Deep Neural Language Models:
The Elicitation of Brand Similarity from Customer Reviews

Abstract: Traditional methods of language modelling in marketing research often rely on an atomistic encoding of brands. This local representation makes it impossible to compare brands numerically. We train deep autoencoders, a dimensionality reduction technique, on customer reviews to circumvent this problem. Our findings suggest, that we can learn distributed representations, which encode information about brands. The representations can then be used to compute brand similarity. An inspection of the brand similarity enables researchers a deeper insight into the market structure.

Keywords*: Deep learning, Brand similarity, Autoencoder*
Track*: Methods, Modelling & Marketing Analytics*

1. Objectives of the Research

User generated content gained notable traction in the marketing discipline. Especially product reviews have shown their capabilities in a set of marketing related analyses, such as sales forecasting (Schneider & Gupta, 2016) or mining of brand perceptions (Culotta & Cutler, 2016). Methods used in marketing research to study customer reviews often rely upon an atomistic representation of words, i.e. a very long vector of zeros with only few active elements. This bag of words assumption is commonly applied to convert textual data for subsequent analyses, like a simple multinomial logistic regression or a latent Dirichlet allocation (Lee, Yang, Chen, Wang, & Sun, 2016; Tirunillai & Tellis, 2014). An atomistic (or local) representation of words, which is also used for brands, is fundamentally problematic if one intends to compare them. If we compare any two given brands using this representational form, they are always equally distant and orthogonal. The brands are dissimilar from each other, although a-priori market knowledge may suggest otherwise.

In this paper, we address the problem of incomparability of brands in a local representation. To compare brands in the given scenario, a function is required to condense the information contained in a review to a compact and more efficient representation. Our approach draws inspiration from the distributional hypothesis, which states that a word defines itself by the words of its surroundings (Firth, 1957). The learned representation should thus ideally derive its knowledge for a single word from all the words in its neighborhood. We use neural networks as a dimensionality reduction technique to estimate this function. Our approach is similar to a principal component analysis, albeit nonlinear, multi-layered, and significantly more powerful. Precisely, we aim to learn a function which condenses a sparse word-count vector from a customer review to a lower dimensional distributed representation (Hinton, 1986), while retaining as much information as possible about the input (Vincent, Larochelle, Lajoie, Bengio, & Manzagol, 2010) . Research shows, that the human brain encodes brand personality in a similar way (Chen, Nelson, & Hsu, 2015). The neural network we are using, an autoencoder (AE), can learn such a distributed representation. An AE is essentially a highly nonlinear dimensionality reduction technique (Bourlard & Kamp, 1988; Vincent, Larochelle, Bengio, & Manzagol, 2008). By leveraging underlying relationships between brands, the AE can derive abstract features from the reviews in the form of latent variables. We then use this representation to compare brands and to derive a similarity matrix for brands in the market for consumer electronics.

Although a single layer representation is powerful, we stack multiple AEs to form a deep AE for even higher performance. We do not know of any previous applications of deeper AEs in the marketing discipline. The absence of deeper neural networks in marketing research is disconcerting, because they have been shown to be very effective (Erhan, Courville, & Vincent, 2010) and highly performant in other research fields (LeCun, Bengio, & Hinton, 2015). We will demonstrate their performance on the given problem. Because the architecture of a neural network is data dependent, we will first outline the data.

2. Data and Preparation

For our analyses we employ a small set of 3.6 million Amazon reviews, which we obtained from He & McAuley (2016). We only consider the market for consumer electronics, because the reviews of electronic products can directly be allocated to a specific brand. We prepare the reviews according to Hinton & Salakhutdinov (2006). Reviews are converted to lowercase and all non-alphabetic characters eliminated. We remove stop-words and stem the remaining words. To convert the reviews into matrices, the occurrences of the first $V = 4096$ words are counted. A vocabulary of this size is common for the given application (Glorot,

Bordes, & Bengio, 2011; Hinton & Salakhutdinov, 2006). To assure that the method can generalize to holdout samples and to prevent overfitting, we split the data into a training (80%), a validation (10%), and a test subset (10%).

