Text Analysis and Classification Model Report
for Women's E-Commerce Clothing Reviews

May 12, 2021

Introduction: E-Commerce clothing has been one of the most successful business nowadays. Since the pandemic started, most individuals tend to prefer online clothes shopping to avoid human contact. Therefore, building a trustworthy online clothing website would attract more customers. To build more confidence in online shopping, most shoppers utilize the used of clothing reviews from previous customers. For example, if most reviews for a certain piece of clothing sound positive, then the customers would be more likely to buy it. However, if most of the reviews sound negative, then the customers would be hesitated to purchase the clothes. Thus, it would be easier for the customers to see the common topics based on all the reviews to make quicker decision to purchase. Therefore, some questions were asked in favor of these issues including: 1) What are the most common terms presented in the reviews? 2) Can topics and labels be generated for each review? And 3) Using a set of historical review data, can the model classify the type of reviews for future entries? This report will attempt to answer these questions using some text analysis methods including natural language processing and language classification models.

Methods: The women's e-commerce clothing reviews dataset was retrieved from *Kaggle*.com with a total of 23,500 reviews. There is one identifier column with clothing ID to identify each review from a specific piece of clothing. Other columns related to the reviewers and the review text such as age of the reviewers, title of the review, if the customer would recommend the product, number of other customers who found the review positive, product high level division, product department name, and product class name. The two most important variables are the review text and rating. The review text would be used to generate topic model; and the rating of the review will be used to generate a classification model with three category good (5-star), so-so (3-star and 4-star), and bad (1-star and 2-star) reviews. One of the biggest assumptions for this dataset would be using only English reviews.

For the topic model, the analysis will use the entire dataset to train the topic. Multiword expressions (MWEs) method will be used to find domain specific terminology that is related to women clothing reviews. Then, the reviews will be tokenized based on the domain specific terms found previously. With the tokenized review texts, topics of each review were detected using a Latent Dirichlet Allocation (LDA) model. This topic model will be evaluate using visualization such as LDA plots and word clouds. The labels of each group of topics will be labeled manually after the topic model is saved. Once a table of labels have been constructed, each review will be assigned a label for further analysis.

For a classifier model, the training and testing data will be split using a random proportion of 0.75 and 0.25, respectively. The target variable is rating group with three factor levels including good, so-so, and bad reviews. The independent variable will be the tokenized review texts. During the model selection processing, six different types of model will be generated to compare including baseline dummy classifier, Bernoulli Naïve Bayes, multinomial Naïve Bayes, SGD classifier, Tfidf transformer, and truncated SVD. Once a model has been selected, a hyper-parameter search process and eliminating marginal scores will be generated to find the best parameters for the model. The model will be evaluated using F1 scores and errors. All analysis will be done using Python with the following basic modules: `numpy`, `pandas`, `cytoolz`, `tqdm`, `spacy`, and `sklearn`.

Topic Model Results: Using the LDA model, 25 different groups of topics were extracted for the entire data set. The parameters for the model were adjusted accordingly to get a robust model that are the most relevant to the clothing reviews. 75 least frequent words were removed from the data set to keep the consistency and quality of the words. In addition, 125 most frequent words were removed from the data to eliminate filter and stop words, see Table 1 for a list of topics that were generated from the clothing reviews. The labels from Table 1 were manually entered based on the words of each bucket.

