Topic Modeling to Extract Information from Nutraceutical Product Reviews

Deena Liz John*†, Ernest Kim*, Kunal Kotian*, Ker Yu Ong*, Tyler White*,
Luba Gloukhova‡, Diane Myung-kyung Woodbridge†, Nicholas Ross†
Data Science Program, University of San Francisco†
Graduate School of Business, Stanford University‡
{djohn,ekim22,kkotian,kong3,tvwhite,dwoodbridge,ncross}@usfca.edu
lvg@stanford.edu

***Abstract*—Consumer purchases of Vitamins and other nutraceuticals have grown over the past few years with most of the growth occurring in on-line purchases. However, general e-commerce platforms, such as Amazon, fail to cater to consumers' specific needs when making such purchases. In this study, the authors design and develop a system to provide tailored information to consumers within this retail vertical. Specifically, the system uses Natural Language Processing (NLP) techniques to extract information from user-submitted nutraceutical product reviews. Using Natural Language Processing, three information streams are presented to consumers (1) a five point rating system for cost, efficacy and service, (2) a summary of topics commonly discussed about the product and, (3) representative reviews of the product. By presenting product-specific information in this manner we believe that consumers will make better product choices.**

***Index Terms*—Natural Language Processing, Topic Modeling, Machine Learning, Mobile Applications, Nutraceuticals**

I. Introduction

As of 2016, more than two-thirds of American adults took dietary supplements each year with the supplement industry contributing \$121.7 billion to the US economy [1]. Among dietary supplement consumers, over 70% take multivitamins and over 30% take fish oil, calcium, Vitamin C or Vitamin D for their health and wellness or to avoid a nutrient deficiency [2]. Experts expect theses numbers will only increase with an aging population, a movement in the health care industry away from disease treatment to prevention and increasing awareness of healthier lifestyle choices [3]. Unsurprisingly, many consumers purchase their dietary supplements on-line with Amazon being the market leader with about 35% of the on-line market share [4][1].

Unfortunately this marketplace is both confusing and rife with fraud [6]. An important contributor to this is that consumers are attempting to extract information specific to nutraceuticals from websites optimized for "general" e-commerce. Consider Amazon, which by nature of its size contains one of the largest collections of user-submitted reviews yet only presents product information in a way that can be generalized to selling every other item on its platform.[2] Consider the case of a consumer looking to purchase "fish oil" on an e-commerce website. On existing e-commerce websites, searching for this product yields a page of highly-rated products with no additional insight into their rating. Since we expect consumers in this space to be specifically interested in a number of aspects of the product (product efficacy, for example), we we propose to use Natural Language Processing ("NLP") techniques to analyze the reviews and then create a rating for that product, along this specific aspect using those reviews.

The developed application leverages our knowledge of consumer motivations in the nutraceutical product class to extract additional useful information from user-submitted reviews. The applied NLP algorithms include tokenization, lemmatization, phrase detection, word embeddings and topic modeling. By using these techniques, more informationally useful measures can be created. In particular, we extract information from user-submitted reviews to present three pieces of information. The first, a five-star rating, focuses on three topics that we believe are the primary motivators behind a consumer's purchasing decision: cost, efficacy and customer service. The second, a list of topics related to the product, employs an information extraction technique based on an unsupervised learning process to identify product-specific topics from consumer reviews. Finally, a set of curated reviews are presented to the consumer.

Using these streams of information we believe that system presented, named "ShopSmart.ml", can assist user consumers when making purchasing decisions in this space.

II. Related Work

The data that we use to assist consumers comes from On-line reviews, which has been heavily studied within the greater management literature. Papers in this area tend to focus on linking reviews to other managerially useful measures, such as profit or sales. For example, Zhu and Zhang demonstrated that sales of less popular products are sensitive to customer reviews

[1]Amazon itself has realized the potential in this market and launched its own private label for supplements [5].

* These authors contributed equally.

