What BERTs and GPTs know about your brand? Probing contextual language models for affect associations

Vivek Srivastava[*], **Stephen Pilli**[*], **Savita Bhat**[*], **Niranjan Pedanekar**[†], **Shirish Karande**
TCS Research, Pune, India
{srivastava.vivek2, stephen.pilli, savita.bhat, n.pedanekar, shirish.karande}@tcs.com

Abstract

Investigating brand perception is fundamental to marketing strategies. In this regard, brand image, defined by a set of attributes (Aaker, 1997), is recognized as a key element in indicating how a brand is perceived by various stakeholders such as consumers and competitors. Traditional approaches (e.g., surveys) to monitor brand perceptions are time-consuming and inefficient. In the era of digital marketing, both brand managers and consumers engage with a vast amount of digital marketing content. The exponential growth of digital content has propelled the emergence of pre-trained language models such as BERT and GPT as essential tools in solving myriads of challenges with textual data. This paper seeks to investigate the extent of brand perceptions (i.e., brand and image attribute associations) these language models encode. We believe that any kind of bias for a brand and attribute pair may influence customer-centric downstream tasks such as recommender systems, sentiment analysis, and question-answering, e.g., suggesting a specific brand consistently when queried for '*innovative*' products. We use synthetic data and real-life data and report comparison results for five contextual LMs, viz. BERT, RoBERTa, DistilBERT, ALBERT and BART.

1 Introduction

Brands play a vital role in marketing strategies. They are essential to company positioning, marketing campaigns, customer relationships, and profits (Lovett et al., 2014). A brand persona is broadly defined by a set of attributes or dimensions; for instance, '*Mountain Dew*' may be recognized by attributes such as '*adventurous*' and '*rugged*'. While Aaker's dimensions (Aaker, 1997) are widely used to define a brand persona, more fine-grained attributes are documented in Lovett et al. (2014). Furthermore, evaluating a brand persona, i.e., how a brand is perceived by various stakeholders such as consumers, competitors, and market analysts has been an active area of research (Culotta and Cutler, 2016; Davies et al., 2018). Following the widespread success of pre-trained word representations, alternatively called Language Models (LMs), consumer-specific downstream tasks such as recommender systems, dialogues systems, and information retrieval engines look to make use of brand persona along with these representations to better fulfill consumer requirements.

Accordingly, we formulate our first research question (**RQ1**) as *Do LMs store implicit associations between brands and brand image attributes?*. To answer this, we look specifically at brands and brand image defined as affect attributes. Since LMs are trained on real-world data; we believe that these representations may be useful in understanding correlations between a brand and its persona attributes. While numerous studies have investigated unintended biases in Natural Language Processing systems (Dev et al., 2020; Dixon et al., 2018; Bolukbasi et al., 2016; Kiritchenko and Mohammad, 2018; Hutchinson et al., 2020), this is probably the first work that explores brand and affect attributes associations in pre-trained LMs.

These LMs are trained in an unsupervised manner on large-scale corpora. The training corpora generally comprise a variety of textual data such as common web crawl, Wikipedia dump, and book corpora. They are optimized to statistical properties of the training data from which they pick up and amplify real-world trends and associations along with biases such as gender and race (Kurita et al., 2019). Some of these biases may be beneficial for downstream applications (e.g., filtering out mature content for non-adult viewers) while some can be inappropriate (e.g., resume sorting system believing men are more qualified programmers than women (Bolukbasi et al., 2016; Kiritchenko and

[*] equal contribution
[†] corresponding author

Proceedings of Deep Learning Inside Out (DeeLIO):
The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pages 119–128
Online, June 10, 2021.

Mohammad, 2018). Marketing applications such as recommender systems and sentiment analysis can also perpetuate and highlight unfair biases, such as consistently showing popular brands as recommendations and not considering uncommon brands with less positive sentiment. With this in mind, we formulate our second research question (**RQ2**) as *Do the associations embedded in LMs signify any bias?* We also investigate *whether these associations are consistent across all LMs* as **RQ3**.

Brand personas are alternatively characterized as brand archetypes in Bechter's work (Bechter et al., 2016). Brand archetypes are widely used as effective branding and marketing strategy. According to Jung (Jung, 1954), archetypes are defined as inherent images within the collective human unconsciousness having universal meaning across cultures and generations. When successfully used, archetypal branding provides a narrative to connect with consumers. We formulate the following research questions: **RQ4** as *Do LMs capture brand personality intended by a brand?* and **RQ5** as *Do LMs capture brand personality as perceived by consumers?* We propose to use brand-attribute associations to understand brand archetypes perceived by LMs.

