Linguistic harbingers of product sales on e-commerce websites

Aparna Gupta, Meghendra Singh, Pratik Anand

Abstract

Millions of users buy regular household items from the Amazon Marketplace every day. Most of these products are not much different from their competitors in terms of brand, pricing or specifications. Various product factors influence a user to choose one product among many similar products, product description and reviews being the two most prominent factors. We want to find out how much these two factors influence a user's decision in choosing a product over its competitors.

Introduction

As opposed to brick and mortar stores, shoppers on e-commerce marketplaces like: Amazon, eBay and Alibaba.com base there buying decisions on product descriptions, customer ratings and reviews instead of touching, feeling and trying out the product. Hence, aspects like product descriptions provide sellers an opportunity to advertise their product and convince potential customers to buy their products. On the other hand, the crowd-sourced customer reviews for a product act like a validation to the claims made by the seller in the product description. Thus, both product description and customer reviews provide important signals informing the purchase decision of a potential customer. Since, both customer reviews and product descriptions are natural language texts, we can extract linguistic features from both of these sources and explore if they play any role in the success of a product. Assuming that a successful product is one with high sales, our research problem boils down to the following questions:

1. Are seller-controlled features like: product description significant predictors of product sales?
2. Do linguistic aspects of product descriptions like: politeness, readability and sentiment play a role in determining product sales?
3. Are crowd-controlled features like: product reviews significant predictors of product sales?
4. Do linguistic aspects of product reviews like: politeness, readability and sentiment play a role in determining product sales?

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

In this work, we try to answer these questions by compiling, augmenting and analyzing a dataset of 8K products from Amazon.Com, containing seller-controlled product descriptions, buyer controlled product reviews and estimated product sales for each product.

Related Work

In the past, there have been multiple studies focusing on predicting sales of products on e-commerce websites using web search data, historical sales data, time series models, log data, product and brand features [15, 17]. Wei et. al. hypothesize that sales on e-commerce websites can be predicted using the frequency of keyword based searches performed by users on the website [15]. They employ an auto-regression model taking into consideration a structured time series model of the search data to predict sales of women's apparel on Taobao.com. Although, they present an interesting analysis aimed at demonstrating the relationship between search data and site-wide sales. While this might be beneficial for the e-commerce platform as a whole, the technique cannot be directly applied individual products. A recent paper by Zhao et. al. tries to make product level sales predictions for 1.8 Million products on CaiNiao.com by combining a host of product and website attributes (historical sales data, number of product listing views, user views, merchandise volume and selling price) into a single data frame[17]. They use these data frames to train a Convolutional Neural Network (CNN) so as to predict sales of individual products. Although, these studies aim to improve aspects like forecast accuracy, they fail to utilize linguistic features of common textual attributes associated with any product (e.g. seller profile, product descriptions and buyer reviews).

There has been some relevant work that links textual features in critique movie reviews with product earnings. Joshi et. al. develop a text regression model that uses textual features (viz. n-grams, part of speech n-grams and dependency relations) obtained from movie reviews along with specific meta-data (viz. running time, country, actors) to predict the opening weekend revenue of the movie [8]. Although the work improves the state of the art, it ignores features like sentiment of the reviews. The paper [10] by Mishne et. al. demonstrates a correlation between revenue earned by a movies and the sentiment of blog posts about the movie. They show that positive sentiment in weblogs about a movie, posted prior to its release, is a good predictor of its success. An interesting paper by Ghose et. al. [7] tries to estimate the economic ("dollar") value associated with buyer feedback postings on Amazon.com. All of these studies employ natural language processing techniques to some aspect of predicting sales, they don't explicitly apply these techniques on attributes such as product descriptions that are inherent to the domain of electronic marketplaces.

Perhaps, the most relevant work to our approach is the recent SIGIR paper by Pryzant et. al. [13]. They extract 90,000 product descriptions from the Japanese e-commerce website, Rakuten and employ a novel neural network architecture to identify textual features predictive of product sales while controlling for factors like: brand loyalty and product pricing strategies. The primary finding reported in the study is that appeals to authority, polite language, and mentions of informative and seasonal language in product descriptions lead to high product sales. One limitation of this approach is that it ignores the buyer feedback which is implicitly present on e-commerce websites in the form of customer reviews. Prior research by t et. al. and Zhu et. al. point to the possibility of customer reviews impacting product sales, especially at e-commerce avenues [12, 18]. Although, these studies don't take into consideration any linguistic features of the buyer reviews in their analysis.