[2]One solution to this problem would be to consult medical professionals for their evaluation of a particular product. Unfortunately, however, hospital systems have yet to come up with a consistent recommendation framework [7].

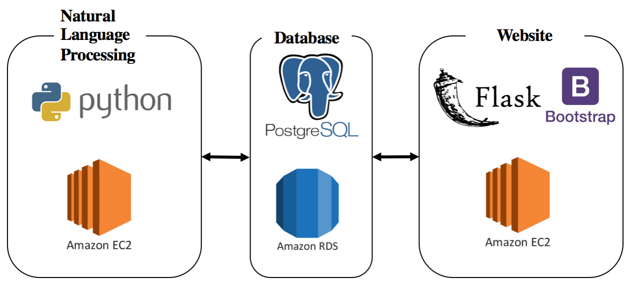

Fig. 1. System Overview

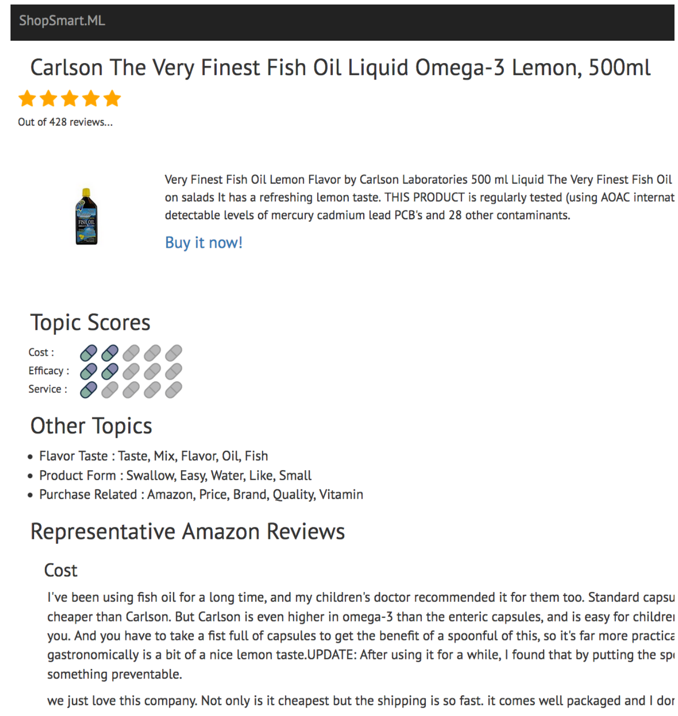

Fig. 2. Example product page

[8] while Hu, Koh and Reddy showed that the sentiment within text reviews has a direct impact on product sales, though the ratings themselves do not. That study also showed that the timing of reviews matters and that reviews voted as most "helpful" are strong sales drivers [9]. Another study also demonstrated that the number of on-line reviews has a high impact on product sales [10].

To more purposefully identify the pathway between reviews and product sales, researchers have also created measures of specific reviews, such as their quality, tone or likelihood of being fake with the goal of understanding what specific features within a product's review affect its sales. Examples of this include [11] which applied machine learning techniques to detect and rank groups of reviewers who leave fake product reviews on Amazon.

While our analysis does not specifically focus on sentiment analysis, on-line reviews are frequently the input into such studies (in some fields this is called "Opinion Mining"). Since on-line reviews are relatively easy to collect, researchers often use them as inputs in order to verify the accuracy of different sentiment analysis techniques. Surveys in this area include [12] and [13].

III. System Overview

ShopSmart.ml itself is a web application where users can search for and purchase nutraceutical products. In this section we present the user-facing portion of the system and then describe the back-end processing (which includes details on the NLP techniques used). Figure 1 provides a short diagram of the technologies used within each part of the system.

