step tested whether the pipeline could sustain accuracy under noisier conditions.

4. **Complete the full four-task profile.** Method classification was added last, forming the most difficult setting involving all four targets.

5. **Evaluate at each stage.** At each step, predictions were evaluated using a hard-match rule, where an instance was correct only if all current task predictions were simultaneously accurate. Both macro F1 and match accuracy were recorded.

This stepwise design offers two advantages. It provides incremental analysis of task interactions, highlighting which combinations remain robust and which introduce instability. It also reflects flexible deployment priorities, allowing organisations to choose whether to predict a single task, a few tasks, or the full trainer profile.

3.9. Summary

This methodology was designed to systematically link structured survey data with unstructured website language. By combining text preprocessing, database normalization, contextual keyword extraction, and semantic feature modelling, the pipeline enabled a clear and scalable approach to mapping linguistic profiles to declared practices. The design focused on modularity, interpretability, and readiness for generalization to other domains or larger datasets.

4. RESULT

4.1. Exploratory Data and Analysis

As a preliminary step to building semantic models, we conducted an exploration of both the survey data and the corpus of trainer websites. This phase was essential for shaping the direction of the analysis and determining which keyphrase signals would be most relevant for modelling language patterns tied to training approaches. By understanding how different trainers describe their methods and the diversity in their language use, we could better inform the design of clustering strategies and downstream classification tasks.

From a total of 91 respondents, it was immediately apparent that the self-reported use of training methods was not balanced. Figure **??** illustrates the proportions across all categories, showing a strong leaning toward the “Introduce Rewards” method, which accounted for over 60 percent of selections. Meanwhile, methods based on aversive strategies, namely “Remove Unpleasant” and “Introduce Unpleasant”, were chosen by far fewer respondents. This imbalance in label distribution had direct implications for our modelling strategy, particularly the choice of macro-averaged F1-score as our main evaluation metric. Since some method categories were underrepresented, relying solely on accuracy would have masked poor performance on minority classes.

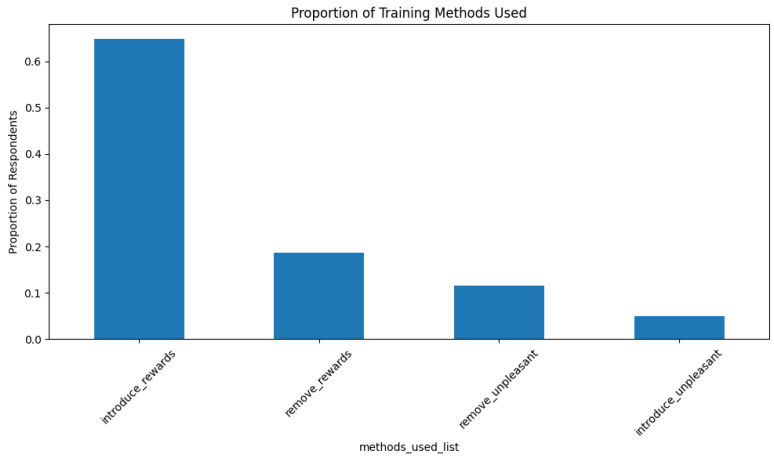

Figure 4.1: Overall distribution of declared training methods across all respondents

Beyond the overall method frequencies, we analysed how these choices were distributed across demographic lines. Specifically, we looked at how gender and professional role appeared to influence a trainer’s preferred approach. Visualised in Figures 4.2 and 4.3, the heatmaps reveal consistent patterns. Female trainers tended to gravitate toward reward-based techniques, especially ”Introduce Rewards”, while male trainers showed a wider spread across different methods, including those involving corrections. Similarly, accredited trainers, those with recognised professional credentials, were more likely to avoid aversive strategies, favouring positive reinforcement and structured guidance instead.

Trainers without formal qualifications displayed more variation in method selection. In addition to the initial exploration of training methods, further analysis of gender, professional role, and training aids provided critical insights that shaped the modelling pipeline.

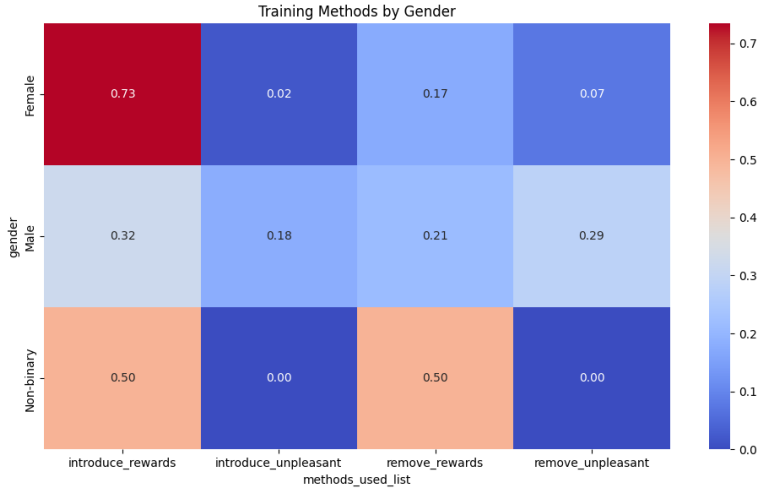

Figure 4.2: Distribution of training methods by gender

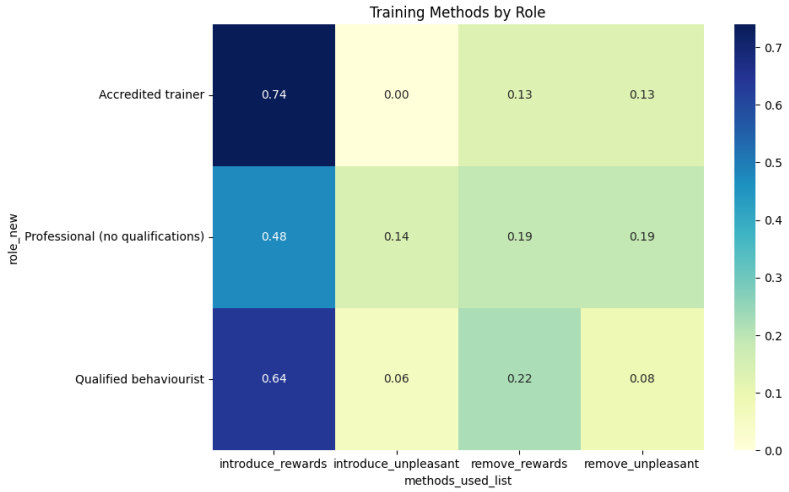

Figure 4.3: Distribution of training methods by professional role

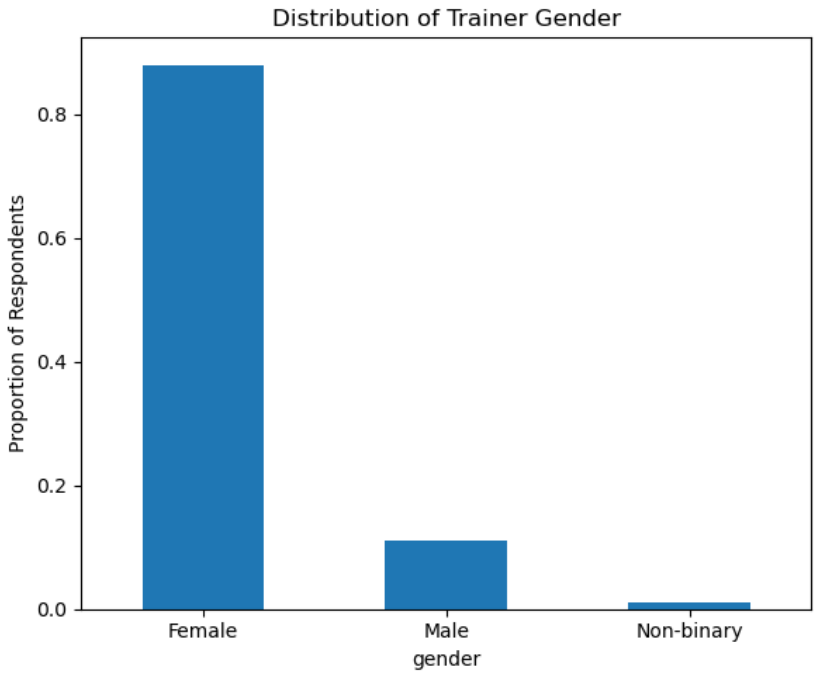

Figure 4.4: Distribution of Trainer Gender

The gender distribution in the survey responses required careful consideration before inclusion in the modelling pipeline. As shown in Figure 4.4, the dataset was heavily skewed toward female trainers, who made up the overwhelming majority of respondents. Male trainers accounted for just over ten percent of the sample, while only a single individual identified as non-binary. This imbalance created a major challenge for classification. A three-way gender split (Female, Male, Non-binary) would have been statistically unstable, since the non-binary category could not be represented consistently across cross-validation folds. With only one data point, any attempt to retain this category independently would risk producing spurious results, and the model would almost certainly overfit to that single instance rather than learning generalisable language patterns. To resolve this, we adopted a reclassification strategy that merged male and non-binary respondents into a single category labelled “Non-Female”. This decision was motivated by the need to preserve both statistical robustness and interpretability. By consolidating the two minority groups, the dataset was reduced to a binary classification problem (Female vs Non-Female), which ensured that both categories contained sufficient examples for training and evaluation. At the same time, the decision acknowledged that the non-binary category could not meaningfully support independent modelling in this dataset. Although this adjustment sacrifices some granularity, it was judged preferable to discarding the non-binary case entirely, as doing so would have erased representation altogether. The reclassification therefore represents a pragmatic compromise: it maintains fairness and inclusivity by retaining the single non-binary case within a broader grouping, while ensuring that the resulting binary task can be modelled reliably without undermining the validity of the analysis.

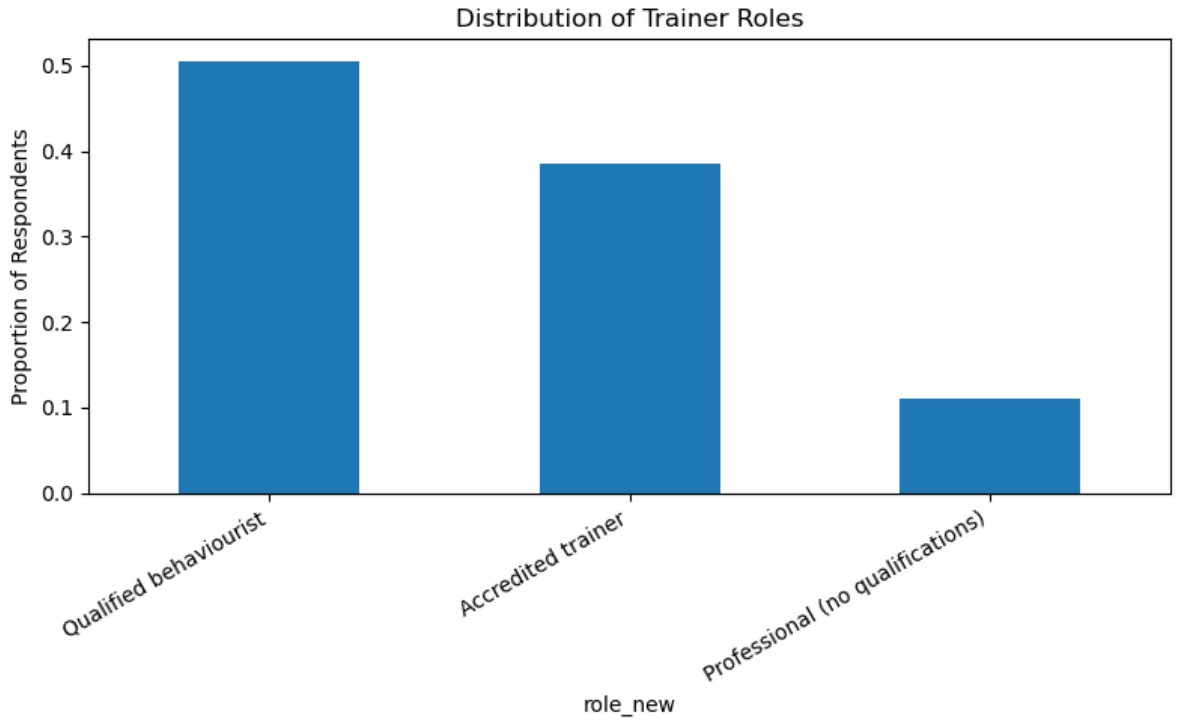

Figure 4.5: Distribution of Trainer Roles

The distribution of professional roles, presented in Figure 4.5, highlights that qualified behaviourists represented over half of the respondents, followed by accredited trainers, while trainers without formal qualifications formed the smallest group. This variation is important, as role-related language proved to be one of the most stable categories during classification. The professional roles are strongly tied to distinctive terminology, with behaviourists frequently using clinical and diagnostic expressions, while accredited trainers referenced structured programmes and certifications. The relatively balanced representation across behaviourists and accredited trainers ensured that role classification could be reliably evaluated without the need for category reduction.

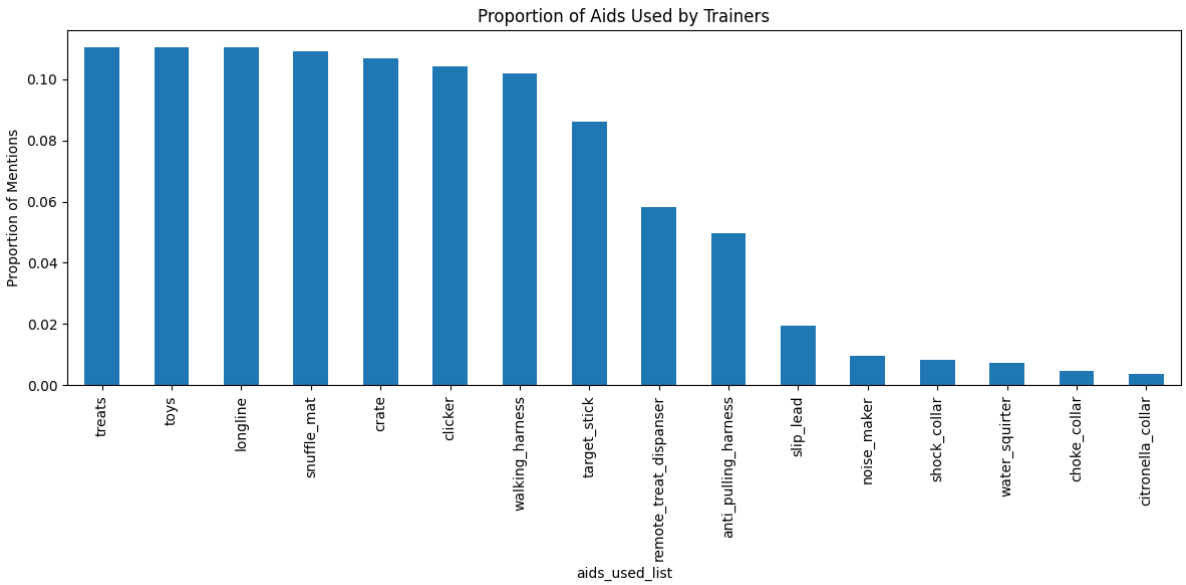

Figure 4.6: Proportion of Aids Used by Trainers

The analysis of training aids, shown in Figure 4.6, indicates that items such as treats, toys, longlines, snuffle mats, crates, clickers, and walking harnesses were frequently mentioned across trainer websites. However, examining these raw frequencies alone would not provide meaningful classification signals, since the list combines aids that differ substantially in training philosophy. For example, positive reinforcement tools such as clickers or snuffle mats appear alongside aversive equipment such as prong collars or shock collars.