A related problem on e-commerce platforms is the abundance of irrelevant customer reviews. In this case it becomes important to rank the reviews so that linguistic features are only extracted from the most relevant product reviews. Although, there has been substantial work on assessing review helpfulness and importance [2, 11, 16], e-commerce platforms handle this problem by crowd-sourcing, where users mark all product reviews that they find helpful and the platforms lists the reviews based on the number of users that found them to be helpful. In our study we only plan to use the most helpful customer review (already tagged by the crowd). Based on our literature survey we did not find any prior work which considers the linguistic aspects (say politeness, sentiment, persuasiveness) of seller controlled product description and buyer controlled product reviews for determining product sales on e-commerce websites.

Methodology

Data

We use Amazon Marketplace product and sales data for our experiments. We retrieve the data from the following sources:

1. Amazon product reviews dataset from [9]. This dataset contains product reviews from Amazon, including 142.8 million reviews spanning May 1996 - July 2014. For the purpose of the pilot study we only consider 500K reviews for products belonging to the "Home and Kitchen" category. These reviews corresponded to 28K products.
2. Amazon product metadata from [9], containing descriptions, titles, images, related products (also bought, also viewed, bought together) and sales rank [1] for 9.4 million products. When we tried to obtain the metadata for the 28K products obtained in the last step, we found that the bare-minimum metadata we required for our analysis (product description and sales rank) was only available for 8K products hence we proceeded with these 8K products for further analysis.
3. While we did have sales rank for the 8K products obtained in the last step, we require actual sales (for e.g., no. of items sold per month) to conduct meaningful analysis. According to Amazon.com, sales rank data is calculated hourly and actual product sales are included in this calculation, however the precise formula to compute sales rank is not available anywhere [1]. We found a handful of websites that provided estimated monthly sales for a product, given the sales rank, although none of them had APIs available for this purpose. To accomplish the conversion of the 8K product sales ranks into estimated sales we ended up writing a Selenium bot which retrieved the estimated sales data from one such website called Jungle Scout [14]. The process was cumbersome and slow, but it seems to be the only way to get estimates of actual sales for a product on Amazon.com
4. Web Scraping is a technique used to extract data from websites. It is an automated process where an application processes the HTML of a Web Page to extract data for manipulation such as Product Review Scraping, Data mining, etc. Selenium is a Web Browser Automation Tool.Primarily, it is for automating web applications for testing purposes, but is certainly not limited to just that. It allows you to open a browser of your choice and perform tasks as a human being would, such as: Clicking buttons,entering information in forms and searching for specific information on the web pages. Selenium uses a web driver which opens the version of the web browser.This has the advantage that the website you are visiting views you basically like any other human surfer allowing you to access information in the same way. For our problem we've used a similar approach where-in we have 'Firefox' web driver to access Jungle Scout APIs like any other web surfer. To fetch estimated sales from Jungle Scout we launched an instance of Firefox browser in incognito mode and extracted estimated sales value.We had to introduced a delay of 20 seconds between subsequent requests to overcome two major blockers: the website runs animations which displays intermediate Estimated Sales value till the final value is being calculated and, to avoid our IP address from being banned from the website because of the frequent requests.

Figure 1 shows a schematic of our data collection and cleaning steps. After retrieving the complete dataset of 8K products, we performed basic cleaning operations on the buyer reviews and product description texts.