Words	Label
off no buttons only worth price sale quality side going	quality
lbs perfectly 5 5'4 for reference 5'5 4 6 5'6 pounds	size reference
pair denim legs stretch pilcro shorts 27 26 skinny jean	bottoms
looked wanted return loved unfortunately however returned disappointed	disappointed
above right hits knee hit below inches 2 hem model	size reference
hips someone who shape curvy tall better however chest body	size reference
compliments wore many received time lots every worn already wearing	complements
jacket coat warm vest sleeves over cardigan cozy wool winter	outerwear
gorgeous unique lovely design absolutely feminine feel quality makes print	complements
sleeves blouse arms shoulders tops boxy fitted longer body nicely	tops
been few now worn years purchased after wearing always two	quality
model person picture online photo what shown better exactly pictured	as described
tight arms shoulders around off chest bust arm across shoulder	tight
fitting loose feel though good cotton want its form thin	thin
fall leggings summer boots wearing casual spring worn dressed pair	bottoms
2 sizes 0 order regular xxs both though smaller try	tight
6 8 m 10 4 s normally l 12 bust	size reference
recommend highly definitely go wardrobe comfy casual piece easy any	recommend
washed dry wash after washing clean only shrink hand shrunk	quality
without being enough feel looking thick while find no heavy	quality
blue white red gray green navy pink purchased both orange	color
saw online try loved went thought sale first decided fell	complements
reviews reviewers said agree another mentioned others based reviewer read	recommend
nicely front side your body right enough around high drapes	complements
underneath bra sheer through tank cami under need white slip	thin

Table 1. Table of topics and labels generated from the entire dataset of clothing reviews.

Some of these words are obvious to assign a label. For example, words like “jacket”, “coat”, “warm”, “vest”, “sleeves”, “cardigan”, “cozy”, “wool”, and “winter” are obviously belonging to an outerwear or winter type of clothing. However, others are more complicated to assign a single bucket. For instance, words like “wanted”, “return”, “loved”, “unfortunately”, and “disappointed” are from the same group. But some words are positive, and some words are negative.

After the topic model, a word cloud was generated for a selected topic. Figure 1 depicts the most 500 frequent words from a topic from Table 1. The bigger the words are the more frequent they appear. For example, words like sleeves, blouse, arms, fitted, shoulders, and tops are most likely belonging to a label related to tops. This indicates that when people talk about a topic related to tops, they will most likely mention sleeves, blouse, arms, fitted, shoulders, and tops in their reviews.

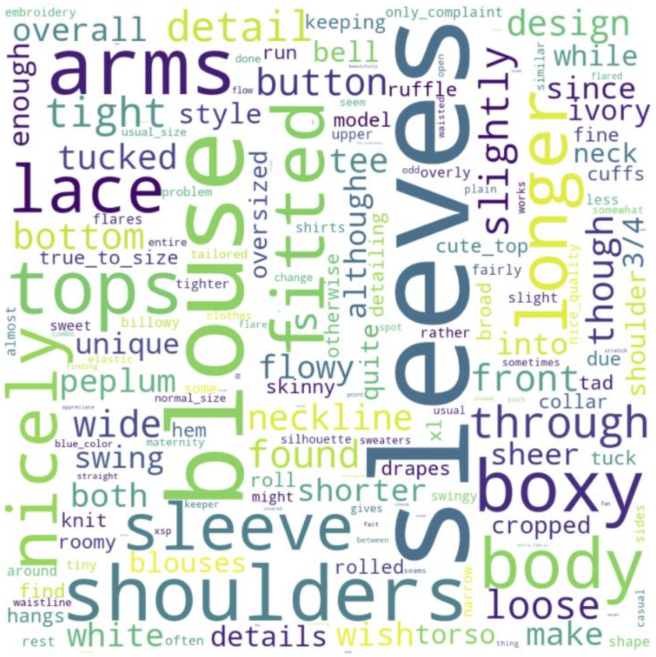

Figure 1. Word cloud of the top 500 most frequent words from topic “tops”.

Once we have completed the topic model development, each label was applied to an appropriate review from the data set. The final topics of the reviews have the following count:

- Complements: 24,937
- Quality: 23,085
- Size reference: 17,336
- Color: 5,816
- As described: 4,997
- Recommended: 10,010
- Disappointed: 9,464
- Outerwear: 3,743
- Bottoms: 9,755
- Thin: 10,131
- Tight: 8,839
- Tops: 4,707

Most of the reviewers complemented the products that they bought. This makes sense because most of their rating are 5-star rating. In addition, the second most frequent topic that reviewers mentioned are about the quality of the products. Following that, lots of reviewers also recommended the product to other customers. On the other hand, topics like outerwear, tops, and color have the lowest frequency in the dataset. This could mean these topics are too specifics to a type of clothing.

