Beyond Words: Natural Language Processing for Profiling Canine Professionals Based on Website Language and Survey Responses

Rezkya Septiani (23070164)

Department of Statistical Science, University College London
Supervisors: Dr. Chak Hei (Hugo) Lo, Dr. Sarah Weidman, Dr. Jana Muschinski
ucakrps@ucl.ac.uk

Abstract. As more people rely on the Internet for canine-related services, the ability to understand and distinguish various aspects of dog training has become increasingly important. This study investigates whether the techniques that dog trainers say they use match the way they present themselves on their websites. Using responses from 91 trainers alongside publicly available website content, we analyzed the consistency between self-reported training styles and the language found online. We extracted method-specific keyphrases by grouping website texts based on survey-derived combinations of training method, gender, and role, leveraging a pre-trained language model (All MPNet Base V2) to extract discriminative keyphrases. These phrases were transformed into semantic feature vectors and used to train a multi-label classification model. The model achieved strong performance, with macro F1-scores above 0.83 for method-only groupings and over 0.93 when gender and role were included. The findings reveal that website language often aligns with the methods trainers reported in the survey. Certain linguistic cues, such as references to empathy, certification, or aversive tools, carry semantic weight that reflects both professional identity and methodological stance. We hope this approach can assist Dogs Trust not only in assessing existing trainers, but also in anticipating the likely orientation of new or unregistered professionals by using our built model and the extracted semantic keyphrases from their website content. This can support more consistent, transparent, and ethical evaluation across the sector.

1 Introduction

In the digital age, dog owners increasingly turn to online sources when choosing training services. A trainer's website often serves as the first point of contact, shaping perceptions and decisions about who to trust with their pets. As a result, website content functions as a key communication channel, conveying not only a trainer's experience and credentials, but also their underlying values and training philosophies.

However, the language used by trainers is not always consistent or transparent. Terms such as *positive*, *balanced*, or *science-based* are frequently used across ideologically divergent training practices, despite substantial differences in the actual techniques involved. This creates semantic ambiguity that may confuse dog owners or obscure the true nature of a trainer's methods. While organisations such as Dogs Trust advocate for humane, evidence-based approaches, their ability to assess trainers based on online presence is hindered by this lack of linguistic clarity.

Even when trainers declare a specific methodological stance, it remains unclear whether this is consistently reflected in their public-facing materials. Prior research has shown that linguistic choices often align with deeper ideological positions. Johnson et al. [5] demonstrated that trainers identifying as force-free tend to use language invoking empathy, ethics, and scientific reasoning, while those aligned with more traditional or mixed approaches often emphasise discipline, structure, or outcomes.

This study explores the alignment between dog trainers' self-reported methodologies and the way they describe themselves on the website. By analysing a combined dataset of structured survey responses and public-facing website content from 91 UK-based trainers, we aim to uncover semantic patterns that reflect each declared method. Our approach consists of two main stages. First, we extract keyphrases by clustering website texts based on survey-derived combinations of training method, gender, and professional role, using a pre-trained language model to identify semantically

discriminative features within each group. Second, we use these semantic representations to train a multi-label classifier that predicts a trainer's declared methodology from their website content.

The methodology draws on recent advances in keyphrase extraction and label-aware clustering. Gibbs et al. [4] demonstrated how contextual embeddings can uncover coherent themes in short-form text. Zhou et al. [8] proposed a generative framework for topic modelling that performs well even when label information is limited or noisy. Building on these insights, we apply semantic clustering and classification to examine how language used online corresponds with survey-reported practices, and to identify where misalignment may occur.

2 Literature Review

Previous research in applied animal behaviour has shown that the language used by trainers often reflects their underlying training philosophy. Studies have reported consistent vocabulary differences across training approaches, with additional variation associated with factors such as gender and certification status [5]. These findings, although generated through manual analysis, suggest that linguistic expression can serve as a proxy for deeper ideological and methodological commitments.

Building on this foundation, recent studies have explored computational ways to detect such differences. Reimers and Gurevych [6] introduced Sentence-BERT, a model for generating dense sentence-level embeddings that improves semantic similarity tasks, including keyphrase extraction. Compared to traditional frequency-based methods, embedding-based approaches better capture contextual meaning and support generalisation across phrasing.

While modern language models are relatively robust to noisy input, text preprocessing remains a critical part of any pipeline. Belinkov and Glass [1] emphasized the continued value of standard preprocessing steps such as stopword removal, normalisation, and structure cleanup to ensure interpretability and consistency, particularly in domain-specific applications. This is especially important in our case, where website content often contains redundant or poorly structured text.