In this work, we probe five different LMs (BERT (Devlin et al., 2018), ALBERT (Lan et al., 2019), RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019) and BART (Lewis et al., 2019)) on affect associations by using Masked Language Model (MLM) head. The choice of LMs was guided by three factors: 1) availability of MLM head, 2) variety in model architectures and 3) type and size of training data used while pre-training. Table 1 summarizes all the five LMs based on the pre-training data and the architecture. We believe that diversity in architectures and training data can influence the affective associations stored in representations. We propose to evaluate word representations based on following dimensions: 1) contextual similarity (Ethayarajh, 2019), 2) statistical implicit association tests (Kurita et al., 2019; Ethayarajh et al., 2019), 3) controlled probing tasks (Talmor et al., 2019) and 4) brand archetypes (Bechter et al., 2016). We observe that LMs do encode affective associations between brands and image attributes (**RQ1**). Some of these associations are consistently observed across multiple LMs (**RQ3**) and are shown to be further enhanced by finetuning thus implying certain bias (**RQ2**). We find that brand images or personality captured by LMs do not concur with either intended or consumer perceived brand personality. We believe that appropriate dataset and more rigor is needed to address **RQ4** and **RQ5**.

LM	Pre-training Data	Architecture
BERT	BookCorpus (800M words), English Wikipedia (2,500M words)	L=24, H=1024, A=16, T=340M
RoBERTa	BookCorpus (800M words), CC-NEWS (63M articles), OpenWebText (8M documents), Stories	L=24, H=1024, A=16, T=355M
DistilBERT	BookCorpus (800M words), English Wikipedia (2,500M words)	L=6, H=768, A=12, T=66M
ALBERT	BookCorpus (800M words), English Wikipedia (2,500M words)	L=24, H=1024, A=12, T=66M
BART	BookCorpus (800M words), CC-NEWS (63M articles), OpenWebText (8M documents), Stories	L=12, H=1024, A=16

Table 1: Variants of LMs. L–total layers, H–hidden size, A–self-attention heads, T–total parameters. We mention the architecture of the *large* version of all the LMs.

2 Related Work

The success of pre-trained word embeddings in achieving state-of-the-art results has sparked widespread interest in investigating information captured in these representations. Typically defined as '*probing task*', a wide variety of analyses have been proposed. For instance, (Hewitt and Manning, 2019) proposes a structural probe to test whether syntax trees are embedded in word representation space. Experiments in (Wallace et al., 2019) are aimed to investigate the numerical reasoning capabilities of an LM. Similarly, (Petroni et al., 2019) presents an in-depth analysis of relational knowledge present in pre-trained LMs. Penha and Hauff (2020) probe the contextual LMs (BERT and RoBERTa) for the conversational recommendation of books, movies, and music. Our work seeks to apply the idea of probing to a relatively unexplored area of affect analysis. To the best of our knowledge, this is the first work that presents a multi-pronged investigation of brands and subjective knowledge like affect attributes represented in contextual representation. Field and Tsvetkov (2019) is the most relevant prior work in terms of affect analysis. They present an entity-centric affective analysis with the use of contextual representations, where they find that meaningful affect information is captured in contextualize word representations but these representations are heavily

biased towards their training data.

A significant effort has been seen in investigating the intrinsic bias in word embeddings. These representations are trained in an unsupervised manner using a large amount of training data typically consisting of common web crawls. As a result, all kinds of biases like gender, race, demography along with trends and preferences get encoded in LMs. Works in (Kurita et al., 2019; Dev et al., 2020; Ethayarajh et al., 2019) propose methodologies to measure and mitigate bias in word representations. Our work is targeted at finding trends and preferences that certain entities have by using a combination of old and new such measures.

