Beauty Concoction: Deep Learning Text Classification on Distinctions in Amazon Product Advertising and Pricing

Christina Chen
christinachen@ischool.berkeley.edu

Justine Taylor
justinetaylor@ischool.berkeley.edu

Abstract

Beauty and Personal Care products are used daily by everyone. We explored price-influencing categories, inspired by Joshua Freedman and Dan Jurafsky's paper, Authenticity in America. We utilized four different datasets, implemented varying NLP techniques, and created ELMO and assorted BERT embeddings through a neural network and CNN to develop an in-depth understanding of product ingredients, marketing strategies, and price. Ultimately, we found that ELMO embeddings with a Neural Network slightly outperformed all other embedding approaches and architectures in predicting price per ounce across both datasets used.

1 Introduction

With the rise of online shopping and delivery as fast as a click of a button, we wanted to investigate the ground truth of products sold on Amazon. In a consumer-focused world, everyday people are deceived by marketing terms and trending items. Common, daily-use beauty products such as men's deodorant, baby soap and women's makeup are often marketed with misleading terms such as "FDA Approved", "All Natural", or "Dermatologist Tested", regardless of containing potentially toxic ingredients or having no evidence-backed claims. We strive to understand the influence of these marketing tactics on price by utilizing various NLP tasks: Cosine Similarity, TF-IDF, Word Embeddings, Neural Networks.

In the scope of this course, we've focused on two main categories: lotions and hair products. In order to detect a decent signal, we analyzed product categories rather than all beauty products at once. We chose categories based on the number of products available in each category and the range of people that use such products. For instance, lotion and shampoo are marketed to all genders, ages, ethnicities, and socio-economic classes.

We evaluated factors that make beauty products have varying price points. First is falutin, which is defined as elaborated language or heightened by artificial or empty means. "Well-marketed products can do well for a prolonged time even if they are not objectively of great efficacy or elegance." (Brucculieri). Second, we looked at Distinction. Brands often use specific language to make their products appear unique such as "premium" or "special". They also compare their product to others using terms such as "better", "more", etc. and use linguistic negation or hyper positivity. Third, we examined Authenticity. We utilized TF-IDF to examine terms that exude authenticity and credibility. Lastly, we examined Health by analyzing product ingredients. We realize formulation, procurement of trade secret ingredients, and processing can have a critical influence on price, it is ultimately not included in our analysis.

The price field in the raw data is the price of the entire product, not the price per unit. We extracted the units (milliliters or ounces) and converted everything to ounces for a standard unit of measure. After each factor computation (falutin, distinction, authenticity, health), we split the products based on the median of the price per unit distribution in order to analyze the difference between lower and higher priced products.

This paper contributes to the analysis of language in the Beauty industry, with the eventual goal to understand the relationship between products, marketing, and price.

2 Background/Related Work

In our research for academic papers, we focused on reviewing papers from 2019 to the present-day that conducted various NLP tasks. From question answering to summarization, we first focused on topic modeling and aspect extraction, which we highlight in 6.1. We ultimately focused on utilizing NLP tasks to apply to Joshua Freedman and Dan Jurafsky's 2011 paper, Authenticity in America: Class Distinctions in Potato Chip Advertising.

The paper highlights the language of food advertising, specifically in potato chips, and they created their own definitions of falutin, distinction, authenticity, and health pertaining to this product. They did their analysis using 12 products (6 low-priced and 6 high-priced), and they did it manually using the text on the front and back of the potato chip bags.

The paper highlights "how price can affect underlying variables in languages" (Word Salad). We seek to find the variables that influence price and expand on its categories.

3 Data

We used four datasets for the project: Hugging Face Amazon Reviews, Jungle Scout, Amazon web scrape, and California Open Data Chemicals in Cosmetics. Hugging Face is our base data set, which we then joined with the rest of the data we obtained.