To improve the coherence and semantic distinctiveness of extracted keyphrases, we adopt a clustering-based strategy that groups documents based on self-reported survey attributes. Gaur et al. [2] proposed community detection on semantic embeddings as a way to organise conceptually similar phrases and improve keyphrase relevance. Similarly, Giarelis and Karacapilidis [3] reviewed deep learning-based methods and found that combining contextual embeddings with unsupervised clustering produced more interpretable output in specialised domains.

Finally, Song et al. [7] offered a comprehensive review of recent developments in keyphrase extraction using pre-trained language models. Their analysis found that embedding-based techniques consistently outperform traditional baselines, particularly in tasks involving categorical distinctions such as those found in ideological or methodological classifications.

3 Methodology

This study followed a structured three-stage pipeline designed to extract discriminative semantic signals from public-facing text and test their predictive alignment with declared training styles. The pipeline consists of (1) structured data preparation and web text cleaning, (2) extraction of keyphrases through survey-guided clustering, and (3) classification based on semantic alignment features.

3.1 System Environment and Dependencies

The entire workflow was implemented in Python 3.12 and executed in a reproducible Conda environment. Data were retrieved from Microsoft SQL Server 2019 via `SQLAlchemy` and `pyodbc`. The following libraries were used for key functionalities:

Language:	Python 3.12 (Conda-managed)
Database:	Microsoft SQL Server 2019
Libraries:	pandas, numpy, matplotlib scikit-learn, sentence-transformers keybert, torch, torchvision
Environment:	SQL Server Management Studio v18

3.2 Stage One: Data Preparation and Cleaning

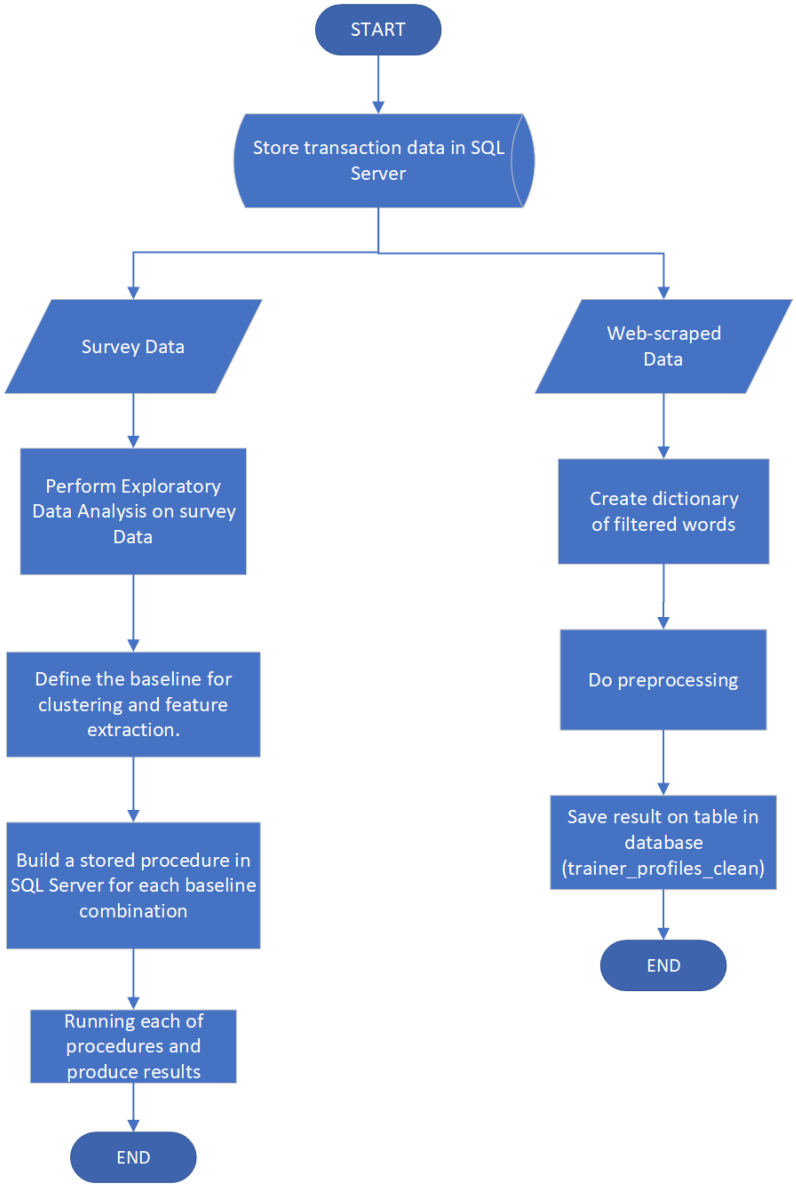

Fig. 1: Stage 1. Cleaning and Structuring of Survey and Web Data

This stage involved assembling and cleaning two linked datasets: structured survey responses from 91 UK-based dog trainers and their associated website content gathered via targeted web scraping. For each trainer, multiple webpages were concatenated into a single text document, forming a one-to-one mapping between survey response and online presence. To reduce structural and noise,

