Predicting Perceptions of Dutch Company Names from Linguistic Features

Hiwa Feizi

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE IN COGNITIVE SCIENCE AND ARTIFICIAL INTELLIGENCE
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES OF TILBURG
UNIVERSITY

STUDENT NUMBER
2094203

COMMITTEE
dr. Giovanni Cassani
dr. Julija Vaitonyte

LOCATION
Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands

DATE
May 10th, 2025

WORD COUNT
7978

Predicting Perceptions of Dutch Company Names from Linguistic Features

Hiwa Feizi

Abstract

How do people form impressions of company names, and can these perceptions be predicted from linguistic features alone? This thesis investigates whether surface-form and semantic features—unigrams, bigrams, and RobBERT embeddings—can model human trait judgments of Dutch company names across four dimensions: femininity, evilness, trustworthiness, and smartness. Using ElasticNet and feedforward neural networks (FFNNs), we evaluate generalization across three experimental setups: in-domain learning, cross-domain transfer from personal names and pseudowords, and a hybrid setting combining both. Results show that simple surface-level features, particularly unigrams, provide the most reliable predictions, while more complex models and high-dimensional embeddings often overfit. Crucially, including even 100 real company names in training significantly improves generalization.

To support further research, we publicly release all trained models and the feature importance rankings from every ElasticNet run, enabling future studies to trace how individual letters, bigrams, or semantic components contribute to perception. Our full codebase is also open source and available at https://github.com/hiwafeizi/thesis, designed to be easily adapted to other name types, traits, or languages.

1.1 *Source/Code/Ethics/Technology Statement*

The dataset used in this thesis was provided by Dr. Cassani, the project supervisor, who owns the rights to the data and granted full consent for its use. The dataset was originally collected from Tilburg University students, with explicit consent obtained from all participants. It includes ratings of Dutch company names across various perceived traits and was downloaded via the secure academic transfer service FileSender by SURF.

All figures and visualizations in the thesis were created by the author using Python-based plotting libraries. No third-party images were included, so no additional permissions were required.

The entire codebase was written from scratch by the author. While no external code was copied, OpenAI's GPT-4 was used as a programming assistant to help debug, refactor, and improve parts of the Python code, particularly during exploratory analysis and model evaluation. All source code was developed in Python 3.13.2 and executed within a virtual environment on Windows 10. The key libraries included scikit-learn (1.4.2), torch (2.7.0), transformers (4.51.3), tokenizers (0.21.1), joblib (1.4.2), pandas (2.2.3), numpy (2.2.4), and matplotlib (3.10.1). A full list of dependencies is provided in the accompanying requirements.txt file for easy environment replication.

No automated paraphrasing tools or academic writing generators were used beyond GPT-4o, which also assisted in brainstorming and editing during writing. The thesis was written and typeset using Google Docs, with grammar and clarity manually reviewed by the author. Reference formatting followed APA style and was applied throughout the document with the use of Scribbr reference management software.

1 Introduction

While the influence of phonological and semantic features on brand name perception is well established, little is known about how these cues operate in the specific context of Dutch company names. This is a notable gap, as company names often diverge from standard linguistic norms by incorporating invented forms, stylized spellings, or foreign elements. Such deviations pose unique challenges for both linguistic interpretation and computational modeling. Understanding how people perceive personality traits in these names can inform not only theories of language and cognition but also practical applications in branding, automated name generation, and marketing analytics. This thesis addresses that gap by systematically investigating the linguistic and semantic factors that shape human impressions of Dutch company names.

One reason names trigger immediate impressions is that they encode patterns—both phonological and semantic—that people intuitively associate with psychological and social traits. These associations are not random; rather, they reflect systematic regularities in how linguistic forms map onto perceived meanings. The following sections outline how both phonological features and semantic content have been shown to influence name perception.

When people encounter a company name for the first time, they often form immediate impressions about the brand's personality. These impressions are not arbitrary; rather, they stem from systematic patterns in the name's linguistic form (Sidhu & Pexman, 2015). Prior research has shown that certain phonological features evoke consistent trait associations. For example, front vowels such as "ee" are typically perceived as soft and feminine, while back vowels like "oo" or "uh" are linked to masculinity and strength (Klink & Athaide, 2011). Similarly, names that are fluent and easy to pronounce tend to be judged more positively, evoking impressions of trustworthiness and familiarity (Alter & Oppenheimer, 2006). These findings suggest that surface-level features—such as individual letters (unigrams), letter pairs (bigrams), and their phonotactic patterns—can carry implicit sound-symbolic meanings.

However, brand name impressions are shaped not only by phonological form but also by semantic content. Many effective names derive their appeal from recognizable morphemes, real words, or familiar concepts that evoke specific associations. *Redbull* and *Booking.com*, for instance, explicitly embed meaningful lexical units—*red* and *bull*, or *booking*—to convey notions of energy, strength, action, or utility. These associations tap into shared conceptual knowledge and are crucial for memorability and perceived relevance. Such effects are difficult to model using surface-level statistics alone. Recent advances in natural language processing address this challenge through contextual embeddings: high-dimensional semantic representations learned from large-scale corpora by transformer-based models. For Dutch, the RobBERT language model offers such embeddings, trained specifically on Dutch-language data (Delobelle, Winters, & Daelemans, 2020). Prior studies have shown that embeddings from

models like BERT align well with human intuitions about meaning, including semantic transparency and lexical dominance in compound words (Buijtelaar & Pezzelle, 2023), making them a promising tool for modeling name-based perception.

The dataset used in this study comprises 600 items spanning three distinct categories: Dutch given names, artificially constructed pseudowords, and invented company names. Each name was evaluated by Dutch-speaking students at Tilburg University along four personality-related traits: femininity, evilness, trustworthiness, and smartness. The original Dutch terms used in the survey were vrouwelijk, slecht, betrouwbaar, and slim, respectively. These traits were selected because they reflect socially and psychologically meaningful dimensions that individuals often project onto names—both in interpersonal settings and branding. Prior research in psycholinguistics and marketing has shown that such impressions play a significant role in judgments of likability, competence, and credibility.

Rather than assigning fixed-scale ratings, participants performed pairwise comparisons, selecting which of two names better reflected a given trait. This format reduces individual bias and increases sensitivity to subtle semantic differences. To infer continuous scores from these comparisons, a Bayesian ranking model was fit separately for each trait and item category. The model generated 10,000 posterior draws per trait–category pair, each representing a plausible ranking consistent with the observed responses. In this thesis, the mean of these draws is used as the target score for each name. Prior to modeling, responses were cleaned by removing participants who failed attention checks or responded implausibly fast or slow. The resulting dataset provides a high-quality foundation for evaluating which types of linguistic features—surface-level (unigrams, bigrams) or contextual (semantic embeddings)—most effectively predict perceived personality traits across different naming domains.

The central aim of this thesis is to investigate how linguistic form and semantic information influence human perceptions of personality traits in Dutch company names. Specifically, the study explores whether surface-level features—such as unigrams and bigrams—and contextual semantic embeddings derived from RobBERT can serve as effective predictors of perceived traits. The analysis focuses on four core questions that address both the predictive power of different feature representations and their generalisability across naming domains and modelling approaches.