3.1 Sources

3.1.1 Hugging Face Amazon

We utilized the Beauty category of the Amazon Customer Reviews from Hugging Face. The size of the downloaded dataset files is 871.73 MB, and the size of the generated dataset is 2286.33 MB. Features included customer_id, helpful_votes, marketplace, product_category, product_id, product_parent, product_title, review_body, review_date, review_headline, review_id, star_rating, total_votes, verified_purchase, vine, but we ultimately only used product id and product title. We also needed to determine the product category (ex. shampoo, makeup, etc.), but the product_category field only indicated "beauty", so we extracted keywords from titles to hard-code the product categories.

3.1.2 JungleScout

JungleScout is a third-party platform providing data to help entrepreneurs and brands grow their businesses on Amazon. We were able to export products related to our 30+ categories of interest, ultimately combining 112 CSVs. The data allowed us to gain insight into each product's: Sales, Revenue, Price, Rank, and Rating. Due to Juggle Scout and its relative newness, we were only able to match 144 products to Hugging Face's product_id. However, we were still able to use the data in our price prediction.

3.1.3 Amazon Web Scrape

We scraped Amazon to further compliment Hugging Face product reviews dataset. The purpose of this dataset is to include details not found in Hugging Face or JungleScout, such as product description, ingredients, directions, safety information, disclaimers, and tags. The features used for our price predictive tasks are: title, price, about, description, and ingredients.

3.1.4 California Chemicals

California Department of Public Health (CDPH) provides open data for the California Safe Cosmetics Program (CSCP). For all cosmetic products sold in California, the California Safe Cosmetics Act requires manufacturers, packers, and/or distributors named on the product label to provide to the CSCP a list of all cosmetic products that contain any ingredients known or suspected to cause cancer, birth defects, or other developmental or reproductive harm. We did not find any datasets with sufficient product ingredient data, therefore we felt scraping this information from Amazon was necessary for our analysis.

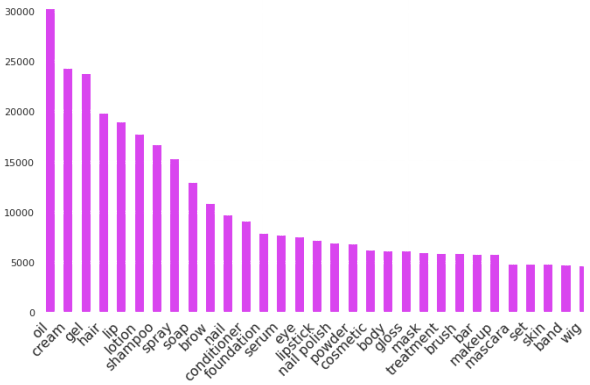
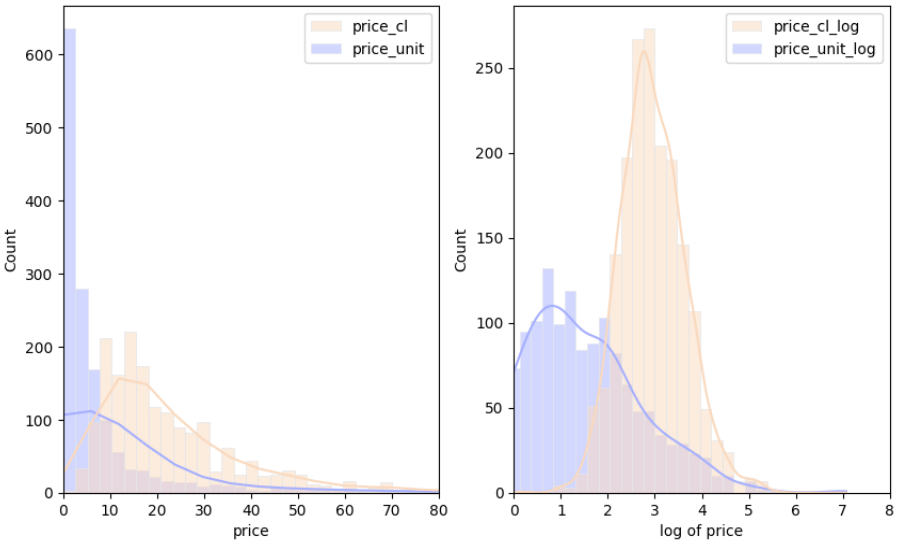