we implemented a rule-based text cleaning pipeline using SQL stored procedures. This involved stripping HTML elements, footers, navigation bars, UK phone numbers, placeholder tokens (such as `[REDACTED]`), and duplicate section headers. A curated dictionary of filtered terms was also used to exclude redundant or generic phrases and elements across websites. These terms were compiled into a filtered word list and iteratively refined to improve the semantic quality of the cleaned dataset. Survey data contained multiple-selection fields, including training methods used, tools used, etc. These fields were normalised and structured into relational tables and masters as part of a broader data modelling strategy. Each set of values was linked back to the main trainer profile through referential joins, enabling flexible querying and efficient subgrouping during downstream analysis. Prior to downstream modelling, we conducted exploratory analysis on the survey data to identify valid trainer subgroups that could serve as interpretable baseline clusters. This step informed the creation of survey-derived combinations (e.g., method-gender-role) which were used to guide semantic keyphrase extraction in the next stage. After this stage, only 89 out of 91 trainer profiles were retained for further processing, as two entries contained no usable website content after cleaning. These cases were treated as missing values and excluded from downstream modelling.

3.3 Stage Two: Keyphrase Extraction and Semantic Vectorisation

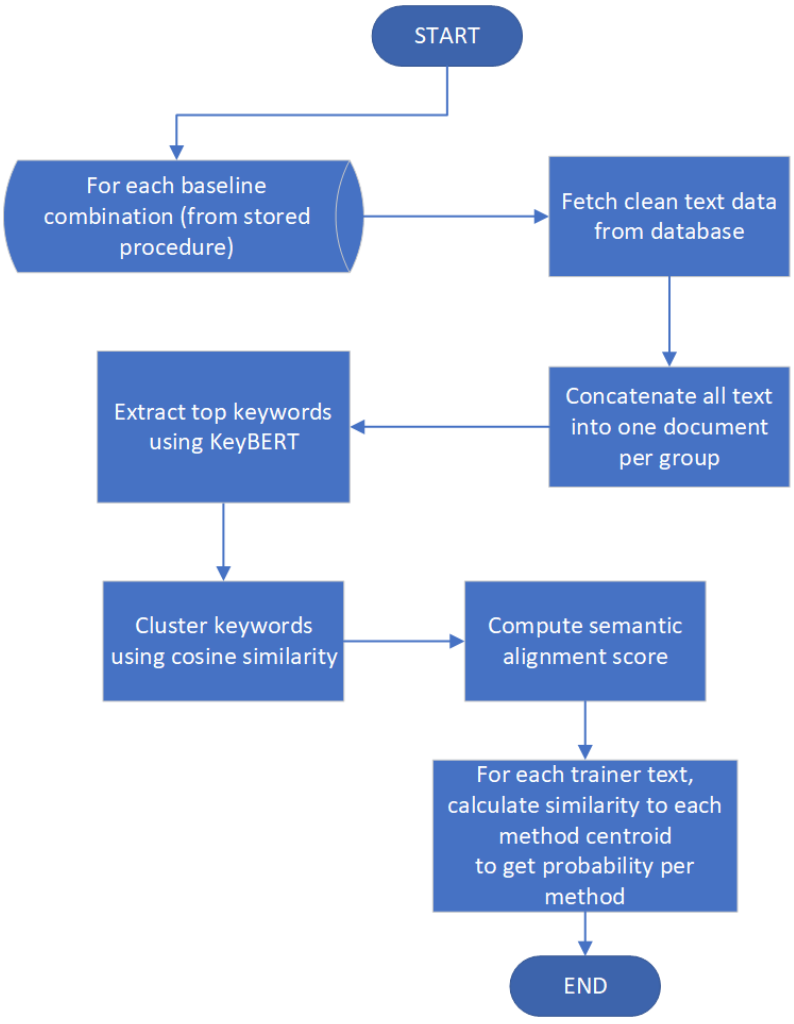

Fig. 2: Stage 2. Extraction of Semantic Keyphrases and Feature Mapping

In this stage, we aimed to extract keyphrases that are semantically discriminative across different trainer subgroups. Trainers were grouped based on baseline combinations derived from their declared method, gender, and role, as defined during the survey preprocessing. Each group was treated as a semantic cluster representing shared language characteristics within that combination.