To address this, the aids were reclassified into three broader categories: Positive (for example, treats, clickers, snuffle mats, remote treat dispensers), Neutral (for example, crates, longlines, anti-pulling harnesses), and Aversive (for example, slip leads, choke collars, shock collars, citronella collars). This reclassification served two main purposes. First, it improved semantic clarity by ensuring that each category aligned with a distinct training orientation, thereby providing stronger discriminative signals for classification. Second, it prevented bias from high-frequency but generic items such as toys or treats, which could otherwise dominate the feature space without offering meaningful information about underlying philosophy.

Table 4.1: Label code mapping for all classification tasks

Label Code	Training Method	Role	Gender	Aid Type
01	Introduce rewards	Accredited trainer	Female	Positive
02	Remove rewards	Qualified behaviourist	Non-Female*	Neutral
03	Remove unpleasant	No qualifications	-	Aversive
04	Introduce unpleasant	-	-	-

**Non-Female group includes both Male and Non-binary respondents, merged due to class imbalance.*

Following the exploratory phase, we proceeded to extract group-specific keyphrases using contextual embeddings. These embeddings captured meaning beyond surface-level word frequency and allowed for detailed comparison between a trainer's website content and the

overall language profile of their assigned group. By aligning trainer texts to their respective subgroup vectors, we could assess how well their online communication reflected their declared training practices. This process formed the backbone of our modelling pipeline and ensured that the classification model was grounded in real, interpretable language use from the trainers themselves. These exploratory insights also anticipated the later modelling outcomes, where role and aid type proved most stable, while gender and method showed greater variability under stricter evaluations.

4.2. Keyphrase and Topic Representation

Disclaimer: It is also important to note that the extracted keywords cannot be reliably matched directly to raw website text outside the NLP pipeline. The pipeline applies several preprocessing steps such as text cleaning, filtering, and n-gram construction. Because of this, some keywords represent composite forms that may not appear verbatim in the original content. For example, a three-word expression may only emerge after tokenisation and the combination of adjacent terms, meaning that attempts to locate it directly in raw text may not succeed. For this reason, interpretation and evaluation should always be performed through the system to ensure validity and consistency.

4.2.1. Training Method

For method code = 01 (Positive Reinforcement), the extracted keyphrases form a coherent and recognisable cluster that strongly reflects reward-based training. Semantic similarity scores generally fall between 0.34 and 0.65, while final scores after gating range from 0.16 to 0.27. High-scoring terms such as ”positive reinforcement reward”, ”heavily positive reinforcement”, and ”positive reinforcement kind” provide direct textual evidence of operant conditioning principles. These are supported by concrete reinforcer-related phrases including ”food treats”, ”bags treats”, and ”gift growl”. The emphasis on tangible rewards makes this lexicon highly discriminative because it is not only semantically cohesive but also behaviourally grounded. Alongside material reinforcers, the cluster also includes more relational terms such as ”affection”, ”encouragement”, and ”nurturing”. These items scored slightly lower but still remained in the mid-range of the distribution, indicating that social and emotional cues are systematically associated with positive reinforcement practice. The presence of context-specific phrases like ”relaxed happy” handling further signals that discourse in this group blends technical references with welfare-conscious framing. The balanced mixture of reward-oriented and emotional expressions makes this set distinctive, and the filtering process has successfully suppressed generic words, leaving a semantically precise and interpretable lexicon.

For method code = 02 (Remove Rewards), the scores are somewhat weaker but still show consistency. Semantic similarity values range from 0.34 to 0.63, with final scores mostly clustered between 0.09 and 0.22. The strongest signals include ”aggression barking”, ”barks attention”, and ”ignoring barks”, which all point to negative punishment strategies where attention or reinforcement is deliberately withdrawn. Other phrases such as ”lunging

behaviour” and ”care tagged biters” highlight problematic canine actions, reinforcing the corrective framing of this cluster. The vocabulary is less emotionally positive than that of positive reinforcement trainers, but its focus on problem behaviour provides reliable discriminative strength. Importantly, mentions of ”chew collar” and ”prong collars” suggest occasional overlap with more aversive methods, hinting at a spectrum of practices rather than a strictly isolated category.

Method code = 03 (Remove Unpleasant) displays a somewhat different profile. Semantic scores here range between 0.31 and 0.52, with final scores concentrated around 0.15 to 0.24. keyphrasess such as ”aggression anxiety barking” and ”behaviour problems barking” highlight a focus on aversive contexts but are framed in language that softens the severity of interventions. Terms like ”handling caring”, ”play biting chewing”, and ”dealt barking jumping” illustrate attempts to present corrective or mildly aversive practices in a more acceptable light. The discriminative value lies in the contrast between this softer vocabulary and the sharper aversive terms of the next cluster.

For method code = 04 (Introduce Unpleasant), the keyphrases are the most unambiguous. Semantic scores span 0.37 to 0.57, with final scores between 0.16 and 0.24. Phrases such as ”prong collar”, ”shock collar”, ”pack leader behaviour”, and ”excessive barking destructive” explicitly reference aversive devices and dominance-oriented methods. Because these items are tangible and distinctive, the filtering and gating processes preserved them as top-ranked indicators. This cluster therefore offers the strongest discriminative power, clearly separating trainers who adopt aversive methods from those who rely on positive reinforcement.

Taken together, the four method codes reflect distinct patterns in how training practices are expressed, ranging from reward-focused to aversive. Code 01 (Positive Reinforcement) is based on reward-based and relational language, with phrases such as “food treats”, “bags treats”, and “nurturing” highlighting positive experiences and welfare-focused interactions. Code 02 (Remove Rewards) is shaped by problem-oriented expressions like “aggression barking”, “ignoring barks”, and “lunging behaviour”, highlighting strategies that manage undesirable behaviours by withdrawing attention or reinforcement. Code 03 (Remove Unpleasant) combines references to problem behaviours such as “behaviour problems barking” with softer language like “handling caring” and “play biting chewing”, indicating a tendency to moderate the tone of correction. Code 04 (Introduce Unpleasant) includes direct and unambiguous phrases such as “prong collar”, “shock collar”, and “pack leader behaviour”, pointing to the use of physical tools and dominance-based approaches. The score distributions support these distinctions: codes 01 and 04 show stronger and more consistent lexical signals, while codes 02 and 03 fall into more variable ranges with mixed framing. The score distributions reinforce this interpretability: higher and more consistent scores in code 01 and 04 highlight the strength of their lexical signals, while codes 02 and 03 fall into intermediate ranges with more nuanced framing.

4.2.2. Role

For Role Code 01 (Accredited Trainer), the keyphrases highlight professionalism, certification, and structured practice. Semantic values range between 0.28 and 0.62, while final scores span 0.07 to 0.33. High-ranking terms include “canines clients” and “instructor animal behaviour”, with final scores of 0.33 and 0.30 respectively, which strongly suggest formal training and academic grounding. Other items such as ”professional development regularly” and ”privacy certified trainer” underline continued education and ethical commitments. Meanwhile, practical references such as ”obedience agility hoopers” or ”heelwork course canine” reflect structured programmes and technical expertise. The combination of certification-focused and practice-oriented phrases supports the discriminative strength of this cluster.

For Role Code 02 (Qualified Behaviourist), the lexicon carries a more clinical and diagnostic emphasis. Scores generally fall between 0.31 and 0.52, with final scores around 0.08 to 0.31. Strong entries such as ”canine specialist mental” (0.31), ”assessments animal” (0.18), and ”vet visits stressful” (0.16) reinforce the therapeutic and diagnostic orientation. Terms like ”psychological”, ”disorders”, and ”behaviour expert” add to this framing, suggesting links with scientific and medical discourse. More specialised cues like ”breed specific enrichment” and ”transition puppies lifeskills” further distinguish behaviourists by highlighting targeted interventions. Collectively, these terms build a profile of evidence-based practice and clinical expertise.

By contrast, Role Code 03 (Professional, no qualifications) is dominated by emotionally engaging and community-oriented language. Semantic scores are slightly higher overall (0.41–0.66), with several final scores exceeding 0.70 (”animal welfare rescues”, ”issues particular breed”, ”patient knowledgeable pups”). Words such as ”furry companion”, ”puppies in families”, and ”gentle kind learning” evoke warmth, empathy, and accessibility. Phrases like ”fun owners environment” and ”trainer brilliant” suggest a marketing-oriented discourse, emphasising relatability rather than certification. The lack of institutional references strengthens their discriminative value, as the lexicon aligns with practitioners who rely more on personal experience and rapport than formal qualifications.

4.2.3. Gender

Gender-related keyphrasess reveal more subtle but still interpretable patterns.

For Gender Code 01 (Female), semantic similarity scores fall between 0.32 and 0.60, with final scores clustered around 0.15-0.24. keyphrasess such as ”management care puppies” (0.23), ”attentive handler” (0.16), and ”trainer school dedicated” (0.16) emphasise themes of care, attentiveness, and responsibility. Other phrases like ”owners friendly” and ”positive reinforcement teach” reinforce a welfare-conscious orientation. Although practical elements are still visible in terms such as ”collar harness” and ”dealing leash”, they are consistently framed within a caring and relational discourse. This cluster therefore reflects a stereotypically nurturing professional identity.

For Gender Code 02 (Non-Female), the lexicon is sharper and more service-oriented. Semantic scores range from 0.29 to 0.56, while final scores lie between 0.10 and 0.22. Strong

signals include ”great trainer polite”, ”trainer service support”, and ”guarantee trainer bark”, which highlight authority, professionalism, and promised outcomes. Terms such as ”house trained best” and ”confident teach” reinforce competence and results-oriented framing. Technical expressions like ”basics collar” and ”behaviour modification” suggest a pragmatic style. Compared with the softer, relational vocabulary of the female-coded group, these terms foreground authority, efficiency, and measurable success.

The discriminative strength of gender-related keyphrasess lies less in individual terms and more in emphasis: female-associated profiles highlight care and relational practice, while non-female discourse leans toward professionalism, confidence, and results.

4.2.4. Aid Type

The aid type category offers the most direct and discriminative signals, as the terms explicitly reference training equipment. Semantic scores range from 0.09 to 0.45, with final scores consistently between 0.08 and 0.15. On the positive reinforcement side, terms such as ”clicker”, ”target stick”, ”snuffle mat”, and ”remote treat dispenser” represent reward-based tools. On the aversive side, items like ”prong collar”, ”shock collar”, ”choke collar”, and ”slip lead” serve as unmistakable markers of aversive practices.

The gating mechanism plays a crucial role here. Common terms such as ”treats” and ”toys” are down-weighted because they lack discriminative power, even though they appear frequently. Instead, distinctive terms like ”shock collar” or ”clicker” retain higher weights because they unambiguously signal orientation. This ensures that the final feature space prioritises items with clear signalling value.

From an interpretive standpoint, aid-related keyphrasess were among the most stable and reliable across the study. The binary nature of equipment, its presence or absence, makes it a strong proxy for the underlying philosophy. The polarity between these two groups gave this category some of the most robust classification performance, confirming the importance of equipment-based vocabulary as a practical and interpretable feature set.

4.3. Robustness Evaluation and Model Consistency

Table 4.2: Robustness evaluation results for all tasks across folds, grouped by metric

Fold	Metric	Method	Role	Gender	Aid Type
Fold 1	Macro F1	0.9998	0.9030	0.8417	0.9630
	Accuracy	0.9998	0.8889	0.8421	0.9444
Fold 2	Macro F1	0.8601	0.9269	0.8036	0.9630
	Accuracy	0.8421	0.8889	0.7895	0.9444
Fold 3	Macro F1	0.9853	0.9582	0.7976	0.9333
	Accuracy	0.9474	0.9444	0.8333	0.8889
Fold 4	Macro F1	0.9093	0.9998	0.9998	0.9333
	Accuracy	0.8947	0.9998	0.9998	0.8824
Fold 5	Macro F1	0.9115	0.9444	0.8952	0.9091
	Accuracy	0.8947	0.8824	0.8889	0.8235
Average	**Macro F1**	**0.9333**	**0.9465**	**0.8676**	**0.9403**
	Accuracy	**0.9158**	**0.9209**	**0.8708**	**0.8967**

To assess model generalisation and stability across varying partitions, five-fold cross-validation was applied to all four classification tasks using oversampled data. The results in Table 4.2 show that the model generalised well overall, with method, role, and aid type classification maintaining stable performance across folds, while gender fluctuated more noticeably. Given the small dataset of 88 records and the use of oversampling to balance labels, macro F1-score is prioritised as the primary metric, with accuracy serving as a complementary perspective.

Overall, the four tasks demonstrated strong generalisation, with method, role, and aid type maintaining balanced performance across folds, while gender fluctuated more and achieved slightly lower averages. The combination of oversampling, class weighting, and semantic features enabled reliable multitask learning, supporting the robustness of this pipeline for scalable auditing of trainer profiles based on online content.

For the method classification task, the model achieved an average macro F1-score of 0.9333 and accuracy of 91.58 percent. Performance showed strong generalisation across folds, with macro F1 ranging from 0.8601 in Fold 2 to 0.9998 in Fold 1. The remaining folds were stable between 0.91 and 0.98, indicating consistently strong performance. These findings confirm that semantic text features provided a robust signal for distinguishing between training methods across different partitions.

The role classification task yielded an average macro F1-score of 0.9465 and accuracy of 92.09 percent. Results were highly consistent, with macro F1 ranging from 0.9030 to 0.9998 and accuracy spanning 88.89 to 99.98 percent. Fold 4 reached particularly high and balanced performance, underscoring the reliability of this classification. Importantly, all folds maintained

class-level stability, showing that role information was strongly encoded in textual descriptions.