This model is beneficial to both the shops/online shops that sell the clothes as well as the customers who are looking to purchase cloths online. With the key words for each review, employers can select the best and trending clothes to accommodate all customers. In addition, employers would have a better idea of the quality of clothes based on customers’ experience. This way, they can order more of the good quality clothes and eliminate bad quality clothes. In terms of benefits for the customers, women will be able to shop faster by reading the most frequent words that previous customers say about a piece of clothing. For example, if the most frequency words for a pair of jeans are "stretch", "true to size", "good quality", and "nice color", then the customers would trust the product more.

Classification Model Results: Once we have the reviews labeled, a classifier model was generated to predict the rating bucket for the review. The purpose of predicting the rating topic for the reviews is to see is a review would be a good review, bad review, or so-so review. Six different types of models were generated to compare their performance. This way, we can pick the best model based on the baseline model. Table 2 shows a comparison of the F1 scores for each model including baseline dummy classifier, Bernoulli Naïve Bayes, multinomial Naïve Bayes, SGD classifier, Tfidf transformer, and truncated SVD.

	F1 Score		
Model Type	**Accuracy**	**Macro Average**	**Weighted Average**
Baseline Dummy classifier	0.55	0.24	0.39
Baseline Bernoulli Naïve Bayes	0.68	0.56	0.67
Baseline Multinomial Naïve Bayes	0.72	0.62	0.71
Baseline SGD classifier	0.69	0.60	0.68
Baseline Tfidf transformer	0.72	0.63	0.71
Baseline Truncated SVD	0.68	0.48	0.64
Optimized Tfidf transformer	0.71	0.64	0.71
Margin Optimized Tfidf tranformer	0.78	0.78	0.79

Table 2. Table of F1 scores for all baseline models and optimized models.

The best models that have highest F1 scores are multinomial Naïve Bayes and Tfidf transformer. These two models have the same accuracy scores of 0.72 and weighted average of 0.71. However, model Tfidf transformer has a higher macro average F1 scores of 0.63 compared to 0.62. The differences are not large, but we will use Tfidf transformer model as a final model. Using this model, a hyper-parameter search process was done to find the best parameters for Tfidf transformer model that would give us the highest mean scores. Unfortunately, the optimized model would generate the same information as baseline model. With a Wilcoxon p-value of 0.99, the baseline model and optimized model are not different from each other. However, we will keep the optimized model as a final model.

To improve F1 scores of the optimized Tfidf transformer model, the margin of the accuracy scores were calculated. Setting the margin can help the model to be more accurate by using more quality data. The accuracy margin was set to be grader than 0.2 for both testing and prediction values. After setting the margin, F1 scores were able to increase to 0.78 for accuracy, .78 for macro average, and 0.79 for weighted average, see Table 2.

The model was evaluated using an error analysis. There were only 16 good reviews that were misclassified as bad reviews. For example, "A nice rayon peasant shirt – like the sleeve length, good for a fuller figure, covers all the bad sports and very comfortable." This review can be misleading to the model because there are some negative words like "peasant" and "bad sports". The model can only have nearly 80% if accuracy, so it cannot predict perfectly. In addition, there were only 91 bad reviews that were misclassified as good reviews. Thus, the number of misclassifications for good and bad reviews are low, which indicate the model is doing quite well.

This classification model also helps both the employers and the customers. Employers can classify the differences between a good and a bad review. With the classification, the shop owners can filter their website to show good reviews first to attract more customers. In addition, this model would also help the customers to filter out the bad or good reviews. For example, shoppers sometimes wonder what are the worse thing that could go wrong with this dress or what are the bad experience that previous customers have with these jeans. Therefore, customers can save more time and filter to read only good or bad review instead of each review at a time.

Future Research:

Both the topic model and classification model can be improved to have more accuracy results. If there are more time, more tunning for topic model would be required to find the best values for removing the most and least frequency words. In addition, sentiment analysis can also be done to improve F1 scores of the model.

References:

Women's E-Commerce Clothing Reviews. (2021). Retrieved 7 May 2021, from https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews