Is *Nike* female? Exploring the role of sound symbolism in predicting brand name gender

Sridhar Moorthy
Rotman School of Management
University of Toronto
moorthy@rotman.utoronto.ca

Ruth Pogacar
Department of Marketing
Haskayne School of Business
University of Calgary
ruth.pogacar@gmail.com

Samin Khan
Computer Science Program
Cognitive Science Program
University of Toronto
samin.khan995@gmail.com

Yang Xu
Department of Computer Science
Cognitive Science Program
University of Toronto
yangxu@cs.toronto.edu

Abstract

Are brand names such as *Nike* female or male? Previous research suggests that the sound of a person's first name is associated with the person's gender, but no research has tried to use this knowledge to assess the gender of brand names. We present a simple computational approach that uses sound symbolism to address this open issue. Consistent with previous research, a model trained on various linguistic features of name endings predicts human gender with high accuracy. Applying this model to a data set of over a thousand commercially-traded brands in 17 product categories, our results reveal an overall bias toward male names, cutting across both male-oriented product categories as well as female-oriented categories. In addition, we find variation within categories, suggesting that firms might be seeking to imbue their brands with differentiating characteristics as part of their competitive strategy.

1 Introduction

When naming humans, clear gender conventions seem to exist in every society. For example, in the English-speaking world, *Jessica*, *Linda*, and *Nancy* are female names, while *John*, *Michael*, and *William* are male names. In turn, decades of gender-stereotyping research suggest that people associate particular genders with particular characteristics. For example, females are viewed as "warm," "expressive," and "emotional," while males are viewed as "assertive," "competent," and "rational" (Broverman et al., 1972; Spence and Helmreich, 1979). We ask whether any of this applies to commercially-traded brands, and is there a gender strategy underlying brand names? If so, what does this strategy say about firms' motivations in making these choices? Is it driven by product category characteristics or is it driven by competitive strategy considerations within each category?

In this paper, we take the first steps toward answering these questions. We start by developing a machine-learning method to predict the gender of human first names. Large labeled data sets of human first names are available from the U.S. and the U.K. to train such an algorithm. Using various linguistic features of such names—for example, "a" ending, sonorant ending (m, n, ng, l, r), etc.—we find that we can predict human gender quite accurately (approximately 80% success rates). We then use this algorithm to predict the masculinity-femininity of brand names (see Figure 1). For this purpose we have identified another data set where a large number of brand names are available, pre-

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1128–1132
Brussels, Belgium, October 31 - November 4, 2018.

classified into many different product categories—detergents, analgesics, health and beauty aids, electronics, etc. Finally, having classified brand names by gender, we ask what the variation in brand name gender, within and across product categories, can tell us about the marketing strategy in brand name choices. By using NLP-methods to analyze brand names, we contribute not only a new application domain for such methods, but also, importantly, advance the branding literature by suggesting that brand name gender may be part of a brand's positioning strategy.

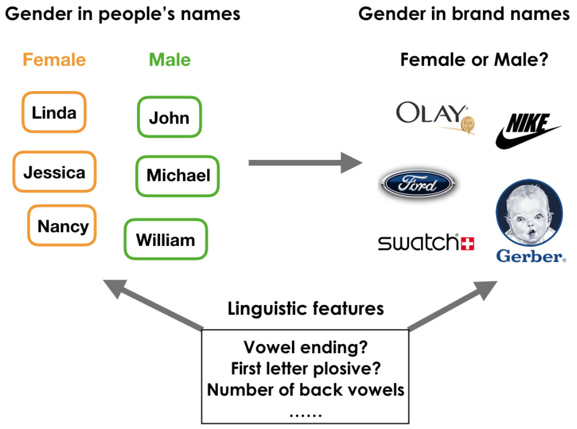

Figure 1: Illustration of our methodology. A gender classifier trained on people's names is applied to brand names.

2 Sound symbolism and gender of names

Existing work from linguistics has suggested that male and female names have different phonological properties. Here we summarize gender-related features of first names identified in the literature: 1) Names ending in the mid-central unstressed vowel sound 'uh' are almost always feminine (Lieberson and Bell, 1992); 2) Names ending in a vowel tend to be feminine (Slater and Feinman, 1985); 3) Names ending in a sonorant (m, n, ng, l, r) may be either masculine or feminine (Barry and Harper, 1995; Lieberson and Bell, 1992; Slater and Feinman, 1985); 4) Names ending in a 'fricative' or 'affricate' produced by restricting airflow through the mouth to create frication (s, z, f, v, h, sh, ch, dj) tend to be masculine (Barry and Harper, 1995); 5) Names ending in a 'plosive' produced by stopping airflow through the mouth (p, b, t, d, k, g) are almost always masculine (Barry and Harper, 1995); 6) Words and names featuring fricatives are more associated with femininity relative to those with plosives, which are associated with masculinity (Folkins and Lenrow, 1966; Guèvremont and Grohmann, 2014; Klink, 2000); 7) The front / back vowel distinction mirrors that of plosives and fricatives. Brand names with front vowels, such as "i," and "ee," seem more feminine than names with back vowels, such as "oh," "oo" (Klink, 2000).