To guide this investigation, the following research questions are addressed:

1. To what extent can linguistic feature sets—unigrams, bigrams, RobBERT embeddings, and their combination—predict Dutch speakers' perceptions of company names when models are trained and tested within the same domain?

2. Can models trained exclusively on structurally similar but domain-distinct data (i.e., personal names and pseudowords) generalise effectively to the domain

of Dutch company names?

3. Does augmenting the training data with a small number of in-domain company names substantially improve generalisation performance to the target domain?

4. Do non-linear models (e.g., feedforward neural networks) offer predictive advantages over linear models when applied to the same linguistic feature sets?

Together, these questions form the foundation for a systematic investigation into how different linguistic representations shape trait perception in brand names, offering new insights for linguistic theory, brand design, and computational modeling.

2 Related work

Understanding how people form impressions of names requires insights from multiple disciplines—psycholinguistics, marketing, and computational linguistics. This chapter reviews foundational work on the phonological and semantic cues that influence name perception, with a focus on features that systematically influence perceptions of social traits such as gender, intelligence, and trustworthiness.. It also highlights gaps in existing literature concerning Dutch commercial naming practices and presents prior methodological approaches that inform this thesis. The review is organized into four parts: (1) phonological and orthographic cues, (2) semantic cues and embeddings, (3) naming conventions in Dutch commercial contexts, and (4) pseudoword modeling as a methodological precedent.

2.1 Phonological and Orthographic Cues

One of the most consistent findings in sound symbolism research is that specific phonetic features evoke reliable associations with abstract social traits across languages. For example, high front vowels like /i/ (“ee”) are often perceived as small, soft, and feminine, whereas low back vowels such as /u/ or /o/ are linked to largeness, strength, and masculinity (Klink, 2000; Klink & Wu, 2013). Consonantal voicing has also been shown to shape perception: voiced stops (/b/, /d/, /g/) tend to be associated with heaviness and power, while voiceless stops (/p/, /t/, /k/) convey lightness or sharpness due to differences in acoustic intensity (Sidhu & Pexman, 2019). These mappings between sound and meaning are thought to be grounded in universal principles of auditory perception, yet their expression can be culturally modulated.

These sound-symbolic effects extend beyond individual phonemes to broader patterns in syllables, stress, and letter combinations. In their study of English names, Cutler, McQueen, and Robinson (1990) found that male names more frequently begin with stressed syllables, while female names tend to begin with unstressed syllables and often contain front vowels—reinforcing perceptions of strength versus softness, respectively. Such findings suggest that relatively

simple phonological and orthographic features—such as syllable structure, vowel quality, or bigram frequencies—can influence judgments of gender and other personality-related traits.

In branding and marketing research, these phonological cues have been shown to affect product evaluation, brand likability, and perceived credibility. Alter and Oppenheimer (2006) found that names that are easier to pronounce are rated as more trustworthy, while Klink and Athaide (2011) showed that vowel and consonant choices influence brand perceptions. These insights suggest that even in artificial or invented brand names, surface-level linguistic features may cue socially meaningful impressions. This is particularly relevant for the current study, which seeks to predict ratings of femininity, trustworthiness, smartness, and evilness based on features such as unigrams, bigrams, and their combinations.

2.2 Semantic Cues and Meaning in Brand Names

While sound symbolism captures how phonetic forms evoke intuitive associations, many modern brand names derive their impact from semantic cues. This typically involves embedding recognizable morphemes or real words that trigger meaningful associations in the mind of the consumer. For example, names like Redbull and Booking.com convey energy, strength, and utility by invoking the concrete meanings of their lexical components (“red,” “bull,” and “booking”). Research shows that such semantically transparent names can enhance memorability, perceived relevance, and consumer trust (Klink, 2001).

Capturing these types of associations, however, is not straightforward—especially when names are stylized, abbreviated, or formed from non-standard language. Traditional surface-form features like character n-grams or morpheme frequency often fail to reflect the deeper meanings evoked by a name. This has motivated the use of contextual semantic models, which embed words within high-dimensional vector spaces based on their usage patterns in large text corpora.

In the Dutch language domain, RobBERT—a RoBERTa-based transformer model trained on extensive Dutch web and news data—offers such representations (Delobelle, Winters, & Daelemans, 2020). These embeddings are context-aware, meaning that they account for word meaning in situational contexts, and have been shown to reflect human intuitions about semantic transparency and lexical compositionality (Cassani et al., 2023). This makes them particularly well-suited for modeling trait perceptions that depend on subtle word associations rather than explicit literal meaning.

Nevertheless, applying contextual embeddings to brand name modeling presents its own challenges. Although transformer-based models like RobBERT are trained on large Dutch corpora, many company names in our dataset are artificially generated and may not occur in natural language usage. As a result,

their semantic representations may be less grounded than those of real words. Additionally, the high dimensionality and opaque nature of transformer-based embeddings complicate interpretability, making it difficult to identify which semantic features drive human trait judgments. These factors underscore the need to empirically evaluate whether contextual embeddings truly offer predictive advantages in the domain of brand perception modeling.

2.3 Brand Naming in Dutch and the Remaining Research Gap

Despite growing interest in sound symbolism and semantic modeling, few studies have addressed how Dutch speakers interpret the linguistic properties of company or brand names. Much of the existing work focuses on English-language stimuli or isolated pseudowords, leaving the Dutch commercial naming context relatively underexplored.

A recent contribution is the Klinkt leuk! project (Cassani, Joosse, & van Kesteren, 2023), which introduces a tool for predicting personality associations with Dutch names, pseudowords, and generated company names. Using best–worst scaling, the authors collected ratings for femininity, trustworthiness, intelligence, and evilness—traits highly relevant to brand perception. They modeled these ratings using FastText and RobBERT embeddings. However, the project has so far only been presented as a conference poster, and no peer-reviewed article or public tool has yet been released. Moreover, their primary focus lies in building a prediction interface rather than in systematically analyzing generalisation across linguistic domains or model architectures.

Another relevant study by Joosse, Kuscu, and Cassani (2023) investigated how formal features of Dutch fictional character names align with perceived attributes such as polarity, gender, and age. Using distributional semantics and machine learning models, the authors demonstrated that names systematically evoke trait inferences in Dutch, echoing similar findings in English. Although the study focuses on fictional rather than commercial names, it supports the broader hypothesis that name form influences social perception—a premise central to this thesis.

To date, few studies have rigorously examined the generalizability of linguistic features—phonological, orthographic, or semantic—across different name categories in Dutch. While earlier work in English has demonstrated that sound-symbolic cues influence brand perception (Klink & Athaide, 2011), such findings have limited applicability to Dutch naming conventions. Dutch company names often include invented forms or foreign words, and many feature stylized spellings with special characters or unconventional capitalization (e.g., *Bol.com*, *Q-Park*, *Phynx*). Although this thesis does not explicitly model such visual or typographic variations, it addresses the broader gap by analyzing how Dutch speakers perceive personality traits in invented company names. In particular, it evaluates which types of linguistic representations—surface-level

features or contextual embeddings—best predict these judgments across domains.