We began by selecting the most helpful buyer review and ignoring all other buyer reviews, for each product. Since, each review for a product on Amazon has some helpful votes associated with it, we select the review which has the highest number of helpful votes as the most helpful buyer review for a product. We found that these most helpful buyer reviews were generally longer than the other reviews, contained factual information about the product and also conveyed the buyers sentiment towards the product after using it. Next, we looked at the length (in number of words) of product descriptions and reviews. Figures 2 and 3 show the distributions of description and review length and we observe that these range from 1 to 1400 words and follow the power law. We decided to remove products that contained very short and very long descriptions and reviews while focusing on products which had reviews and descriptions ranging from 20 to 300 words. We selected this interval as it covered almost 80% of the reviews and descriptions while eliminating outliers that might bias our analysis. Figures 4 and 5 show the truncated distributions for descriptions and most helpful reviews respectively. This left us with 6K products for further analysis. We did basic text cleaning operations like conversion to lowercase, removal of HTML tags, special characters and stop words on the description and most helpful review text for each of the 6K products. We also create Bag of words (BOW) for the description and most helpful review text for all the products, although we haven't used the BOWs for any analysis at the time of writing this report.

The next step involved finding sets of similar products for which impact of descriptions and reviews on sales can be analyzed. It is important that the products be relatively similar so that a difference in there sales can be attributed only to descriptions and reviews and not the function or frequency of consumption of the product. Consider for example, the sales of a shampoo is only comparable with other shampoos and not backpacks. Clearly, shampoo is a consumable product, that might be bought multiple times by an individual over the course of a few months, whereas the same individual might only buy a backpack once in a couple of years. This implies that consumables would naturally have very high sales as compared to other products and hence consumables like shampoos should only be grouped with other shampoos and non-consumables like backpacks should only be grouped with other backpacks, for any analysis where sales is the dependent variable. Therefore, to identify groups (buckets) of similar products in our 6K "Home and Kitchen" products, we follow a simple heuristic based approach. For each product (p) we begin by looking at the set of related products (R) present in it's metadata. Figure 6 shows the distribution of the related products in our dataset. In Figure 6, we see that there are 583 products which don't have any related products and there is one product which has 85 related products.

Each related product $r \in R$ was either viewed and/or bought by customers who bought the product p. We found that these related products R were very similar to the original product p and hence the right candidates to be grouped together with p. However, there might be a set of related products $Q \subset R$, which is not present in our dataset (i.e., we don't have complete metadata about products in Q). In this case we can consider our bucket of similar products to be $B = p \bigcup (R - Q)$. We can continue to look at the related products for each product $b \in B$ which are not already in B until we stop getting new products. At the end of this procedure we will have one bucket B of very similar products. For the purpose of the pilot study we take the product which has 85 related products and compute a bucket of similar products using these related products. This bucket of similar products had 385 different products and Figure ... shows the distribution of sales among these 385 products. Almost all of these products are variants of shelves and organizers. The minimum and maximum number of sales per month for this bucket of products are 5 and 18437 and they follow the power law.

Analysis and Algorithms

For the purpose of answering our research questions, we begin by computing four linguistic features for the description and review of each of the 385 products in the similar products bucket. These features are as follows:

1. **Empath**: A tool to generate and validate new lexical categories on demand from a small set of seed terms (like "bleed" and "punch" to generate the category violence). Empath draws connotations between words and phrases by deep learning a neural embedding across more than 1.8 billion words of modern fiction [5]. We used empath to generate 192 features each for the product descriptions and reviews.
2. **Vader**: A lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains [3]. We use Vader to compute the positive, negative, neutral and compound sentiment features for the product descriptions and reviews.
3. **Stanford Politeness API**: A computational framework for identifying linguistic aspects of politeness with relation to the context [4]. We use this API to compute the politeness score of descriptions and reviews. This politeness score serves as another linguistic feature.
4. **Flesch-Kincaid Readability Score**: A formula that focuses on the average number of syllables per word and words per sentence [6]. We use this formula to compute the readability score of descriptions and reviews.The score lies in the range of 0-100. The readability scores serve as another linguistic feature.

   90-100 : Very Easy 80-89 : Easy 70-79 : Fairly Easy 60-69 : Standard 50-59 : Fairly Difficult 30-49 : Difficult 0-29 : Very Confusing

Experiments and Results

We ended up with 400 features for each product, that captured important linguistic features computed from it's description and most helpful review. Therefore, we now had a feature vector along with an estimated sales for all of the 385 products in our similar products bucket. We treat the estimated sales as a dependent variable and all of the linguistic features as independent variables and begin our analysis. Figure ... gives an overview of our experiment and analysis approach.