Despite the rich literature on sound symbolism and gender, sparse attempts have been made to systematically examine the contribution of different linguistic features in gender prediction of names at scale (Bird et al., 2009). Although recent work has examined gender issues in web-crawled data (Zhao et al., 2017) and historical corpora (Garg et al., 2018), there has been no study on exploiting linguistic features to predict gender of brand names.

3 Computational methods

Table 1 summarizes the full set of features that we used for modeling. We took orthographic forms of names as a proxy for phonetic forms due to practical scalability. We started by considering names ending in the sound 'uh' as ending with the letter 'a', because 'a' accounts for the vast majority of instances of this sound. After this initial step, we dummy-coded names ending in a vowel with the letters 'a', 'e', 'i', 'o', 'u', and 'y' (i.e., any name ending in a vowel was coded 1 on this feature, whereas all other names were coded 0; see Table 1 for an illustration). We coded names ending in fricatives and plosives with the letters 'p', 'b', 't', 'd', 'k', 'g', 'f', 'v', 's', 'z', 'th', 'sh', 'ch', and 'dge'. We coded sonorant endings based on the letters 'm', 'n', 'ng', 'r', 'and 'l'.

We also considered word-initial occurrences of all the previously mentioned sounds, making symmetric hypotheses that word-initial vowels will predict femininity, word-initial plosives will predict masculinity, and that sonorants and fricatives may be predictive in some way. Coding was based on the same letters with some minor variations (the letters 'w' and 'y' were included with sonorants, and the letter 'j' was included with fricatives).

Finally, we considered the total number of occurrences of all the previously mentioned sounds, hypothesizing that the total number of vowels will predict femininity, total number of plosives will predict masculinity, and total number of sonorants and fricatives may be predictive in some way. Additionally, we consider two distinct categories of vowels: front vowels, represented by the letters 'i', and 'e', which we hypothesize will predict femininity, and back vowels, represented by the letters ‘o’ and ‘u’, which we hypothesize will be more predictive of masculinity. We excluded features such as stressed syllable and number of syllables that have been linked with name gender (Slater and Feinman, 1985) because they require manual coding and do not scale. However, we considered total number of letters as a proxy for name length.

Table 1: Features for gender prediction of people’s and brand names, with example values for *Linda*.

Linguistic feature	Value of *Linda*
A ending?	1
Vowel ending (a, e, i, o, u, y)?	1
Fricative ending (f,v,th,s,z,sh,ch,dge)?	0
Sonorant ending (m,n,ng,l,r)?	0
Plosive ending (p,b,t,d,k,g)?	0
1st letter vowel (a, e, i, o, u)?	0
1st letter fricative (f,v,th,s,z,sh,ch,j)?	0
1st letter sonorant (m,n,l,r,w,y)?	1
1st letter plosive (p,b,t,d,k,g)?	0
Number of letters	5
Number of front vowels	1
Number of back vowels	0
Number of vowels (a, e, i, o, u)	2
Number of fricatives (f,v,th,s,z,sh,ch,j)	0
Number of sonorants (m,n,l,r,w,y)	2
Number of plosives (p,b,t,d,k,g)	0

We used three simple models for name gender classification based on the features we have described. For all methods, we considered binary classification (female/male) and excluded “gender-neutral” names that are ambiguous. 1) To examine how each feature contributes to name gender prediction individually, we first considered a single-feature logistic model that determines gender of a name y from one of the features x: $\log \frac{p(y=female)}{p(y=male)} = \beta_0 + \beta_1 x$. 2) To examine how features combine to name gender prediction, we considered a multivariate version of the logistic model. We applied a sparsity constraint with automatic relevance determination (Yamashita et al., 2008) to explore the minimal set of features necessary for gender determination, taking into account the fact that the features are over-complete and not interdependent, e.g. “a ending” entails “vowel ending.” We used the default settings on the hyper-parameters in the open-source Python package. [1] 3) We considered random forest as an alternative multivariate classifier, based on decision trees using the Gini impurity criterion and bootstrapped subsamples in ensemble averaging. We used the Python $sklearn$ package for this model.