A. User-facing Component

The web application itself was written using Python and Flask, deployed on an Amazon Web Service (AWS) Elastic Compute Cloud (EC2) instance with data stored in PostgreSQL. Flask is a web framework written in Python, providing a web server gateway interface and template engine [14]. We used Bootstrap, an open source toolkit for developing with HTML, CSS, and JavaScript as the user interface. Bootstrap is a tool for building responsive web sites which simplifies the rendering of pages across a variety of devices, screens and window sizes [15]. AWS EC2 provides a cloud-based virtual machine instance with auto-scaling features, allowing the application to use additional resources under load [16].

The front page of the application contains a subset of popular products and a search bar. Once a product is selected (or searched for via the search bar), a product page appears, one of which can be found in Figure 2. The product page begins with high-level title information taken directly from Amazon. This includes the star-rating, a picture of the product, the product's description and a link to purchase the product directly from Amazon.

Following the title information, three additional pieces of information are provided. The first is a a "Topic Score", which is a one-to-five rating for each of cost, efficacy and service. The second, "Other Topics" is a list of additional topics, specific to that product, that occur within that product's reviews. The final section, "Representative Amazon Reviews", contains a set of curated reviews.

The three pieces of information at the bottom (Topic Scores, Other Topics and Representative Customer Reviews) are created by using NLP techniques, as described in the following section.

Data

We utilized Amazon product and product review data collected between May 1996 and July 2014 by McAuley [17], [18]. This dataset has also been used in research projects for developing a recommender system [19], sentiment analysis

TABLE I
AMAZON HEALTH AND PERSONAL CARE PRODUCT AND REVIEW DATA

Data Set	Details	Size
Product	Single-compound products	26,818
Product Review	Filtered 5-star reviews	217,530

on product reviews [20], large-scale data analysis using distributed computing [21], [22] and fake review detection [11].

Among these reviews, we focus on review data from single-compound products in the Health and Personal Care category (Table I), excluding sports nutrition, multivitamins and weight-loss products and including 5-star reviews, which represent over half the rerviews. Before we created our three data streams, we processed the data in the manner described below.

B. Back-end Processing

Figure 3 presents an overview of the back-end system used to process the reviews. Before creating our three information streams, we transformed our review using the process below.

Filtering: To conduct this analysis we limit our dataset to those observations corresponding to single-compound products. This is done to avoid handling multiple classification problems. For example, if a person is taking a supplement with multiple compounds (for example, fish oil and Vitamin C), reviews may state that one compound "works" for a particular problem, but the other does not. We thus remove these products to focus on a smaller, single focused, problem. To identify single compound nutraceutical products, a list of substrings that were part of the relevant products' category names, which are available for each product, was constructed. Using regular expressions, only those products whose category names matched a substring in this list were retained, while the rest of the products were ignored.

We also removed extreme reviews: those with a title but without any text and those having more than 2,000 words, which was about the 1% of the remaining reviews. These reviews tended to either contain little information or too much information that was not specific to the problem at hand.[3]

Our final two filtering steps were to focus only on 5-star reviews (about 61% of the reviews in total) and remove reviews which were tagged as "EDITED" by the user. We removed those reviews which were edited over time by the user since they generally contained multiple, conflicting sentiments. Rather than attempt to parse these statements we focused our analysis on more consistent, information dense reviews, namely 5-star ratings. Our analysis focused on 5-star reviews because users tend to only purchase those goods with high-ratings. Filtering this way allows us to identify the factors causing a consumer to positively react to a product.

[3]A particularly common occurrence is consumers writing up their motivation for taking the product in incredible detail.

Pre-processing: We then added the title of the review to the review body as a separate sentence as the title often contained information informing the main review.

This text is then processed using spaCy, a Python library for parsing text. Using the spaCy parser's English language parser, each review was tokenized, or broken down into a list of constituent words, with all punctuation and unnecessary white space removed. The resulting constituent words were lemmatized and tagged with their part-of-speech. Lemmatization is a normalization process that groups and transforms words into non-inflected and non-derivate forms. For example, words such as "good" and "better" tend to represent the same idea and thus they are grouped together using this technique. The result of this process was a set of tokens associated with each review. Finally, any token consisting of a webpage URL was removed.