The gender classification task obtained an average macro F1-score of 0.8676 and accuracy of 87.08 percent. Accuracy values remained moderately high (between 78.95 and 99.98 percent), but macro F1 varied more substantially, from 0.7976 in Fold 3 to 0.9998 in Fold 4. Compared with the other tasks, gender classification fluctuated more across folds and showed slightly lower scores on average. Nonetheless, the results indicate that the model was still able to identify gender reliably from text, even if less stably than method, role, or aid type.

For the aid type classification task, the model achieved an average macro F1-score of 0.9403 and accuracy of 89.67 percent. Fold-level scores were stable, with macro F1 between 0.9091 and 0.9630 and accuracy between 82.35 and 94.44 percent. Compared with gender, this task demonstrated steadier results, confirming that aid-related textual features were well aligned with the supervised labels.

4.4. Supervised Semantic Prediction of Trainer Profiles

To evaluate how well semantic features derived from trainer websites could reflect self-declared training practices, we implemented a multitask classification pipeline. This approach allows the model to simultaneously learn multiple targets, specifically the declared training methods, gender, professional role, and the preferred type of training aid. All targets were treated as independent labels within a unified classification framework using shared semantic features, making the task both multi-label and multitask in nature.

Each prediction task was implemented using one or more binary classifiers wrapped into a multitask model, where the same set of semantic keyphrases features was shared across all outputs. These features were constructed by computing the semantic similarity between trainers' website texts and sets of keyphrases that had been extracted and filtered based on declared survey labels. Since each trainer may be associated with multiple categories simultaneously, a multitask approach not only reflects the real-world structure of the data, but also allows for better generalisation and reuse of language features.

This section presents the performance results in three structured parts: single-task classification, multitask evaluation, and robustness assessment. For clarity, each part is summarised in a table, followed by detailed interpretation.

4.4.1. Single Task Results

Table 4.3: Single-task classification results based on survey-aligned labels

Task	Macro F1-score	Accuracy
Training Method (4 labels)	0.983	92.1
Role (3 labels)	0.974	96.6
Gender (2 labels)	0.891	94.3
Aid Type (3 labels)	0.984	97.7

The single-task results show robust performance across all label types. Training method prediction achieved macro F1 of 0.983 and accuracy of 92.1 percent, suggesting that even nuanced method distinctions are reflected in public-facing language. Gender classification performed slightly lower, with macro F1 of 0.891, though accuracy remained strong at 94.3 percent, indicating that some demographic cues were more ambiguous than others. Role classification reached 96.6 percent accuracy, with macro F1 of 0.974, reflecting the distinct ways that different trainer roles present themselves online. Aid type classification continued to perform well, with macro F1 of 0.984 and accuracy of 97.7 percent. This strong performance is likely supported by the earlier reclassification of the original sixteen tools into three functional categories: positive, neutral, and aversive, based on their intended use in training. These categories were constructed directly from the survey-reported list of specific aids, and the model used this grouping as the basis for its aid type features. This restructuring reduced label sparsity and clarified semantic boundaries between aid types, enabling stronger and more consistent signal capture. Overall, each task produced clear signals when evaluated independently, validating the semantic separability of the target categories.

4.4.2. Independent Multitask Results

Table 4.4: Hard match evaluation across all task combinations

Task Combination	Macro F1-score	Match Accuracy
Method + Gender	0.9522	87.5
Method + Role	0.9790	88.6
Method + Aid Type	0.9835	90.9
Gender + Role	0.9404	92.1
Gender + Aid Type	0.9467	92.1
Role + Aid Type	0.9789	94.3
Method + Gender + Role	0.9593	85.2
Method + Gender + Aid Type	0.9628	86.4
Method + Role + Aid Type	0.9805	87.5
Gender + Role + Aid Type	0.9568	89.8
All Tasks (Method + Gender + Role + Aid Type)	0.9655	84.1

Combining task predictions introduces a stricter evaluation scenario by requiring simultaneous correctness across multiple dimensions. As expected, overall performance declined slightly as more labels were included, reflecting the increased difficulty and potential for mismatch. Among two-task combinations, the strongest result was obtained for Role and Aid Type, with a macro F1 of 0.9789 and match accuracy of 94.3 percent. This suggests that linguistic patterns related to trainer role and tool use are both distinct and frequently co-occurring in a structured way. The combination of Method and Aid Type also performed well (F1 = 0.9835), likely because website descriptions of training tools tend to reflect the underlying training orientation. In contrast, combinations involving Gender showed lower F1 scores, which may be due to the subtler and more indirect linguistic cues associated with demographic identity. The most challenging condition required a correct match across all four targets. This produced a macro F1 of 0.9655 and a match accuracy of 84.1 percent,

indicating that full profile reconstruction remains feasible from text alone, though with reduced precision as complexity increases.

4.4.3. Greedy Stepwise Evaluation

Table 4.5: Greedy stepwise task selection results (sorted by added task)

Stepwise Combination	Macro F1-score	Match Accuracy
Aid Type	0.9841	97.7
Aid Type + Role	0.9789	94.3
Aid Type + Role + Gender	0.9568	89.8
Aid Type + Role + Gender + Method	0.9655	84.1

While the multitask results demonstrated the feasibility of reconstructing trainer profiles, we further explored a stepwise evaluation to understand how each additional task contributes to predictive reliability. This analysis begins with the single most reliable task and sequentially incorporates additional dimensions, allowing observation of how each layer impacts overall prediction reliability. Such a procedure mirrors a potential deployment scenario, in which an organisation may choose to prioritise certain dimensions before expanding to full profile reconstruction.

The results confirm that Aid Type provides the most stable foundation, with nearly perfect match rates on its own. Adding Role preserved high performance, suggesting that tool use and professional title are often expressed in linguistically consistent ways. Incorporating Gender reduced match accuracy more sharply, reflecting the weaker and subtler nature of demographic cues. The addition of Method produced the full four-task combination, which lowered performance further but still maintained a macro F1 above 0.96. This pattern validates the modular design of the pipeline and shows how organisations could flexibly decide which dimensions to prioritise.

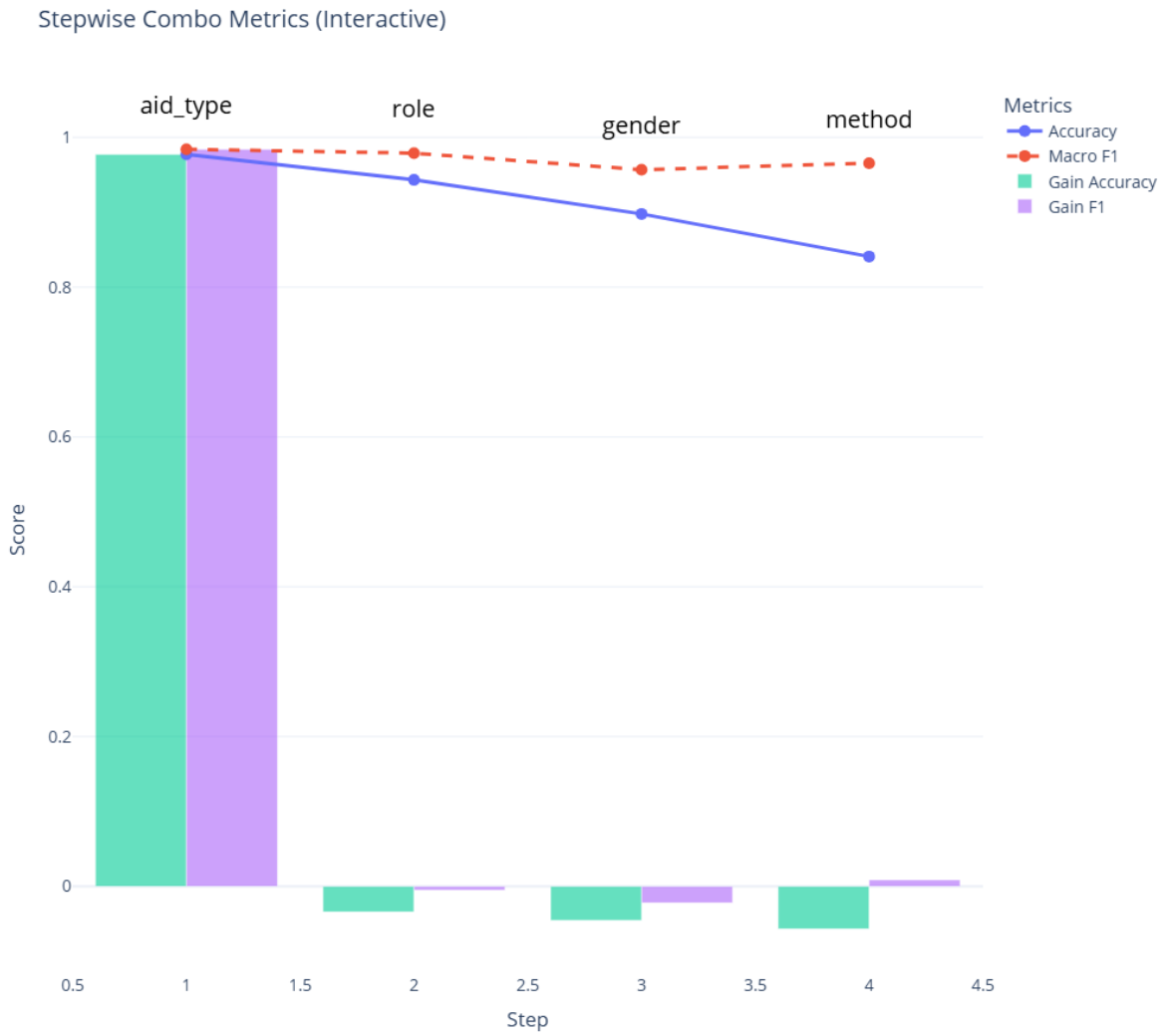

Figure 4.7: Greedy stepwise evaluation metrics across incremental task combinations

4.4.4. Alignment Outcomes

To move from aggregate metrics toward individual-level interpretation, we examined alignment outcomes between predictions and self-reports.

Table 4.6: Summary of Alignment Outcomes (Hard-Match Evaluation)

Outcome	Number of Trainers (N=88)
Fully aligned across all four dimensions	74
At least one mismatch	14
Most stable dimensions	Role, Aid Type
Most variable dimensions	Gender, Method

Beyond aggregate metrics, it is equally important to examine how predictions translate into concrete alignment between website language and trainer self-reports. To address Objective 2, the model outputs were compared with survey responses at the individual trainer level. Under the strict hard-match criterion, where all four labels (method, gender, role, and aid type) must

be predicted correctly at once, the system reproduced full profiles for 74 out of 88 trainers. This represents approximately 84 percent of the sample achieving complete alignment.

The remaining 14 trainers showed at least one area of mismatch. Inspection of the results stored in the evaluation database indicates that these discrepancies were not randomly distributed. Instead, they concentrated in particular categories. Gender predictions were the most variable, with several cases where the model assigned the opposite label despite high confidence in other dimensions. This supports earlier findings that gender-related language cues are subtle, less distinctive, and therefore more difficult to classify reliably. Method labels also contributed to some mismatches, particularly where websites contained a mixture of positive reinforcement terms alongside references to corrective or aversive strategies. In such cases, the ambiguity of the website narrative may have genuinely reflected hybrid practice rather than a clear misrepresentation.

By contrast, role and aid type classifications proved more stable. Trainers identified as accredited or behaviourist were usually predicted correctly, reflecting the discriminative power of professional terminology such as “certified,” “assessment,” or “behaviour expert.” Similarly, aid type benefited from explicit lexical markers, with tools like clicker or prong collar providing unambiguous evidence of orientation. The combination of these two categories also yielded the highest pairwise alignment rates, consistent with their strong individual stability observed in earlier single-task evaluations.

The interpretation of alignment outcomes suggests two key insights. First, systematic mismatches can be informative rather than merely errors. They point to areas where trainers’ public messaging diverges from their declared practice, which is precisely the type of inconsistency that Dogs Trust may wish to investigate. Second, the modular design of the system allows stakeholders to calibrate which dimensions to prioritise. For example, relying on method and aid type predictions may provide a more robust initial screening, while adding role and gender can tighten the alignment check at the cost of reduced stability.

Taken together, these findings confirm that the pipeline does more than deliver aggregate accuracy. It provides a structured mechanism for flagging trainers whose public websites suggest potential misalignment with self-reported practices. While the majority of trainers displayed strong consistency, the subset of mismatched cases demonstrates the practical value of this tool for supporting transparency and accountability in the sector.

These results collectively demonstrate that semantic signals extracted from trainer websites can robustly reproduce declared practices while surfacing meaningful cases of misalignment, providing both methodological validation and practical utility for stakeholder-driven screening.

4.5. Exploring Language-Based Inference Without Survey Data

This section presents an exploration of how the developed pipeline performs when applied to external language data in the absence of structured survey responses. The aim is to evaluate whether the setup can generate relevant and structured outputs from naturally occurring text, using the same inference pipeline built and validated in our methodology stages. The process is designed to test the technical feasibility of language-based inference methods using practitioner-authored website content as a potential input source, and to

consider their applicability for future research contexts.

To support this exploration, trainer website text was considered as prospective input to the existing pipeline. As outlined in the preprocessing workflow, the content from each source would first be cleaned to prepare it for embedding. Multiple pages could be combined into a single document, and standardised elements such as repeated headers, menus, and template-based sections would be removed. Generic promotional phrases would also be filtered out using a predefined list to help retain only the language that is likely to reflect the trainer's own communication style. The result is a refined textual representation suitable for the next stage of inference.

