Aspect Based Sentiment Analysis for Fashion Products

Springboard Data Science Career Track

Introduction

The fashion industry is characterized by rapidly changing trends and consumer preferences. Consumers express their emotions through product reviews and social networks such as Instagram, Facebook and Twitter. The ability to detect consumer satisfaction from user-generated content online can be very useful for fashion brands to swiftly respond to customer needs accordingly.

Sentiment analysis is a natural language processing (NLP) technique used to distinguish positive and negative opinions from textual data. It is often adopted by businesses for social listening and monitoring brand and product sentiment in real time.

While sentiment analysis can help identify the sentiment behind an opinion or statement, there might be several aspects that have triggered the identified sentiment. Aspect Based Sentiment Analysis (ABSA) is a technique that takes into consideration the underlying aspects and identifies the sentiment associated with each aspect.

Reviews for fashion products can incorporate sentiments towards different aspects such as design, quality, sizing, etc.. With the implementation of ABSA, fashion brands will be able to have a more in-depth look into customer feedback and specific needs of the market.

Problem Statement

The main objective of this project is to develop a novel ABSA pipeline for identifying aspects from fashion product reviews and predicting the sentiment toward each aspect.

The pipeline includes three main components: text preprocessing, aspect extraction, and sentiment prediction.

Typical ABSA requires labeled data containing aspect terms and aspect categories for each statement along with its sentiment score. Since there is no readily available labeled data for the fashion industry, the aspect extraction step utilizes word embedding methods in an unsupervised manner.

For the sentiment prediction step, we use the supervised approach and train a classification model to predict positive, neutral, or negative sentiment for each aspect extracted.

Fashion brands can use this novel ABSA pipeline to extract aspects from product reviews and predict the sentiment toward each aspect, which allows them to acquire more targeted and actionable insights for a fast moving industry.

Data Source

Amazon Fashion is the apparel department on Amazon that focuses on fashion products and services. The dataset contains product reviews and metadata from Amazon Fashion, including reviews spanning 2002 - 2018. This dataset includes reviews (ratings, text, helpfulness votes, etc.) and product metadata (descriptions, category information, price, brand, and image features). The dataset was retrieved from Amazon Review Data (2018). The review dataset contains 883,636 rows and 12 columns. The product dataset contains 186,627 rows and 18 columns.

The variables of the product review dataset include:

- asin - ID of the product, e.g. 0000013714
- image - images that users post after they have received the product
- overall - rating of the product
- reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B
- reviewerName - name of the reviewer
- reviewText - text of the review
- reviewTime - time of the review (raw)
- style - a dictionary of the product metadata, e.g., "Format" is "Hardcover"
- summary - summary of the review
- unixReviewTime - time of the review (unix time)
- verified - whether or not the review is verified, e.g. True, False
- vote - helpful votes of the review

Data Wrangling

Duplicates and Missing Values

When a customer reviews a product, the review is copied over to other product options under the same parent product, which results in duplicated reviews in the dataset. For example, a t-shirt listing (parent product) can have multiple color options such as black, white and blue (child products). When a customer leaves a review for the white color option (child product), the review is copied over to the black and blue color options

(child products) because they are under the same t-shirt listing (parent product). We remove the duplicated reviews and are left with 853,805 rows.

We also remove rows containing null values for the reviewText column that provide no value for the study. After removing duplicate rows and rows with missing review text, we are left with 852,636 rows.

Natural Language Processing

The initial steps of natural lagrange processing contain:

- Remove urls using regular expressions
- Remove html tags using the BeautifulSoup package
- Remove extra white spaces
- Remove accented characters using the Unidecode package
- Expand contractions using the Contractions package
- Convert all characters to lowercase

After the initial steps, tokenize the review text using the spaCy package and the trained pipeline en_core_web_lg, then perform below steps:

- Remove special characters
- Lemmatize tokens to their base forms
- Extract nouns and adjectives for the aspect extraction step

We keep stop words, punctuations and numbers because they may provide additional information for the sentiment prediction step. For example, if the review says “I am not happy with my purchase” and we remove the stop word *not*, the resulting sentence becomes “I am happy with my purchase”, which completely changed the sentiment of the sentence.