3 Dataset

In this work, we evaluate affect information captured in the LMs for different brands. Accordingly, the selected brands should have large volumes of online data to get significant representation in the LMs. We choose 697 major US national brands reported in (Lovett et al., 2014). These brands are categorized into 16 different product categories. To analyze affect associations, we refer to surveys conducted by Young and Rubicam (Y&R) (Lovett et al., 2014) to measure a broad array of perceptions and attributes for a large number of brands. We choose 40 affect attributes listed as a part of '*Brand Image*' in (Lovett et al., 2014). We also manually map (see Table 8 in supplementary material and Bechter et al. (2016)) these attributes to one of the five Aaker's dimensions of brand personality. We restrict our analysis only to positive affect attributes since '*Arrogant*' and '*Unapproachable*' were the only two negative affect attributes observed in Y&R surveys. We understand the analysis with negative attributes is essential to explore the complete brand perception and we intend to pursue this in future. We consider three different data sources for our experiments as tabulated in Table 2. We choose appropriate datasets based on experiments' requirements. We describe the datasets in detail in supplementary material.

4 Experimental Setup

We outline our approach for exploring answers to the research questions stated above.

- **RQ1, RQ3**: Understanding brand and attribute word association at different layers of the LMs (see contextual geometry in Section 4.1).
- **RQ1, RQ2, RQ3, RQ4, RQ5**: Analyzing closeness between the brand and attribute words using statistical tests (see implicit association test in Section 4.2).
- **RQ1**: Probing for the association as well as the influence of brand name and the surrounding context on the attribute word (see probing task in Section 4.3).
- **RQ4**: Examining brand perceptions in terms of archetypes and affect attributes (see brand archetype in Section 4.4).

4.1 Contextual Geometry

Taking inspiration from (Ethayarajh, 2019), we use geometrical analysis to understand associations between brands and brand image attributes. Ethayarajh (2019) analyzes geometry of contextual representations across different layers. We follow the same approach to specifically analyze representations for brands and affect attributes. We use two metrics introduced in (Ethayarajh, 2019): *self-similarity* and *intra-sentence similarity*. Additionally, we use a similar methodology to define associations among brand words and affect words. We consider Ads. Dataset data for these experiments.

Let *bw* be a *brand word* and *aw* be an *attribute* or *affect word* appearing in sentences $\{s_1, s_2, ..., s_n\}$ at positions $\{i_1, i_2, .., i_n\}$ and $\{j_1, j_2, .., j_n\}$ respectively. Accordingly, $bw = s_1[i_1] = s_2[i_2] = .. = s_n[i_n]$ and $aw = s_1[j_1] = s_2[j_2] = .. = s_n[j_n]$ with i_k and j_k representing positions in sentence s_k. In other words, a brand word *bw* is the i_1^{th} word in sentence s_1 and attribute word *aw* is the j_1^{th} word in sentence s_1. Let $f_l(s, i)$ be a function that maps $s[i]$ to its representation in layer *l* of language model f (Ethayarajh, 2019). Then,

4.1.1 *affect-similarity*

The *affect-similarity* between *bw* and *aw* in layer *l* is defined as the average cosine similarity between contextualized representations of brand and attribute across *n* unique contexts.

$$AffSim_l(bw, aw) = \frac{1}{n} \sum_k cos(f_l(s_k, i_k), f_l(s_k, j_k))$$

Dataset	Data	Example	Brand	Attribute
Ads. Dataset (Hussain et al., 2017)	35k Action Reason pairs	"I should buy Converse shoes because they are stylish."	Converse	stylish
BCD (Roy et al., 2019)	1962 sentences from webpages containing both brand and affect attributes	"Verizon is a global leader delivering innovative communications solutions."	Verizon	innovative
Synthetic (Table 16 in Supplementary Material)	40 hand crafted sentences	"Apple is a trendy brand."	Apple	trendy

Table 2: Representative examples from three different datasets.

4.1.2 *intra-brand similarity*

The *intra-brand similarity* between a pair of brand words in layer *l* is

$$IntraBrandSim_l(bw_i, bw_j) = \frac{1}{n(n-1)} \sum_k \sum_{p \neq k} cos(f_l(s_k, i_k), f_l(s_p, j_p))$$

In other words, the *intra-brand similarity* provides average cosine similarity between representations of two brands across *n* different contexts. This measure captures how close the two brands are in the vector space.

4.1.3 *intra-attribute similarity*

Similarly, we define the *intra-attribute similarity* between a pair of attributes in layer *l* as the average cosine similarity between two attributes across *n* different contexts. This measure helps us understand the association between different affect words in the vector space and can be used while defining and analyzing brand persona.