Linear Regression and LASSO We did a 75:25 training and test set split and employ linear regression to predict estimated sales using only the computed linguistic features. This resulted in a root-mean-squared error (RMSE) of

Figure 1: Data Collection and Cleaning Pipeline

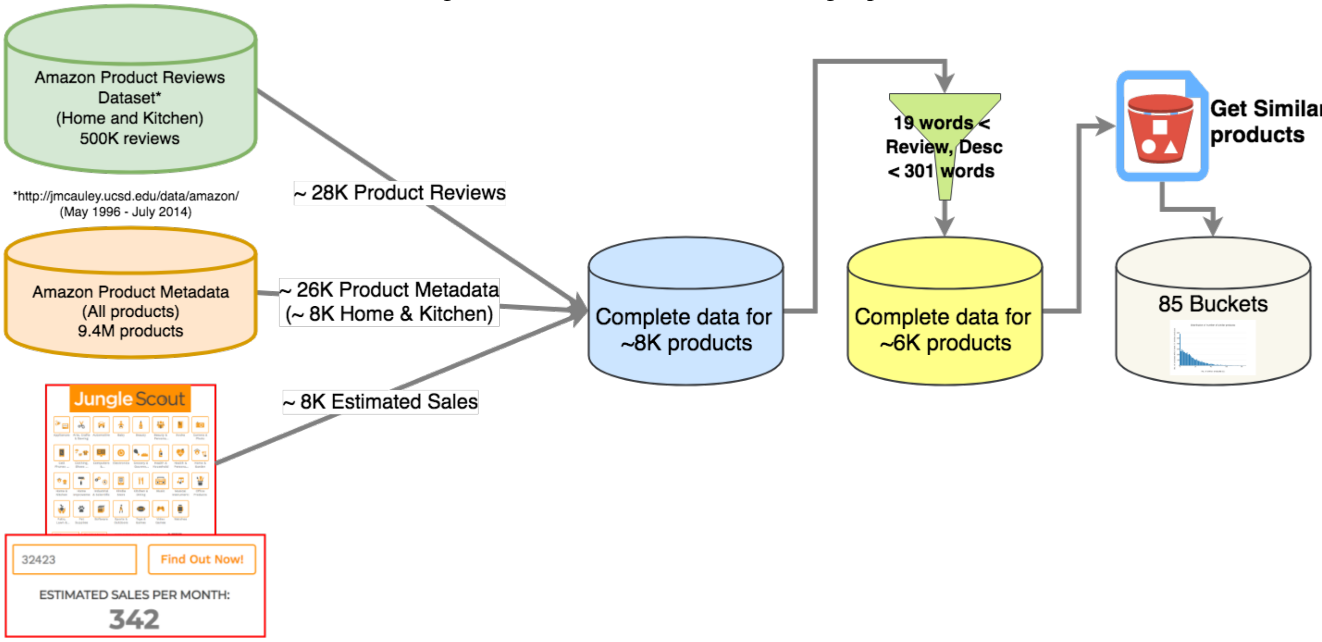


Figure 2: Description length distribution

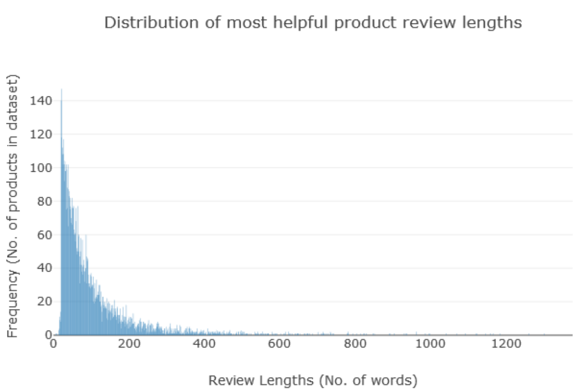
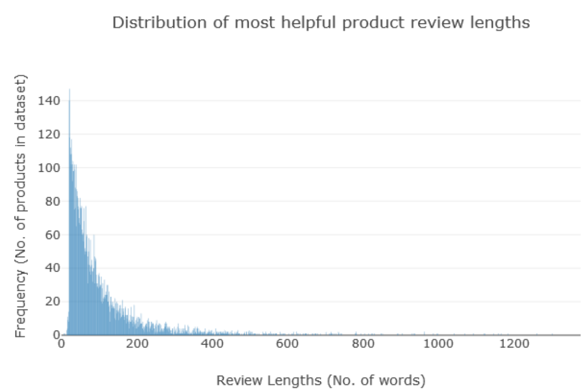