For each group, the corresponding website content was merged and processed separately. We applied a pre-trained sentence transformer model (All MPNet Base V2) to generate semantic representations of the text. Candidate keyphrases were extracted using a contextual keyword extraction model and grouped to eliminate redundancy and enhance interpretability. A predefined blacklist was applied to remove generic expressions and ensure that each group retained unique and meaningful keyphrases.

The final output of this process was a final set of discriminative keyphrases representative of each baseline combination. These phrases were then compared against each trainer's website to evaluate how strongly their language aligned with different subgroup profiles. These semantic representations were later used as input for classification.

3.4 Stage Three: Classification and Evaluation

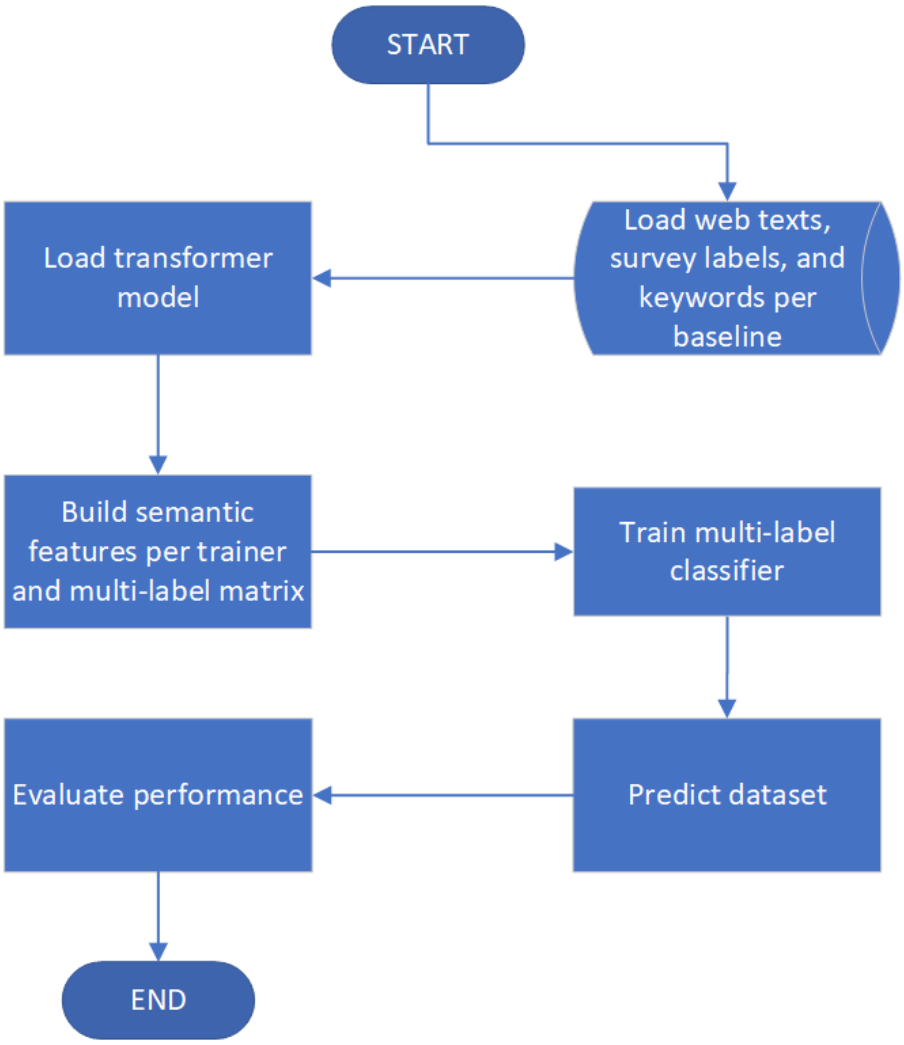

Fig. 3: Stage 3. Classification and Evaluation Using Semantic Features

These patterns support our modelling strategy, where method-only labels are used for baseline comparisons, and richer combinations of method, gender, and role serve to extract more discriminative semantic features. The multi-label structure of the task is further reinforced by the observation that many trainers selected more than one method, necessitating a classification framework that accommodates overlapping categories.

Rather than apply a generic model across all trainers, we structured the semantic extraction process around these survey-derived groupings. Each subgroup was treated as a linguistically meaningful cluster, enabling the identification of keyphrases that are semantically discriminative for each configuration. This setup provided the foundation for extracting contextual features and aligning individual trainer texts with the linguistic profiles of their respective groups.