4.2 Implicit Association Tests

The Implicit Association Test (IAT) (Greenwald et al., 1998) in its purest form measures association between two target concepts with respect to an attribute. This test has enabled the examination of unconscious thought processes and implicit biases among people in different contexts (Sleek, 2018). We believe that a variety of implicit biases and associations may be encoded in LMs. We use two interpretations of IAT (viz. WEAT and RIPA) to investigate brand and attribute associations in LMs.

The *Word Embedding Association Test* (WEAT) (Caliskan et al., 2017) for non-contextual word embeddings shows implicit biases captured in these representations. May et al. (2019) extend this test to sentence embeddings for contextual LMs. Since our focus is on words; we follow the approach used in (Kurita et al., 2019) to adapt WEAT for words. We also consider the new measure, *log-probability bias score*, introduced in (Kurita et al., 2019). This test follows a similar approach to WEAT except for the cosine similarity computation between target word and attributes is replaced by log-probability.

The work in (Ethayarajh et al., 2019) proves that any embedding model that implicitly does matrix factorization, subspace projection under certain conditions, can be considered as debiasing the embedding vectors. Accordingly, they propose a new method of the association called *relational inner product association* (RIPA) that uses the subspace projection method. We adapt RIPA measure for brands and attribute words.

Both *log-probability* and *RIPA* have been proposed as an alternative to the basic WEAT association test. We detail the experimental structure for these tests below.

4.2.1 WEAT

The WEAT test simulates the human implicit association test for word embeddings, measuring the association between two equal-sized sets of target concepts and two sets of attributes (May et al., 2019). Specifically, in our case, we consider high-level brand categories as target concept sets and Aaker's dimensions as attribute sets. Specific details about test statistics along with permutation test and effect size can be found in (Caliskan et al., 2017; May et al., 2019; Kurita et al., 2019).

4.2.2 Log-probability score

We consider the same set of broad categories for brands and Aaker's dimensions for attributes as target and attribute sets respectively for finding log-probability score. Similar to (Kurita et al., 2019), we compute the mean log probability bias score for each attribute and permute the attributes to measure statistical significance with the permutation test.

For both *WEAT* and *log-probability* test, we use synthetic data generated by appropriate handcrafted templates. We apply these tests to all combinations of brand categories and Aaker's dimensions. We apply these tests on combinations of all brand categories except '*Food and Dining*' and 5 Aaker's

LM	Brand pair		Attribute pair		Brand-attribute pair	
	MS	LS	MS	LS	MS	LS
BERT	Chrysler-Jeep	ESPN-Wilson	Safe-Secure	Innovative-Reasonable	Disney-Magical	Toyota-Reasonable
RoBERTa	Dodge-Jeep	BBC-Sonic	Bright-Vibrant	Tough-Responsible	Disney-Magical	Target-Kind
DistilBERT	Chrysler-Volkswagen	Fox-Honda	Nice-Wonderful	Fun-Robust	IBM-Innovate	Microsoft-Popular
ALBERT	Honda-Toyota	Sprint-IBM	Lovely-Charming	Funny-Bright	Volkswagen-Excellent	Samsung-Best
BART	Dodge-Lincoln	Intel-Nokia	Strong-Efficient	Friendly-Lovely	Jeep-Simple	Intel-Efficient

Table 3: Affect associations across different LMs for least similar (LS) and most similar (MS) brands and attributes.

affect dimensions. We use the pairwise ranking to rank these combinations.

4.2.3 RIPA

For our affect analysis formulation, we define RIPA as the projection of the affect word vector i.e. attribute onto the bias subspace defined by a pair of brands. We use handcrafted templates to generate sentences corresponding to 40 attributes combined with brand words. Thus, we get 40 representations for every brand and 697 representations for every attribute. Final brand and attribute vectors are computed by taking an average of corresponding vector sets. RIPA score between each attribute word and a pair of brand words is then calculated by taking the inner product of the first principal component of the subspace defined by the pair of brand words and attribute word. For a brand pair (x,y) and an attribute word w, a positive RIPA score suggests the relatively more association of w with the brand x and vice-versa.

4.3 Probing Tasks

A large body of research comprising of probing tasks is dedicated to exploring what is captured by contextual LMs. We define two probing tasks that are essentially cloze tasks to analyze brand and affect attributes associations. In the simplest form, we consider MLM setup: given a sentence with brand and masked attribute word, we use pre-trained LM with MLM head to predict words at the masked position. If a model predicts the correct attribute in the top-5 position, then we infer that the model representations have captured the corresponding affect association. Additionally, to understand the behavior after fine-tuning, we introduce MLP with a 1-hidden layer to the MLM setup to train the LMs as discussed in (Talmor et al., 2019); we call this setup MLP-MLM.