Figure 3: Review length distribution



6510.31 and an R-squared value of 0.998. The high RMSE and R-squared values signify a bad fit and multicollinearity issues. To address this problem we performed LASSO regularization in the linear regression model which lead to better RMSE and R-squared values of 2475.27 and 0.785 respectively. The regularization shrunk the coefficients of 247 features to 0 and resulted in 153 features with non-zero coefficients. We use these 153 features in our subsequent modeling and analysis experiments. Additionally, we examined the pair-plots of each of the linguistic features versus the dependent variable (estimated sales) and found that almost none of them seemed to be correlated with the dependent variable. This implied that linear regression may not be best model for our dataset.

Random Forest For the next analysis we used Random Forest Regressor on the dataset, using only the 153 features which were selected using LASSO to predict estimated sales for the bucket of similar products. We used 100 estimators for this model and this resulted in further improvement of the RMSE and R-squared values to 1475.58 and 0.83 respectively.

Analysis of Results

We found that 143 out of the 153 features selected by LASSO regression were empath features. Out of these 81

Figure 4: New description distribution

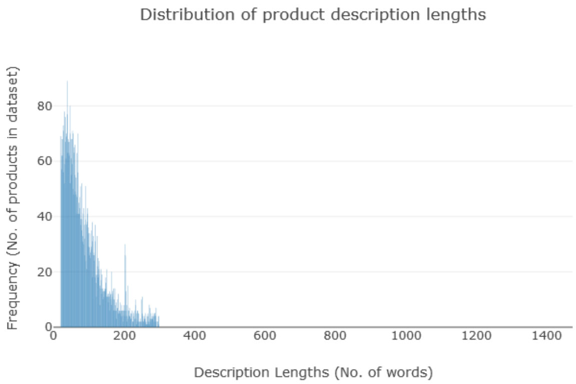
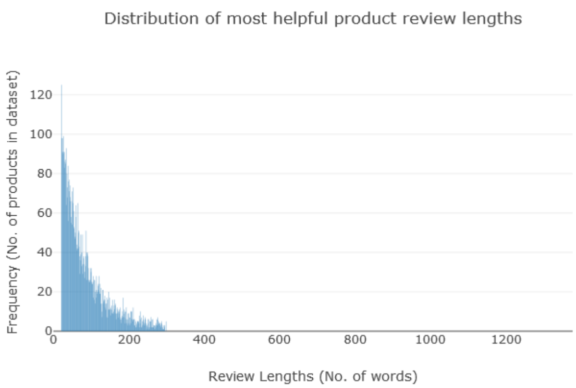


Figure 5: New review distribution



were associated with the most helpful product review and 62 were associated with the product description. Six of the remaining 10 features were associated with the most helpful product review and the remaining were associated with the product description. Some other observations regarding the features selected by LASSO are as follows:

- The readability index of reviews as well as descriptions are good predictors of sales.
- Positive, Negative and Compound sentiments in product reviews as well as descriptions are good predictors of sales.
- Politeness scores of both product review and description are good predictors of sales.

These results indicate that linguistic aspects of both seller and crowd/buyer controlled features are good predictors of sales of a product.