4 Result

4.1 Exploratory Data and Analysis

Before building semantic models, we conducted exploratory analysis on both the survey responses and website corpus to guide modelling decisions and feature extraction strategies. Across the 91 respondents, the distribution of declared training methods showed substantial imbalance. As visualised in Figure 4, the "Introduce Rewards" approach dominated, with over 60% of trainers selecting this option. In contrast, aversive techniques such as "Introduce Unpleasant" or "Remove Unpleasant" were used far less frequently, contributing to skewed class distributions. This justified the use of macro-averaged F1-score as the primary evaluation metric during classification.

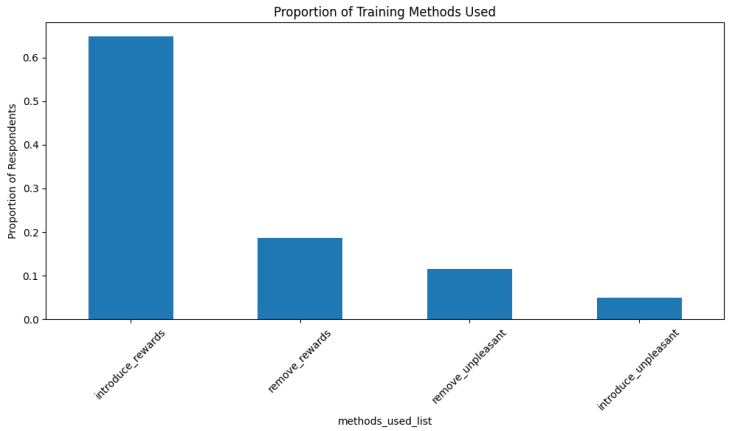

Fig. 4: Overall distribution of declared training methods across all respondents

Beyond overall frequencies, we examined how method selection varied by demographic subgroup. Figures 5 and 6 show that gender and professional role appear to shape training orientation. For instance, female trainers were more likely to adopt reward-based methods, while male trainers showed a broader spread across different approaches. Similarly, accredited trainers displayed stronger preference for reward-based practices compared to those without formal qualifications. These observations motivated our decision to include demographic attributes as part of the baseline clustering for keyphrase extraction.

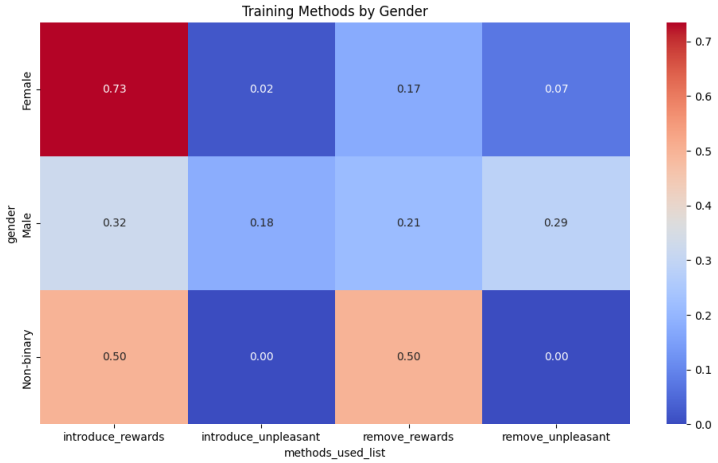

Fig. 5: Distribution of training methods by gender

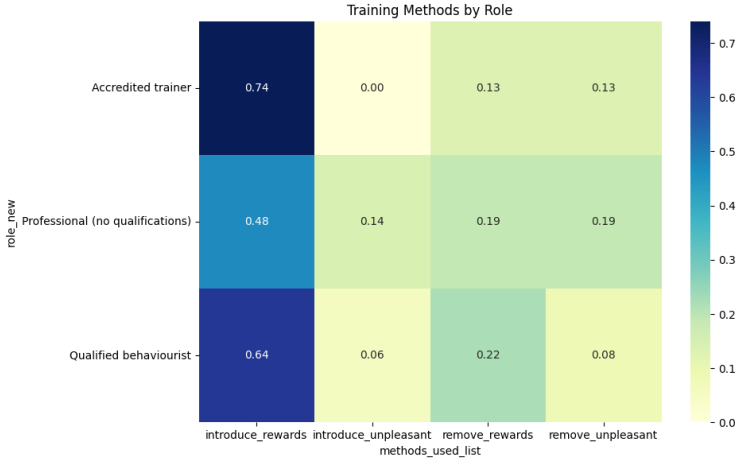

Fig. 6: Distribution of training methods by professional role

These patterns support our modelling strategy, where method-only labels are used for baseline comparisons, and richer combinations of method, gender, and role serve to extract more discriminative semantic features. The multi-label structure of the task is further reinforced by the observation that many trainers selected more than one method, necessitating a classification framework that accommodates overlapping categories.