2.4 Methodological Inspiration from Pseudoword Modeling

A relevant methodological inspiration for this study is the work by Gatti, Raveling, Petrenco, and Günther (2024), which explored how linguistic features predict human valence judgments of pseudowords—novel, non-lexical items with no inherent meaning. Their study combined surface-level orthographic features, such as bigram frequencies and orthographic neighborhood density, with semantic embeddings from fastText to model perceived emotional valence. Using regression-based evaluation, they found that surface-form features were generally more predictive than semantic embeddings, suggesting that perceptual judgments often arise from shallow linguistic cues rather than deeper lexical semantics.

While our research diverges in scope and focus, their framework offers conceptual and methodological parallels. Like Gatti et al. (2024), we examine how form and meaning-based features contribute to human judgments, but our study targets a broader set of perceived traits—femininity, trustworthiness, smartness, and evilness—and applies these models to company names, a more ecologically grounded and socially relevant domain. We also adopt a similar regression-based evaluation approach, using R^2 to assess how well different feature representations explain human trait ratings.

Instead of fastText, we utilize RobBERT, a Dutch transformer-based language model, to investigate whether contextual embeddings capture trait-relevant information in the stylized domain of commercial names. While our feature sets and target traits differ, the Gatti et al. (2024) study reinforces the value of combining surface and semantic cues and provides a benchmark for comparing the predictive power of these approaches in novel linguistic domains.

3 Methods

This chapter outlines the methodological framework used to investigate how linguistic and semantic features of Dutch company names predict human judgments of personality-related traits. Building on prior work in sound symbolism and semantic modeling, the study evaluates the predictive utility of surface-form features (unigrams, bigrams) and contextual semantic embeddings (RobBERT) across four social traits: femininity, evilness, trustworthiness, and smartness. To assess both in-domain performance and generalizability, we implement a series of modeling setups using linear and non-linear algorithms, including Elastic Net regression and feedforward neural networks (FFNNs). The following sections describe the dataset, feature extraction process, experimental design, modeling techniques, and evaluation metrics in detail, providing a comprehensive account of how trait perception was computationally modeled

and assessed.

3.1 Data Collection and Preprocessing

This section outlines the structure and preparation of the dataset used to model perceived personality traits in Dutch names. The goal is to provide a clear overview of how the raw data—originally formatted for ranking-based analysis—was transformed into trait scores suitable for predictive modeling. We describe the types of items included, the nature of the trait annotations, and the specific steps taken to standardize, translate, and format the data for downstream analysis.

The dataset, originally collected by Dr. Giovanni Cassani and colleagues, includes 600 word items evenly distributed across three categories: personal names (real Dutch first names), pseudowords (non-lexical but pronounceable strings), and company names (artificially constructed brand-like names). Dutch-speaking participants evaluated these words by performing best–worst comparisons for four traits: femininity, evilness, trustworthiness, and smartness. Rather than providing scalar ratings, participants selected the most and least representative word from a given set, reducing individual bias and increasing sensitivity to fine-grained distinctions.

A Bayesian ranking model was applied to the comparison data for each trait–category pair, producing 10,000 posterior samples that reflect plausible rankings of all items. To convert these rankings into trait scores suitable for regression, we computed the mean rank for each word across all 10,000 draws. This averaging approach is widely used in psycholinguistics to produce stable, fine-grained estimates of subjective ratings while minimizing noise from individual comparisons. The resulting trait scores served as the target variables for all downstream modeling tasks.

To facilitate data exploration and model development, the original Parquet files were converted into CSV format. File and column labels were translated from Dutch to English, and one unified CSV file was created per trait. These four files—each containing the averaged trait scores for all 600 items—served as the primary inputs for feature extraction and evaluation throughout the pipeline.

3.2 Feature Extraction

To model human trait judgments from brand-like names, we extracted three types of linguistic features for each word: character-level unigrams, bigrams, and contextual word embeddings. These features were selected to capture complementary information—ranging from surface-level orthographic structure to high-dimensional semantic cues.

Unigram Frequencies (27 features)

Each word was encoded as a 27-dimensional vector representing the absolute frequency of individual characters. The feature set includes all lowercase letters from a to z, plus the Dutch-specific character ë. This basic orthographic encoding captures character composition and reflects sound-symbolic associations previously linked to perceptions of gender, size, and valence.

Bigram Frequencies (479 features)

To capture sequential structure and phonotactic patterns, we constructed padded bigram representations of each word using a special start and end symbol (e.g., <w, wo, or, rd, d>). Based on a character set of 28 symbols (26 lowercase letters, ë, and a padding symbol), the full bigram space consists of 784 possible combinations. However, only 479 bigrams were observed in the dataset and retained as features. Unobserved bigrams were excluded to reduce feature sparsity and prevent the model from learning noise from rare or unrepresented transitions.

RobBERT Embeddings (768 features)

To model semantic and stylistic properties not evident from surface form, we used contextualized word embeddings from the RobBERT v2 Dutch transformer model (Delobelle, Winters, & Daelemans, 2020). Each word was passed through RobBERT's tokenizer and model, and we computed the mean of all token embeddings from the final hidden layer. This produced a 768-dimensional vector per word, encoding distributional semantic properties learned from large-scale Dutch corpora. These embeddings provide a deeper representation of connotative meaning and stylistic nuance, complementing the character-based features.

Combined Representation (1274 features) and Normalization

For our main experiments, we concatenated the three feature sets—unigrams (27), bigrams (479), and RobBERT embeddings (768)—into a single 1274-dimensional feature vector per word. This hybrid representation enabled the models to jointly learn from low-level orthographic patterns and higher-level semantic information. Only the combined representation was z-score standardized prior to modeling. This ensured that all features contributed on a comparable scale while maintaining their original distributional characteristics when used separately.

3.3 Experimental Design:

This section outlines the experimental configurations used to evaluate whether linguistic features can predict perceived personality traits in Dutch company names. The experiments were designed to test generalization, domain transfer, and model complexity across four controlled setups. All models were trained and evaluated independently for each of the four traits: femininity, evilness, trustworthiness, and smartness. To ensure reproducibility, data splits were created using a fixed random seed (random_state=42).

3.3.1 Overview of Experimental Configurations

To investigate how linguistic features support trait prediction in Dutch company names, we designed four experimental configurations. Each setup isolates a specific aspect of the modeling problem, targeting different forms of generalization and model complexity.

The first configuration served as a baseline for in-domain prediction. Elastic Net regression models were trained and tested exclusively on company names, with 150 items used for training and 50 held out for testing. This setup evaluated whether trait ratings could be learned using linguistic features alone, without any external data from other domains.

The second configuration tested out-of-domain generalization. Again using Elastic Net, the models were trained on 400 items consisting of 200 pseudowords and 200 Dutch personal names—structured yet domain-distinct stimuli. The models were then evaluated on all 200 company names to determine whether representations learned from more regular linguistic items could transfer to stylized commercial names.

In the third configuration, we investigated whether adding a small amount of in-domain data would improve model generalization. Using Elastic Net, the training set included the same 400 pseudowords and personal names as in the previous setup, but with 100 additional company names added. The remaining 100 company names were held out as a test set. This setup enabled us to assess the value of limited exposure to in-domain company names in bridging the domain gap.