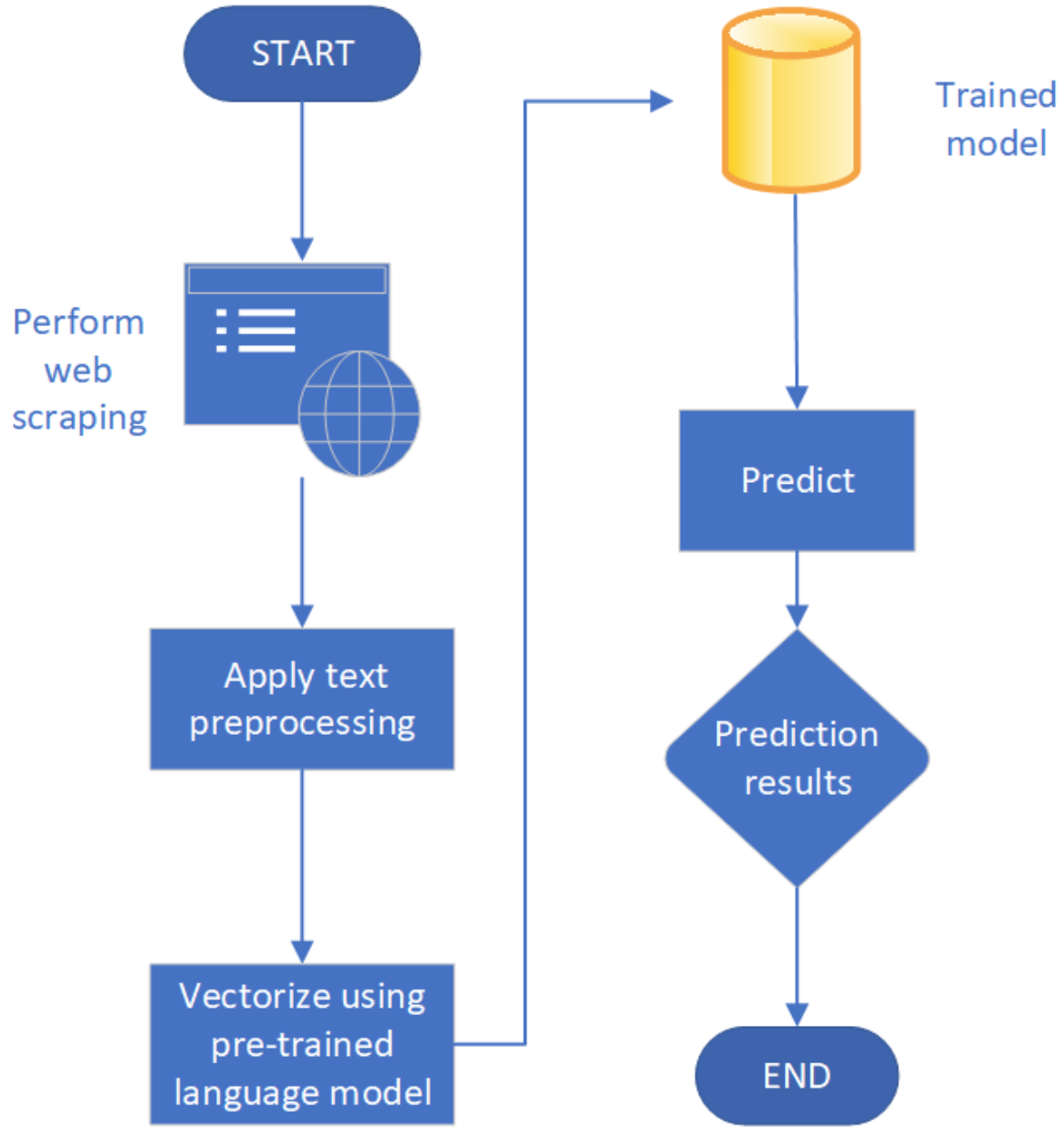

Figure 4.8: Exploring Prediction Without Survey Data

The cleaned text would then be converted into numerical vector representations using a pre-trained language model. These embeddings capture the semantic structure of the content

and enable comparison with semantic features developed during Objective 1. Specifically, the embedded text would be aligned with validated keyphrase sets associated with four prediction targets: training method, professional role, aids used, and gender. Each keyphrase is linked to weighting parameters such as semantic score, semantic prompt score, gate factor, and final score, each derived through conceptual alignment and statistical validation.

Cosine similarity between the embedded input and each keyphrase vector would then be calculated, with similarity values adjusted by their corresponding weights. This produces gated similarity scores that form the basis of the feature vector for each prediction label. Additional components such as the average similarity, the highest match score, and an indicator for the presence of strong matches are also included.

These constructed feature vectors would be passed into the final classification models developed in Objective 2, which had been trained and validated using the complete internal dataset. To mirror a typical inference workflow, the inputs are treated as unseen examples and passed through the model without any additional training or adjustment. The predictions are generated using the same thresholds defined during model validation. This step reflects the continuity of the logic established across Objectives 1 and 2, leveraging validated semantic features to explore how the pipeline might respond to new inputs.

The purpose of this exploration is twofold: to assess the technical feasibility of applying the inference workflow to new, unstructured inputs, and to inform future research directions where language-based alternatives to survey methods may be valuable. This aligned inference workflow may be particularly relevant in research settings that require large-scale evaluation but face practical constraints in collecting direct responses. The results highlight the potential of using natural language inputs to support structured inference, especially when survey-based data collection is costly or resource-intensive.

5. DISCUSSION, LIMITATION, AND FUTURE RESEARCH DIRECTIONS

5.1. Discussion

Overview

This discussion examines the key technical and conceptual contributions of the study, with a focus on the effectiveness of the semantic feature pipeline, the interpretability of extracted keyphrases, and the ability of the classification framework to reflect trainer identity through language. It begins by considering how well the extracted features capture meaningful patterns across trainer types, using classification as a sanity check rather than a predictive goal. The role of keyphrase design and feature construction is then explored in greater depth, followed by an evaluation of robustness under real-world constraints such as class imbalance and label sparsity.

The second half of the discussion turns to the issue of misalignment between website content and survey-reported practice. It assesses the extent to which inconsistencies can be detected through modular classification and strict hard-matching. Finally, broader implications are considered, including how these findings may inform future research in contexts where direct survey data are unavailable, and how NLP-based pipelines can support transparency in under-regulated professional sectors.

5.2. Objective 1: Identifying Linguistic Patterns Across Trainer and Using Classification as a Sanity Check

The first objective was not to deploy a classifier for practical use, but rather to evaluate whether meaningful and interpretable keyphrases could be extracted from website language. In this context, the classification task served as a semantic sanity check, testing whether the extracted features, built from trainer-declared categories, carried signal strong enough to support consistent prediction. Overall, the results confirm that such patterns do exist and can be captured especially for training method, professional role, and aid type. Robustness testing under cross-validation further showed that the model generalises well despite real-world constraints like class imbalance and label sparsity. These findings validate the quality of the semantic features and provide a foundation for building interpretable resources that help organisations and dog owners make sense of linguistic cues in an unregulated training landscape. Importantly, the success of this stage also established a validated feature space that underpins the misalignment analysis in Objective 2.

5.2.1. Pipeline Design and Evaluation

The pipeline developed in this study was implemented in a reproducible environment, ensuring that each step from data ingestion and preprocessing to embedding, feature construction, and classification could be reliably repeated and adapted in future analyses. This reproducibility was important not only for validating the modelling process, but also for examining how the approach might generalise to previously unseen inputs, as considered in the inference stage.

An additional consideration in the design was its sustainability, allowing the pipeline to be extended in an exploratory phase where predictions could be generated without the support of survey responses. This continuity reflects the stability of the feature engineering process and suggests its relevance for future work seeking to derive structured insights from natural language independently of self-reported information.

Two main challenges emerged during the development of the system. First, standard text cleaning alone did not isolate useful linguistic signals. Many generic terms appeared across trainer types, making it harder to distinguish between classes based on language use. To address this, a filtered dictionary was developed to remove non-discriminative words and retain only those that were more likely to reflect each trainer's specific orientation.

Second, the limited dataset size necessitated stronger semantic representation. Thus, contextual embeddings from a pretrained sentence transformer model were used to enrich the input space, capturing deeper conceptual meaning and improving feature generalisability.

5.2.2. Keyphrase Extraction and Feature Interpretability

The overall success of the project depended on the extraction and refinement of keyphrases, which constituted the most technically demanding and conceptually significant stage of this study. This process functioned as the central bridge between unstructured, often ambiguous website language and structured, interpretable features that could be used for semantic classification. It formed the backbone of the model's predictive capability and played a foundational role in ensuring that the final outputs carried both contextual relevance and discriminative power. Without accurate and well-calibrated keyphrases, subsequent classification models would not have had access to sufficiently informative signals to support either objective. Thus, this component was pivotal not only for the semantic validation of features in Objective 1, but also for the validity of the misalignment analysis in Objective 2.

All downstream inference depended on the reliability and distinctiveness of these keyphrases, making this step the true cornerstone of the project and a determining factor in the overall success of the study. Each keyword was linked to a specific label grouping derived from survey responses and then evaluated for relevance, distinctiveness, and exclusivity. Terms shared across label types were excluded to avoid noise. The process relied on both algorithmic filtering and human-in-the-loop refinement to ensure interpretability.

The final features were interpretable vectors based on gated scores combining semantic similarity, alignment with prompt descriptions, and inter-group divergence. Feedback from Dogs Trust confirmed that the selected terms aligned with professional expectations, further validating the feature space.

5.2.3. Robustness as a Sanity Check for Keyphrase Extraction

To assess model robustness, five-fold cross-validation was applied alongside fold-specific guided oversampling and classifier regularisation. This process was not intended to optimise the final model used for inference, but to create a testing environment that realistically mirrored the kinds of imbalance, sparsity, and underrepresentation that models face in real-world settings. The underlying principle was clear: if the classifier could still produce consistent predictions under stress, then the keyphrases it relied on must reflect meaningful semantic distinctions. Oversampling was used solely within training folds that would otherwise contain no examples for a given label, treating this step with care to avoid distorting the label distribution. Regularisation was also applied to the Random Forest classifier to mitigate overfitting, which was especially important given the small dataset and the presence of class sparsity. By isolating this five-fold cross-validation phase from the final prediction pipeline, the analysis maintains methodological integrity while providing evidence that the extracted features can generalise reliably.

The classifier showed consistently high macro F1 scores across the primary tasks. Training method, professional role, and aid type yielded averages of 0.93, 0.94, and 0.94, respectively, demonstrating that the semantic features retained their predictive value even under artificially balanced conditions. Gender performance fluctuated more substantially, with scores ranging from 0.79 to 0.99. This variability aligns with earlier findings that gender-related cues in professional language are often subtler, more culturally sensitive, and sometimes entangled with stylistic rather than functional language choices. In light of this, gender predictions should be interpreted with caution and not overgeneralised as fixed identity markers.

Taken together, these results confirm that Objective 1 has been achieved. The classifier not only performed reliably under stress, but also validated the semantic features extracted from trainer-authored text. These findings provide a solid foundation for the misalignment evaluation in Objective 2, demonstrating that key linguistic signals are both consistent and robust under real-world constraints.

5.3. Objective 2: Evaluating Misalignment Between Website and Self-Reported Practice

The second objective focused on evaluating whether predictions based on website content aligned with the profiles declared by trainers in the survey. To achieve this, the pipeline was designed in a modular way, allowing each task to be evaluated independently or in combination. This flexibility reflects practical needs, since Dogs Trust may choose to focus on individual categories such as training method, or on complete trainer profiles that integrate multiple dimensions.

The evaluation of misalignment was based on a modular design of the classification system. Instead of forcing all tasks into a single framework, the pipeline was built as independent models that can be evaluated separately or combined depending on the needs of the stakeholder. This design choice was made deliberately to reflect that Dogs Trust or other organisations may have

different priorities in practice. For example, if the focus is only on training method, the single-task evaluation already provides high confidence. If the interest is in full trainer profiling, then multiple dimensions such as role and aid type can be combined under a stricter assessment.

This modular approach is also justified by the varying robustness of different tasks. The robustness testing showed that method, role, and aid type produced consistently strong performance, while gender was less stable. Based on this, Dogs Trust can choose a modular evaluation strategy: they may rely on method and aid type when stability is the priority, or combine multiple dimensions when a stricter full-profile assessment is required.

To operationalise this flexibility, both single-task and independent multitask evaluations were implemented. Hard-matching was introduced in the multitask setting, requiring all predicted labels to be correct simultaneously. This created a stricter alignment check that highlights potential misalignment when trainers present inconsistent signals across multiple dimensions. The relatively higher performance of single-task models is expected, since each classifier concentrates only on one label group, reducing the complexity of decision boundaries. This demonstrates that alignment between survey data and website language is strong when assessed dimension by dimension.

Single-task evaluation confirmed this, with training method, role, and aid type achieving very high results while gender remained weaker. When tasks were combined under independent multitask evaluation, performance naturally declined under the hard-match criterion, yet results still showed that full trainer profiles can be reconstructed from text. The greedy stepwise evaluation highlighted that aid type carried the most stable signals, largely due to the use of a predefined list of recognised training tools that anchored each category. This structured mapping provided clear semantic distinctions, enhancing the model's ability to extract consistent features. As the label space expanded, adding role maintained robustness, but the inclusion of gender reduced stability, and the final inclusion of method lowered strict-match accuracy though still at a reliable level.

The hard-match evaluation across all four labels showed that 74 out of 88 trainers were aligned with their survey profiles, while 14 exhibited at least one mismatch. This represents around 16 percent of the sample and provides a concrete measure of misalignment beyond aggregate F1 performance.

Notably, most mismatches involved either the gender or method categories, aligning with the earlier observation that certain dimensions are less semantically stable. This reinforces the pipeline's utility in flagging inconsistencies that merit closer scrutiny by the organisation.

Important note: The weaker performance of gender in the single-task evaluation directly confirms the findings from robustness testing in Objective 1, where gender also showed greater variability across folds. This consistency across analyses indicates that gender-related cues in trainer websites are inherently weaker and more culturally sensitive than other categories. As such, gender predictions should be treated with greater caution, while confidence in method, role, and aid type classifications remains strong.

In summary, these results confirm that the second objective has been achieved. The pipeline successfully reproduced survey-based categories from website content and identified systematic cases of misalignment under stricter evaluation. Its modular design provides Dogs Trust with the flexibility to focus on single dimensions or to assess complete trainer profiles,

making the system a practical and adaptable tool for supporting transparency in the sector.

5.4. Limitation

While the study provides valuable contributions, some limitations must be noted. The reliance on general-purpose pretrained embeddings, built from broad non-specialised corpora, may not fully capture domain-specific language in dog training. Contextual terms, newer expressions, brand-related language, or informal trainer jargon might be misrepresented in the embedding space. Without fine-tuning on domain-relevant texts, the model risks underrepresenting subtle linguistic cues that are important in practice.

Secondly, using a master list of aid types ensured consistency in keyphrase extraction but may reduce adaptability as trainers introduce new terms or reframe aids more subtly. Softer or creative descriptions of aversive tools may fall outside fixed categories. As the field evolves, a more flexible or data-driven approach could improve long-term relevance.

Another methodological consideration is the manual refinement of keyword dictionaries. While this step preserved interpretability and semantic alignment, it introduced subjectivity. Human involvement in curating and validating keywords may cause inconsistencies or bias and could be difficult to sustain when adapting to new domains or evolving terminology.

From a modelling perspective, separate classifiers for each label group support modularity and targeted evaluation, but they do not capture relationships between labels. A trainer's declared method may influence their role, gender, or choice of aids, and overlapping keyphrases can appear across these groups. Ignoring such interdependencies may limit the expressiveness of the extracted features.