Phrase Detection Algorithms: After the reviews were processed the resultant data was analyzed to identify naturally occurring phrases. Upon identification, a phrase was treated as a single token within the review.

Phrase detection was undertaken using the Gensim package in Python which includes a robust topic modeling toolkit [23]. In particular, we calculated the Normalized Pointwise Mutual Information (NPMI) score for each set of words [24] which can be found in the formula below (x and y are particular words within a review):

$$NPMI(x,y) = \left(\ln \frac{p(x,y)}{p(x)p(y)}\right) / -\ln p(x,y) \qquad (1)$$

The idea behind this method is that for each set of words two numbers are calculated: the number of times those words appear as a phrase and the number of times they appear in isolation. Larger values of this ratio indicate a higher likelihood of those words being their own phrase. All pairs of words that had a score larger than a specified threshold (0.6) were treated as phrases. Before calculating this number, we removed common terms such as "and, "of, "with", "without", and "or".

The phrase detection algorithm was run twice on the entire corpus of reviews. The first pass identified bigrams (two-word phrases) and the second pass identified additional bigrams and trigrams (three-word phrases).

Our final phrase detection step filters out bigrams and trigrams which fail to conform to commonly used part-of-speech patterns. Part-of-speech (POS) tagging is a tagging method identifying each word in a text as noun, verb, pronoun, preposition, adverb, conjunction, participle, or article [25]. Johnson's study showed that part-of-speech tagging and template matching can extract more useful phrases from bigrams and trigrams [26]. In our study, bigrams are only accepted if they follow the pattern [noun or adjective][noun] and trigrams are only accepted if they follow the pattern [noun or adjective][any part-of-speech][noun or adjective]. If a potential phrase failed to conform to the patterns above the words were separated and re-tokenized as single words, rather than as

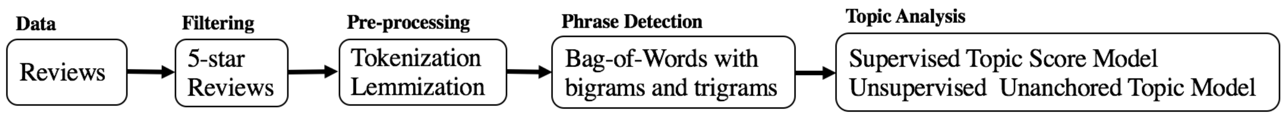

Fig. 3. System Diagram

phrases. After this process, the three most frequent phrases were found to be fish_oil, side_effect, and high_quality.

After phrase detection was complete, the token list was filtered one final time. First all common English stop words (such as "the", "is" or "at") were removed along with all tokens which were identified as pronouns. Single character and number tokens were also removed. Finally, very rare tokens (those that appeared in less than 10 reviews) and very common tokens (those appearing in more than 50% of all reviews) were also ignored.

Topic Analysis: After the phrase detection process was concluded two topic analysis processes were undertaken. The first process consisted of identifying reviews concerned with efficacy, cost and service ("Topic Score Model") concerns. The second topic identification process identified eight topic areas, such as those dealing with taste and product form (pill vs. powder, etc.).

a) Topic Score Model: We choose to provide a 5-point scale for three important factors that we believe would be of primary interest to purchasers of nutraceuticals: (1) Efficacy, (2) Cost, and (3) Service. A primary concern of the literature in marketing and marketing psychology is understanding how consumers *value* a product when making purchasing decisions. While there are many different models suggested by this literature, their is still no precisely defined theory [27]. Models of this choice, however, tend to include a notion of "product value" or how a consumer feels about the precise price-object trade-off and also a notion of "shopping value" which values the experience that a consumer has while making the purchase [27] [28]. In the case of product value, a consumers perception of a nutraceutical is clearly linked to its efficacy and cost. Since all purchases are made through a single e-tailer (Amazon), variations in the shopping value achieved by a customer are going to be primarily driven by the service they receive from the company fulfilling the order. We thus focus on cost, efficacy and service as the primary drivers of the value that a customer receives for making a purchase and hence why they would make a particular nutraceutical purchase.

For each topic-product combination we created a score, representing the extent to which that topic was present in the specific product's reviews. To create our score, two token lists were created for each of the topics, one using a Word2Vec Model and another using an Anchored Topic Model.

The first token list above was based on a custom Word2Vec model trained over the corpus of reviews. The Word2Vec algorithm transforms a text corpus into vector representations (embeddings) in an n-dimensional space. In this case, we trained the model to generate 100-dimensional embeddings for each word token [29][30].

The trained model was then used to find "similar" words to a specified word, such as those that appeared in similar contexts across the review corpus. Once similar words were identified, they were reviewed to verify their relationship, with words that appeared to be unrelated to the rest of the words in that topic being removed. Note that this list of similar words found using Word2Vec also helped identify and group together misspelled versions of words in the reviews (e.g. the misspelled word "noticible" was identified as being similar to its correct form "noticeable" and was added to the list of target word tokens).

A second word list was created using an Anchored Topic Model for each of the topics from the Topic Score Model. This second word list model was intended to boost the performance of the previous model by providing an additional set of tokens. Anchored Correlation Explanation (CorEx) used for this topic model allows controlling the nature of the topics that emerge from a text corpus by specifying the words that may be a part of those topics [31][32]. The Anchored CorEx is for optimizing the following objective function where X, Y are groups of random variables, TC and I represent total correlation and mutual information respectively. In this equation, x is an anchor word.

$$Maximize \quad TC(X;Y) + \beta \sum I(x;y) \tag{2}$$

As before, we directed the model to the topics of cost, efficacy and service.

In order to anchor topics to the specific ideas of efficacy, cost, and service, we found words related to those ideas occurring in the review text corpus using the trained Word2Vec model mentioned earlier. This method of finding "similar" words was used to construct a small set of words that were used to train the anchored CorEx topic model. After the anchored CorEx model was trained, it yielded a set of words representing each of the three topics. These were the words that model 'learned' as being representative of efficacy, cost, and service based on the review corpus. These lists of representative words for the three anchored topics also included a few unusual/non-representative words (e.g. first names of people) which were removed based on a manual review.

The final cleaned lists of representative words from the anchored topic model were used to boost the sets of words the previous model looked for while assigning topic scores to reviews.

For each product, the proportion of the reviews which contained at *at least one token* from the three lists of representative words (for efficacy, cost, and service) was calculated.

Importantly, we previously restricted our analysis to 5-star reviews, meaning that if the token appeared in a particular

TABLE II
THE LENGTH OF CREATED TOKEN LISTS USING WORD2VEC AND ANCHORED TOPIC MODEL

Category	Word2Vec	Anchored Topic Model
Efficacy	73	9
Cost	48	8
Service	89	5

review it was likely that it was being referred to in a positive way.

b) Unanchored Topic Model: The "Other Topics" section of the product page was created to find topics outside of cost, efficacy and service which were motivating buyers. To find these additional topics, an unanchored topic model was trained, once again using CorEx.

The trained unanchored topic model was then used to assign topic labels to each review based on the tokens within the review. These topic labels were manually generated based on each topics representative words identified by the CorEx model. Note that each review could be associated with zero or more topics labeled above.

For each review and every Other Topic, the trained unanchored topic model gives binary indicators denoting whether that review belongs to the Other Topic. A topic score was determined for each Other Topic by calculating the proportion of reviews of a given product that belonged to that specific Other Topic. For each product, the top three Other Topics were identified as those with the highest three topic scores. Within an Other Topic, the top five words associated with that other topic, which also appeared in that product's reviews are also displayed on the product's page.

IV. RESULTS

After initial filtering, we were left with 217,530 reviews on single-compound products. The preprocessing with NMPI over 0.6 identified 14,441,052 distinct bigrams and trigrams.

The number of tokens after applying part-of-speech templates and removing stop words was 7,580,786. The number of tokens after removing tokens with no alphabets and single character tokens was 133,384. Finally, the number of tokens after filtering tokens based on frequency (very rare tokens, ie. those appearing in less than 10 reviews and removal of very common tokens, i.e. those appearing in more than 50% of reviews in the corpus) was 13,850. [4]

For topic modeling, Table II shows the number of tokens that the Word2Vec model generated and the Anchored CorEx yielded beyond the initial sets of words from the Word2Vec model. In particular, the final list included 82 tokens associated with Efficacy. We then calculated the proportion of reviews which contained a single one of these 82 tokens.

To find these additional topics, an unanchored topic model was trained with the number of expected topics set to 13. The topics learned by this model were examined by reviewing the list of words associated with them. Out of the 13 topics, we identified coherent themes related to the sets of words for 8 topics. The coherent topics identified were labeled as follows:

1) Common Ailments
2) Flavor/Taste
3) Workout-related
4) Chronic Ailments
5) Purchase-related
6) Appearance-related
7) Product Form
8) Gut Health

The topics for which coherent themes could not be assigned were simply ignored.[5]

V. CONCLUSIONS

This paper presents the development of "Shopsmart.ml", a web application that assists consumers in purchasing nutraceutial products. The purpose of this project is to create a web site which caters to consumers' informational needs when making these specific purchases. In particular, our system uses NLP techniques on user-submitted review in order to rank each product on a 5-point scale along the dimensions of Cost, Efficacy and Service, as well as using an unsupervised topic model to identify additional specific features of a product that consumers have identified as important in their purchasing decision.

A key reason for building this system is that general e-commerce platforms, such as Amazon, fail to cater to the specific needs of these consumers. These platforms tend to only provide a single five-point scale to consumers, which makes sense if you are trying to build an information system that generalizes across all product categories. However, when looking at a single product category then product-specific information systems can be devised which will provide additional value to consumers (as well as assisting them in making more efficient purchasing decisions).

Of future interest would be to identify which parts of this system provide additional value to consumers. For example, are there other characteristics (such as taste) that drive consumer purchases decisions strongly enough to warrant directly highlighting this aspect directly for all products? Or are there additional sources of information which could be used to buttress the current, user-submitted, reviews, such as those from an expert source? With the growth of the nutraceutical market, these types of additional features could provide significant consumer value.

REFERENCES

[1] "Dietary supplement industry contributes $122 billion to u.s. economy," *Neutralceutical World*, vol. 19, July–August 2016.
[2] A. Dickinson, J. Blatman, N. El-Dash, and J. C. Franco, "Consumer usage and reasons for using dietary supplements: report of a series of surveys," *Journal of the American College of Nutrition*, vol. 33, no. 2, pp. 176–182, 2014.

[4]From this point forward, we use the terms "token" and "word" interchangeably to reference the output of this analysis.

[5]A number of the non-coherent topics included phrases relating to the personal history of the person writing the review.

[3] M. Pandey, R. K. Verma, and S. A. Saraf, "Nutraceuticals: new era of medicine and health," *Asian J Pharm Clin Res*, vol. 3, no. 1, pp. 11–15, 2010.

[4] "ecommerce maintains top spot in the $12.8 billion vitamin category," *Neutralceutical World*, vol. 19, July–August 2016.

[5] S. Perez, "Amazons private label elements expands for first time in years with invite-only vitamins and supplements," March 2017.

[6] A. O'Connor, "Knowing What's in Your Supplements," *Well*, Feb. 2015. [Online]. Available: https://well.blogs.nytimes.com/2015/02/12/107141/

[7] M. H. Cohen, A. Hrbek, R. B. Davis, S. C. Schachter, and D. M. Eisenberg, "Emerging credentialing practices, malpractice liability policies, and guidelines governing complementary and alternative medical practices and dietary supplement recommendations: a descriptive study of 19 integrative health care centers in the united states," *Archives of Internal Medicine*, vol. 165, no. 3, pp. 289–295, 2005.

[8] F. Zhu and X. Zhang, "Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics," *Journal of marketing*, vol. 74, no. 2, pp. 133–148, 2010.

[9] N. Hu, N. S. Koh, and S. K. Reddy, "Ratings lead you to the product, reviews help you clinch it? the mediating role of online review sentiments on product sales," *Decision support systems*, vol. 57, pp. 42–53, 2014.

[10] A. Babić Rosario, F. Sotgiu, K. De Valck, and T. H. Bijmolt, "The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors," *Journal of Marketing Research*, vol. 53, no. 3, pp. 297–318, 2016.

[11] A. Mukherjee, B. Liu, J. Wang, N. Glance, and N. Jindal, "Detecting group review spam," in *Proceedings of the 20th international conference companion on World wide web*. ACM, 2011, pp. 93–94.

[12] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications," *Knowledge-Based Systems*, vol. 89, pp. 14–46, 2015.

[13] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, "Sentiment analysis: A review and comparative analysis of web services," *Information Sciences*, vol. 311, pp. 18–38, 2015.

[14] A. Ronacher, "Flask," *Pocoo,[Online]. Available: http://flask. pocoo. org/.[Använd 22 02 2018]*, 2016.

[15] T. Bootstrap. (2018) Bootstrap. the most popular html, css, and js library in the world. [Online]. Available: https://getbootstrap.com/

[16] Amazon Web Service. (2018) Amazon ec2. [Online]. Available: https://aws.amazon.com/ec2/

[17] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel, "Image-based recommendations on styles and substitutes," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015, pp. 43–52.

[18] J. McAuley, R. Pandey, and J. Leskovec, "Inferring networks of substitutable and complementary products," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 785–794.

[19] S. Zhang, L. Yao, and A. Sun, "Deep learning based recommender system: A survey and new perspectives," *arXiv preprint arXiv:1707.07435*, 2017.

[20] A. Radford, R. Jozefowicz, and I. Sutskever, "Learning to generate reviews and discovering sentiment," *arXiv preprint arXiv:1704.01444*, 2017.

[21] L. Chen, R. Li, Y. Liu, R. Zhang, and D. M.-k. Woodbridge, "Machine learning-based product recommendation using apache spark," *IEEE UIC International Workshop on Data Science and Computational Intelligence (DSCI)*, 2017.

[22] E. Jonas, Q. Pu, S. Venkataraman, I. Stoica, and B. Recht, "Occupy the cloud: Distributed computing for the 99%," in *Proceedings of the 2017 Symposium on Cloud Computing*. ACM, 2017, pp. 445–451.

[23] R. Rehurek and P. Sojka, "Gensim–python framework for vector space modelling," *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.

[24] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," *Proceedings of GSCL*, pp. 31–40, 2009.

[25] D. Jurafsky and J. H. Martin, *Speech and language processing*. Pearson London, 2014, vol. 3.

[26] D. Johnson, V. Malhotra, and P. Vamplew, "More effective web search using bigrams and trigrams," *Webology*, vol. 3, no. 4, p. 35, 2006.

[27] R. Sánchez-Fernández and M. Á. Iniesta-Bonillo, "The concept of perceived value: a systematic review of the research," *Marketing theory*, vol. 7, no. 4, pp. 427–451, 2007.

[28] B. J. Babin, W. R. Darden, and M. Griffin, "Work and/or fun: measuring hedonic and utilitarian shopping value," *Journal of consumer research*, vol. 20, no. 4, pp. 644–656, 1994.

[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[31] R. J. Gallagher, K. Reing, D. Kale, and G. V. Steeg, "Anchored correlation explanation: topic modeling with minimal domain knowledge," *arXiv preprint arXiv:1611.10277*, 2016.

[32] K. Reing, D. C. Kale, G. V. Steeg, and A. Galstyan, "Toward interpretable topic discovery via anchored correlation explanation," *arXiv preprint arXiv:1606.07043*, 2016.