The final configuration evaluated whether model complexity influenced predictive performance. A feedforward neural network (FFNN) was trained on the same 400 pseudowords and personal names used in the previous models. From the 200 company names, 100 were used as a validation set for hyperparameter tuning and early stopping, while the remaining 100 served as a held-out test set. This setup allowed us to assess whether a non-linear architecture could better capture trait-relevant patterns or whether it led to overfitting compared to the linear baseline.

3.3.2 Data Splits

The data splits for each experimental configuration are summarized in Table 1. All models were trained and evaluated on items drawn from the same 600-item dataset, which includes 200 personal names, 200 pseudowords, and 200 company names. Each experiment used a distinct partition of this dataset, depending on the specific research objective.

Across all configurations, test sets were drawn exclusively from the 200 company names. This ensured a consistent evaluation protocol focused on

generalization to the commercial name domain. All splits were stratified and kept consistent across traits using a fixed random seed (random_state=42).

Table 1: Overview of data splits for each experimental configuration.

Experiment	Train	Validation	Test	Notes
1	150 company names	—	50 company names	In-domain learning
2	200 pseudowords + 200 personal names	—	200 company names	No in-domain data
3	200 pseudowords + 200 personal names + 100 company names	—	100 company names	Few-shot setup
4	200 pseudowords + 200 personal names	100 company names	100 company names	Used for FFNN only

3.4 Modeling Approach

To evaluate how well linguistic features predict perceived personality traits in Dutch brand-like names, we implemented two types of supervised regression models: Elastic Net and feedforward neural networks (FFNNs). Each model type was trained separately for the four traits—femininity, evilness, trustworthiness, and smartness—using four distinct feature sets: unigrams, bigrams, RobBERT embeddings, and a combined representation. This resulted in 16 models per experiment. Elastic Net was used in Experiments 1–3 to establish linear baselines under different data configurations, while FFNNs were used in Experiment 4 to assess whether non-linear modeling could improve generalization. We maintained this structure to ensure methodological consistency and comparability across feature types and architectures.

3.4.1 Elastic Net Regression

Elastic Net regression was used in Experiments 1, 2, and 3 to test trait predictability under linear constraints. For each experiment, we trained one Elastic Net model per trait–feature pair, yielding 48 models in total. The models were implemented using scikit-learn's ElasticNet class, with hyperparameters tuned via 5-fold cross-validation. To ensure domain balance across folds, we stratified by item type (pseudoword, personal name, company name).

The tuning grid included alpha values [0.01, 0.05, 0.1, 0.2, 1, 5, 10, 100] and l1_ratio values [0.1, 0.3, 0.5, 0.7, 0.9]. A fixed random seed (random_state=42) ensured reproducibility. Unigram and bigram features were used as raw frequency counts. RobBERT embeddings were left unprocessed. The combined feature set, which concatenated all three, was z-score normalized to account for differences in scale and magnitude across feature types.

Elastic Net was selected because its blend of L1 and L2 regularization is

particularly well suited for high-dimensional, correlated input spaces. Its linearity also allows for interpretability, making it a valuable diagnostic tool in addition to being a predictive model.

In Experiment 3, we tested both unweighted and weighted versions of the training set, where company names were upsampled by a factor of four. The weighted models showed superior generalization and are reported in the main results. The unweighted variant is included in Appendix Figure a.1 for comparison.

3.4.2 Feedforward Neural Networks (FFNNs)

In Experiment 4, we evaluated whether non-linear architectures could better capture trait-relevant interactions than linear models. We trained one FFNN for each combination of trait and feature set (16 total), using scikit-learn's MLPRegressor. This configuration matched the feature setups of the Elastic Net models for direct comparison.

Each model was trained on 400 items (200 pseudowords + 200 personal names), validated on 100 company names, and tested on a separate 100-company-name set. All input features were standardized using scikit-learn's StandardScaler within a pipeline, except the combined feature set, which had been pre-normalized before training.

FFNNs used the ReLU activation function and employed early stopping based on validation loss to mitigate overfitting. We tested multiple architectures—single-layer networks with 20 to 200 units and a two-layer (20, 20) setup—and varied the L2 regularization parameter (alpha) across six values: [0.1, 1, 10, 100, 1000, 10000]. The best configuration per model was selected based on validation R^2 and then evaluated on the held-out test set.

While FFNNs offer greater flexibility than linear models, their performance under few-shot learning conditions remained inconsistent, as discussed in the results.

3.4.3 Evaluation Metric

We adopted the coefficient of determination (R^2) as our primary evaluation metric due to its interpretability, scale independence, and suitability for comparing model performance across traits and feature sets. In our regression tasks involving subjective human ratings, even modest R^2 values can indicate meaningful predictive structure. A score of 0.0 corresponds to a model that simply predicts the mean trait score for all items, while positive values reflect explained variance beyond that naive baseline.

R^2 was computed separately for the training, validation, and test sets in all experiments. It served as the key metric for performance reporting and for selecting the best models during hyperparameter tuning.

Although Mean Squared Error (MSE) was also computed during development for diagnostic purposes, we do not report it in the main results. R^2 was prioritized because it offers a clearer benchmark for evaluating model success relative to a baseline prediction.

3.5 Software and Computational Environment

All experiments were conducted using Python 3.13.2 in a venv-managed virtual environment on a Windows 10 machine. Code development was carried out in Visual Studio Code, and Git was used for version control to track changes, manage model iterations, and ensure reproducibility. Exploratory analysis, preprocessing, and evaluation were performed primarily in Jupyter Notebooks.

ElasticNet and FFNN models were implemented with scikit-learn (v1.4.2), with joblib (v1.4.2) used for model persistence and parallelization. Neural models and contextual embeddings were handled using PyTorch (v2.7.0), along with Hugging Face's transformers (v4.51.3) and tokenizers (v0.21.1). For semantic features, we used RobBERT (Delobelle et al., 2020), a Dutch BERT-based model. Data processing relied on pandas (v2.2.3) and numpy (v2.2.4), and all visualizations were created with matplotlib (v3.10.1).

4 Results

This chapter presents the results of four experiments, each designed to answer one of the research questions introduced in the Introduction section. These experiments tested the extent to which linguistic features can predict perceived traits in Dutch company names under varying training and evaluation setups.

Each experiment consisted of 16 models—one for each combination of the four traits (femininity, evilness, trustworthiness, smartness) and four feature sets (unigrams, bigrams, RobBERT embeddings, and a combined representation).

Because test set sizes ranged from 50 to 200 and feature dimensionality varied widely (e.g., 768 for RobBERT), we interpret test R^2 scores above 0.10 as potentially meaningful indicators of generalizable structure. Scores below this threshold are approached more cautiously, as they may reflect overfitting or noise in low-data settings.

Each subsection includes both quantitative outcomes and brief interpretations, with a focus on how different modeling configurations affect generalizability, overfitting risk, and the linguistic encoding of perceived brand traits.

4.1 In-Domain Learning with ElasticNet (Experiment 1)

We first evaluated whether linguistic features could predict perceived traits when both training and testing were conducted within the company name

domain. For each of the four traits—femininity, evilness, trustworthiness, and smartness—Elastic Net regression models were trained on 150 company names and tested on 50, using four distinct feature sets: unigrams, bigrams, RobBERT embeddings, and their combination.