As the last step of the data wrangling, we remove rows with null values for the processed reviews after the above steps and are left with 852,589 rows.

Exploratory Data Analysis

Review Word Clouds

To have a better intuition of the review text data, we remove stop words and punctuations, group reviews by products and create word clouds for the top 5 products with the longest reviews. From the word clouds (Figure 1), we can clearly see what are the most frequently used words in the reviews for a specific product.

For the shoe insole, the words with the highest frequencies are *foot, shoe* and *insole*. Some other words that are frequently used in the reviews are *good, support, great* and *arch*. For the yoga pants, the words with the highest frequencies are *fit, love, great* and *pant*. Some other words that are frequently used in the reviews are *color, wear* and *pair*.

Figure 1. Word Clouds for a Shoe Insole product (Left) and a Yoga Pants product (Right)

Review Word Count Distribution

We are also interested in the distribution of the review word count since it can provide intuitions for the LSTM model for sentiment prediction. From the histogram below (Figure 2), we can see that the majority of the reviews have a word count of 50 or less, while extreme outliers can have word counts of up to 2250.

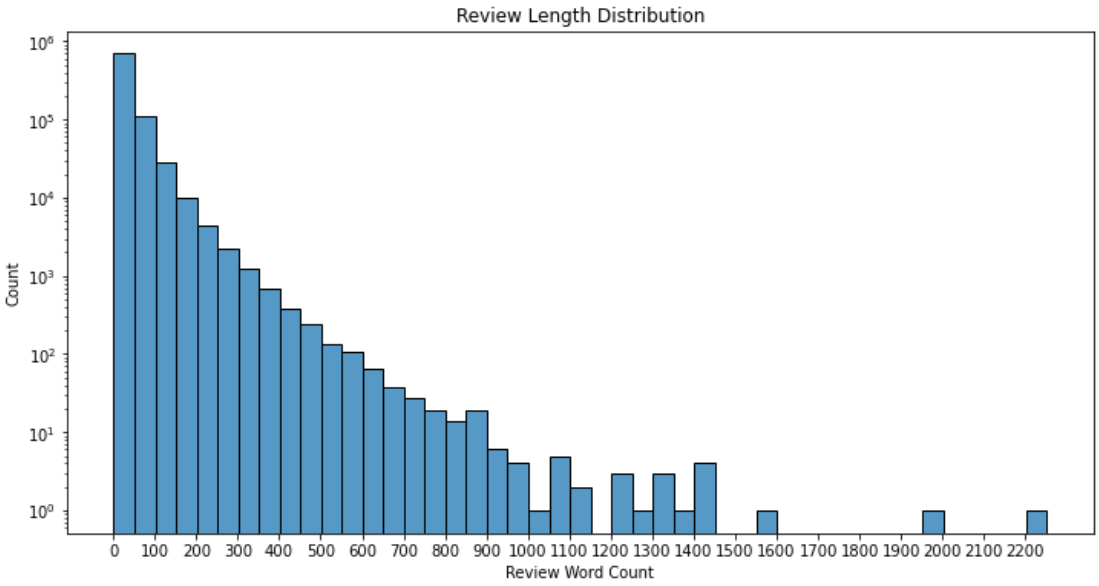

Figure 2. Histogram for Review Word Count

Review Sentiment Distribution

To have a general understanding of the distribution of the review sentiment, we map reviews with ratings of 1 or 2 stars as Negative, 3 as Neutral, 4 or 5 as Positive.

We can see from the pie charts (Figure 3) of the sentiment distribution that Amazon Fashion only received a total of 5 reviews in 2002, among which 80% (4 reviews) are positive. Fast forward to 2018, Amazon Fashion received 54,304 reviews, among which 72.21% (39,212 reviews) are positive.