Table 3: Lotions: price and price unit distribution

3.2 Processing



The task to gather data and process it required a lot of work. For the Amazon scrape, we created a script using request_html, selenium, BeautifulSoup, and tor to manually render javascript per page, scrape the necessary data, and parse text from HTML output. Amazon pages are not standardized and ingredients can exist in different columns of our output, we only extracted information out of the column it's most likely to be in. To pre-process our data, we removed duplicate products, standardized units of measurements to ounces, created price per unit, labeled 'category' as -PROD- in our text columns, lower-cased all text, removed punctuation, removed stopwords, edited stop words as deemed necessary, removed numbers, tokenized, stemmed, and lemmatized. We removed null prices and removed outliers that exceeded a z-score of 1.71. We joined products that were able to match. We realized a benefit in removing brand names from our text, we decided not to go forth in the interest of time.

Sample sentence:

> **Inexpensive:** eucerin origin heal prod bodi prod extrem dri skin compromi skin emmoli enrich bodi prod fragranc free dermatologist recommend brand ounc jar pack emolli enrich prod leav sooth layer skin lock moistur dri skin relief long last rich bodi prod help heal dri compromi skin leav feel smooth dermatologist recommend year eucerin pioneer skincar innov today trust one lead recommend brand product design protect support moistur enhanc look skin gentl daili prod daili moistur bodi prod also use hand prod dri skin relief gentl skin prod fragranc dye paraben free size ounc pack eucerin origin heal prod timet formula help heal protect extrem dri sensit compromi skin thick rich formula provid inten moistur replac dri skin origin heal prod bind water skin provid effect moistur help prevent moistur loss replenish skin ' moistur barrier origin heal ' irrit skin noncomedogen free fragranc dye year eucerin pioneer skincar innov today recogn trust dermatologist one lead recommend brand
>
> **Expensive:** skinceut emolli rich restor prod normal dri skin jar fl oz rich restor prod normal dri skin formul exclu select natur extract oil perfect use dri skin ideal high altitud cold dri climat hydrat nourish skin exclu combin lipid marin extract vitamin e optim dri skin three nutrientrich brazilian sea alga nourish hydrat skin oil grape seed rise hip macadamia restor maintain moistur emolli ideal high altitud cold dri climat

4 Methods

4.1 Dan Jurafsky Replication: Class Distinctions in Potato Chip Advertising

We adapted Freedman and Jurafsky paper on two categories: lotions and hair products. Throughout the categories below, we examined the data through Textual Complexity, Cosine Similarity, TF-IDF, Word Embeddings, and Neural Networks. It is important to note, that for each product and feature, we segmented the data based on the median price for this analysis.

High-Falutin. The paper distinguishes differences in high-price and low-priced potato chips on readability score, number of words, commonality. We used python's textstat module to calculate the flesch kincaid grade level, readability ease score and lexicon count. We also hand-curated a list of marketing terms (lemmatized) and added the raw count of marketing terms present in the product text as a feature as a proxy for word commonality.

Lotion Results: The lower priced products had higher flesch-kincaid grade, lower reading ease, higher number of words, and slightly more marketing terms on average in comparison to the higher priced products.

Hair Results: The lower priced products had a slightly lower flesch-kincaid grade, higher readability, lower number of words and a lower number of marketing terms on average in comparison to the higher priced products.

The results were not consistent across both datasets, however it is possible that the target audience for lotion and hair product consumers differs slightly. It is also important to note that terms in the product text are very rare and sometimes complex terms due to the nature of the beauty and personal care domain.