In addition, the study assumes that survey responses accurately reflect ground truth. While surveys are valuable for capturing self-reported identity, there is no guarantee that responses are always reliable. Trainers may interpret questions differently, respond based on aspirations rather than actual practice, or provide incomplete information. These inconsistencies introduce noise that may affect both training and evaluation.

From an ethical standpoint, the use of web-scraped data should be approached with caution, particularly for sensitive attributes such as gender or professional credibility, to avoid reinforcing bias or misrepresentation. Inference from such data should be limited to authorised parties to ensure contextual appropriateness and ethical integrity.

Lastly, the dynamic nature of language and professional practice in this domain presents an ongoing challenge. Changes in website content, evolving terminology, and shifting norms may gradually reduce the model's effectiveness if not addressed through continuous retraining and evaluation.

5.5. Future Research Directions

Looking ahead, several promising directions could build upon this work. One particularly important consideration is how the system can remain effective over time as trainer language evolves. For this reason, concept drift emerges as a suitable and timely candidate for future development.

5.5.1. Preparing for Concept Drift and Retraining

One key consideration for future development is the fact that professional language is not static. In the dog training sector, the ways trainers describe their methods, tools, and philosophies evolve over time, shaped by new research, shifting norms, and changing audience expectations. As a result, models built on current linguistic patterns may gradually lose their ability to reflect future usage. In this context, concept drift becomes a natural and important direction for future research.

The strength of the current approach lies in the continuity between its stages. The semantic features developed and validated in Objective 1 were not constructed for a one-off classification task, but as reusable signals that can support inference across new and unseen inputs. Objective 2 demonstrated that these features could reliably capture patterns of alignment or misalignment when applied to real-world data. Both objectives relied on the same logic of semantic validation, ensuring that any forward-looking adaptation remains grounded in the structure already established.

To maintain relevance over time, the model should be equipped with retraining mechanisms that allow it to respond to emerging trends in language. This may involve collecting updated website content at regular intervals, reprocessing it through the existing feature extraction pipeline, and retraining the classification models based on refreshed inputs. Because the system was built with modularity and reproducibility in mind, this retraining does not require changing the underlying architecture. Instead, it simply refreshes the training data and adjusts model weights to reflect current linguistic usage.

Such updates can be scheduled routinely or triggered by specific indicators, such as performance degradation or the appearance of novel vocabulary. Monitoring tools can be developed to detect when certain keywords begin to lose predictive value, prompting targeted updates to the feature space. When needed, new keyphrases can be incorporated and outdated ones removed, all within the same semantic scoring framework established earlier.

By embedding this adaptability into the original design, the study ensures that its contributions remain meaningful even as the language of the sector evolves. For organisations such as Dogs Trust, this opens the possibility of maintaining an evidence-informed resource that stays relevant over time, helping dog owners interpret trainer language in a sector where regulation is limited and linguistic clarity is critical.

5.6. Summary

In conclusion, this study shows that trainer website language carries semantic signals aligned with self-reported practices. Using structured feature extraction, keyphrase-based classification, and rigorous validation, these patterns were found to be consistently interpretable across method, role, and aid use.

The findings highlight how professional claims are reflected in public language and demonstrate that inconsistencies between language and practice can be systematically identified. While categories such as gender remain more variable, the results provide a foundation for future academic and sector-level work on transparency and ethical engagement with professional discourse.

Ultimately, this research offers dog owners a principled way to interpret trainer language and assess its consistency with claimed practice in a largely unregulated sector.

6. CONCLUSION

This study introduced an interpretable NLP-based framework for examining how UK dog trainers communicate their professional identity through website content. Aimed at supporting dog owners in making informed decisions, the framework identifies consistent patterns between public language and self-declared training practices. Grounded in survey data and validated through a structured semantic pipeline, it offers both technical rigour and practical relevance.

The research was guided by two primary objectives. The first was to evaluate whether a modular NLP pipeline could identify consistent and meaningful linguistic patterns in trainers' website content. Classification was used as a semantic sanity check to validate whether the extracted keyphrases captured these patterns reliably. The second objective focused on comparing the predictions to trainers' self-reported profiles to identify consistency or misalignment. These objectives shaped the framework design and guided its interpretative goals in an unregulated sector.

While the application focused on dog training, the framework may extend to other NLP tasks that require assessing alignment between declared categories and public language. Such challenges are common in domains where interpretability, consistency, and transparency are critical. The framework's broader potential is grounded in four key contributions detailed in this study:

First contribution: reversed workflow

One of the key methodological contributions of this project is its reversed workflow. Unlike standard classification studies that infer categories from unlabelled text, this pipeline begins with structured self-reported categories from survey data and examines how these identities are reflected in website language. Classification is used as a consistency check to assess whether linguistic patterns align with declared practices.

This reversed setup reshaped every stage of the pipeline. Instead of extracting features from raw text and assigning categories post hoc, the system traced back from known categories to identify distinctive and interpretable phrases that could carry semantic signal. In doing so, the project elevated classification from a technical task into a form of validation. The classifier became a lens for examining alignment between identity and expression, not merely a label generator.

Few applied NLP pipelines adopt this inverted approach, especially with strong grounding in real-world self-report data. Anchoring the model in declared professional claims ensures that outputs remain robust, interpretable, and relevant for end users. This structure supports the development of transparent tools that help dog owners interpret online communication and make informed decisions in an unregulated sector.

Second contribution: developing a scoring mechanism for extracting discriminative keyphrases

A second methodological novelty was the design of a keyphrases scoring algorithm that could meaningfully represent how keywords align with both website content and the underlying training philosophies. Instead of relying on default similarity metrics or

frequency-based rankings, the scoring function in this study was carefully constructed from three key components: semantic alignment with web content, conceptual alignment with predefined prompts, and topical exclusivity measured through distributional divergence. These elements were combined using fixed weights, guided by recent findings in supervised keyword extraction and semantic ranking.

To prevent misleading contributions from vague or overly generic terms, a gating function was also introduced. This function adjusted the influence of each keyword depending on how well it aligned with the conceptual definition of a given category. Keywords with low prompt alignment were downweighted, while stronger candidates were preserved and scaled proportionally. This gating design was not drawn from any existing pipeline directly, but emerged through iterative testing and manual evaluation of keyword relevance across tasks.

Each keyword was then represented not by a single score, but by a block of semantic and conceptual indicators, including final gated scores, prompt relevance, and average document similarity. These scores formed the foundation of the semantic feature matrix, which served as the input to the classification models.

By combining semantic similarity, concept-specific alignment, and exclusivity in a single gated score, the resulting features maintained both discriminative power and interpretability. This scoring strategy was essential for translating unstructured website content into structured representations that could reflect trainer identity across multiple dimensions. The formulation itself was calibrated through multiple rounds of refinement, drawing from both theoretical references and empirical testing during model development. The result is a flexible and meaningful representation space that anchors the entire modelling pipeline.

Third contribution: modular classification design

The third key contribution of this study is the modular classification workflow that supports both single-task evaluation and multitask inference. Instead of a unified model, four independent classifiers were trained for method, role, gender, and aid type. This design improves interpretability, targeted assessment, and adaptability.

In single-task evaluation, results were consistently strong: method (F1 0.983, accuracy 92.1%), role (F1 0.974, accuracy 96.6%), and aid type (F1 0.984, accuracy 97.7%) all performed well, while gender was less stable (F1 0.891, accuracy 94.3%).

The modular setup also enabled a combined multitask evaluation under a hard-match condition, where a prediction was considered correct only if all predicted labels matched the ground truth simultaneously. As expected, overall accuracy declined with increased task complexity. The full four-task configuration resulted in a macro F1-score of 0.9655 and match accuracy of 84.1 percent. This stricter evaluation criterion provides a conservative but informative check on alignment, highlighting where trainer communication may be less aligned with their declared profile, offering insight into how information is conveyed online.

To assess each task's contribution, we conducted a stepwise evaluation under the hard-match criterion. Starting with aid type, the most reliable task, the model achieved 97.7% match accuracy (F1 = 0.9841). Adding role preserved high performance at 94.3% (F1 = 0.9789). Incorporating gender reduced performance to 89.8% (F1 = 0.9568), while including method produced the full profile at 84.1% (F1 = 0.9655). This progression shows aid type and role as stable anchors, while gender and method introduce greater variability.

Final evaluation showed 74 of 88 trainers (84%) fully aligned with survey data, while 14 showed partial mismatches, mainly in gender and method. This highlights the system's consistency and its ability to capture where linguistic signals vary more.

This modular workflow also offers clear practical advantages. Stakeholders can emphasise specific aspects of a trainer's profile, such as method or aid use, while leaving aside more variable dimensions like gender. Because each classifier operates independently, individual components can be updated (for example, adapting the aid type classifier to new terminology) without affecting the others. This ensures long-term adaptability and supports use in dynamic professional contexts.

In sum, the modular workflow enables precise, dimension-specific classification without sacrificing the ability to perform comprehensive profile evaluations. It balances robustness with interpretability and allows stakeholders to adjust the strictness of evaluation depending on their operational goals. The workflow is not only technically effective but also strategically aligned with the needs of organisations seeking transparent and scalable profiling tools.

Fourth contribution: extending predictions beyond survey data

The fourth contribution of this study is showing that structured predictions can be generated directly from trainer websites, even without survey data. This demonstrates the pipeline's technical viability in more scalable, data-limited contexts.

In conclusion, this study has introduced an interpretable and modular NLP framework for linking professional claims with their public communication. Through the reversed workflow design, gated semantic scoring, and modular classification, the approach emphasises interpretability, transparency, and practical value. While grounded in the context of dog training, the framework may be transferable to other domains where professional self-presentation guides decision-making.

Most importantly, the results provide not only a practical tool for supporting informed decision-making among dog owners, but also a flexible framework for future NLP research that addresses alignment, identity, and transparency across other domains.

7. APPENDIX

7.1. Reproducibility of Code and SQL Pipeline Access

To ensure transparency and full reproducibility, the complete database schema and backend Python scripts used in this project have been made available at the following public GitHub repository:

- **Repository:** `https://github.com/rezkyaps/NLP_SEMANTIC.git`

The repository contains the entire NLP profiling pipeline, including data cleaning, keyword extraction, semantic scoring, and multi-task classification. It also includes SQL schema files and setup scripts to recreate the backend system.

System Setup (summarised from README)

1. **Install dependencies:** Use Anaconda to set up the Python environment (`conda env create -f environment.yml`).
2. **Set up SQL Server 2019:** Run `generate_sql.sql` from SSMS to create all `m_` and `t_` tables, stored procedures, and logic views.
3. **Configure connection:** Edit `db.py` and relevant scripts to update your local `connection_string`.
4. **Run pipeline:** Execute `main.py` to launch full cleaning, scoring, and multi-task prediction.

Database Schema Structure

- **Master tables (m_):** Define fixed reference values for methods, roles, gender, aids, and filtering terms.
- **Transaction tables (t_):** Store cleaned web text, survey input, keyword scores, and normalized multi-selects.
- **Stored procedures:** Implement core logic for batch scoring, prediction alignment, and prompt-based filtering.

This SQL backend ensures consistent data handling across experiments, supports semantic scoring via prompt alignment and exclusivity, and enables traceable, task-specific profile prediction.

For further instructions and step-by-step details, users are advised to consult the `README.md` file in the GitHub repository.

7.2. Keyphrases Extraction Result

Table 7.1: Top Representative keywords for Role Code = 01 (Accredited Trainer)

Keyword	Semantic	Prompt	JSD	Raw	Gate	Final
canines clients	0.621	0.621	1.000	0.792	0.417	0.330
control exercises book	0.294	0.273	0.322	0.305	0.25	0.076
vegan treats	0.290	0.137	0.322	0.289	0.25	0.072
lives teach	0.284	0.245	1.000	0.602	0.25	0.151
force fear	0.283	0.068	0.322	0.279	0.25	0.070
kind ethical	0.282	0.081	1.000	0.585	0.25	0.146
collaborative approach	0.278	0.255	0.459	0.357	0.25	0.089
trainer customize	0.276	0.357	0.322	0.305	0.25	0.076
privacy certified trainer	0.273	0.399	1.000	0.613	0.25	0.153
instructors	0.273	0.378	0.523	0.396	0.25	0.099
trainers pact	0.273	0.323	1.000	0.605	0.25	0.151
scent detection	0.288	0.195	0.604	0.421	0.25	0.105
instructor animal behaviour	0.615	0.591	1.000	0.786	0.385	0.303
canine enrichment	0.598	0.569	0.538	0.568	0.362	0.206
heelwork course canine	0.555	0.636	0.322	0.459	0.435	0.199
train giving confidence	0.477	0.452	1.000	0.710	0.269	0.191
neutering really behaviour	0.555	0.384	1.000	0.738	0.250	0.185
obedience agility hoopers	0.489	0.407	1.000	0.711	0.251	0.178
prepare pup	0.380	0.432	1.000	0.664	0.259	0.172
socialising puppies	0.555	0.581	0.322	0.453	0.374	0.170
trainers passionate	0.385	0.406	1.000	0.664	0.251	0.166
needs behaviour issues	0.419	0.229	1.000	0.661	0.250	0.165
drive natural behaviours	0.412	0.189	1.000	0.654	0.250	0.164
canines reinforcing positive	0,397	0,518	1.000	0,681	0,316	0,215
effective raising confident	0.342	0.258	1.000	0.630	0.250	0.157
teach recall	0.322	0.283	1.000	0.623	0.250	0.156
professional development regularly	0.301	0.362	1.000	0.622	0.250	0.155
privacy certified trainer	0.273	0.399	1.000	0.613	0.250	0.153
fear free experience	0.290	0.135	1.000	0.594	0.250	0.148
proven scientific	0.272	0.159	0.322	0.283	0.250	0.071
certified separation anxiety	0.329	0.295	0.322	0.322	0.250	0.081

Table 7.2: Top Representative keywords for Role Code = 02 (Qualified Behaviourist)