These exploratory insights reinforced our decision to structure the semantic extraction process around baseline groupings derived from the survey, both at the method-only level and in combination with gender and professional role. Rather than apply a generic model across all trainers, we treated each subgroup as a linguistically meaningful cluster, allowing for the identification of keyphrases that are semantically discriminative for each configuration.

The next stage focused on extracting these group-specific keyphrases using contextual embeddings, followed by vector-based alignment between trainer texts and subgroup profiles. This allowed us to quantify the extent to which individual trainers' website content semantically reflects their declared practices.

4.2 Keyword Extraction and Semantic Differentiation

Keyword Extraction by Method To assess linguistic patterns distinctive to each training philosophy, we performed guided keyphrase extraction anchored on the four primary training method groups declared in the survey. Each phrase was assigned a semantic score, reflecting its alignment with the corresponding method's embedding profile.

Table 1: Top semantically relevant keywords per training method and inferred topic

Method Code	**Keyword**	**Score**
01 (Introduce Rewards)	obedience vets testimonials, heelwork course canine, faq bark busters, barking, tricks just treats	0.677-0.470
02 (Remove Rewards)	behaviour vets puppies, behaviourist animal instructor, barking learn, trainer behaviourist, bark motivation persistence	0.653-0.546
03 (Remove Unpleasant)	positive paws, decoding discomfort canine, breed behaviour, barking destructive behaviour, fit pup needs	0.598-0.498
04 (Introduce Unpleasant)	local behaviourist trainer, therapist bark, wolves captivity learning, excessive barking destructive, general temperament improved	0.635-0.498

The *semantic score* quantifies how representative a keyword is of the method group's linguistic profile, based on cosine similarity within embedding space. Values near 1.0 indicate high contextual alignment. Here, top terms across methods range between 0.47-0.68, showing effective capture of semantically relevant discourse.

To evaluate the quality and interpretability of the extracted keywords, we manually examined the top five terms for each of the four training methods, guided by semantic scores and contextual alignment. Overall, the results indicate strong internal coherence and support for the underlying survey labels, suggesting that our semantic extraction approach is meaningful and robust.

Method 01 : Positive Reinforcement & Basic Obedience: This category produced keywords such as *"obedience vets testimonials"*, *"heelwork course canine"*, and *"tricks just treats"*, which are highly consistent with introductory-level, reward-based training. The phrase *"faq bark busters"* further grounds this category in well-known positive training brands. Together, these terms suggest a focus on trust-building, early obedience, and owner engagement. Words like *"animal instructor"* and *"canine connections"* hint at structured group environments. These alignments support the interpretation of this cluster as beginner-oriented, treat-driven instruction grounded in praise and repetition.

Method 02 : Behavioural Intervention: The top keywords for this group, including *"behaviour vets puppies"*, *"trainer behaviourist"*, and *"bark motivation persistence"*, reflect a more structured, intervention-style training. The use of professional terms (e.g., *"behaviourist animal instructor"*) implies expert involvement, especially in guiding puppy behavior. While there is no direct indication of aversive methods, the absence of reward-centric language (e.g., treats, play) distinguishes this method from Method 01. Instead, phrases suggest persistent, goal-oriented shaping of behavior, often used when casual reinforcement is insufficient.

Method 03 : Positive Training Tailored to Dog Needs: This cluster includes keywords such as *"positive paws"*, *"conditioning"*, *"trainer agility positive"*, and *"fun welcome breeds"*, which convey a customisable, health-informed approach. The presence of terms like *"canine nutrition ni"* and *"fit pup needs"* suggests a holistic philosophy that emphasizes well-being over coercion. Importantly, while behaviour correction is still present (e.g., *"barking destructive behaviour"*), it is handled with positive reinforcement and breed-sensitive planning. This supports the notion of low-stress training adapted to specific dog types or issues.

Method 04 : Strong Correction and Behaviour Control: The most distinctive set of keywords was observed in this group. Terms such as *"wolves captivity learning"*, *"therapist bark"*, *"excessive barking destructive"*, and *"general temperament improved"* suggest that this training method targets more serious or ingrained behavioural challenges. Unlike other methods, this cluster

also includes references to clinical framing (*"residential book session"*) and emotional stress (*"stress local behaviourist"*). These linguistic cues align well with aversive or corrective philosophies, where behavioural control is achieved via structured discipline rather than motivation alone.