4 Data

We draw data from three primary sources, two for people’s names and one for brand names. For people’s names, we relied on databases of U.S. and U.K. names available at `https://github.com/OpenGenderTracking/globalnamedata/tree/master/assets`. The U.S. data come from the yearly birth records maintained by the U.S Social Security Administration from 1880 to 2013; the U.K. data come from the UK Office of National Statistics, the Northern Ireland Statistics and Research Administration, and the Scotland General Register Office. After removing names that are labeled as both male and female, we ended up with 97102 unique English names (60984 female, 36118 male) to work with. For brand names, we relied on Kantar Media’s Stradegy database. This database documents U.S. advertising spending by brands in virtually every product category. [2] In this case, after removing multi-word names that are derivative brands (e.g., *Ford Escort*) and a small number (66) of names that are common English words (e.g., *Coach*) based on the $\sim$5000 most frequent words in the British National Corpus, [3] we ended up with 1021 brand names in 17 product categories. We represented each name as a 16-dimensional vector based on the features described, and we made all data available in the supplementary materials.

5 Results

Evaluation. We tested all gender classifiers on the English dataset in a supervised setting. Table 2 summarizes the model performances in a five-fold cross validation with the data initially randomized.

Overall, all models predicted gender of people’s names substantially above chance (50%). In particular, the multivariate models performed better than the single-feature model though the difference is small, suggesting that gender information is likely encoded in a restricted set of features. Figure 2 confirms this finding by showing the fitted weights on different features from the sparse logistic classifier. The top four features with the highest weights are “a ending,” “plosive ending,” “sonorant ending,” and “fricative ending,” suggesting the dominance of gender information in the ending of names. The same four features also yielded the highest predictive accuracies in the single-feature model except for the feature of “vowel ending” - 75.1%: a end-

[1] https://github.com/KamitaniLab/sml

[2] https://www.kantar.com/

[3] http://www.kilgarriff.co.uk/BNClists/lemma.al

ing 65%; plosive ending - 66%; sonorant ending - 72%; fricative ending - 66%.

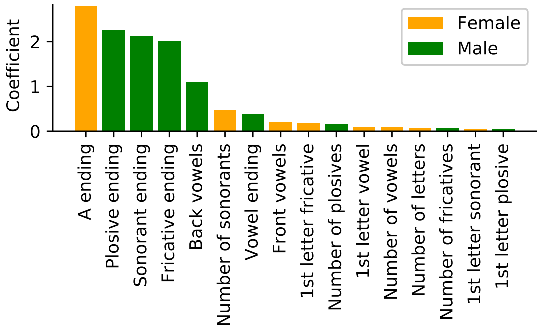

Figure 2: Feature coefficients in logistic regression.

We observed a small difference in predictive accuracy between the random forest model and the sparse logistic model, suggesting the features are relatively robust to classifier variation. We also observed that predictive accuracies on male names tend to be poorer than those on female names, possibly because some male names also end with vowels 27.1%, e.g., *Joshua*), but female names predominantly end with vowels 76.4%).

Table 2: Model accuracies on name gender prediction.

Model	Acc. %	Female %	Male %
Single-feature	75.1	76.4	72.9
Sparse logistic	80.5	83.9	74.9
Random forest	81.8	85.4	75.7

Brand name gender prediction. We applied the three classifiers to the brand name database, by using the weights estimated from the English name database. We used a conservative criterion in determining gender of brand names. In particular, we took majority vote from prediction of the three classifiers for any given brand name as opposed to relying on prediction from a single classifier. [4]

Although it is difficult to fully evaluate the accuracy of our classifiers on brand name gender, we identified "true" gender of a small set of brand names where etymology can be found from the Oxford English Dictionary, summarized in Table 3. We found that our procedure of gender classification yielded an accuracy of 83.3% on this small set, which is consistent with the accuracies we obtained with people's names. We refrained from evaluating our models against human judgments on brand name gender, because people's conceptions might be biased or primed given the products associated with brands (e.g., cosmetic brand names might be perceived as female). Figure 3 visualizes the brand names in the feature space we considered, with predicted gender and annotated example brands.

Table 3: Brand names with gender identified from the OED and model-predicted gender (Male vs. Female).

Brand	Etymology	True	Pred.
Amazon	Female warrior (1398)	F	M
Titan	Helio's father (1413)	M	M
Pandora	Woman (1581)	F	F
Hermes	Zeus' son (1605)	M	M
Nike	Goddess (1846)	F	F
Lincoln	US president (-)	M	M

Brand name gender distribution. We analyzed gender distributions across and within the 17 product categories. Across categories, we observed a strong asymmetry in frequency between male and female brand names. The scatter-plot in Figure 3 illustrates that male and female brand names separate in sound-attribute space. tSNE (Maaten and Hinton, 2008) with 17 feature attributes was used to generate this plot, which shows that male and female names separate along the first dimension, but not in the second dimension. This suggests that these names share commonalities, but they are also different.

We observed that the gross number of male brand names is significantly greater than the number of female brand names (binomial $p < 0.0001$). Several factors could contribute to this bias. For example, many present-day brands originated long before gender equality was valued, and the brand names that emerged from these male-dominated eras tend to skew masculine. It should also be noted that many brands are named after company founders (e.g., *Ford*), and surnames may tend to be more masculine than first names.