Keyword	Semantic	Prompt	JSD	Raw	Gate	Final
canine specialist mental	0.451	0.637	1.000	0.717	0.436	0.312
assessments animal	0.465	0.444	1.000	0.704	0.265	0.186
transition puppies lifeskills	0.523	0.268	1.000	0.712	0.250	0.178
issues breed	0.427	0.413	1.000	0.683	0.252	0.172
behaviour changes	0.414	0.407	0.322	0.372	0.251	0.093
barking getting	0.382	0.388	1.000	0.661	0.250	0.165
association study animal	0.346	0.381	1.000	0.644	0.250	0.161
vet visits stressful	0.316	0.426	1.000	0.635	0.257	0.163
behaving	0.335	0.256	0.555	0.426	0.250	0.107
disorders	0.321	0.333	0.389	0.353	0.250	0.088
psychological	0.315	0.216	0.495	0.386	0.250	0.097
prey drive	0.319	0.147	0.483	0.376	0.250	0.094
poo masters qualified	0.350	0.088	1.000	0.616	0.250	0.154
dangerous act improved	0.323	0.169	1.000	0.612	0.250	0.153
providing essential mental	0.326	0.147	0.322	0.306	0.250	0.077
breed specific enrichment	0.363	0.334	0.322	0.342	0.250	0.085
transitioning	0.396	0.045	0.322	0.328	0.250	0.082
behavourial school	0.310	0.189	0.322	0.304	0.250	0.076
fastrack school behaviour	0.310	0.182	0.322	0.303	0.250	0.076
hyperactivity	0.308	0.303	0.556	0.419	0.250	0.105
threat	0.302	0.105	0.511	0.376	0.250	0.094
trainer behaviourist lengthy	0.437	1.000	0.181	0.262	0.25	0.692
behaviour expert	0.306	0.771	0.141	0.141	0.25	0.565
behavioural group	0.169	0.322	0.085	0.085	0.25	0.340
behaviour assessment	0.190	0.470	0.102	0.102	0.25	0.407
education behavioural	0.218	1.000	0.161	0.161	0.25	0.643
psychology	0.206	0.655	0.114	0.114	0.25	0.454
mental challenges	0.167	0.322	0.075	0.075	0.25	0.300

Table 7.3: Top Representative keywords for Role = 03 (Professional, no qualifications)

Keyword	Semantic	Prompt	JSD	Raw	Gate	Final
canine companions	0.570	0.533	0.586	0.574	0.329	0.188
furry companion	0.414	0.377	0.660	0.130	0.250	0.521
companion animals	0.528	0.429	0.751	0.160	0.258	0.618
issues particular breed	0.478	0.423	1.000	0.181	0.256	0.708
animal welfare rescues	0.498	0.430	1.000	0.185	0.258	0.717
patient knowledgeable pups	0.664	0.698	1.000	0.420	0.512	0.819
canine coaching based	0.640	0.746	0.322	0.294	0.579	0.508
hounds hamble	0.627	0.347	1.000	0.192	0.250	0.767
welcome hamble hounds	0.607	0.354	1.000	0.190	0.250	0.759
terrier friendly encouraging	0.579	0.588	1.000	0.293	0.250	0.769
skills ideal puppies	0.567	0.571	1.000	0.278	0.250	0.762
breed rescue organisations	0.549	0.499	1.000	0.224	0.300	0.747
puppies in families	0.538	0.418	1.000	0.186	0.254	0.734
temperament soundness adopting	0.456	0.384	1.000	0.173	0.250	0.694
gentle kind learning	0.372	0.323	1.000	0.162	0.250	0.649
patient understanding friendly	0.305	0.284	1.000	0.154	0.250	0.616
fun owners environment	0.310	0.236	1.000	0.153	0.250	0.613
trainer brilliant	0.353	0.459	1.000	0.179	0.273	0.655
children regardless breed	0.347	0.200	1.000	0.157	0.250	0.626
pooch know feeling	0.449	0.483	1.000	0.701	0.25	0.289
terrier friendly encouraging	0.579	0.588	1.000	0.769	0.25	0.381
hyperactivity separation	0.316	0.117	1.000	0.604	0.250	0.151

Table 7.4: Top Representative keywords for *method_code = 01* (Positive Reinforcement)

Keyword	Semantic	JSD	Prompt	Raw	Gate	Final
positive reinforcement reward	0.4571	1.0000	0.5921	0.7149	0.3859	0.2759
heavily positive reinforcement	0.5446	1.0000	0.4731	0.7424	0.2819	0.2093
positive reinforcement kind	0.5248	1.0000	0.4835	0.7345	0.2890	0.2123
way teach behave	0.4112	1.0000	0.4191	0.6769	0.2543	0.1721
treats	0.4534	1.0000	0.3944	0.5629	0.2500	0.1407
relaxed happy handling	0.4267	1.0000	0.3994	0.6819	0.2500	0.1705
affection	0.3437	1.0000	0.3535	0.6090	0.1523	0.9312
encouragement	0.3779	1.0000	0.4219	0.5445	0.2553	0.7382
nurturing	0.4286	1.0000	0.3803	0.5432	0.2500	0.6940
kind limited kindness	0.3674	1.0000	0.2904	0.6444	0.2500	0.1611
food treats	0.4018	1.0000	0.4679	0.4263	0.2785	0.4416
gift growl	0.3809	1.0000	0.2536	0.6468	0.2500	0.1617
approach journeys cats	0.4617	1.0000	0.2032	0.6781	0.2500	0.1695
bags treats	0.3675	1.0000	0.2515	0.6405	0.2500	0.1601
teach associate positive	0.3684	1.0000	0.2881	0.6446	0.2500	0.1611
care reactive rescue	0.3845	1.0000	0.2254	0.6456	0.2500	0.1614
behaviour relaxed happy	0.4532	1.0000	0.4859	0.7025	0.2906	0.2042
behaved canine	0.6157	1.0000	0.3493	0.7577	0.2565	0.1894
veterinary behaviour support	0.6556	1.0000	0.3710	0.4488	0.2500	0.1122
new positive reinforcement	0.5080	1.0000	0.1900	0.6585	0.4111	0.2592
award socialisation care	0.3531	1.0000	0.2896	0.6378	0.2500	0.1595
behaving way	0.3332	1.0000	0.2702	0.6270	0.2500	0.1567
behaviour good	0.4646	0.4421	0.4561	0.4536	0.2714	0.1231
cats cooperative care	0.4068	1.0000	0.2195	0.6550	0.2500	0.1638
feeding	0.3758	0.6507	0.3679	0.4988	0.2500	0.1247
food treats	0.4018	0.4416	0.4679	0.4263	0.2785	0.1188
fostering	0.4000	0.7457	0.3489	0.5505	0.2500	0.1376
happy mannered makes	0.3544	1.0000	0.3155	0.6410	0.2500	0.1603
navigate obedience behavioural	0.5588	1.0000	0.4180	0.7433	0.2539	0.1887
obedience behavioural issues	0.5305	1.0000	0.3766	0.7264	0.2500	0.1816
positive reinforcement	0.5132	0.6545	0.7207	0.5975	0.5430	0.3245
positive techniques	0.3903	0.5379	0.5010	0.4678	0.3018	0.1412
reinforcement does work	0.4836	1.0000	0.5507	0.7227	0.3444	0.2489
teach behave	0.4646	1.0000	0.4346	0.7025	0.2604	0.1829
using positive reinforcement	0.5385	0.4741	0.7058	0.5262	0.5228	0.2751

Table 7.5: Top Representative keywords for *method_code = 02* (Remove Rewards)

Keyword	Semantic	Prompt	JSD	Raw	Gate	Final
aggression barking	0.5642	0.5005	1.0000	0.7539	0.3014	0.2272
aggression noisy scared	0.3734	0.2987	0.4421	0.3968	0.2500	0.0992
aggressive fearful defensive	0.3816	0.2720	1.0000	0.6489	0.2500	0.1622
assertive practiced	0.3458	0.1032	1.0000	0.6159	0.2500	0.1540
barks attention	0.5722	0.3943	1.0000	0.7469	0.2500	0.1867
barks lunges	0.4564	0.2762	1.0000	0.6830	0.2500	0.1708
breeds behaviour history	0.5807	0.3709	0.4421	0.4974	0.2500	0.1243
breeds problems	0.5065	0.2767	1.0000	0.7056	0.2500	0.1764
busters behaviour	0.3492	0.2939	1.0000	0.6365	0.2500	0.1591
care tagged biters	0.3544	0.2398	1.0000	0.6334	0.2500	0.1584
categories reactive breeds	0.4612	0.3048	1.0000	0.6880	0.2500	0.1720
chew collar	0.4643	0.3793	1.0000	0.6968	0.2500	0.1742
chew collar grabs	0.3873	0.3505	1.0000	0.6593	0.2500	0.1648
comition obedience lifeskills	0.3681	0.2492	0.4421	0.3895	0.2500	0.0974
dealt barking	0.4594	0.3224	1.0000	0.6890	0.2500	0.1722
ignoring barks	0.5701	0.4643	1.0000	0.7530	0.2763	0.2081
lunging behaviour	0.3978	0.1925	0.4421	0.3972	0.2500	0.0993
lungy barky	0.3617	0.2131	0.4421	0.3830	0.2500	0.0957
needs cause bite	0.3775	0.2480	1.0000	0.6447	0.2500	0.1612
obedience flits reactivity	0.3954	0.2945	1.0000	0.6574	0.2500	0.1643
obedience lifeskills	0.3481	0.2328	0.4421	0.3788	0.2500	0.0947
obedience puppies reactivity	0.6393	0.4926	0.4421	0.5359	0.2954	0.1583
persistent barking letting	0.5624	0.4711	1.0000	0.7502	0.2806	0.2105
problem aggression behaviour	0.3736	0.3649	1.0000	0.6546	0.2500	0.1637
prong collars	0.3996	0.3140	0.4988	0.4356	0.2500	0.1089
puppies ask stop	0.4814	0.3842	1.0000	0.7051	0.2500	0.1763
puppies reactivity socialising	0.5833	0.3987	0.4421	0.5013	0.2500	0.1253
puppies tagged grabbing	0.4242	0.3379	1.0000	0.6747	0.2500	0.1687
reactive breeds behaviour	0.6051	0.4638	1.0000	0.7687	0.2760	0.2122
reactivity tagged barking	0.4358	0.3636	1.0000	0.6825	0.2500	0.1706
rearing experiences wrong	0.3772	0.2187	1.0000	0.6416	0.2500	0.1604
strangers change behaviour	0.3421	0.1935	1.0000	0.6233	0.2500	0.1558
tamed	0.3789	0.2092	1.0000	0.6414	0.2500	0.1604
teaching know cues	0.3489	0.1584	1.0000	0.6228	0.2500	0.1557

Table 7.6: Top representative keywords for *method_code* = *03* (Remove Unpleasant)

Keyword	Semantic	JSD	Prompt	Raw	Gate	Final
affection spoiling	0.3721	1.0000	0.3418	0.6516	0.2500	0.1629
aggression anxiety barking	0.4987	1.0000	0.4786	0.7223	0.2856	0.2063
anxious reactive	0.3101	1.0000	0.2709	0.6166	0.2500	0.1542
areas troublesome companions	0.3774	1.0000	0.1819	0.6380	0.2500	0.1595
barking destructive behaviour	0.5006	0.6270	0.5133	0.5588	0.3115	0.1741
behaviour prevention amazing	0.4550	1.0000	0.3730	0.6921	0.2500	0.1730
behaviour problems barking	0.5068	1.0000	0.5424	0.7323	0.3367	0.2466
behaviour toileting	0.3444	0.6286	0.3208	0.4699	0.2500	0.1175
behavioural issues watch	0.4252	1.0000	0.3173	0.6731	0.2500	0.1683
care room cooperative	0.3130	1.0000	0.0215	0.5930	0.2500	0.1482
care zoo	0.4726	1.0000	0.2003	0.6827	0.2500	0.1707
caring sorts animals	0.4708	1.0000	0.2597	0.6878	0.2500	0.1720
chew toy	0.3058	0.7962	0.3596	0.5319	0.2500	0.1330
collar lead	0.3199	1.0000	0.2794	0.6219	0.2500	0.1555
consent select paw	0.3416	1.0000	0.3036	0.6341	0.2500	0.1585
cooperative care	0.3416	0.4732	0.1187	0.3786	0.2500	0.0946
dealt barking jumping	0.4740	1.0000	0.3763	0.7009	0.2500	0.1752
discomfort canine	0.5259	1.0000	0.4701	0.7337	0.2800	0.2054
discomfort canine companion	0.5768	1.0000	0.4764	0.7572	0.2841	0.2151
ensure grows mannered	0.3234	1.0000	0.2331	0.6188	0.2500	0.1547
good observation behaviour	0.3669	1.0000	0.3783	0.6530	0.2500	0.1632
handler understand	0.3280	1.0000	0.2792	0.6255	0.2500	0.1564

Table 7.7: Top representative keywords for *method_code* = *03* (Remove Unpleasant) (continue)

Keyword	Semantic	JSD	Prompt	Raw	Gate	Final
handling caring	0.3952	1.0000	0.2359	0.6514	0.2500	0.1629
muzzle	0.3060	0.5182	0.2338	0.3943	0.2500	0.0986
nervous rescues	0.4258	0.6282	0.3074	0.5050	0.2500	0.1263
nervous rescues energetic	0.4120	1.0000	0.2919	0.6646	0.2500	0.1662
nervous unsocialised years	0.2943	1.0000	0.1832	0.6007	0.2500	0.1502
neutered	0.3315	0.7582	0.2635	0.5167	0.2500	0.1292
nipping canine	0.3950	1.0000	0.3995	0.6677	0.2500	0.1669
obedience residential	0.4382	1.0000	0.3499	0.6822	0.2500	0.1705
overexcited jumpy pup	0.4503	1.0000	0.3510	0.6878	0.2500	0.1719
people house aggression	0.3032	1.0000	0.2125	0.6077	0.2500	0.1519
play biting chewing	0.3035	1.0000	0.2928	0.6159	0.2500	0.1540
reactive people	0.3013	1.0000	0.2878	0.6144	0.2500	0.1536
reactive year old	0.4042	1.0000	0.2383	0.6557	0.2500	0.1639
select paw	0.3852	1.0000	0.2850	0.6519	0.2500	0.1630
signs aggression	0.3088	1.0000	0.1964	0.6086	0.2500	0.1521
stress local behaviourist	0.4344	1.0000	0.3577	0.6813	0.2500	0.1703
terribly yappy reactive	0.3457	1.0000	0.2405	0.6296	0.2500	0.1574
treat chew	0.3772	1.0000	0.3800	0.6578	0.2500	0.1644
troublesome companions	0.3962	1.0000	0.2696	0.6553	0.2500	0.1638
yappy reactive	0.3291	1.0000	0.2088	0.6190	0.2500	0.1547

Table 7.8: Top Representative keywords for *method_code = 04* (Introduce Unpleasant)