3. Research Method

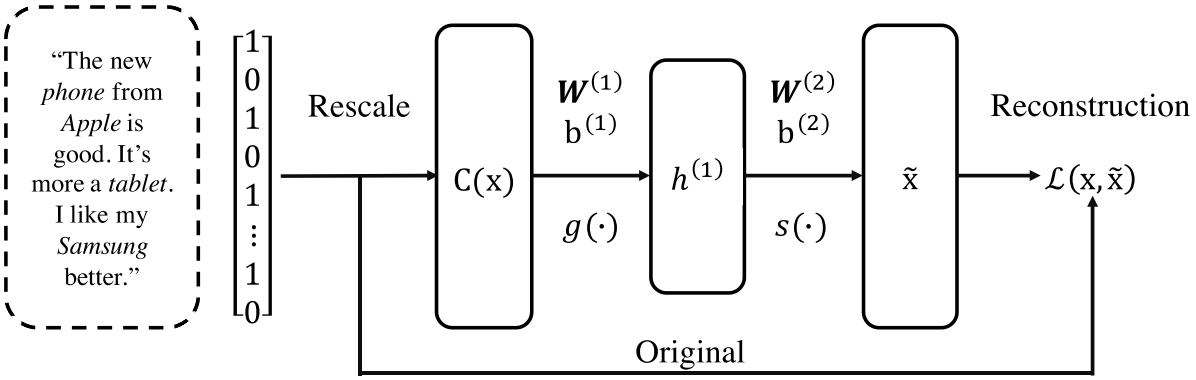

Figure 1: Single Layer Autoencoder.

Figure 1 depicts our application of an AE on customer reviews, which can be formalized as follows:

$$h^{(1)}(\mathrm{x}) = g(\boldsymbol{W}^{(1)}\mathrm{x} + \mathrm{b}^{(1)}) \quad (1)$$

$$f_\theta(\mathrm{x}) = s(\boldsymbol{W}^{(2)}h^{(1)}(\mathrm{x}) + \mathrm{b}^{(2)}) \quad (2)$$

Where $\boldsymbol{W}^{(1)}$ denotes a $M \times V$ weight matrix and $\mathrm{b}^{(1)}$ the M dimensional bias vector for layer one. The input x is a vector with $V = 4096$ dimensions, which contains the word counts of the review. M is the number of neurons of the hidden layer and θ is the set of all learnable parameters. Because input and output of an AE must have the same dimension, we tie the weights, so that $\boldsymbol{W}^{(2)}$ is the transpose of $\boldsymbol{W}^{(1)}$. For the remainder of this paper, we define the reconstruction $f_\theta(\mathrm{x})$ as x̃. For our computations we are using denoising AEs, which have been shown to be superior to standard AEs (Vincent, Larochelle, Lajoie, Bengio, & Manzagol, 2010), because they learn more robust features. Initially, every i-th word in a review vector is rescaled by equation (eq.) 3 so that each review becomes a probability distribution, which sums to unity.

$$\mathrm{x} = \frac{\mathrm{x}}{\sum_{i=1}^{V} \mathrm{x}_i} \quad (3)$$

After rescaling, we keep two versions of each review: The original and a corrupted review. A stochastic corruption process $C(\mathrm{x})$ is applied to the latter, which randomly retains a previously defined fraction of the words, e.g. 90%. The corrupted review is then used to compute the activation value $h^{(1)}$ using eq. 1. The activation function $g(\cdot)$ is an element-wise transformation of every neuron in the hidden layer. Past marketing research often used the sigmoid activation $g(a) = 1/(1 + \exp(-a))$ (Briesch & Rajagopal, 2010; Thieme, Song, & Calantone, 2000; West, Brockett, & Golden, 1997). The sigmoid is an old activation and suffers from several drawbacks (Goodfellow, Bengio, & Courville, 2016). As activation function we choose the rectified linear unit (ReLU), a recently developed and highly performant function from deep learning literature (Glorot et al., 2011; Nair & Hinton, 2010). It is essentially the identity for positive inputs and zero otherwise:

$$g(\mathrm{a}) = \max(0, \mathrm{a}) \quad (4)$$

The resulting hidden layer vector $h^{(1)}$ is a lower-dimensional representation of the input, which should retain a significant amount of information about each review. Eq. 2 then

computes the reconstruction from the hidden layer representation, which ought to be close to the original input. This reconstruction is constrained by the softmax function (eq. 5).

$$s(\mathrm{a}) = \frac{\exp(\mathrm{a})}{\sum_{i=1}^{V} \exp(\mathrm{a}_i)} \tag{5}$$

Applying the softmax on the output enforces the reconstruction to be a probability distribution, just as the input. To optimize the given system, we minimize the average reconstruction error over all T examples in our dataset between the uncorrupted input and the reconstruction. Although past marketing research often used the mean squared error as loss function when applying neural networks (Agrawal & Schorling, 1996; Briesch & Rajagopal, 2010; Juan, Hsu, & Xie, 2017), we use the binary cross entropy between the individual words of a review.

$$\arg\min_{\theta} \frac{1}{T} \sum_{t}^{T} \left(- \sum_{i}^{V} \left[\mathrm{x}_{t,i} \log\left(\tilde{\mathrm{x}}_{t,i}\right) + \left(1 - \mathrm{x}_{t,i}\right) \log\left(1 - \tilde{\mathrm{x}}_{t,i}\right) \right] \right) \tag{6}$$

The binary cross entropy is usually applied when the inputs are probabilistic (Vincent, Larochelle, Bengio, & Manzagol, 2008). The architecture of the AE is hence resembling a non-linear logistic regression with 4096 outcomes. As we have formalized the problem, a simple gradient based optimization technique can be employed to minimize eq. 6.