Conclusion and Future Work

In this pilot study we explored what are the important linguistic features in seller and crowd/buyer controlled texts as-

Figure 6: Similar products

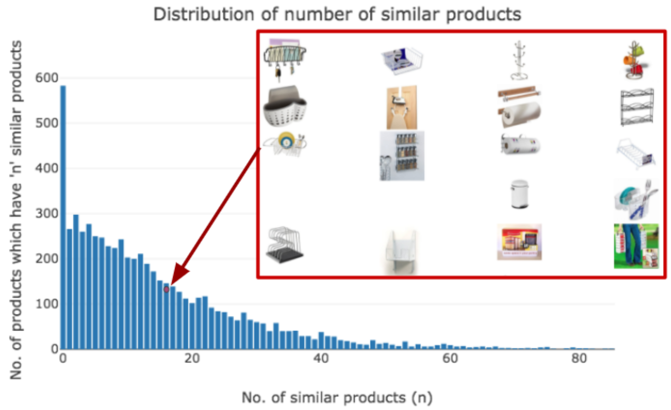
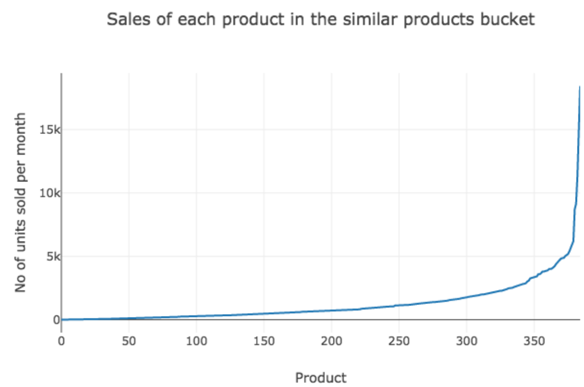


Figure 7: Distribution of sales in the bucket of similar products being analyzed



sociated with products in online marketplaces. The initial results seem to suggest that there is some degree of association between such features and product sales. We emphasize that we spent a large amount of time in the data collection and pre-processing steps due to various challenges encountered in these steps. As a consequence, we weren't able to perform a large-scale and well structured analysis on the dataset.

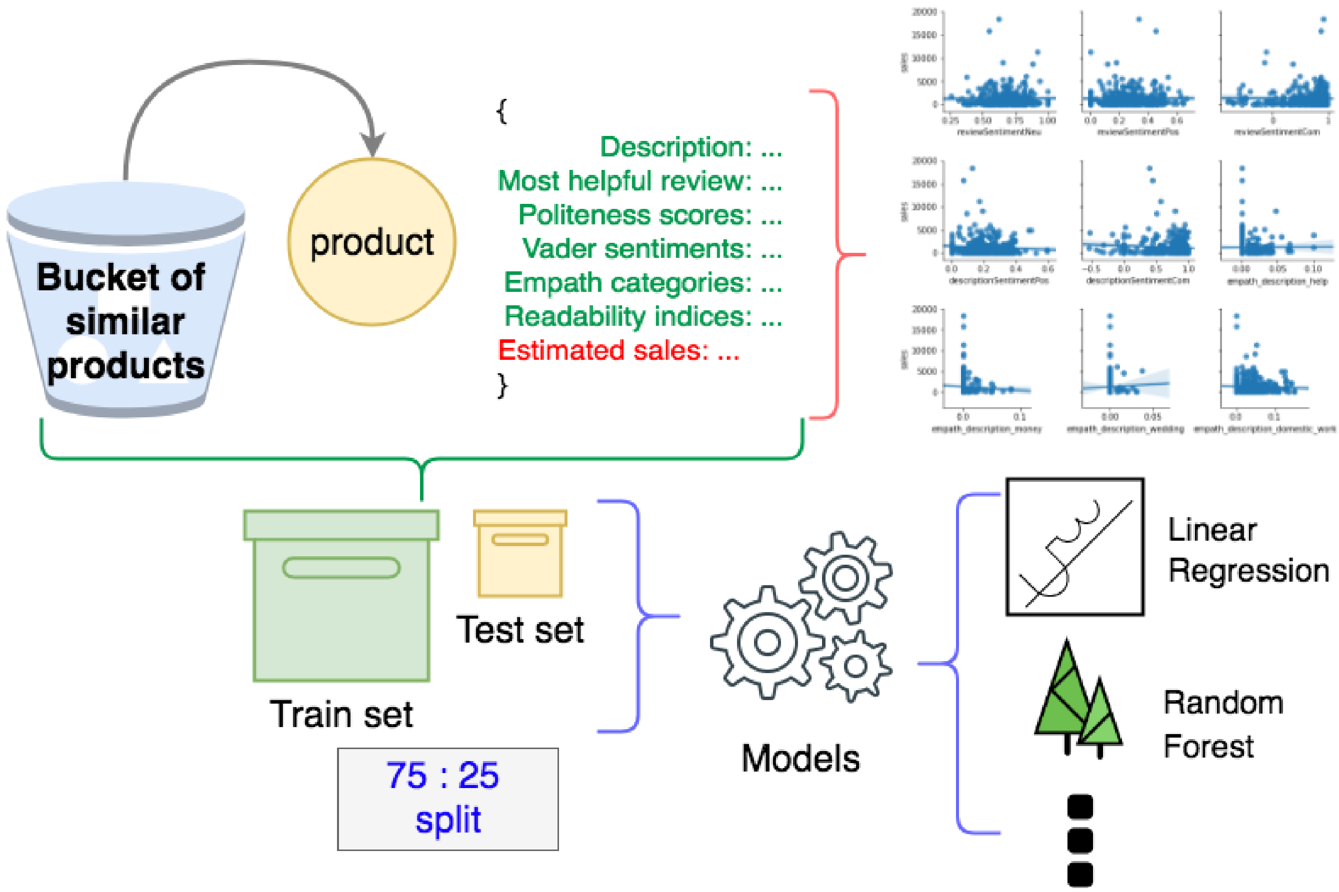
For our next steps, we intend to experiment with more predictive models such as bag of words and deep neural networks and test more controls like: review ratings, lengths of descriptions and reviews and other product metadata. We also intend to repeat these experiments across multiple buckets of similar products and different product categories. This would allow us to report more general patterns and answer our research questions with greater confidence.