A more fine-grained analysis in the individual product categories revealed nuanced patterns in gender distribution (Figure 3). Although the male bias we observed at the broad level applies to 14 of 17 categories (binomial $p < 0.0001$) including some intuitive ones such as "power tools," "military," and "baseball equipment," this bias is surprisingly weak or non-existent in "car / trucks" and "men haircare" where we expect more male as opposed to female consumers. Second, brands where

[4] We also considered a more stringent criterion by analyzing only the subset of brand names where all three classifiers agreed on gender prediction ($N = 702$ out of 1021 names), and our results are similar based on that subset of brand names.

we expect more female consumers such as "jewelry," "cosmetics," and "womens underwear" have a relatively high proportion of male-oriented names. These patterns suggest that although male brand names might be an overall preferred convention, gender-polarized brands might have adopted a reversal strategy in naming with the opposite gender.

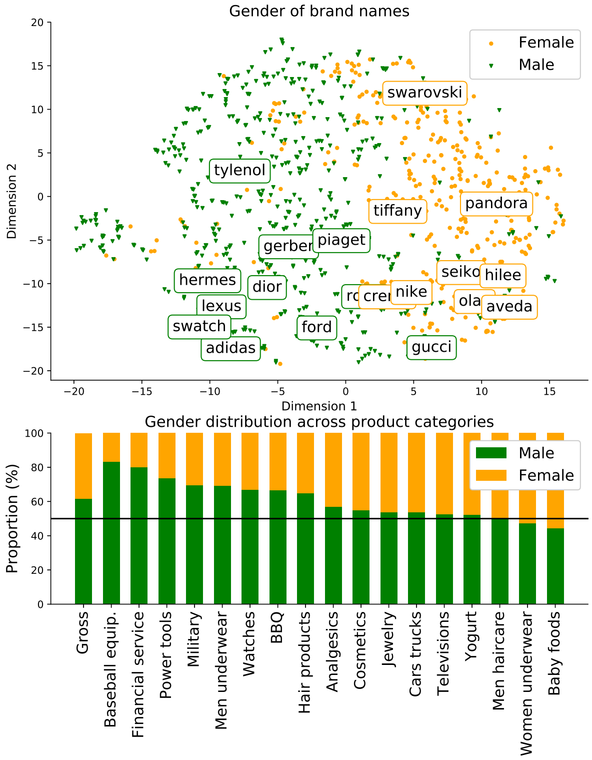

Figure 3: Predicted gender of 1021 brand names given 16 features and mapped a 2D space via tSNE. Bottom panel shows gender proportions in gross and 17 individual product categories.

6 Discussion

While the notion of brand name gender is an intriguing concept in theory, its practical measurement has proved elusive. This paper has shown that NLP-methods can be used fruitfully to get a handle on this problem. Our results suggest that brand names are more male-oriented than female-oriented overall. However, under this broad result, there are several interesting nuances. First, the overall preference for male names applies not only in categories that are primarily male-oriented, but also in some categories that are primarily female-oriented. Second, there is considerable within-category variation in brand name gender. Understanding these nuances is an important topic for future research.

References

Herbert Barry and Aylene S Harper. 1995. Increased choice of female phonetic attributes in first names. *Sex Roles*, 32(11-12):809–819.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media.

Inge K Broverman, Susan Raymond Vogel, Donald M Broverman, Frank E Clarkson, and Paul S Rosenkrantz. 1972. Sex-role stereotypes: A current appraisal. *Journal of Social issues*, 28(2):59–78.

Carlyle Folkins and Peter B Lenrow. 1966. An investigation of the expressive values of graphemes. *The Psychological Record*, 16(2):193–200.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Amélie Guèvremont and Bianca Grohmann. 2014. Can good news be bad? the role of brand communication strategy and brand commitment in the announcement of product improvements. *Journal of Marketing Communications*, 20(5):352–365.

Richard R Klink. 2000. Creating brand names with meaning: The use of sound symbolism. *Marketing Letters*, 11(1):5–20.

Stanley Lieberson and Eleanor O Bell. 1992. Children's first names: An empirical study of social taste. *American Journal of Sociology*, 98(3):511–554.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Anne Saxon Slater and Saul Feinman. 1985. Gender and the phonology of north american first names. *Sex Roles*, 13(7-8):429–440.

Janet T Spence and Robert L Helmreich. 1979. *Masculinity and femininity: Their psychological dimensions, correlates, and antecedents*. University of Texas Press.

Okito Yamashita, Masa-aki Sato, Taku Yoshioka, Frank Tong, and Yukiyasu Kamitani. 2008. Sparse estimation automatically selects voxels relevant for the decoding of fmri activity patterns. *NeuroImage*, 42(4):1414–1429.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.