In summary, the extracted semantic features not only demonstrate strong internal validity (as reflected by semantic scores ranging from 0.47 to 0.68), but also yield interpretable clusters that align with expert intuition and prior EDA. This reinforces the notion that training method labels correspond to distinct discursive patterns within trainer websites.

Keyword Extraction by Combined Method, Gender, and Role To explore how language use varies across different trainer profiles, we re-ran the keyword extraction process based on combinations of training method, gender, and professional role. Keywords were ranked by their semantic relevance, with the aim of identifying distinctive phrases that reflect how specific groups position themselves online.

Table 2: Top 5 filtered keywords and semantic score range per method-gender-role combination

Cluster Code	**Top 5 Keywords (Filtered)**	**Semantic Score Range**
01-01-01	canine trainers knowledge, animal behaviour council, agility canine, herding breeds, canine companion	0.525–0.675
01-01-02	consultations group muttamorphosis, behaviour owners vets, educators canine, testimonials team paws4teaching, masters qualified animal	0.474–0.540
01-01-03	hamble hounds trainer, pup teaching, canine coaching based, courses older pups, pupstars behavioural support	0.546–0.719
01-02-01	labrador excellent trainer, experience working breeds, rescue frenchie behavioural, experienced local trainer, session lot trainers	0.375–0.604
01-02-02	trainer specialising, trainers behaviourists, experienced canine behaviourist, trainee handler bit, vets nurses trainers	0.512–0.638
01-02-03	aggression bark busters, stressful pup disruptive, therapist trainer bark, chippy cotswolds, education bark busters	0.457–0.520
01-03-01	compliments trained works, works trainers encyclopaedic, confident pup approach, committed professional growth, emotional level works	0.440–0.662
02-01-01	paws train pre, heavily positive reinforcement, animal trainer proud, consultations beginner obedience, animal care	0.505–0.593
02-01-02	behaviour musings, obedience shows, socialisation, consultation make animal, obedience lifeskills	0.389–0.460
02-01-03	clever canines working, clever canines ages, behavioural clever canines, house advanced obedience, suitable breed	0.547–0.805
02-02-01	obedience trainers, trainer great paw, experience sizes breeds, struggling english bull, rescue frenchie behavioural	0.477–0.604
02-02-02	trainers behaviourists, scent detection course, experienced canine behaviourist, trainer specialising, vets nurses trainers	0.448–0.638
02-02-03	trainer independent business, leadership role, organisations, giving guidance, behaviour therapist business	0.360–0.506
03-01-01	private agility arena, paddock professional trainer, fun agility practice, agility areana, knowledgeable trainer daughter	0.477–0.652
03-01-02	crate reactivity aggression, rspca animal, temperament improved, stress local behaviourist, local behaviourist trainer	0.434–0.542
03-02-01	british college canine, canine behaviour shop, limited electric collars, microchipped canine, great trainer polite	0.380–0.582
03-02-02	mastering canine, behavioural framing, excessive barking, wolves captivity learning, minds canine cognition	0.463–0.641
04-01-02	remote collars, trainer behaviour manual, collars uses importance, behaviourist trainer resolve, behaviourists hand advice	0.461–0.573

Note: Cluster codes are structured as `MM-GG-RR`, where:

- `MM` indicates the training method:
  - 01 = introduce rewards
  - 02 = remove rewards
  - 03 = remove unpleasant
  - 04 = introduce unpleasant
- `GG` indicates gender:
  - 01 = Female
  - 02 = Male
  - 03 = Non-binary
- `RR` indicates professional role:
  - 01 = Accredited trainer
  - 02 = Qualified behaviourist
  - 03 = Professional (no qualifications)

A total of 371 keywords were extracted from the 17 combined clusters of training method, gender, and professional role. From these, we selected the top five keywords in each cluster based on semantic distinctiveness. This approach aimed to capture how different trainer profiles linguistically present themselves online.

When comparing clusters with the same method but different genders, such as 01-01-03 (female, no qualifications) and 01-02-03 (male, no qualifications), a noticeable contrast emerges. While both reflect the same reward-based philosophy, female trainers use phrases that evoke education and care (e.g., pup teaching, courses older pups), whereas male trainers employ more authoritative or corrective language (e.g., aggression bark busters, therapist trainer bark).