Unigram models produced consistent results across traits, with test R^2 scores ranging from 0.12 to 0.29. Femininity and smartness were the most successfully predicted traits, suggesting that even simple character-level statistics carry meaningful signal for trait perception in brand names.

Bigram models showed meaningful predictive power only for femininity (R^2 = 0.19). For the other three traits, no bigram features were retained during training—suggesting that, with only 150 training items, the model lacked sufficient evidence to learn stable associations from the bigram space. This likely reflects a combination of data sparsity and inconsistent bigram patterns in short, stylized brand names.

RobBERT embeddings showed strong fit on the training set, but weaker generalization. Test R^2 peaked at 0.13 for evilness, while other traits produced near-zero scores. The high dimensionality of RobBERT (768 features) combined with limited data likely contributed to overfitting and poor out-of-sample performance.

The combined feature set yielded the strongest result overall: an R^2 of 0.31 for femininity. This suggests that combining surface-level and semantic features can enhance predictive performance in some cases. However, improvements were not consistent across traits, and the increased model complexity also introduced overfitting risk.

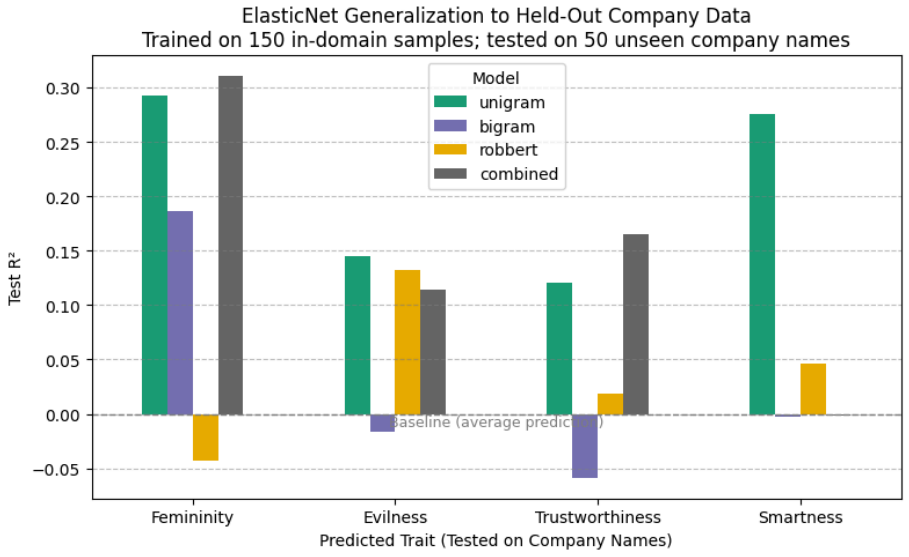

Figure 1

Figure 1 summarizes test R^2 scores for all 16 models (4 traits × 4 feature

sets), enabling direct comparison across experimental conditions. These results demonstrate that linguistic features—particularly unigrams—encode interpretable patterns linked to social trait perception in brand names. In the next experiment (Experiment 2), we test whether these patterns generalize beyond the company name domain.

4.2 Cross-Domain Generalization with ElasticNet (Experiment 2)

Experiment 2 evaluated whether models trained on domain-distinct but linguistically structured data—personal names and pseudowords—could generalize to company names. The models were trained on 400 items (200 personal names + 200 pseudowords) and tested on all 200 company names. This setup probes the extent to which linguistic patterns learned from name-like stimuli transfer to stylized brand names.

Overall, performance declined relative to the in-domain setup in Experiment 1. Most test R^2 scores fell below 0.10, particularly for high-dimensional feature sets, indicating weak generalization and model instability under domain shift.

Unigram models produced low but consistently positive test R^2 scores (0.02 to 0.10), with slightly better results for femininity and evilness. These results suggest that simple character-level patterns learned from natural names offer limited but non-random generalization to brand names.

Bigram models performed best for femininity ($R^2 = 0.28$), indicating that some bigram patterns generalized across domains in this trait. For the remaining traits, however, test performance was near zero or negative, and only a small number of bigrams were retained—suggesting that learned patterns may have been sparse or unstable given the input structure and domain gap.

RobBERT embeddings showed weak generalization across all traits, with test R^2 values generally below 0.05. The high dimensionality of the embeddings and the mismatch between pretrained semantic representations and stylized brand names likely contributed to these results.

The combined feature set yielded the highest overall test R^2 (0.30 for femininity), but offered limited improvement for the other traits. For femininity, only 31 out of 1274 features were retained during training. This suggests that the model identified a focused subset of informative features rather than overfitting to the full input space. The sparse solution may have helped the model generalize more effectively, despite the increased dimensionality of the combined representation. For the remaining traits, test performance remained low, indicating that combining features does not automatically lead to better generalization across the board.

Figure 2 summarizes test R^2 performance across all 16 models. While cross-domain generalization was limited, the relatively better performance for femininity (especially with bigrams and combined features) suggests some traits

may be more learnable from domain-adjacent data. Experiment 3 builds on this by introducing limited in-domain examples to test whether even small amounts of company name data improve generalization.

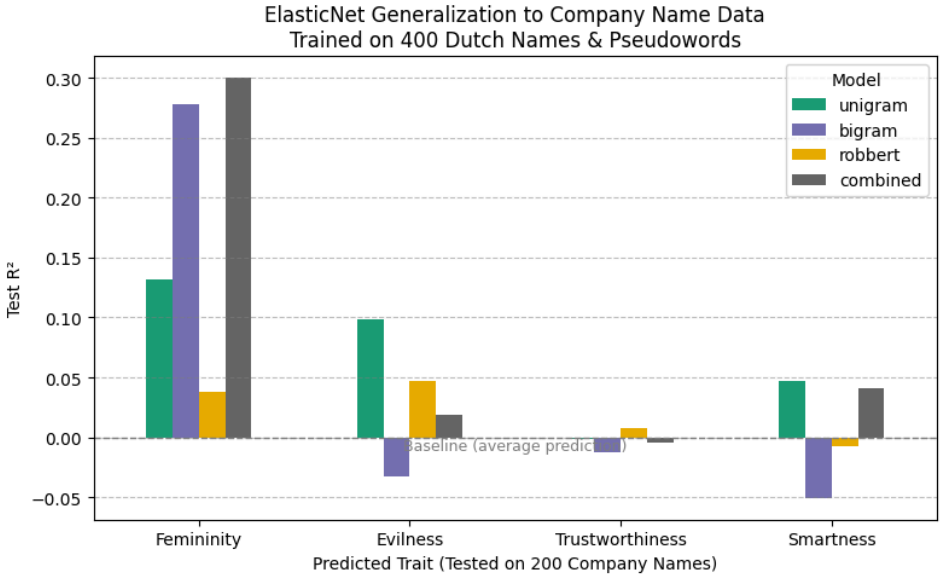

Figure 2

4.3 Few-Shot Generalization with Elastic Net (Experiment 3)

Experiment 3 evaluated whether adding a limited number of in-domain examples improves generalization to company names. Each model was trained on 400 items (200 personal names and 200 pseudowords) plus 100 company names, with the remaining 100 company names held out for testing. To examine the effect of domain weighting, we trained two versions of each model: one with uniform sampling, and one in which the 100 company names were upsampled by a factor of four to match the size of the external data sources.

