Topic Modeling Amazon Product Reviews

Hathaway Zhang*
Leeds School of Business, Boulder, CO, 80302, U.S.

This report analyzed meta-data about Amazon products that are in categorized as *Clothing, Shoes* & *Jewelry* and their corresponding product reviews. More specifically, this report focused on the information of Timberland. Within the analysis, detailed marketing insights were given through performing topic modeling and data clustering.

Nomenclature

ASIN Amazon Standard Identification Number
LDA Latent Dirichlet Allocation Clustering

I. Topics

For this project, I took a close look at the product information about Timberland and performed topic modeling based on the data I extracted. Table 1 demonstrated the most popular topics I found through **k-means clustering**. I performed the clustering for multiple times and finally chose 33 clusters for the optimal result. Even though some topics seemed duplicated, this model did the best job of covering most key points and reveling some defects of the products. Since the majority of my topic models are positive reviews, I segmented the raw data of Timberland based on sentiment and generated another models for negative reviews. Table 2 visualized the top 30 topics after filtering the original data.

0: steel toe boots work	17: husband loves boots great
1: good price boots fit	18: shoe great comfortable shoes
2: day feet hours work	19: pair boots years second
3: right box boots comfortable	20: wide width feet boots
4: warm snow boots feet	21: waterproof boots boot great
5: big size run little	22: color great brown boots
6: boots great comfortable fit	23: narrow wide size boots
7: size big smaller half	24: product great good quality
8: stars five four good	25: boots good comfortable great
9: work boots great boot	26: nice boots comfortable fit
10: son loves boots great	27: small size run fit
11: 34 boots boot great	28: sandals b005fi1ve6 great sandal
12: best ever boots owned	29: socks thick wear fit
13: gift christmas loves bought	30: shoes great comfortable pair
14: comfortable like boots great	31: boots boot fit calves
15: love boots great fit	32: poor quality boots shoes
16: wallet cards pocket good	

Table 1. Top 33 topic models

*hathawayzhang@gmail.com

0: bungee close shoe place	15: nice boots boot really
1: large size shoes 10	16: feet hurt shoes wet
2: shoes buy product one	17: boat shoes classic shoe
3: love boots fit pair	18: sandals son old year
4: quality poor shoes product	19: wallet cards nice card
5: narrow width wide boot	20: big size run smaller
6: socks thick wear sock	21: muy de la el
7: five stars love excellent	22: stars four three nice
8: tight little fit size	23: wide width enough boots
9: pair years second another	24: color picture brown wrong
10: size smaller small ordered	25: loves husband daughter boots
11: good shoes quality product	26: 34 boots little like
12: boot boots size foot	27: boots fit time would
13: shoe size like lace	28: small size boots fit
14: like shoes look feel	29: made well fit boots

Table 2. Topics after segmenting

II. Topics Description

After getting my topics through k-means clustering, I performed topic classification on the documents to cluster my review data by topics. To have a better understanding of the topics, I inspected the most critical reviews and summarized their meaning along with detailed product information. The following list expounds the most representative topics in the top 33 topic models.

- Topic 0: Best fitting work boots. Lasted a long time. [Timberland PRO Men's Pitboss 6″ Steel-Toe Boots]
- Topic 4: Works perfectly in snow days. Keeping feet warm in winter conditions. [Timberland Men's White Ledge Mid Waterproof Ankle Boots & Timberland Men's Chocorua Trail Gore-Tex Mid Hiking Boots]
- Topic 5: Very comfortable but run a bit large in the sizing. Should be one size smaller. [Timberland Men's Earthkeeper Laceup Boots]
- Topic 10: As a gift for son. Stylish boots. [Timberland Men's Classic 6″ Waterproof Boot & Timberland Men's White Ledge Mid Waterproof Ankle Boots]
- Topic 11: Very comfortable and well made. Most customers said these shoes exceeded their expectations considering the price. [Timberland Men's White Ledge Mid Waterproof Ankle Boots]
- Topic 13: Give this as Christmas gift. Affordable and really love it. [Timberland Men's Classic 6″ Waterproof Boot]
- Topic 16: Gift for father/husband/boyfriend. Great quality with a lot of space. [Timberland Deep Cognac Passcase Bifold Leather Wallet & Timberland Men's Block Island Trifold Wallet]
- Topic 17: Husband loves it. Fits great and is suitable for long hours work. [Timberland Pro Men's Pitboss 6″ Soft-Toe Boots]
- Topic 20: Fits tight and narrow. Order extra wide width if you have wide feet. [Timberland Men's White Ledge Mid Waterproof Ankle Boots]
- Topic 21: Good for trails and hiking. Waterproof with long lasting sole. [Timberland Men's White Ledge Mid Waterproof Ankle Boots & Timberland Men's Chocorua Trail Gore-Tex Mid Hiking Boots]
- Topic 23: Beautiful looking brown but color is slight different from the picture. [Timberland Men's Earthkeeper Original Marron, Timberland Men's Piper Cove Fg Boat Shoe & Timberland Men's Classic 6″ Waterproof Boot]

- Topic 28: Super comfortable and durable sandals. Perfect summer sandals for little. [Timberland Adventure Seeker Two-Strap Sandal for Kids & Timberland Men's Altamont Fisherman]
- Topic 29: Keep its shape after washing. Durable and super comfortable. [Timberland 3-Pk Crew Socks & Timberland Men's 4 Pack Outdoor Leisure Crew Assorted Colors]
- Topic 31: Have narrow calf. Fit tight on the calf. Leather is thick and gets soften after wearing. [Timberland Women's Savin Hill Tall Boots]

My second topic models put more attention on negative sentiment; thus, it contains more complaints and negative reviews. The following list describes some interesting points in Table 2.