Health. The paper touches upon health factors in chips of different price points, but does not go into details other than expensive chips had lack of trans fat or lack of some other supposed unhealthy elements. In contract, our analysis of beauty products provided a vast resource into ingredients and health-related terms used. Furthermore, we added a layer of complexity to it by incorporating toxic ingredients from the CDPH. We found that although items may be priced higher, and/or priced higher per unit, we found no difference in the number of toxic ingredients present. If anything, we found more toxic ingredients in higher priced products.

Results: Across both datasets, the number of toxic ingredients was consistent on average. If anything, higher priced products had a slightly higher count of toxic ingredients. Lower priced items included ingredients knowledgeable to common people, as shown below, whereas more expensive items included ingredients that were more abstract or beauty chemicals.

Common ingredients in lotion products:

Inexpensive: lavend(er), avocado, almond, alo(e), oil, shea, coconut, oatmeal

Expensive: percent, spf (uvb, uva, uv), sunscreen, acid, retinol, hyaluron, glycerin, vitamin, peptide

Common ingredients in hair products:

Inexpensive: argan, shea, phathalat

Expensive: bergamot, keratin, extract, vitamin

Distinction. The paper categorized distinction as a divider between the lower and upper class. More expensive and "exotic" chips for the higher-class, popular chips for the working class. Freedman and Jurafsky investigated 3 characteristics of distinction: superlative words such as most and best, words with suffixes like -er in **better**, as well as negation of words, all to assert a chip's seemingly higher quality.

Similarly, we wrote a list of hand-curated "uniqueness" and "comparator" words (lemmatized). We took the raw count of the number of occurrences of these terms in the product as a baseline. We improved upon it using two methods: unigram, bigram and trigram custom word embeddings and pre-trained glove embeddings. For the n-gram custom embedding approach, we took the unigram, bigram and trigrams for each product text, connected bigrams and trigrams by underscores and trained a custom Word2Vec model. For both approaches, we took the cosine similarity of each word in the product text to each unique/comparator term in the hand curated list. If the value was greater than 0.5, it was considered in the "count" of unique/comparator terms present in the product text.

In assessing sentiment, we used nltks SentimentIntensityAnlyzer and hypothesized that both negative and positive scores could potentially provide some signal.

Lotion Results: The baseline approach (raw count) achieved little to no signal, so it has been discarded and we focus on the improvements for the analysis between lower and higher priced products. The custom unigram, bigram, trigram method demonstrated that lower priced products had, on average, fewer instances of "uniqueness" or "comparator" words. Interestingly, the pre-trained GLOVE vector approach showed the opposite.

For sentiment, we found that lower priced products had a higher compound sentiment score indicating that sentiment, both positive and negative, plays a stronger role in lower priced products.

Hair Care Results: The baseline approach (raw count) achieved little to no signal, similar to the lotion dataset so, again, it has been discarded and we focus on the improvements for the analysis between lower and higher priced products. The custom unigram, bigram, trigram method demonstrated that lower priced products had, on average, fewer instances of "uniqueness" or "comparator" words, which is consistent with the results from the lotion dataset. However, unlike the lotion dataset, this one was consistent with the pre-trained GLOVE vector approach, albeit slightly.

For sentiment, we found that the hair care results were consistent with the lotion results.

Common unique/comparator words in lotion products:

Inexpensive: bodi, (no) paraben, colorsaf(e), sulfat(e), paraben free(e)

Expensive: Sleeker, rather, wrinkle, anti wrinkle, sulfate free

Common unique/comparator words in hair products:

Inexpensive: brand

Expensive: type, without, color, pure

Authenticity. The paper highlights authenticity as high quality, consistent, location-bound, history, and down-play of commercial motivations. In chips, the categories included naturalness, ingredient and processing, historical elements, and locality. In beauty products, we found it to be a similar method, with some additions. Naturalness, ingredients, and locality is consistent with the potato chip analysis and we found no historicity for products. We found a new category, **Confirmation** and **Anti-Marketing Marketing**, in a product's authenticy. Confirmation contained words that provided safety to a consumer such as "certified", "dermatologist", etc. Anti-marketing marketing contained phrases such as "no b.s." that alluded to the feel of transparency from a brand.