A similar pattern is observed when contrasting roles within the same method and gender, such as 01-01-01 (female, accredited trainer) versus 01-01-03 (female, no qualifications). The former favours technical or institutional terms (e.g., behaviour council, canine trainers knowledge), whereas the latter opts for softer and more educational language that appeals to general audiences.

Gender-matched comparisons across roles also show divergence. For instance, in clusters 02-01-01 and 02-01-03 (both female), the accredited trainers emphasise structured reinforcement (beginner obedience, animal care), while those without formal credentials adopt branding-like language (clever canines, suitable breed).

These differences indicate that tone and public presentation are not solely tied to method but are co-shaped by gender and professional identity, influencing how trainers balance authority, empathy, and marketing appeal in their online profiles.

5 Model Evaluation Using Semantic Features

To evaluate whether language use can help predict a trainer's stated methods, we trained a multi-label Random Forest model using the extracted semantic features. The model was trained on 80% of the data with internal cross-validation and tested on the remaining 20%.

Metric	**Method Only**	**Method + Gender + Role**
Macro F1-score	0.8310	**0.9352**
Accuracy	0.8421	**0.8571**

Table 3: Comparison of classification metrics based on semantic features from two grouping strategies

The evaluation shows that incorporating gender and professional role alongside training method significantly improves the model's ability to generalise across classes. While both models performed strongly, the macro F1-score, which is particularly important for imbalanced multi-label data, increased from 0.8310 in the method-only setup to 0.9352 when using the combined grouping.

This substantial improvement suggests that including additional contextual information allows the model to capture more consistent linguistic patterns related to how trainers describe their practices.

Although the overall accuracy also increased slightly, from 84.21% to 85.71%, the gain in macro F1-score highlights a better balance between precision and recall across all classes, including those that are less frequently represented. In other words, the model trained with method-gender-role combinations was not only more accurate, but also more sensitive to varied language styles and professional perspectives.

These results confirm that the extracted phrases reflect more than surface-level style. They carry meaningful signals about how different types of trainers communicate their approach, making semantic features from web content a useful basis for identifying training philosophies.

6 Conclusion

This study demonstrates that the language used by dog trainers on their websites carries significant signals about their approach to training. Through exploratory analysis, semantic keyword extraction, and multi-label classification, we found that certain words and phrases tend to align with specific self-reported methods.

By grouping the data based on training method, gender, and professional role, we were able to extract keywords that not only describe surface branding or tone, but also reflect deeper methodological orientation. The classification model trained on these semantic features achieved strong performance, with a macro F1-score of 0.9352 using the combined grouping, compared to 0.8310 for method-only clusters.

Importantly, these discriminative keywords now serve as reliable signals for inferring which type of training a professional is likely to favour. This allows the model to be used beyond the current dataset, including for new or unregistered trainers. By analysing their website content alone, the system can provide an initial prediction of whether a trainer's language reflects reward-based, mixed, or aversive tendencies.

This approach offers practical value for organisations such as Dogs Trust. It enables early and low-effort screening of trainer orientation, supports consistent evaluation across diverse professionals, and helps promote safer and more transparent decision-making in the dog training sector.

References

1. Belinkov, Y., Glass, J.: Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics **7**, 49–72 (2019)
2. Gaur, M., Alambo, A., Sheth, A.: Keyphrase extraction using semantic embeddings and community detection. In: Proceedings of The Web Conference 2022 (WWW). pp. 3127–3137 (2022)
3. Giarelis, N., Karacapilidis, N.: Deep learning and embeddings-based approaches for keyphrase extraction: a literature review. Expert Systems with Applications **192**, 116387 (2024)
4. Gibbs, A., Zhao, Y., Kim, L.: Gibbs-bertopic: A hybrid approach for short-text topic modeling. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2025), `https://www.researchgate.net/publication/389929907_Gibbs-BERTopic_A_Hybrid_Approach_for_Short_Text_Topic_Modeling_February_2025`
5. Johnson, L.M., Feuerbacher, E.N., Wynne, C.D.L.: Training dogs with science or with nature? an exploration of trainers' word use, gender, and certification across dog-training methods. Journal of Veterinary Behavior **62**, 1–10 (2023). `https://doi.org/10.1016/j.jveb.2023.05.004`
6. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. pp. 3982–3992 (2019)
7. Song, M., Feng, Y., Jing, L.: A survey on recent advances in keyphrase extraction from pre-trained language models. Findings of the Association for Computational Linguistics (EACL) pp. 2108–2119 (2023)
8. Zhou, J., Lee, M., Fernandes, A.: Topicgen: Topic modeling for small data using generative llms. arXiv preprint arXiv:2211.12878 (2024), `https://arxiv.org/abs/2211.12878`