- Topic 0: Ordered for no-tie ease but received one with tie. Different color and kids don't like it. [Timberland Trail Force Bungee Sneaker]
- Topic 1: Way too large. Need to order a size down. [Timberland Men's Piper Cove Fg Boat Shoes & Timberland Men's Earthkeeper Laceup Boots]
- Topic 2: Comfortable but broke quickly. Used to last for years. [Timberland Men's White Ledge Mid Waterproof Ankle Boots]
- Topic 4: Very poor quality and doesn't meet the quality standard for this manufacturer. [Timberland Men's White Ledge Mid Waterproof Ankle Boots & Timberland Ekkiawahby]
- Topic 6: Comfortable and thick socks. [Timberland 3-Pk Crew Socks]
- Topic 9: Repurchase. Have this shoes 2/6/10 years before and purchase again. [Timberland Men's White Ledge Mid Waterproof Ankle Boots & Timberland Men's Chocorua Trail Gore-Tex Mid Hiking Boots]
- Topic 16: Made with quality material but very hard with poor fit. Had blisters all over. [Timberland Pro Men's Pitboss 6" Soft-Toe Boots]
- Topic 21: Reviews in Spanish.
- Topic 24: Lose color after wearing. Looks darker than the picture online. [Timberland Men's Classic 6" Waterproof Boot]
- Topic 25: My daughter/son loves this sandal. Solid and durable. [Timberland Mad River 2-Strap sandal]
- Topic 27: Wear socks to avoid ankle rubs. Gets better as time goes on. [Timberland Pro Men's Pitboss 6" Soft-Toe Boots]

III. Preprocessing Steps

The preprocessing section involves two steps. First, filtering the metadata by a certain brand and extracting a list of ASINs along with corresponding reviews. Second, training Word2Vec model with the data and segmenting data by sentiment. To begin with, I obtained data which is categorized in clothing, shoes & jewelry. After loading this JSON file into a dictionary, I found this data is still too large to perform a specific topic modeling. In this case, I focused on the meta-data of Timberland and identified the ASINs associated with products from Timberland. I also extracted the reviews that match those ASINs for further analysis. With this information, I was able to perform my first data clustering.

Using Word2Vec to filter negative sentiment is more complicated. In order to train Word2Vec, I split the sentences into words and divided the original data into 30% and 70%. After training with 70% data, the Word2Vec model can tell the relationship between words. For example, if I input `wont` into the model, Word2Vec will return `("won't", 0.926), ('cant', 0.849), ("wouldn't", 0.830), ('cannot', 0.813)` automatically. This information will be able to help us identifying negative reviews by searching negative attitude words. In this case, I built a sentiment classifier through creating a list of positive words and a list of negative words. I also inspected the word list to avoid including some common words like `him` and `had`. In

general, I have options to segment the data: (1) only keeping reviews that contain any words in the negative word list, (2) removing reviews that contain any words in the positive word list, and (3) removing reviews that match both conditions. After three attempts, I found option (2) had the best performance. Keeping reviews that contain negative words seem to be the most straightforward approach; however, since most reviews are positive, the reviews would still contain many positive words after clustering. It is not obvious to discover negative topics in this way. Option (2) and option (3) almost have the same result, so I prefer option (2) to have more flexibility. With this information, I was able to perform my second topic clustering.

IV. Methodology

In this project, I used k-means clustering for both topic models. For the topic modeling section, I used Scikit Learn package to enable the functionality of k-means. A tf-idf-transformer is applied to the words matrix through `TfidfVectorizer`. Then, I performed k-means cluster through `KMeans` function and printed out the topics. Since I used k-means clustering, I need to decide the number of clusters at the first stage and select the initial clustering centers randomly. Initially, I chose 50 clusters and found a large amount of duplicated words, such as "comfortable," "great," and "good." To eliminate this situation, I picked a relatively small number to test the model; however, I found many helpful keywords are missing while duplicated words still existing. After several attempts, I decided to keep 33 clusters which contains relatively less similar topics while keeping most distinct topics. I applied the same methodology in the second clustering.

Fast and efficiency are the major reason that drove me to choose k-means instead of LDA. However, the advantages of LDA are not neglectable. LDA is powerful in document clustering as it helps semantic mining and information retrieval. Moreover, LDA can assign a document to a mixture of topics and provide a helpful interactive clustering plot. Because my target documents are reviews from multiple buyers, I preferred k-means over LDA since the relationship between each sentence are less logical in my case. K-means can partition the data in distinct clusters, which could help me identify positive and negative reviews more efficiently.