Common authenticity words in lotion products:

Inexpensive: moistur, dermatologist test, approve, african, korean, japan, mediterranean,

Expensive: women, anti, dermatologist, fda, treatment, cruelty-free, physician, manuka

Common authenticity words in hair products:

Inexpensive: moroccan, gurantee

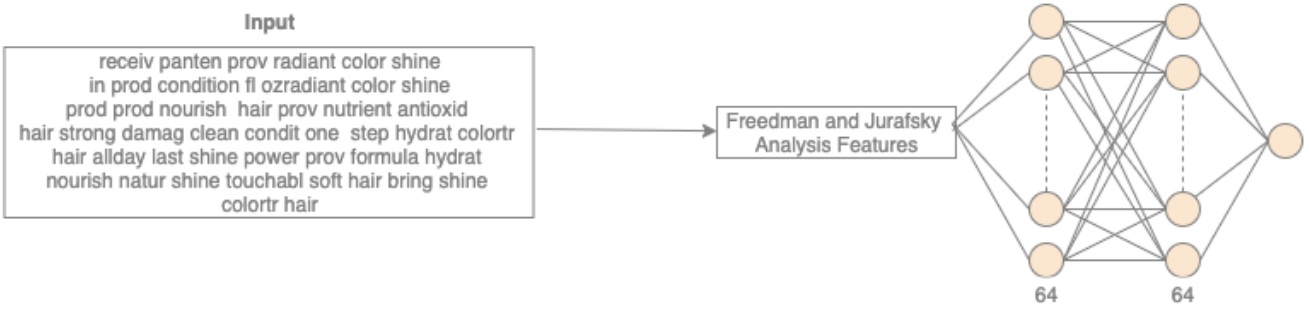
Expensive: treatment, safe, recommend

4.2 Models

In an effort to expand on the analysis done in Freedman and Jurafsky's paper, we utilized the features we extracted from that analysis in a neural network to try to predict price per ounce. We then took it a step further, and created various embeddings for the product text and passed those through both a neural network and a convolutional neural network to try to predict price per ounce. The motivation for using embeddings comes from the hypothesis that the features extracted from the Freedman and Jurafsky adaptation are ideally captured by embeddings.

For both the baseline and the improvement, we used a three (dense) layer neural network with a 0.1 dropout rate. We also limited the input text size to at most 100 words for efficiency. Given more time and resources, we would have considered scaling this upwards as our longest sequence was about 500 words. Additionally, for all models, we examined both mean squared error and mean absolute error. We did not solely focus on mean squared error because our output has a slight skew and therefore is not entirely gaussian.