The results below reflect the weighted models, which consistently outperformed their unweighted counterparts (see Appendix Figure a.1). This confirms that increasing the influence of in-domain examples can improve generalization.

Unigram models showed clear gains over previous setups, with test R^2 scores between 0.19 and 0.26 across traits. These results indicate that even simple character-level features benefit from targeted in-domain exposure.

Bigram models showed weaker generalization than in Experiment 2. For femininity, performance declined from 0.28 to 0.22, despite the model retaining over 100 features. This suggests that while in-domain data led the model to rely on a broader set of bigram patterns, the added complexity may have reduced generalization. For evilness (0.01) and trustworthiness (–0.02), performance remained low, reinforcing the risk of overfitting in high-dimensional spaces with limited signal.

RobBERT embeddings achieved high training performance but generalized inconsistently. Test R^2 scores ranged from –0.01 to 0.22, with strongest performance for evilness, but near-zero or negative values for smartness and trustworthiness. This again reflects the challenges of applying high-dimensional semantic features in few-shot scenarios.

The combined feature set reached a test R^2 of 0.27 for femininity—slightly lower than in Experiment 2 (0.30), but with a much larger number of selected features (149 vs. 31). This suggests that the model became more confident in using a broader set of features, though the added complexity may have reduced generalization slightly. For evilness, performance improved modestly (R^2 = 0.16), while trustworthiness (–0.03) and smartness (0.08) remained weak. These results suggest that additional features and training data did not consistently improve generalization across traits.

Figure 3 summarizes test R^2 scores for all 16 weighted models. The results highlight that even small amounts of in-domain data can enhance generalization, especially when appropriately weighted. Nonetheless, improvement varies by trait and feature type, pointing to the limitations of linear models in capturing non-linear patterns in few-shot conditions—an issue we explore further with feedforward neural networks in Experiment 4.

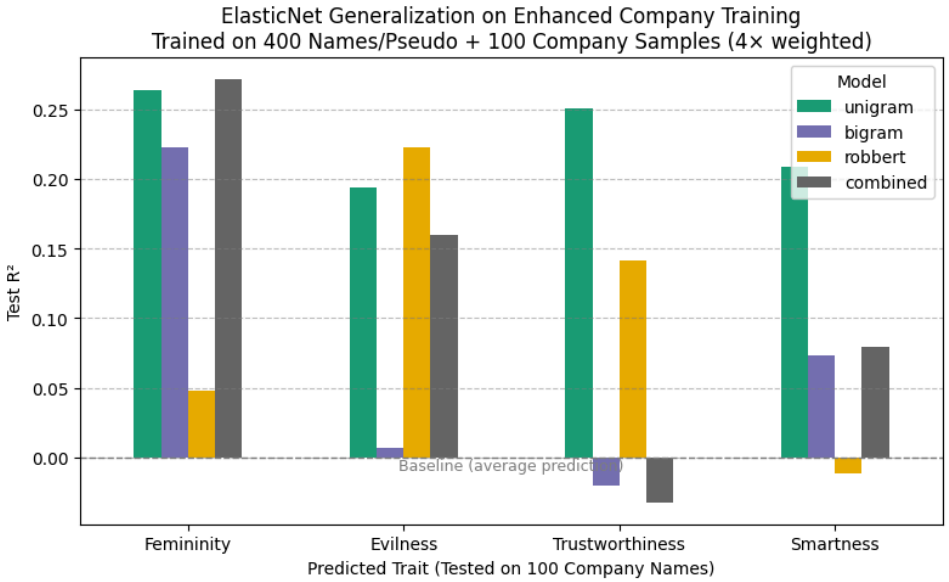

Figure 3

4.4 Few-Shot Generalization with Feedforward Neural Networks (Experiment 4)

Experiment 4 evaluated performance using feedforward neural networks (FFNNs) trained on 400 pseudowords and personal names, with 100 company names used for validation and another 100 held out for testing. Models were trained separately for each trait and feature set, and selected based on validation R^2. While FFNNs offer greater modeling flexibility than Elastic Net, they did not

outperform the linear baseline in any consistent way. Several models achieved strong training scores, but test R^2 values were often low, near zero, or even negative—indicating poor generalization and increased susceptibility to overfitting. Compared to Model 3, performance generally declined across traits, confirming that model expressiveness alone was not sufficient to overcome the constraints of limited data and high input dimensionality.

Unigram models achieved test R^2 scores ranging from 0.11 (femininity) to –0.15 (smartness). In all cases, performance was lower than in Elastic Net models, despite the simplicity of the input features. The notably negative score for smartness suggests that FFNNs may be prone to instability under few-shot conditions—even when using low-dimensional input.

Bigram models again performed best for femininity (0.23) and showed small gains for smartness (0.02). For evilness (0.02) and trustworthiness (~0.00), performance remained low but above zero.

RobBERT-based models did not generalize well. Test R^2 scores remained weak across all traits (≤ 0.08), despite reasonable training performance. As in earlier models, the semantic richness of RobBERT embeddings did not translate into reliable predictions—likely due to the limited data and the stylized, out-of-distribution nature of many brand names.

Combined models performed worst overall. Test R^2 scores remained low across all traits, peaking at just 0.03 for femininity and dropping as low as –0.30 for evilness. These results show that increasing input complexity did not improve generalization and may have amplified overfitting under limited data conditions.

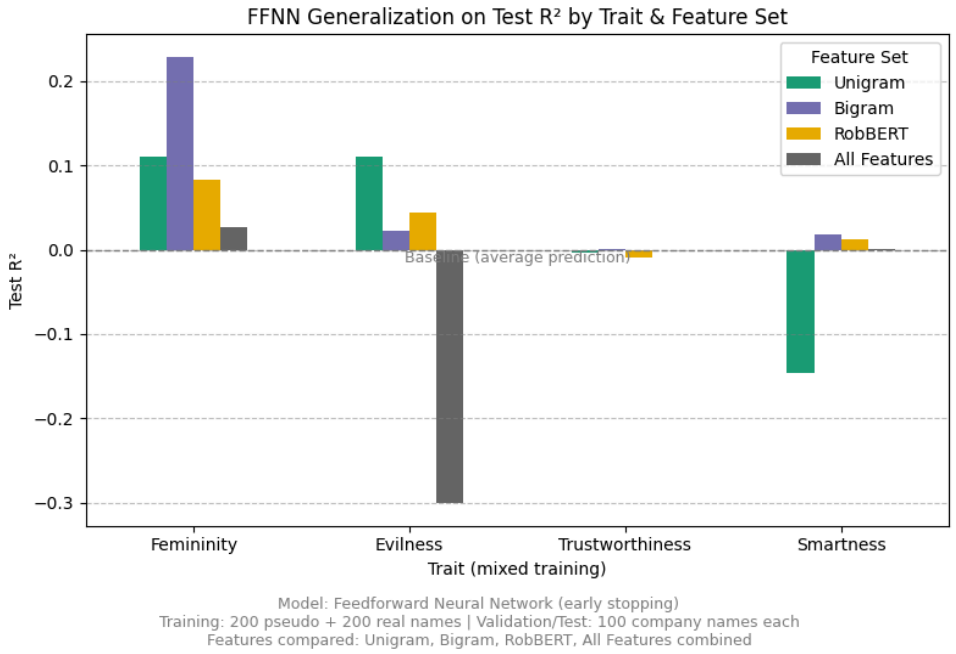

Figure 1

Figure 4 summarizes test R^2 scores for all 16 FFNN models. Overall, Experiment 4 shows that non-linear architectures offer no consistent advantage under few-shot conditions and may be more susceptible to overfitting than simpler baselines. These results suggest that generalization depends more on data coverage than on model complexity.