4. Major Results

We trained a three layer AE with 500-500-50 neurons for 30 epochs. A three layer network was chosen, due to the advantages of depth (Bengio, 2009). The number of neurons and other parameters (e.g. learning rate) were chosen based on the performance of several architectures on the validation and testing set. We trained each layer individually, which is referred to as pretraining (Erhan, Courville, & Vincent, 2010). After the pretraining of the first layer is finalized, the layer is used to transform the raw input. The second layer then learns from the transformed input. After all layers completed the training phase, they are concatenated to become a deep AE. In line with the literature we set the fraction of kept words to 90%, although the AE is fairly insensitive to the amount of corruption (Vincent et al., 2010). The corruption is only applied during training. We then input a review with just one word, the brand name, in the stacked AE and obtain a numerical distributed representation. This representation c_i for a brand i is the vector derived from the last hidden layer with 50 neurons, which we then use to calculate the Pearson correlation $r\left(\mathrm{c}_i, \mathrm{c}_j\right)$ between brands. To investigate the representation for any given brand, we compute the similarity for some of the most frequent brands in our data set. Table 1 depicts the similarity measurements. Judging from the results we can derive exploratory knowledge about the market structure.

Consider for example Nikon and Canon ($r = .91$), which exhibit the highest similarity in the given set. Both are manufacturers of photography equipment. The AE is technically trying to mimic what customers would write, given the set of latent variables. Hence the similarity of the representation encourages the viewpoint, that Canon and Nikon share a very similar set of latent features. Nikon also shares ties with Sony (r =.81), as the latter is a supplier for cameras as well. Sony offers a much broader range of products, like home entertainment or car electronics. Their product range overlaps with Panasonic. Panasonic also sells a wide variety of home entertainment products, photo and video equipment, and kitchen appliances. Hence, the relationship between Panasonic and Sony is also remarkably strong (r =.88). We also observe, that Samsung is similar to Sony ($r = .84$). Interestingly, we would assume that

		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1.	Logitech	1									
2.	Sony	.67	1								
3.	Canon	.53	.76	1							
4.	Nikon	.64	**.81**	**.91**	1						
5.	Samsung	.62	**.84**	.74	.78	1					
6.	Microsoft	.67	.60	.61	.67	.67	1				
7.	Apple	.70	.69	.63	.70	.74	.71	1			
8.	Asus	.71	.70	.67	.70	**.79**	**.79**	**.80**	1		
9.	Panasonic	.67	**.88**	.74	.78	**.78**	.55	.69	.69	1	
10.	Klipsch	.69	.61	.49	.56	.57	.55	.55	.63	.69	1
11.	JBL	.64	.61	.39	.46	.49	.39	.50	.52	.61	.70

Table 1: Similarity Scores between Brands. The eight highest similarities are in bold.

Apple and Samsung would be much more related than $r = .74$, because they compete closely in the market for smartphones. Let us consider Apple, which exhibits strong similarity to Asus (r =.80). Both companies offer laptops. Asus is also in the vicinity of Microsoft (r = .79) and Samsung ($r = .79$). We also observe the appearance of less prominent relationships. Logitech for example is somewhat related to most of the other brands, maxing out at $r = .71$ for the relationship with Asus. JBL, a manufacturer of audio equipment, is only somewhat related to Logitech, but has an overall dissimilar set of latent variables compared to other brands, like Canon, Samsung, Nikon, or Apple. This behavior also emerges for the speaker and headphone supplier Klipsch. Note that Klipsch and JBL are similar with an $r = .70$. which encourages the viewpoint that pure audio manufacturers exhibit a different set of latent variables. The lowest similarity $r = .25$ in our analyses was exhibited by Lowepro (backpacks for camera) and Apple.

5. Implications and Limitations

Deriving distributed representations of brands is a difficult task. The proposed method allows marketing researchers to compress data to a severe degree and to obtain insight into what consumers talk about. This paper demonstrates the effectiveness of deep neural networks as a tool researchers can employ to derive knowledge from large quantities of unstructured information. We introduced the denoising AE as a dimensionality reduction technique, which can be stacked multiple times to effectively reduce the size of our data. The fundamental strength of AEs, and hence of neural networks, is, that they are broadly applicable to a wide variety of problems. Marketing practitioners can benefit from the application of neural networks to model the language their customers use. The proposed method allows them to analyze, whether the companies' perception of competition is aligned with the customers' perception of competing brands. If divergence is observed, companies can align strategic decisions to account for the companies that are perceived by customers as competitive brands.

The method is nonetheless subject to limitations. First, the approach requires a lot of data. We trained the model on 3.6 million reviews. Hence, for a brand to appear in the vocabulary, it has to be mentioned frequently by customers. Second, we only address the limitations of a local representation, but not of word ordering. Our approach similarly ignores the ordering in which the words appear. A comparison of brands, which can convey positive or negative connotations, is basically interpreted as similar. Third, our approach requires a lot of tuning.

Especially the choice of layer size and depth is a critical problem to be address by extensive hyperparameter tuning. We argue that the absolute values of similarity are rather inexpressive and depend on layer size. These values, thus, have to be interpreted relatively as presented in this study. Fourth, validating the results is a crucial aspect. A validation could consist of computing multiple models and averaging their similarity measurements. Alternatively, one could use different models on the same dataset, which could then be compared. Despite these limitations and considering the findings of this study, we expect deep learning to gain ground as a research tool for marketing researchers in the near future.

6. References

Agrawal, D., & Schorling, C. (1996). Market share forecasting: An empirical comparison of artificial neural networks and multinomial logit model. *Journal of Retailing*, *72*(4), 383–407. https://doi.org/10.1016/S0022-4359(96)90020-2

Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends® in Machine Learning*, *2*(1), 1–127. https://doi.org/10.1561/2200000006

Bourlard, H., & Kamp, Y. (1988). Auto-Association by Multilayer Perceptrons and Singular Value Decomposition. *Biol. Cybern*, *59*, 291–294. https://doi.org/10.1007/BF00332918

Briesch, R., & Rajagopal, P. (2010). Neural network applications in consumer behavior. *Journal of Consumer Psychology*, *20*(3), 381–389. https://doi.org/10.1016/j.jcps.2010.06.001

Chen, Y.-P., Nelson, L. D., & Hsu, M. (2015). From "Where" to "What": Distributed Representations of Brand Associations in the Human Brain. *Journal of Marketing Research*, *52*(4), 453–466. https://doi.org/10.1509/jmr.14.0606

Culotta, A., & Cutler, J. (2016). Mining Brand Perceptions from Twitter Social Networks. *Marketing Science*, *35*(3), 343–362. https://doi.org/10.1287/mksc.2015.0968

Erhan, D., Courville, A., & Vincent, P. (2010). Why Does Unsupervised Pre-training Help Deep Learning? *Journal of Machine Learning Research*, *11*, 625–660. https://doi.org/10.1145/1756006.1756025

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*, 1–32.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. *AISTATS '11: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, *15*, 315–323. https://doi.org/10.1.1.208.6449

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

He, R., & McAuley, J. (2016). Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 507–517). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2872427.2883037

Hinton, G. (1986). Learning Distributed Representations of Concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (pp. 1–12). Hillsdale, NJ: Erlbaum.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507. https://doi.org/10.1126/science.1127647

Juan, Y. K., Hsu, Y. H., & Xie, X. (2017). Identifying customer behavioral factors and price premiums of green building purchasing. *Industrial Marketing Management*, *64*, 36–43. https://doi.org/10.1016/j.indmarman.2017.03.004

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436. http://dx.doi.org/10.1038/nature14539

Lee, A. J. T., Yang, F. C., Chen, C. H., Wang, C. S., & Sun, C. Y. (2016). Mining perceptual

maps from consumer reviews. *Decision Support Systems*, *82*, 12–25. https://doi.org/10.1016/j.dss.2015.11.002

Nair, V., & Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. *Proceedings of the 27th International Conference on Machine Learning*, (3), 807–814. https://doi.org/10.1.1.165.6419

Schneider, M. J., & Gupta, S. (2016). Forecasting sales of new and existing products using consumer reviews: A random projections approach. *International Journal of Forecasting*, *32*(2), 243–256. https://doi.org/10.1016/j.ijforecast.2015.08.005

Thieme, R. J., Song, M., & Calantone, R. J. (2000). Artificial Neural Network Decision Support Systems for New Product Developmen Project Selection. *Journal of Marketing Research*, *37*(4), 499–507.

Tirunillai, S., & Tellis, G. J. (2014). Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. *Journal of Marketing Research*, *51*(4), 463–479. https://doi.org/10.1509/jmr.12.0106

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning - ICML '08*, 1096–1103. https://doi.org/10.1145/1390156.1390294

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P.-A. (2010). Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.*, *11*, 3371–3408.

West, P. M., Brockett, P. L., & Golden, L. L. (1997). A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice. *Marketing Science*, *16*(4), 370–391.