References

[1] Amazon.com. Faq: What does the sales rank mean?, 2018.

[2] Cristian Caraballo and Richard Addo Hanna Kim. Predicting the helpfulness of amazon reviews.

Figure 8: Data Analysis Steps



[3] E.E. Gilbert C.J. Hutto. Vader: A parsimonious rule-based model for sentiment analysis of social media. 2014.

[4] Dan Jurafsky Jure Leskovec Christopher Potts Cristian Danescu-Niculescu-Mizil, Moritz Sudhof. A computational approach to politeness with application to social factors. 2013.

[5] Michael S. Bernstein Ethan Fast, Binbin Chen. Empath: Understanding topic signals in large-scale text. 2016.

[6] Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

[7] Anindya Ghose, Panagiotis Ipeirotis, and Arun Sundararajan. Opinion mining using econometrics: A case study on reputation systems. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 416–423, 2007.

[8] Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A Smith. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 293–296. Association for Computational Linguistics, 2010.

[9] UCSD Julian McAuley. Amazon review data, 2014.

[10] Gilad Mishne, Natalie S Glance, et al. Predicting movie sales from blogger sentiment. In *AAAI spring symposium: computational approaches to analyzing weblogs*, pages 155–158, 2006.

[11] Susan M Mudambi and David Schuff. Research note: What makes a helpful online review? a study of customer reviews on amazon. com. *MIS quarterly*, pages 185–200, 2010.

[12] Hulisi Öğüt and Bedri Kamil Onur Taş. The influence of internet customer reviews on the online sales and prices in hotel industry. *The Service Industries Journal*, 32(2):197–214, 2012.

[13] Reid Pryzant, Young-joo Chung, and Dan Jurafsky. Predicting sales from the language of product descriptions. 2017.

[14] Jungle Scout. Jungle scout free sales estimator, 2018.

[15] Dai Wei, Peng Geng, Liu Ying, and Li Shuaipeng. A

prediction study on e-commerce sales based on structure time series model and web search data. In *Control and Decision Conference (2014 CCDC), The 26th Chinese*, pages 5346–5351. IEEE, 2014.

[16] Zhu Zhang and Balaji Varadarajan. Utility scoring of product reviews. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 51–57. ACM, 2006.

[17] Kui Zhao and Can Wang. Sales forecast in e-commerce using convolutional neural network. *arXiv preprint arXiv:1708.07946*, 2017.

[18] Feng Zhu and Xiaoquan Zhang. Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of marketing*, 74(2):133–148, 2010.