BASELINE



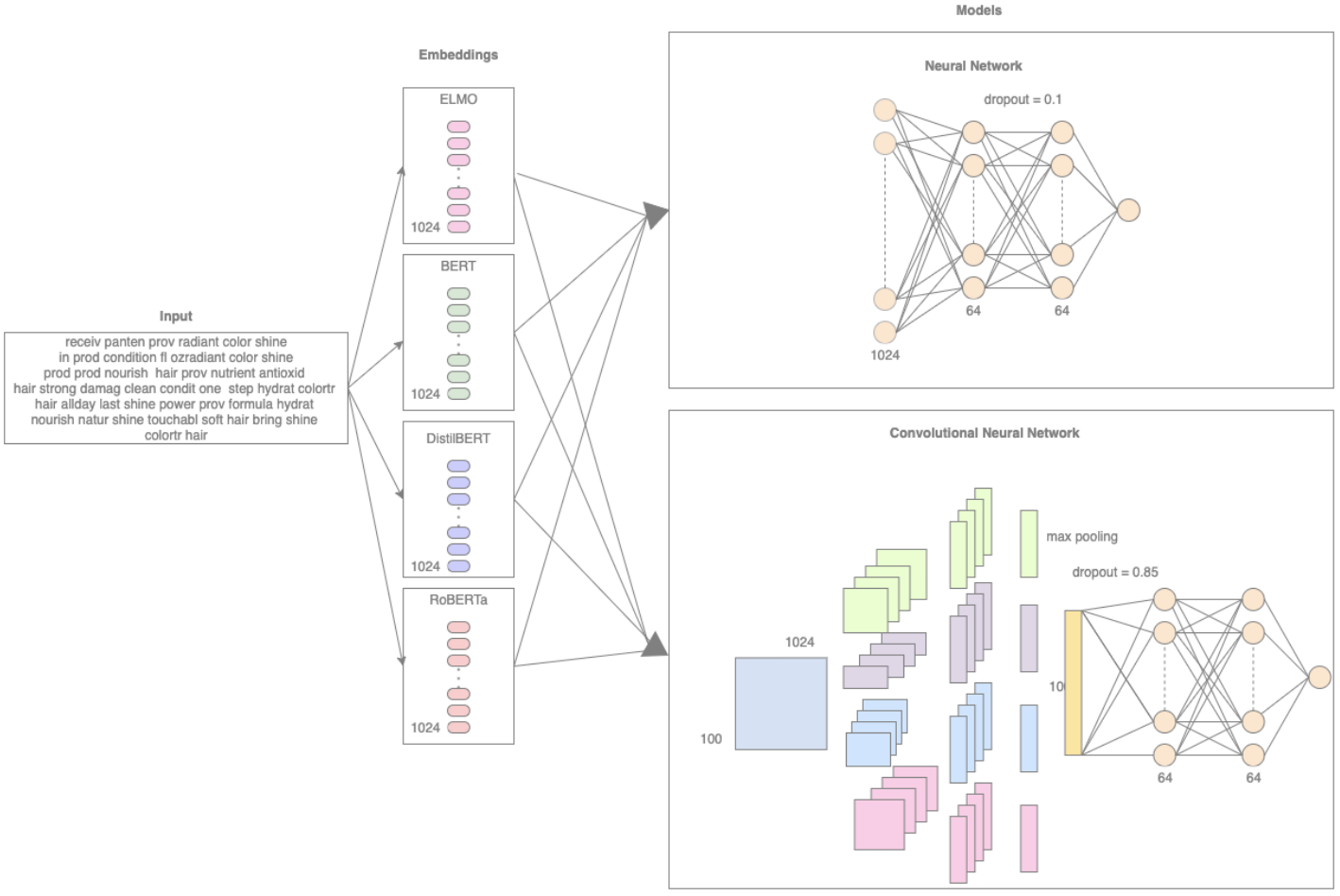
For our baseline, we experimented with the features extracted from the aforementioned analysis. We tried using the raw values, scaled values and variations of the following features: number of marketing terms (raw), flesch kincaid grade, reading ease score, number of words, number of marketing terms (based on cosine similarity using GLOVE), sentiment scores and number of toxic chemicals.

Baseline Results

Mean Absolute Error: 1.473

Mean Squared Error: 9.056

Improvement



Embeddings: ELMO

We chose ELMO as a word embedding due to its contextual advantages in assigning embeddings. We realize in beauty products, many words change meaning depending on context in titles, descriptions, and in the future reviews.

Embeddings: DistilBERT, RoBERTa, BERT

We wanted to utilize transformers because they are newer and could potentially yield better results, however we recognize that this kind of task is not necessarily what BERT variations are meant for. Nonetheless, we experimented with DistilBERT, RoBERTa and BERT.

Neural Network

We used the same Neural Network architecture for both the baseline and the improvement. The only difference was the input layer dimensions which was dependent on the feature size (6-7 features) or embedding size (768 for BERT, 1024 for ELMO).

CNN

For the convolutional neural network, the size of the input was 100 words by the length of the embeddings (either 768 or 1024). Through hyperparameter tuning, we used 4 filters for each kernel, kernels of size 4,2,3 and 4 and a dropout rate of 0.85. We followed the filters with max pooling and a 3 (dense) layer neural network.

Lotion Results:

	Neural Network	Convolutional Neural Network
ELMO	Mean Absolute Error: 1.357 Mean Squared Error: 8.916	Mean Absolute Error: 1.470 Mean Squared Error: 9.116
DistilBERT	Mean Absolute Error: 1.508 Mean Squared Error: 8.691	Mean Absolute Error: 1.467 Mean Squared Error: 9.115
RoBERTa	Mean Absolute Error: 1.778 Mean Squared Error: 12.314	Mean Absolute Error: 1.471 Mean Squared Error: 9.153
BERT	Mean Absolute Error: 1.506 Mean Squared Error: 9.488	Mean Absolute Error: 1.470 Mean Squared Error: 9.106

We were surprised to see little difference across the various embedding approaches and modeling approaches. However it is clear that in this case, ELMO embeddings with a Neural Network provided the best output.

Hair Care Results:

	Neural Network	Convolutional Neural Network
ELMO	Mean Absolute Error: 1.467 Mean Squared Error: 6.371	Mean Absolute Error: 1.338 Mean Squared Error: 6.559
DistilBERT	Mean Absolute Error: 1.423 Mean Squared Error: 6.645	Mean Absolute Error: 1.337 Mean Squared Error: 6.567

RoBERTa	Mean Absolute Error: 1.415 Mean Squared Error: 8.206	Mean Absolute Error: 1.338 Mean Squared Error: 6.586
BERT	Mean Absolute Error: 1.371 Mean Squared Error: 6.581	Mean Absolute Error: 1.339 Mean Squared Error: 6.562

Once again, consistent with the lotion dataset, ELMO embeddings with a neural network provided the best output with slight variations across all embedding types and architectures.

6 Future Additions

In our analysis for products, we realized the stark contrast in language within our product price points, especially in lotions. In the future, rather than programmatically extracting product categories, we'd like to segment products into further categories using machine learning techniques so that we could analyze language within products such as eye "cream", suntan "lotion" with higher granularity. We would also like to incorporate more features in predicting the products' prices, such as the number of reviews, brand and location. Additionally, after we pivoted (post-midpoint feedback), we no longer utilized reviews. We feel the review data held a great deal of interesting and valuable insights into the product such as packaging and quality. For this, we would like to improve upon our KeyBERT baseline for this task using Attention Based Aspect Extraction. Last, we hope to find the Influence of other parts of speech aside from adjectives on product description, especially adverbs.

7 References

Alammar, J. (n.d.). The illustrated bert, elmo, and co. (how nlp cracked transfer learning). Retrieved April 11, 2021, from http://jalammar.github.io/illustrated-bert/

Bennani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., & Jaggi, M. (2018). Simple unsupervised keyphrase extraction using sentence embeddings. arXiv preprint arXiv:1801.04470.

Brucculieri, J. (2019, January 24). Why that $325 Moisturizer might not be any better than a $12 One. Retrieved April 11, 2021, from https://www.huffpost.com/entry/face-moisturizer-expensive_l_5c48813de4b083c46d649020

Chemicals in cosmetics. (n.d.). Retrieved April 11, 2021, from https://data.chhs.ca.gov/dataset/chemicals-in-cosmetics

Chahuneau, V., Gimpel, K., Routledge, B. R., Scherlis, L., & Smith, N. A. (2012, July). Word salad: Relating food prices and descriptions. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 1357-1367).

Freedman, J., & Jurafsky, D. (2011). Authenticity in America: Class distinctions in potato chip advertising. Gastronomica: The journal of food and culture, 11(4), 46-54.