5 Discussion

This thesis examined how linguistic and semantic features shape perceptions of Dutch company names across four traits—femininity, evilness, trustworthiness, and smartness. By investigating both surface-level linguistic cues (unigrams, bigrams) and high-dimensional semantic embeddings (RobBERT), the study specifically targeted an under-explored intersection: stylized, invented Dutch brand names. The findings shed light not only on predictive accuracy but also on theoretical and methodological considerations critical for interpreting name perception.

These findings directly address the four research questions. For RQ1, models trained and tested within the company name domain showed that unigram features were the most consistent predictors across traits, confirming the value of surface-form cues. RQ2 revealed limited cross-domain generalization overall, though femininity showed notably higher transferability. RQ3 demonstrated that adding even a small number of in-domain examples improved generalization substantially, especially for linear models. RQ4 found that non-linear models such as feedforward neural networks did not offer reliable improvements and were more prone to overfitting in low-data conditions. Together, these outcomes show that simple, interpretable linguistic features—particularly unigrams—remain robust predictors, even in stylized and artificial naming contexts.

A core finding was the robust predictive performance of simple character-level unigram features. This strongly supports prior research indicating that basic phonological cues such as vowel quality and letter frequency carry intrinsic symbolic meanings linked to personality traits (Klink, 2000; Sidhu & Pexman, 2015). The stability of unigram predictors, even in cross-domain generalization experiments, suggests that participants relied heavily on fundamental phonological patterns when forming trait judgments.

Notably, femininity emerged as particularly robust across domains, suggesting that it is encoded in phonological patterns that generalize well, even from pseudowords and personal names to invented company names. This aligns with the idea that certain front vowels and soft phoneme clusters are consistently associated with feminine traits across naming domains. Feature importance coefficients from the cross-domain Elastic Net model reinforce this interpretation. Among the top positively weighted features were the vowel ‘e’ (+0.145) and ‘a’ (+0.110)—both frequently linked to softness and familiarity—as well as terminal vowel bigrams like ‘a>’ (+0.266) and ‘e>’ (+0.168), which

emphasize open, vowel-ending structures typical of names perceived as feminine. Supporting this pattern, the bigram ‘li’ (+0.066), associated with syllables in diminutive or affectionate forms, also had a positive contribution. In contrast, negatively weighted features included sharp or voiceless consonants such as ‘b’ (–0.135), ‘k’ (–0.127), and ‘t’ (–0.094), which are commonly linked with strength or harshness rather than softness. These findings not only corroborate prior sound-symbolism literature (e.g., Klink & Wu, 2013) but also demonstrate that interpretable phonological features extracted from artificial or non-lexical names can reliably signal femininity, even without meaningful semantic content.

Contrary to the expectations outlined in the introduction, RobBERT’s semantic embeddings performed poorly overall, showing only limited success in specific cases such as evilness under few-shot learning. This raises questions about the suitability of pretrained language models for tasks involving stylized or invented names. RobBERT was trained on natural Dutch language data (Delobelle, Winters, & Daelemans, 2020), which means it learns semantic relationships based on how words appear in context. However, the company names in our dataset are often artificial, shortened, or creatively spelled—making them poorly represented in RobBERT’s training distribution. Although RobBERT embeddings have been shown to reflect semantic transparency and compositional meaning in Dutch compounds (Buijtelaar & Pezzelle, 2023; Cassani et al., 2023), these capabilities do not transfer well to brand perception, where judgments are influenced more by sound, structure, and style than by meaning in context. This highlights a broader limitation: contextual embeddings are not well suited to domains where form-driven associations—such as phonological symbolism—play a central role.

The primary novel contribution of this thesis lies in its targeted exploration of cross-domain linguistic generalization specifically within Dutch company naming practices—an area previously underrepresented in sound-symbolism research. The systematic evaluation of linguistic feature sets across multiple experimental setups (in-domain, cross-domain, few-shot scenarios, and linear vs. non-linear modeling) enhances methodological rigor and provides clear guidance for future researchers on feature selection and model complexity trade-offs. Moreover, publicly releasing the trained models, feature importance rankings, and reproducible codebase significantly extends the practical value and methodological transparency of this work, facilitating future comparative research and domain-specific fine-tuning.

These findings carry clear implications for psycholinguistic theory, particularly reinforcing the significance of phonological symbolism in social trait perception. Practically, they suggest simple, interpretable models based on character-level cues could serve as valuable tools for brand evaluation, enabling marketers and linguists to predict social perceptions quickly and inexpensively. Conversely, the limited success of RobBERT embeddings emphasizes a critical gap in applying complex semantic modeling to artificial naming domains—a

challenge future research must address if embeddings are to become practically useful in branding.

5.1 Limitations

This study faces several limitations that may affect the generalizability and interpretability of its findings. First, the dataset used for model training and evaluation was relatively small, consisting of only 200 in-domain company names. This limited sample size restricts statistical power, increases variability in model performance, and may reduce the robustness of conclusions drawn from test set results.

Second, all trait judgments were collected from Dutch-speaking students at Tilburg University. While this population is linguistically appropriate for the task, it may not represent the broader Dutch-speaking public. Perceptions of brand traits such as femininity or trustworthiness can be influenced by age, education, cultural background, or familiarity with branding, and the homogeneity of the participant pool may introduce bias that limits external validity.

Third, the company names in the dataset were provided in a simplified form using only lowercase Latin letters, with no inclusion of numerals, special characters, or capitalization. As a result, typographic and stylistic elements commonly found in real-world brand names (e.g., “Bol.com”, “Q-Park”) were not represented. This lack of visual and structural variation may reduce the ecological validity of the models, as these cues are often central to how people interpret and differentiate brand names.

Lastly, although Elastic Net regression offers built-in interpretability through feature coefficients, this study did not fully analyze which specific features were most predictive across traits. While the feature importance scores were saved for all models and traits, a detailed analysis was omitted due to space constraints. As a result, the current discussion provides only limited insight into the specific linguistic patterns driving model predictions, which constrains the depth of theoretical interpretation. These outputs remain available for future research and analysis.

5.2 Future Work

This thesis provides a strong foundation for several future research directions. One immediate opportunity lies in the detailed analysis of the stored Elastic Net coefficients. Since these weights are already available for all models and traits, future studies could use them to identify which specific letters, bigrams, or embedding dimensions drive perception of traits like femininity or trustworthiness. This would move the focus beyond predictive performance toward a more interpretable, theory-driven understanding of how linguistic form

shapes social impressions.