Keyword	Semantic	JSD	Prompt	Raw	Gate	Final
address bark	0.4156	1.0000	0.2045	0.6575	0.2500	0.1644
address barking	0.4627	1.0000	0.2569	0.6839	0.2500	0.1710
aggression animals people	0.3729	1.0000	0.3891	0.6567	0.2500	0.1642
aggression separation	0.3293	0.6286	0.3751	0.4686	0.2500	0.1171
aggression separation anxiety	0.3712	0.6286	0.2979	0.4797	0.2500	0.1199
aggressive owners	0.3760	1.0000	0.3490	0.6541	0.2500	0.1635
areas local therapist	0.3483	1.0000	0.0994	0.6167	0.2500	0.1542
aspects behaviour aggression	0.4205	1.0000	0.4157	0.6808	0.2532	0.1724
assessment session	0.3259	0.9278	0.2259	0.5867	0.2500	0.1467
bark lunge alarmingly	0.4272	1.0000	0.3074	0.6730	0.2500	0.1682
barking afternoon	0.4285	1.0000	0.2815	0.6710	0.2500	0.1677
barking afternoon attacked	0.3823	1.0000	0.2038	0.6424	0.2500	0.1606
behaviour aggression	0.4091	0.2592	0.4673	0.3475	0.2782	0.0966
behaviour aggression eased	0.4799	1.0000	0.5127	0.7172	0.3110	0.2231
behaviour causing stress	0.3414	1.0000	0.2936	0.6330	0.2500	0.1582
behaviour modification near	0.5720	1.0000	0.4983	0.7572	0.2997	0.2270
behaviour problems routine	0.5185	1.0000	0.4230	0.7256	0.2556	0.1855
behaviour problems train	0.5774	1.0000	0.4443	0.7543	0.2651	0.1999
behaviour separation anxiety	0.4248	1.0000	0.2730	0.6684	0.2500	0.1671
behaviours aggression	0.4035	0.9579	0.4503	0.6577	0.2682	0.1764
biting behavioural issues	0.3931	1.0000	0.3529	0.6622	0.2500	0.1655
boisterous behaviour	0.3680	0.9780	0.3049	0.6362	0.2500	0.1590
canine relationship reduce	0.5235	1.0000	0.2866	0.7142	0.2500	0.1786
contented volunteers therapy	0.4146	1.0000	0.2832	0.6649	0.2500	0.1662
excessive barking destructive	0.5536	0.6282	0.4179	0.5736	0.2539	0.1456
general temperament improved	0.4979	1.0000	0.3110	0.7052	0.2500	0.1763
grooming bad behaviour	0.5269	1.0000	0.5370	0.7408	0.3318	0.2458
hardy dealt barking	0.3798	1.0000	0.2191	0.6428	0.2500	0.1607
house aggression	0.3377	0.6286	0.3430	0.4692	0.2500	0.1173
house aggression separation	0.3979	0.6286	0.2902	0.4910	0.2500	0.1227

Table 7.9: Top representative keywords for *method_code = 04* (Introduce Unpleasant)(continue)

Keyword	Semantic	JSD	Prompt	Raw	Gate	Final
leader behaviour modification	0.4546	0.6286	0.3774	0.5252	0.2500	0.1313
local therapist	0.3536	1.0000	0.0635	0.6154	0.2500	0.1539
manners control	0.4016	1.0000	0.3597	0.6667	0.2500	0.1667
manners overcoming	0.3536	1.0000	0.2510	0.6342	0.2500	0.1586
manners overcoming fears	0.3468	1.0000	0.2867	0.6347	0.2500	0.1587
manners recall	0.3407	1.0000	0.2560	0.6289	0.2500	0.1572
nipping canine misbehaviour	0.4368	1.0000	0.3959	0.6862	0.2500	0.1715
obedience work	0.4112	0.7232	0.3691	0.5474	0.2500	0.1368
pack leader behaviour	0.3431	0.6286	0.2289	0.4602	0.2500	0.1150
seeing behaviour improved	0.4627	1.0000	0.3872	0.6969	0.2500	0.1742
separation anxiety understanding	0.3620	1.0000	0.1652	0.6294	0.2500	0.1574
stopped excessive barking	0.5739	1.0000	0.4159	0.7498	0.2532	0.1899
successful behavioural modification	0.5473	0.9804	0.4615	0.7336	0.2746	0.2014
temperament improved	0.5428	1.0000	0.3304	0.7273	0.2500	0.1818
temperament improved significantly	0.5234	1.0000	0.2594	0.7115	0.2500	0.1779
temperaments nervous	0.3488	1.0000	0.2198	0.6289	0.2500	0.1572
therapist trainer	0.3583	1.0000	0.1023	0.6215	0.2500	0.1554
therapists register	0.3277	1.0000	0.0839	0.6059	0.2500	0.1515
therapists trainers	0.4654	1.0000	0.1724	0.6767	0.2500	0.1692
therapy temperament	0.5292	1.0000	0.3243	0.7206	0.2500	0.1801
therapy temperament assessor	0.4340	1.0000	0.2335	0.6686	0.2500	0.1672
uk therapists	0.3860	1.0000	0.0672	0.6304	0.2500	0.1576
wolves captivity learning	0.5533	1.0000	0.3470	0.7337	0.2500	0.1834
working wolves captivity	0.5233	1.0000	0.2199	0.7075	0.2500	0.1769
years working wolves	0.3368	1.0000	0.1187	0.6134	0.2500	0.1534

Table 7.10: Top representative keywords for Gender Code 01 (Female)

Keyword	Semantic	Prompt	JSD	Raw	Gate	Final
communicate important barking	0.5526	0.5265	1.0000	0.7513	0.3226	0.2424
management care puppies	0.5495	0.5149	1.0000	0.7488	0.3129	0.2343
canine care	0.6015	0.5624	0.4421	0.5258	0.3556	0.1870
human canine	0.4808	0.4695	1.0000	0.7133	0.2795	0.1994
care puppies	0.4446	0.4427	1.0000	0.6944	0.2642	0.1835
positive reinforcement teach	0.4879	0.3278	1.0000	0.7024	0.2500	0.1756
reinforcement teach	0.4460	0.2903	1.0000	0.6797	0.2500	0.1699
reinforcement teach appropriate	0.4433	0.2866	1.0000	0.6782	0.2500	0.1695
positive reinforcement does	0.4485	0.2483	1.0000	0.6766	0.2500	0.1692
behaviour management care	0.4328	0.2681	1.0000	0.6716	0.2500	0.1679
trainer school dedicated	0.4006	0.3350	1.0000	0.6638	0.2500	0.1659
dealing leash ambush	0.4133	0.2787	1.0000	0.6639	0.2500	0.1660
trainer positive expanded	0.4019	0.3211	1.0000	0.6630	0.2500	0.1657
handler attentive	0.4008	0.3577	1.0000	0.6661	0.2500	0.1665
attentive handler	0.3801	0.3454	1.0000	0.6556	0.2500	0.1639
owners friendly	0.3862	0.3163	1.0000	0.6554	0.2500	0.1639
institute modern trainers	0.3856	0.3122	1.0000	0.6547	0.2500	0.1637
dealing leash	0.3965	0.3099	1.0000	0.6594	0.2500	0.1649
collar harness	0.3933	0.3354	1.0000	0.6605	0.2500	0.1651
words hear vet	0.3625	0.3443	1.0000	0.6476	0.2500	0.1619
hear vet	0.3668	0.3181	1.0000	0.6469	0.2500	0.1617
just treats make	0.3632	0.2750	1.0000	0.6409	0.2500	0.1602
posted obedience	0.3755	0.2359	1.0000	0.6425	0.2500	0.1606
confidence excellent teach	0.3325	0.3006	1.0000	0.6297	0.2500	0.1574
change positive motivational	0.3384	0.2116	1.0000	0.6235	0.2500	0.1559
behave way	0.3210	0.2900	1.0000	0.6235	0.2500	0.1559
positive motivational techniques	0.3294	0.2381	1.0000	0.6221	0.2500	0.1555
cross submit hound	0.3250	0.1741	1.0000	0.6137	0.2500	0.1534
growl eats	0.3152	0.1958	1.0000	0.6114	0.2500	0.1529

Table 7.11: Top representative Keywords for Gender Code 02 (Non-Female)

Keyword	Semantic	Prompt	JSD	Raw	Gate	Final
accredited experienced canine	0.511	0.500	1.000	0.730	0.301	0.220
great trainer polite	0.529	0.476	1.000	0.736	0.284	0.209
guarantee trainer bark	0.418	0.515	1.000	0.689	0.313	0.216
trainer service	0.568	0.381	1.000	0.743	0.250	0.186
trainer service support	0.522	0.335	1.000	0.718	0.250	0.180
feet train educate	0.402	0.294	1.000	0.660	0.250	0.165
focusing companion animals	0.410	0.381	1.000	0.672	0.250	0.168
train educate matter	0.422	0.353	1.000	0.675	0.250	0.169
form local trainer	0.476	0.345	1.000	0.699	0.250	0.175
excellent trainer great	0.441	0.403	1.000	0.689	0.250	0.172
pulling lead barking	0.294	0.447	1.000	0.627	0.266	0.167
trainer click	0.408	0.288	1.000	0.662	0.250	0.166
care animal professionals	0.394	0.365	1.000	0.664	0.250	0.166
house trained best	0.405	0.494	1.000	0.681	0.296	0.202
individual trainer	0.494	0.437	1.000	0.716	0.261	0.187
experienced canine	0.449	0.562	0.442	0.457	0.355	0.162
basics collar	0.324	0.426	1.000	0.638	0.257	0.164
barking hyperactivity	0.288	0.412	1.000	0.621	0.252	0.157
jumping barking	0.287	0.410	1.000	0.620	0.252	0.156
confident support guide	0.305	0.367	1.000	0.624	0.250	0.156
confident teach	0.277	0.380	1.000	0.613	0.250	0.153
dealt barking pulling	0.283	0.369	1.000	0.614	0.250	0.154
barking separation	0.283	0.343	0.730	0.490	0.250	0.123
barking destructive	0.281	0.339	0.595	0.428	0.250	0.107
clearly expert	0.313	0.213	1.000	0.612	0.250	0.153
hyperactivity prominent fed	0.322	0.162	1.000	0.611	0.250	0.153
appointment come trainer	0.469	0.391	1.000	0.700	0.250	0.175
issued kennel	0.302	0.353	1.000	0.621	0.250	0.155
barking separation	0.283	0.343	0.730	0.490	0.250	0.123
aggression bark	0.293	0.452	0.442	0.376	0.269	0.101
way behaviour modification	0.341	0.314	0.442	0.384	0.250	0.096

Table 7.12: Top representative Keywords for Aid Type

Keyword	Semantic	Prompt	JSD	Raw	Gate	Final
clicker	0.0979	0.3407	0.6415	0.3668	0.25	0.0917
remote treat dispanser	0.1882	0.2313	1.0000	0.5578	0.25	0.1395
snuffle mat	0.1121	0.1193	0.5631	0.3158	0.25	0.0789
target stick	0.0670	0.2170	1.0000	0.5019	0.25	0.1255
toys	0.2761	0.4324	0.7946	0.5250	0.2594	0.1362
treats	0.4534	0.4205	0.7646	0.5901	0.2547	0.1503
walking harness	0.2354	0.4092	1.0000	0.5968	0.2514	0.1501
anti pulling harness	0.1991	0.3231	1.0000	0.5719	0.25	0.1430
crate	0.2540	0.3352	0.7073	0.4661	0.25	0.1165
longline	0.0958	0.2376	0.5631	0.3202	0.25	0.0801
choke collar	0.2442	0.4327	0.5631	0.4065	0.2595	0.1055
citronella collar	0.2835	0.3020	0.5631	0.4112	0.25	0.1028
noise maker	0.1518	0.2177	1.0000	0.5401	0.25	0.1350
prong collar	0.3312	0.2625	0.6446	0.4653	0.25	0.1163
shock collar	0.3960	0.4986	0.5631	0.4814	0.3000	0.1444
slip lead	0.0565	0.0776	0.6322	0.3177	0.25	0.0794
water squirter	0.0825	0.1881	0.5631	0.3093	0.25	0.0773

7.3. Results of One-to-One Misalignment

Table 7.13: Prediction vs Ground Truth (Method-Gender-Role-Aid)

ID	Pred (M-G-R-A)	True (M-G-R-A)	Probability (M;G;R;A)	Match
R_1	0-0-0-1	1-0-0-0	0.84;0.27;0.32;0.47	0
R_2	1-0-0-0	1-0-0-0	1.00;0.03;0.10;0.23	1
R_3	1-1-0-0	1-0-0-0	0.90;0.43;0.24;0.17	0
R_4	1-0-0-0	1-0-0-0	0.97;0.20;0.19;0.17	1
R_5	1-0-1-0	1-0-1-0	1.00;0.00;0.60;0.19	1
R_6	1-0-1-0	1-0-1-0	0.97;0.02;0.91;0.06	0
R_7	1-0-1-0	1-0-1-0	0.91;0.04;0.88;0.18	1
R_8	1-0-0-0	1-0-0-0	0.87;0.06;0.18;0.38	1
R_9	1-0-0-0	1-0-0-0	1.00;0.20;0.17;0.07	1
R_10	1-0-1-0	1-0-1-0	1.00;0.10;0.71;0.23	1
R_11	1-0-0-0	1-0-0-0	1.00;0.19;0.24;0.35	1
R_12	1-0-1-0	1-0-1-0	1.00;0.00;0.81;0.18	1
R_13	1-0-0-0	1-0-0-0	0.94;0.08;0.23;0.21	1
R_14	1-0-1-0	1-0-1-0	0.97;0.25;0.78;0.36	1
R_15	1-0-0-0	1-0-0-0	1.00;0.05;0.32;0.18	1
R_16	1-0-0-0	1-0-0-0	1.00;0.29;0.16;0.23	1
R_17	1-1-0-0	1-0-1-0	0.94;0.50;0.39;0.41	0
R_18	1-0-1-0	1-0-1-0	1.00;0.03;0.67;0.04	1
R_19	1-0-1-0	1-0-1-0	0.97;0.24;0.64;0.14	1
R_20	1-0-0-0	1-0-0-0	1.00;0.03;0.27;0.19	1
R_21	1-0-0-0	1-0-0-0	1.00;0.12;0.29;0.10	1
R_22	0-0-0-1	1-0-0-1	0.68;0.36;0.06;0.69	0

Table 7.14: Prediction vs Ground Truth (Method-Gender-Role-Aid) (continue)