To further analyze sensitivity to context, we define perturbed language control, where we introduce nonsensical words into the sentences. We observe if there is any effect of nonsense words to affect associations. MLM setup is used to experiment on all LMs using Ads. Dataset and BCD datasets, whereas MLP-MLM uses only Ads. Dataset and is experimented on all the LMs except BART.

4.4 Brand Archetypes

Brand archetypes provide a relatable connection between brands and consumers. We consider implicit and explicit perceptions of archetypes. We use Lovett's data (Lovett et al., 2014) to understand people's tacit perceptions about brand archetypes in terms of affect attributes. We believe that training data used for pre-training LMs may record impressions about the brand in the wild. Accordingly, we consider pre-trained LMs to investigate the explicit perceptions for archetypes. We consider 12 archetypes (Jung, 1954) for this analysis. We manually map every archetype to a set of affect attributes from Lovett's attributes (Lovett et al., 2014) with the help from (Bechter et al., 2016) (see Table 8 and 10 in Supplementary Material).

To understand the brand archetype information captured in the LMs, we take the intersection of the top attributes obtained using the brand-attribute affect similarity and the attributes for a given archetype (obtained after manual mapping). First, we identify the top-5 attributes for a given brand using the affect similarity score and then we take the percentage overlap with the list of attributes corresponding to each of the archetypes. The percentage overlap suggests the degree of brand archetype-related knowledge instilled in the LMs. To better evaluate our results qualitatively we choose five brands (*Adidas*, *Apple*, *GAP*, *Pepsi*, and *Porsche*) from different brand categories.

5 Discussion

We present a battery of analyses aimed at finding how much knowledge do the off-the-shelf LMs capture about brands and affect attributes.

5.1 Affect Association

We believe that *brand persona* can be succinctly defined by a set of affect words, namely attributes.

We make use of *intra-attribute similarity* to understand which of the attributes are closer to each other in embedding space. Using *intra-brand similarity*, we also examine how the brands of a category are positioned in the vector space. Additionally, the *affect similarity* helps us find the correlation between brand and affect words. We argue that a brand persona can be identified by combining results from these three measures. It should be noted that some of these associations of brands and attributes are indeed consistent across all LMs (**RQ1, RQ3**). Table 3 reports some of the most similar and least similar associations. By far, brands of category '*Cars*' are seen to have high similarity among themselves consistently across all LMs. In some instances, brands of categories '*Technology*' and '*Telecommunication*' are found to have a close association. Similarly, *cliques* of attributes are observed such as *elegant*, *lovely*, *fashionable*, *popular* in BERT and *reliable*, *efficient*, *helpful*, *convenient* in DistilBERT. These clusters of attributes can further be beneficial in defining a brand persona. Using the *affect-similarity*, we found interesting associations between brands and attributes. For instance, brand '*Disney*' is associated most with attributes , '*magical*' and '*fun*' across all LMs whereas brand '*IBM*' is highly associated with '*innovative*' and '*intelligent*'. These positive associations help understand the brand persona. We also observe the least similar relations across all LMs. There are some surprising results, such as brands '*Intel*' and '*Samsung*' not being '*efficient*' and '*Best*' respectively. Such associations may not be what brand marketing teams would want to portray for their brands. We believe that these negative associations are also important in identifying the perception of a brand.

5.2 Contextual Representation

The *self-similarity* metric provides a measure to evaluate the contextualization of a word. Following (Ethayarajh, 2019), lower self-similarity is observed when the representations are more contextualized. We compare the average self-similarity of a representative brand and attribute words for each layer of selected LMs. For all five models, self-similarity is lower in upper layers or final layers i.e. the word representations are more context-specific. Out of five LMs, RoBERTa representations have the lowest self-similarity. Furthermore, it should be noted that different words have different levels of context specificity in different LMs.