V. Marketing and Product Insights

V.A. Attributes That People Like & Dislike

We need to accentuate the factors that are impressed and praised by customers in our advertisements. In this case, positive reviews are critical in advertising our products. Topic 0 in Table 1 shows that Timberland PRO Men's Pitboss 6" Steel-Toe Boots is a top choice as boots for long hours working. "Comfortable" and "Durable" are the key characteristics that have been mentioned multiple times in Topic 0, 2, 3, 6, 9, 14, and 25. Not only for work and casual wear, customers also like wearing Timberland's boots for special occasions, such as hiking and trails. According to Topic 21, most people agree that Timberland's boots are definitely "waterproof" and suitable for outdoor-sporting. Many buyers also agree that Timberland's boot is good for snow condition. The boots are able to keep feet warm and isolate cold. Across all the topics, people tend to think that most Timberland shoes have great quality and exceed their expectation considering the price. In Topic 13, a lot of people said Timberland is an affordable present. Moreover, Timberland's boots are highly praised for their style and color. Topic 10 tells that Timberland Men's Classic 6" Waterproof Boots are very stylish. Topic 22 also shows that Timberland's signature brown is very popular. Besides shoes, the wallet of Timberland is also a good choice of gift and has good quality with a lot of space.

Table 2 generates more negative reviews and tells attributes that people dislike. According to Topic 1 in Table 2 and Topic 20 in Table 1, improper size and fit is the most common issue of Timberland's boots. Timberland Men's White Ledge Mid Waterproof Ankle Boots is tight and narrow. Customers need to order extra width if they have wide feet. Timberland Men's Piper Cove Fg Boat Shoes and Timberland Men's Earthkeeper Lace-up Boots, on the other hand, are too large for most people. Besides, many people complain that the boots used to have good quality but their recent purchase are not good enough. Topic 16 suggests that the boots are very hard and cause blisters all over. Furthermore, Topic 29 in Table 1 shows Timberland Women's Savin Hill Tall Boots have thick leather and narrow calf. Boots fit relatively tight on the calf. Other than fit, Topic 24 also tells the boots start losing color after wearing, and the actual color is different from the picture.

V.B. Purchase Occasions

Topic 9 and 13 in Table 1 show that customers often buy Timberland's products as a gift, particularly for their sons and husbands. The purpose of purchasing is changing with recipients and buyers. Parents would like to purchase Timberland Classic 6" Boot for sons as birthday gifts. Wives prefer to purchase Timberland White Ledge Mid Waterproof Ankle Boots and Timberland Chocorua Trail Hiking Boot for husbands as a birthday present and anniversary present. Girlfriends like to purchase Timberland Classic 6" Boot for boyfriends as an anniversary present. Men would also purchase Timberland's boots for work or repurchase because of the good quality. For those who purchase Timberland's products for daily use, these boots are an excellent selection for work, outdoor activities, and winter snowshoes substitution.

V.C. Product Improvement and Pricing Suggestions

According to the topics above, people tend to agree that Timberland's products are affordable and worth the price. However, I don't recommend Timberland to raise the price. Base on the review, most people are purchasing Timberland because it has great quality like many other famous brands while offering a relatively lower rate. At this stage, Timberland needs to maintain its reputation and stops selling products that haven't meet its quality standard. For some products that have unusual fit, Timberland should emphasis special fit instruction on the product description page.

V.D. What's Not in the Data

The data doesn't contain too much diversity. According to the review data, most products are hiking/ankle boots and boat shoes. As a famous manufacturer focusing on footwear, Timberland also sells apparel, such as clothes, watches, sunglasses, and other leather goods. Nevertheless, our data only contains reviews of boots and small leather goods. If we could obtain more feedback on all the products, Timberland might be able to improve customers' satisfaction in all categories and introduce more product lines.

V.E. Further Recommendation

The first topic models revels several popular Timberland products: Timberland men's white ledge mid waterproof ankle boots, Timberland men's Chocorua trail gore-tex mid hiking boot, Timberland pro men's pit boss 6" steel-toe boots, Timberland men's earthkeeper lace-up boots, Timberland men's classic 6" waterproof boot, Timberland deep cognac passcase bifold leather wallet, Timberland adventure seeker two-strap sandal, Timberland 3-Pk crew socks, and Timberland women's savin hill tall boots. Cross these products, we could summarize the rank of Timberland's potential customers as men ¿ boys/women ¿ kids/girls. Even though male buyers are the most competitive customers, Timberland could also advertise on female customers to expand its marketing share. Classic 6" Waterproof Boots are the most classical product of Timberland. To expand Timberland's market, the manufacturer could promote classic 6" boots for women and kids. Timberland could also improve its classic design and release some retro or vintage products. Moreover, Timberland could collaborate with other high fashion brands to release some limited version, which helps Timberland attract some luxury consumers.