ID	Pred (M-G-R-A)	True (M-G-R-A)	Probability (M;G;R;A)	Match
R_23	1-0-0-0	1-0-0-0	0.88;0.07;0.96;0.05	1
R_24	1-0-1-0	1-0-1-0	1.00;0.32;0.49;0.73	0
R_25	1-0-0-0	1-0-0-0	0.94;0.10;0.39;0.20	1
R_26	1-0-1-0	1-0-1-0	1.00;0.06;1.00;0.04	1
R_27	1-0-1-0	1-0-1-0	0.97;0.04;0.98;0.07	1
R_28	1-0-1-0	1-0-1-0	1.00;0.13;0.50;0.05	1
R_29	1-0-1-1	1-0-1-1	1.00;0.06;0.96;0.62	1
R_30	1-1-1-0	1-0-0-0	1.00;0.48;0.22;0.09	0
R_31	1-0-0-0	1-0-0-0	0.94;0.20;0.31;0.26	1
R_32	1-0-0-0	1-0-0-0	1.00;0.21;0.35;0.23	1
R_33	1-0-0-0	1-0-0-0	0.94;0.10;0.44;0.19	1
R_34	1-1-0-0	1-1-0-0	0.97;0.55;0.13;0.08	1
R_35	1-1-0-0	1-1-0-0	0.97;0.52;0.35;0.11	1
R_36	1-0-1-0	1-0-1-0	0.97;0.03;0.74;0.03	1
R_37	1-0-1-0	1-0-1-0	1.00;0.28;0.60;0.12	1
R_38	1-0-0-0	1-0-0-0	0.97;0.09;0.21;0.05	1
R_39	1-0-1-0	1-0-1-0	1.00;0.22;0.86;0.34	1
R_40	1-0-0-0	1-0-0-0	1.00;0.25;0.43;0.29	1
R_41	1-0-1-0	1-0-1-0	0.88;0.17;0.25;0.13	0
R_42	1-1-1-1	1-1-1-1	0.90;0.54;0.61;0.59	1
R_43	1-1-0-0	1-1-0-0	1.00;0.57;0.41;0.14	1
R_44	1-0-1-0	1-0-1-0	1.00;0.15;0.79;0.21	1

Table 7.15: Prediction vs Ground Truth (Method-Gender-Role-Aid) (continue)

ID	Pred (M-G-R-A)	True (M-G-R-A)	Probability (M;G;R;A)	Match
R_45	1-0-0-0	1-0-0-0	0.91;0.19;0.33;0.31	1
R_46	1-0-1-0	1-0-1-0	1.00;0.06;0.94;0.07	1
R_47	1-0-1-0	1-0-1-0	0.90;0.13;0.62;0.28	1
R_48	1-0-0-0	1-0-0-0	0.94;0.14;0.31;0.18	1
R_49	1-0-1-0	1-0-1-0	1.00;0.06;0.64;0.12	1
R_50	1-0-1-0	1-0-1-0	0.90;0.19;0.79;0.17	1
R_51	1-1-0-0	1-1-0-0	1.00;0.69;0.34;0.17	1
R_52	1-0-0-0	1-0-0-0	0.87;0.16;0.19;0.09	1
R_53	1-0-0-0	1-0-0-0	1.00;0.16;0.14;0.03	1
R_54	1-1-0-0	1-1-0-0	1.00;0.77;0.30;0.20	1
R_55	1-0-0-0	1-0-0-0	0.91;0.11;0.21;0.18	1
R_56	1-0-1-0	1-0-1-0	0.97;0.11;0.82;0.28	1
R_57	1-0-0-0	1-0-0-0	1.00;0.08;0.15;0.12	1
R_58	1-0-0-0	1-0-0-0	1.00;0.02;0.27;0.03	1
R_59	1-1-1-1	1-1-1-1	0.97;0.56;0.38;0.54	1
R_60	1-1-0-0	1-1-0-0	1.00;0.70;0.15;0.06	1
R_61	1-1-1-1	1-1-1-1	1.00;0.75;0.78;0.71	1
R_62	1-0-0-0	1-0-0-0	0.97;0.24;0.12;0.00	1
R_63	1-1-0-0	1-1-0-0	0.94;0.10;0.12;0.06	1
R_64	1-1-0-0	1-1-0-0	1.00;0.18;0.11;0.00	1
R_65	1-0-0-0	1-0-0-0	0.97;0.17;0.22;0.08	1
R_66	1-0-1-0	1-0-1-0	0.94;0.16;0.74;0.02	1

Table 7.16: Prediction vs Ground Truth (Method-Gender-Role-Aid) (continue)

ID	Pred (M-G-R-A)	True (M-G-R-A)	Probability (M;G;R;A)	Match
R$_{67}$	1-1-0-0	1-1-0-0	1.00;0.68;0.29;0.20	1
R$_{68}$	1-1-1-1	1-1-1-1	1.00;0.78;0.72;0.71	1
R$_{69}$	1-0-0-0	1-0-0-0	0.91;0.09;0.17;0.09	1
R$_{70}$	1-1-1-1	1-1-1-1	0.87;0.79;0.56;0.70	1
R$_{71}$	1-1-0-0	1-0-0-0	0.87;0.11;0.22;0.22	0
R$_{72}$	0-1-0-0	1-1-0-0	0.65;0.66;0.73;0.72	0
R$_{73}$	0-1-0-0	1-0-0-0	0.75;0.38;0.24;0.10	0
R$_{74}$	1-0-0-0	1-0-0-0	0.97;0.15;0.18;0.15	1
R$_{75}$	1-0-0-0	1-0-0-0	0.97;0.03;0.15;0.00	1
R$_{76}$	0-1-0-0	1-0-0-0	0.81;0.10;0.12;0.12	0
R$_{77}$	1-1-1-0	1-1-1-0	0.87;0.75;0.73;0.00	1
R$_{78}$	1-0-1-0	1-0-1-0	1.00;0.08;0.64;0.12	1
R$_{79}$	1-0-1-0	1-0-0-0	0.97;0.18;0.44;0.07	0
R$_{80}$	1-0-0-0	1-0-0-0	0.97;0.08;0.12;0.03	1
R$_{81}$	1-0-1-0	1-0-1-0	0.90;0.19;0.79;0.17	1
R$_{82}$	1-0-1-0	1-0-1-0	0.97;0.16;0.57;0.07	1
R$_{83}$	1-0-0-0	1-0-0-0	1.00;0.10;0.24;0.12	0
R$_{84}$	0-1-0-0	0-1-0-0	0.27;0.81;0.14;0.09	1
R$_{85}$	1-0-1-0	1-0-1-0	1.00;0.12;0.61;0.06	1
R$_{86}$	1-0-0-0	1-0-0-0	0.94;0.08;0.17;0.03	1
R$_{87}$	1-0-0-0	1-0-0-0	1.00;0.18;0.33;0.00	1
R$_{88}$	1-1-1-1	1-1-1-1	1.00;0.75;0.70;0.74	1

BIBLIOGRAPHY

Breiman, L. (2001). “Random Forests”. In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

Choi, M., Gwak, C., and Kim, S. (Dec. 2023). “SimCKP: Simple Contrastive Learning of Keyphrase Representations”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, pp. 3003–3015. DOI: 10.18653/v1/2023.findings-emnlp.199.

Demidova, L. and Ivkina, M. (2019). “Defining the Ranges Boundaries of the Optimal Parameters Values for the Random Forest Classifier”. In: *2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA)*. IEEE, pp. 518–522. DOI: 10.1109/SUMMA48161.2019.8947569.

Fernandez, E. J. (2024). “The Least Inhibitive, Functionally Effective (LIFE) Model: A New Framework for Ethical Animal Training Practices”. In: *Journal of Veterinary Behavior* 71, pp. 63–68. DOI: 10.1016/j.jveb.2023.12.001.

Gabrielsen, A. M. (2017). “Training Technologies. Science, Gender and Dogs in the Age of Positive Dog Training”. In: *Nordic Journal of Science and Technology Studies* 5.1, pp. 5–16. DOI: 10.5324/njsts.v5i1.2251. URL: https://doi.org/10.5324/njsts.v5i1.2251.

Ghimire, A. and Amsaad, F. (2024). “A Parallel Approach to Enhance the Performance of Supervised Machine Learning Realized in a Multicore Environment”. In: *Machine Learning and Knowledge Extraction* 6.3, pp. 1840–1856. DOI: 10 . 3390 / make6030090. URL: https://doi.org/10.3390/make6030090.

Goutte, C. and Gaussier, E. (2005). “A probabilistic interpretation of precision, recall and F-score, with implication for evaluation”. In: *Advances in Information Retrieval: Proceedings of ECIR 2005*. Springer, pp. 345–359. DOI: 10.1007/978-3-540-31865-1_25.

Greenebaum, J. B. (2010). “Training Dogs and Training Humans: Symbolic Interaction and Dog Training”. In: *Anthrozoös* 23.2, pp. 129–141. DOI: 10.2752/175303710X12682332909936. URL: https://doi.org/10.2752/175303710X12682332909936.

Han, S., Williamson, B. D., and Fong, Y. (2021). “Improving random forest predictions in small datasets from two-phase sampling designs”. In: *BMC Medical Informatics and Decision Making* 21.1, pp. 1–13. DOI: 10.1186/s12911-021-01688-3.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer. DOI: 10.1007/978-0-387-84858-7. URL: https://link.springer.com/book/10.1007/978-0-387-84858-7.

Hiby, E. F., Rooney, N. J., and Bradshaw, J. W. S. (2004). “Dog training methods: their use, effectiveness and interaction with behaviour and welfare”. In: *Animal Welfare* 13.1, pp. 63–69. DOI: 10.1017/S0962728600026683.

Johnson, A. C. and Wynne, C. D. L. (2023). “Training Dogs with Science or with Nature? An Exploration of Trainers’ Word Use, Gender, and Certification Across Dog□Training Methods”. In: *Anthrozoös* 36.1, pp. 35–51. DOI: 10.1080/08927936.2022.2062869.

Karaman, I. H., Koksal, G., Eriskin, L., and Salihoglu, S. (2024). “A Similarity-Based Oversampling Method for Multi-label Imbalanced Text Data”. In: *arXiv preprint arXiv:2411.01013*. URL: https://arxiv.org/abs/2411.01013.

Khvatskii, G., Moniz, N., Doan, K. D., and Chawla, N. V. (2025). “Class-aware contrastive optimization for imbalanced text classification”. In: *Discover Data* 3.27. DOI: 10.1007/s44248-025-00064-0.

Kim, E., Shim, K., Chang, S., and Yoon, S. (2024). “Semantic Token Reweighting for Interpretable and Controllable Text Embeddings in CLIP”. In: *arXiv preprint*

arXiv:2410.08469. DOI: 10.48550/arXiv.2410.08469. URL: https://arxiv.org/abs/2410.08469.

Kocev, D., Vens, C., Struyf, J., and Džeroski, S. (2013). "Tree ensembles for predicting structured outputs". In: *Pattern Recognition* 46.3, pp. 817–833. DOI: 10.1016/j.patcog.2012.09.023. URL: https://doi.org/10.1016/j.patcog.2012.09.023.

Meisenbacher, S., Schopf, T., Yan, W., Holl, P., and Matthes, F. (2024). *An Improved Method for Class-specific Keyword Extraction: A Case Study in the German Business Registry*. arXiv preprint. arXiv: 2407.14085 [cs.CL].

Moreno-Torres, J. G., Sáez, J. A., and Herrera, F. (2012). "Study on the Impact of Partition-Induced Dataset Shift on K-fold Cross-Validation". In: *IEEE Transactions on Neural Networks and Learning Systems* 23.8, pp. 1304–1312. DOI: 10.1109/TNNLS.2012.2199516.

Mujahid, M., Kına, E., Rustam, F., Villar, M. G., Alvarado, E. S., Diez, I. D. L. T., and Ashraf, I. (2024). "Data Oversampling and Imbalanced Datasets: An Investigation of Performance for Machine Learning and Feature Engineering". In: *Journal of Big Data* 11, p. 87. DOI: 10.1186/s40537-024-00943-4.

Nisar, M. A., Shirahama, K., Irshad, M. T., Huang, X., and Grzegorzek, M. (2023). "A Hierarchical Multitask Learning Approach for the Recognition of Activities of Daily Living Using Data from Wearable Sensors". In: *Sensors* 23.19, p. 8234. DOI: 10.3390/s23198234. URL: https://www.mdpi.com/1424-8220/23/19/8234.

Nomoto, T. (2022). "Keyword Extraction: A Modern Perspective". In: *SN Computer Science* 4.92. DOI: 10.1007/s42979-022-01481-7.

Oniani, D., Chandrasekar, P., Sivarajkumar, S., and Wang, Y. (2023). "Few-Shot Learning for Clinical Natural Language Processing Using Siamese Neural Networks: Algorithm Development and Validation Study". In: *JMIR AI* 2.1, e44293. DOI: 10.2196/44293.

Papazis, S., Giotis, A. P., and Nikou, C. (2025). "Enhancing Keyword Spotting via NLP-Based Re-Ranking: Leveraging Semantic Relevance Feedback in the Handwritten Domain". In: *Electronics* 14.14, p. 2900. DOI: 10.3390/electronics14142900. URL: https://doi.org/10.3390/electronics14142900.

Probst, P., Wright, M. N., and Boulesteix, A.-L. (2019). "Hyperparameters and tuning strategies for random forest". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.3, e1301. DOI: 10.1002/widm.1301. URL: https://doi.org/10.1002/widm.1301.

Reimers, N. and Gurevych, I. (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 3982–3992. DOI: 10.18653/v1/D19-1410.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q. V., Hinton, G., and Dean, J. (2017). "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer". In: *arXiv preprint arXiv:1701.06538*. DOI: 10.48550/arXiv.1701.06538. URL: https://arxiv.org/abs/1701.06538.

Shyrokykh, K., Girnyk, M., and Dellmuth, L. (2023). "Short Text Classification with Machine Learning in the Social Sciences: The Case of Climate Change on Twitter". In: *PLoS ONE* 18.9, e0290762. DOI: 10.1371/journal.pone.0290762.

Sokolova, M. and Lapalme, G. (2009). "A systematic analysis of performance measures for classification tasks". In: *Information Processing & Management* 45.4, pp. 427–437. DOI: 10.1016/j.ipm.2009.03.002.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). "Attention Is All You Need". In: *arXiv preprint arXiv:1706.03762*. DOI: 10.48550/arXiv.1706.03762.

Xu, D., Shi, Y., Tsang, I. W., Ong, Y.-S., Gong, C., and Shen, X. (2019). “A Survey on Multi-output Learning”. In: *arXiv preprint*. URL: https://arxiv.org/abs/1901.00248.

Ye, H., Sunderraman, R., and Ji, S. (Sept. 2024). “MatchXML: An Efficient Text-Label Matching Framework for Extreme Multi□Label Text Classification”. In: *IEEE Transactions on Knowledge and Data Engineering*. DOI: 10.1109/TKDE.2024.3374750.

Zhang, L. et al. (2024). “IWF-TextRank Keyword Extraction Algorithm Modelling”. In: *Applied Sciences* 14.22. DOI: 10.3390/app142210657.