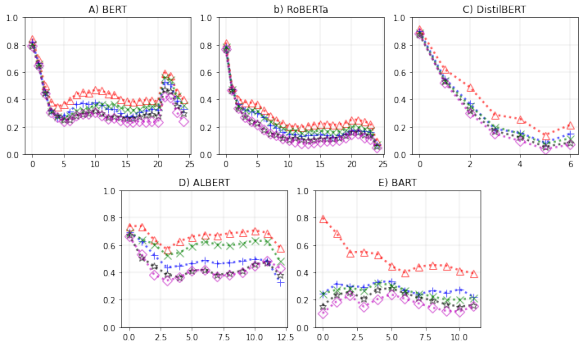

Figure 1: *Self-Similarity* for brand and attribute words '*Google*' (+), '*Gymboree*' (△), '*good*' (∗) , '*exceptional*' (x) and '*bad*' (◇).

Ethayarajh (2019) observes that the variety of context is important for having variations in representation and common words or popular words like '*the*', '*of*' and '*to*' generally have larger variation in their representations. We believe that popular brands have the diverse contexts in the training data used for pre-training the LMs and hence are more contextualized. As can be seen in Figure 1, representations for *Google* are more context-specific as compared to those for *Gymboree*. Affect words '*good*','*bad*' and '*exceptional* also have different context specificity implying a certain kind of inequality in the encoded knowledge corresponding to different words. This pattern is observed across all LMs implying that variation in representations is consistent irrespective of the amount of training data used while pre-training.

5.3 Implicit Association Tests

In WEAT as well as in Log Probability, the null hypothesis is that there is no significant difference between the two sets of brand categories in terms of their relative similarity to the two sets of Aaker's dimensions. The polarity of the effect size indi-

LM	Brand Category	Aaker's Dimensions	WEAT	LOG PROB
BERT	Sports/Health	sincerity/ ruggedness	-0.5244	-1.1856
RoBERTa	Media/Finance	excitement/ sincerity	0.63615	0.6602
DistilBERT	Childrens/ Dept. Stores	competence/ excitement	-0.9681	-1.1161
ALBERT	Tech./Beauty	sophistication/ competence	0.3396	-0.6067

Table 4: Effect-size of *WEAT* and *Log Probability* (at p-value < 0.01)

.

LM	Media & entertainment		Technology product & stores		Cars	
	Fun	Original	Original	Reasonable	Traditional	Worthy
BERT	Disney	HBO	Sony	Microsoft	Volvo	Volvo
RoBERTa	YouTube	CNBC	IBM	Apple	GM	GM
DistilBERT	YouTube	MTV	Apple	Samsung	GM	Jaguar
ALBERT	YouTube	MTV	Pioneer	Sharp	Buick	Buick

Table 5: Top brand and attribute associations for three different brand categories using RIPA association test.

cates that the categories and dimensions are directly or inversely related. For example, consider, the *Sports*/*Health* in brand category and *sincerity*/*ruggedness* in Aaker's Dimensions from Table 4 the polarity of effect size indicates that they are inversely related, which means *'Sports'* is more associated with *'ruggedness'* similarly *'Health'* is to the *'sincerity'* (**RQ2**). Since we are considering the permutation test, the p-value indicates the significance of their association. Most of these associations are consistently observed across all LMs (**RQ1, RQ3**). This has intrigued us to further examine which LM is better at capturing brand personality as perceived by consumers. The pairwise ranking is applied to all the combinations of brand categories and Aaker's dimensions (Aaker, 1997). The resultant ranked dimensions of all the categories are assessed against the ground truth values/ consumers perception (please refer Table 9 in Supplementary Material) in Lovett's data (Lovett et al., 2014). Using the same procedure, all the LMs are ranked independently for each brand category (refer to Table 15 in Supplementary Material). We observe that BERT has better agreement with consumers' perceptions of brand personality amongst all the language models in both WEAT and Log Probability (**RQ5**). Though RoBERTa did follow, other LMs agree equally likely in Log Probability. Furthermore, DistilBERT has a consistently poor agreement in Log Probability. One interesting observation is that WEAT and Log Probability give the same ranking for all LMs in the '*Cars*' brand category.

RIPA test measures the word embedding association using the subspace projection method (Ethayarajh et al., 2019). A positive score suggests that brand x is more associated with attribute word w than brand y for a given brand pair (x,y) and attribute word w. We combine this score for a brand with all attributes to compute a preference score for a brand. Based on this preference score, we found the most associated brands for every attribute word. Representative results are presented in Table 5. We observe that the predictions across different LMs for a given category are occasionally consistent (e.g., YouTube being associated as a fun brand in RoBERTa, DistilBERT, and ALBERT) (**RQ3**). This could be attributed to the perception of brands being captured by the various LMs. Also, we see the diversity in the predictions for different attribute words (e.g., BERT and RoBERTa has different brand association across different categories) which also signifies that the brand associations being captured by the LMs vary with the context (**RQ1**).