A second avenue involves testing the generalizability of these findings across languages and naming contexts. Given that femininity showed strong cross-domain transfer in this study, it could serve as a useful test case for examining phonological symbolism in other cultural or linguistic settings. Applying the modeling pipeline to domains such as political party names, product brands, or online usernames could reveal whether the observed patterns reflect broader cognitive or cultural regularities.

Finally, expanding the dataset is essential for improving both ecological validity and model performance. The current models were limited by the lack of typographic and stylistic diversity in the input—real brand names often include numerals, symbols, uppercase letters, or non-Latin scripts. A larger, more representative dataset would enable the use of higher-capacity models, such as fine-tuned transformer architectures, and support more realistic assessments of how people interpret commercial names in the wild.

6 Conclusion

This thesis examined whether four perceived traits of Dutch company names—femininity, evilness, trustworthiness, and smartness—can be inferred from linguistic form. In total, approximately 150 candidate models were trained; of these, 64 (16 in each of four experiments) were evaluated on held-out test sets that ranged from 50 to 200 company names.

Two main conclusions can be drawn. First, low-dimensional character statistics are the most dependable predictors: unigram-based Elastic Net models produced the best results for trustworthiness ($R^2 = 0.25$) and smartness ($R^2 = 0.28$) and showed no signs of over-fitting. Second, richer representations help only in trait-specific circumstances: RobBERT embeddings improved evilness ($R^2 = 0.22$), and a combined surface-plus-embedding model yielded the top femininity score ($R^2 = 0.31$), but neither approach enhanced the other traits. In general, larger feature spaces and feed-forward neural networks did not surpass well-regularised Elastic Net models and often reduced out-of-sample accuracy.

Because all code, trained weights, and feature–importance files are publicly released, future work can scale the corpus, fine-tune transformer embeddings on naming data, and probe the phonosemantic patterns that underlie these trait impressions—steps that are likely to yield more stable, generalisable models for automated brand-name evaluation.

references

Alter, A. L., & Oppenheimer, D. M. (2006). Predicting short-term stock fluctuations by using processing fluency. Proceedings of the National Academy of Sciences, 103(24), 9369–9372.

https://doi.org/10.1073/pnas.0601071103

Buijtelaar, L., & Pezzelle, S. (2023). A psycholinguistic analysis of BERT's representations of compounds. arXiv preprint arXiv:2302.07232.

Cutler, A., McQueen, J., & Robinson, K. (1990). Elizabeth and John: sound patterns of men's and women's names. Journal of Linguistics, 26(02), 471. https://doi.org/10.1017/s0022226700014754

Cassani, G., Bianchi, F., Attanasio, G., Marelli, M., & Guenther, F. (2023, October 11). Meaning Modulations and Stability in Large Language Models: An Analysis of BERT Embeddings for Psycholinguistic Research. https://doi.org/10.31234/osf.io/b45ys

Cassani, G., Joosse, A., & van Kesteren, E.-J. (2023). Klinkt leuk! A tool to predict associations with names and (non)words in the Dutch language. Poster session presented at Computational Linguistics in The Netherlands , Antwerp, Belgium. https://clin33.uantwerpen.be/abstract/klinkt-leuk-a-tool-to-predict-associations-with-names-and-nonwords-in-the-dutch-language/

Delobelle, Pieter & Winters, Thomas & Berendt, Bettina. (2020). RobBERT: a Dutch RoBERTa-based Language Model. 3255-3265. 10.18653/v1/2020.findings-emnlp.292.

Gatti, D., Raveling, L., Petrenco, A., & Günther, F. (2024). Valence without meaning: Investigating form and semantic components in pseudowords valence. Psychonomic Bulletin & Review. https://doi.org/10.3758/s13423-024-02487-3

Joosse, A. Y., Kuscu, G., & Cassani, G. (2024). You sound like an evil young man: A distributional semantic analysis of systematic form-meaning associations for polarity, gender, and age in fictional characters' names. Journal of Experimental Psychology Learning Memory and Cognition. https://doi.org/10.1037/xlm0001345

Klink, R. R. (2000). Creating Brand Names with Meaning: The Use of Sound Symbolism. Marketing Letters, 11(1), 5–20. https://doi.org/10.1023/a:1008184423824

Klink, R. R. (2001). Creating Meaningful new Brand Names: A study of semantics and sound Symbolism. The Journal of Marketing Theory and Practice, 9(2), 27–34. https://doi.org/10.1080/10696679.2001.11501889

Klink, R. R., & Athaide, G. A. (2011). Creating brand personality with brand names. Marketing Letters, 23(1), 109–117. https://doi.org/10.1007/s11002-011-9140-7

Klink, R. R., & Wu, L. (2013). The role of position, type, and combination of sound symbolism imbeds in brand names. Marketing Letters, 25(1), 13–24. https://doi.org/10.1007/s11002-013-9236-3

Sidhu, D. M., & Pexman, P. M. (2015). What's in a name? sound symbolism and gender in first names. PLoS ONE, 10(5), e0126809. https://doi.org/10.1371/journal.pone.0126809

Sidhu, D. M., & Pexman, P. M. (2019). The sound symbolism of names. Current Directions in Psychological Science, 28(4), 398–402. https://doi.org/10.1177/0963721419850134

appendix a

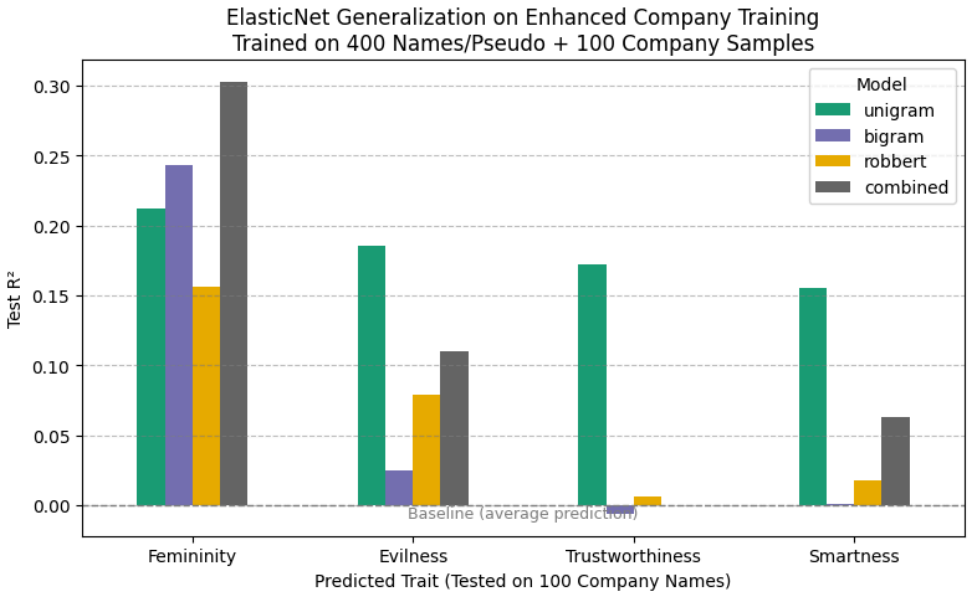

Figure 1