5.4 Impact of fine-tuning

Comparing the LMs off-the-shelf gives us an idea of how affect-related attributes are represented in LMs. From Table 6, we find that BART and RoBERTa have the better brand and attribute associations amongst the LMs on the Ads. Dataset and the BCD datasets (**RQ1**). Further, to understand the impact of fine-tuning, we employ techniques proposed by (Talmor et al., 2019) to measure the language mismatch. In this exercise, we fine-tune the LM with examples from Ads. Dataset; high performance indicates that the LM was able to overcome the language mismatch with a very small number of samples. Trends in the Figure 2 conveys that BERT and RoBERTa achieve high performance with a limited number of samples, in turn indicating that their internal representations are well suited for any downstream tasks related to brand personality. On the other hand, ALBERT has the least performance improvement of 8.08%, meaning ALBERT has poor internal representation and needs more samples to overcome the language mismatch. BERT outperforms all LMs with 22.28% improvement followed by RoBERTa with 20.06%.

5.5 Sensitivity to context

To understand the context-dependency of the attributes related to affect, we employ perturbed language control as discussed by (Talmor et al., 2019). This control task gives us an idea of how well the pre-trained representation of the words in context can influence the affect association. For exam-

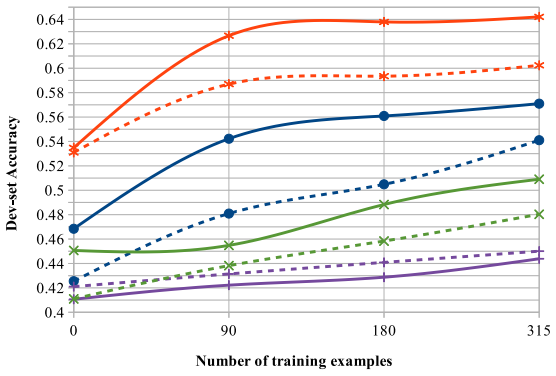

Figure 2: MLP-MLM with (- -) and without perturbation (–) for different LMs- BERT (•), ALBERT (+), DistilBERT (x), RoBERTa (*)

ple, consider the statement *"I should play Nintendo because it is [MASK] ."* and its perturbed version *"I snap play Nintendo ya it is [MASK] ."*. If *'fun'* from the set of attributes is persistently perceived to be in top-5 predictions irrespective of perturbation, we say that context doesn't influence attributes. In either of the setups discussed in ***Controlled probing task***, the drop in accuracy after perturbation indicates that the affect attributes are context-dependent. Our observations on *MLM* setup (Table 6) and *MLP-MLM* setup (Figure 2) indicate that the attributes are moderately influenced by the context. We need more samples to comment on ALBERT.

LM	Zero-shot		Perturbed	
	Ads. Dataset	BCD	Ads. Dataset	BCD
BERT	0.51	0.66	0.47	0.64
RoBERTa	0.58	0.77	0.51	0.77
DistilBERT	0.46	0.55	0.45	0.54
ALBERT	0.42	0.55	0.41	0.53
BART	0.65	0.61	0.59	0.61

Table 6: MLM setup with and without perturbation on the Ads. Dataset and BCD datasets.

LM	Top archetype(s) based on the attribute overlap
BERT	Creator, Jester, Outlaw, Magician, Hero, Sage, Explorer, Innocent
RoBERTa	Creator, Jester, Outlaw, Magician, Hero, Sage, Explorer, Innocent
DistilBERT	Ruler, Everyman, Magician, Sage, Innocent
ALBERT	Creator, Jester, Outlaw, Magician, Hero, Sage, Explorer, Innocent
BART	Ruler, Everyman, Magician, Sage, Innocent

Table 7: Archetype information extracted from the LMs for the brand *Adidas*.

5.6 Archetypes

We investigate implicit perceptions about brands using data collected in a survey (Lovett et al., 2014). Table 7 shows the result of the top archetype(s) extracted from the various LMs for the brand *Adidas*. The actual archetype of *Adidas* is Creator[1]. We make three major observations about the brand archetype extracted from different LMs (**RQ4**). First, we observe the same prediction of the top archetype across various LMs. For instance, we get the same set of top archetype(s) prediction with BERT, RoBERTa, and ALBERT for the brand *Adidas*. This behavior could be attributed to the absence of explicit brand archetype-related information in the LMs. Next, we observe multiple top archetypes with the same degree of attribute overlap which suggests that LMs does not capture the brand archetype information distinctly. Lastly, we observe that the degree of attribute overlap for the top archetypes is consistently very low (i.e., an overlap of only one out of five attributes) for all the five brands across all the five LMs. This low degree of attribute overlap is also suggestive of the absence of archetype-related information in the LMs. The actual archetype of a brand can not be distinguished in any of the LMs. We make similar observations for other brands as well (see Table 11 to 14 in Supplementary Material). The current observation that the LMs do not reflect the expected perception of the brand's archetype needs to be investigated further with archetype-specific datasets.

6 Conclusion

In this paper, we presented a series of exploration setups to address research questions pertaining to associations between brands and brand image attributes.

Our analyses were able to tease out varied responses even from the models having identical training data and pre-training learning objectives. We observed that there exists a definite association between brands and attribute affect words across all LMs (**RQ1**). This impression is observed across a range of abstraction i.e. from individual brands and broader categories to attributes and Aaker's dimensions.

In all our experiments, some categories such

[1] https://report.adidas-group.com/2019/en/group-management-report-our-company/corporate-strategy/adidas-brand-strategy.html

as '*Cars*' and '*Technology product & stores*' and brands such as '*Disney*' and '*Intel*' are found to have consistent associations across all LMs (**RQ3**). However, it is interesting to note that these biases do not concur with both consumer perceptions and intended perceptions of the brand (**RQ4** and **RQ5**).

Lastly, it is seen that perturbations in sentence moderately influences the association between brands and affect words. Improved performance in fine-tuning implies that affect associations are enhanced (**RQ2**). Since we do not have enough data, it remains to be seen how additional training data changes the landscape.

This work documents an initial investigation of brand and attribute associations in different LMs. With enough task-specific data, we plan to evaluate how the affect associations are enhanced. We also intend to use these observations in further defining brand-persona and brand-archetype definitions. These impressions can help understand perceptions about a brand. Furthermore, this can be extended in investigating impressions about iconic entities such as sports teams, celebrities, and politicians.

References

Jennifer L Aaker. 1997. Dimensions of brand personality. *Journal of marketing research*, 34(3):347–356.

Clemens Bechter, Giorgio Farinelli, Rolf-Dieter Daniel, and Michael Frey. 2016. Advertising between archetype and brand personality. *Administrative Sciences*, 6(2):5.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv:1606.06121*.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Aron Culotta and Jennifer Cutler. 2016. Mining brand perceptions from twitter social networks. *Marketing science*, 35(3):343–362.

Gary Davies, José I Rojas-Méndez, Susan Whelan, Melisa Mete, and Theresa Loo. 2018. Brand personality: theory and dimensionality. *Journal of product & brand management*.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7659–7666.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. *arXiv preprint arXiv:1908.06361*.

Anjalie Field and Yulia Tsvetkov. 2019. Entity-centric contextual affective analysis. *arXiv preprint arXiv:1906.01762*.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1705–1715.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*.

CG Jung. 1954. Psychological aspects of the mother archetype. collected works 9/1.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mitchell Lovett, Renana Peres, and Ron Shachar. 2014. A data set of brands and their characteristics. *Marketing Science*, 33(4):609–617.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628.

Gustavo Penha and Claudia Hauff. 2020. What does bert know about books, movies and music? probing bert for conversational recommendation. In *Fourteenth ACM Conference on Recommender Systems*, pages 388–397.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Soumyadeep Roy, Niloy Ganguly, Shamik Sural, Niyati Chhaya, and Anandhavelu Natarajan. 2019. Understanding brand consistency from web content. In *Proceedings of the 10th ACM Conference on Web Science*, pages 245–253.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Scott Sleek. 2018. The bias beneath: Two decades of measuring implicit associations. *APS Observer*, 31(2).

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. olmpics–on what language model pre-training captures. *arXiv preprint arXiv:1912.13283*.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do nlp models know numbers? probing numeracy in embeddings. *arXiv preprint arXiv:1909.07940*.