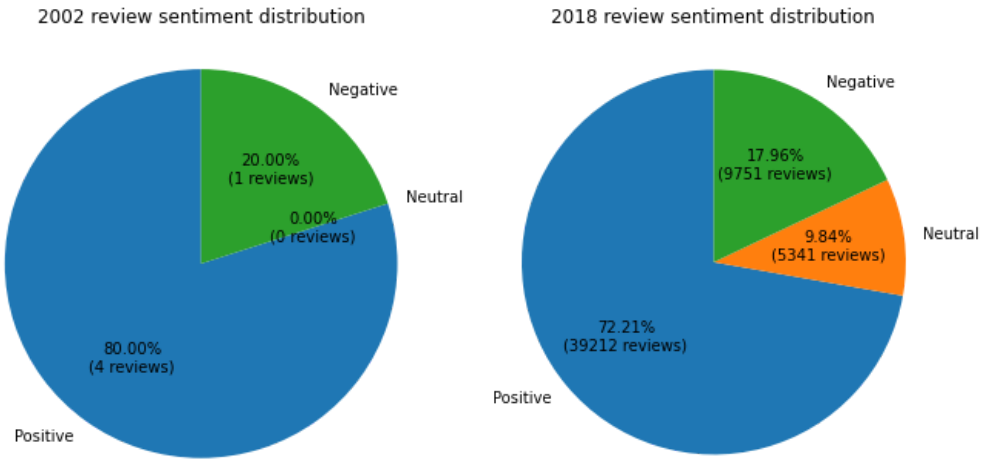

Figure 3. Pie Charts for Review Sentiment Distribution for Year 2002 (Left) and 2018 (Right)

We are also interested in any trends or changes in the review sentiment distribution over the years. From the bar chart below (Figure 4), we can see that the largest lift in the positive review percentage took place between 2005 and 2006, an increase from 62% to 76% to be specific.

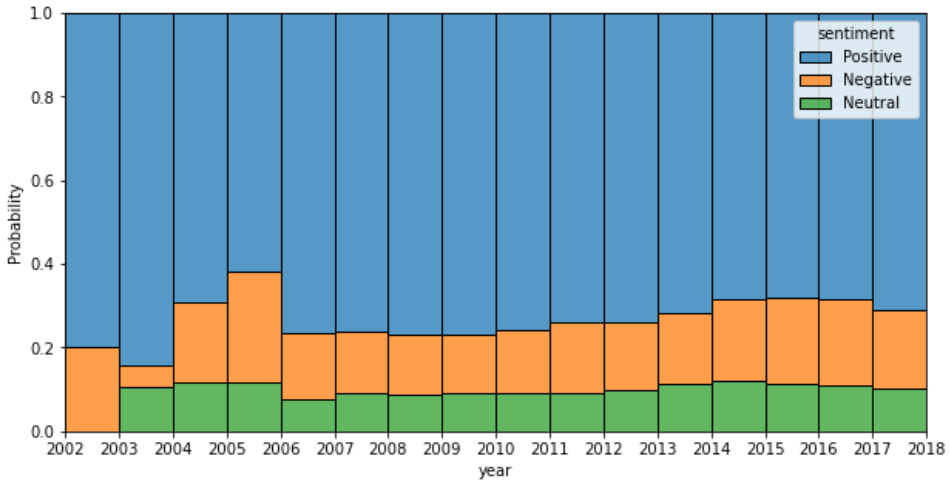

Figure 4. Stacked Bar Chart for Review Sentiment Distribution from 2002 to 2018

Hypothesis Testing

We are interested in testing if the changes in the positive review percentage is due to chance. We select 5% as the significance level α and use two proportions z-test to test our hypothesis that for each year from 2003 to 2018:

H_0: The positive review percentage for the current year is lower than or equal to the previous year ($p_{year} \leq p_{year - 1}$)
H_1: The positive review percentage for the current year is higher than the previous year ($p_{year} > p_{year - 1}$)

The table below (Table 1) summarizes the results of the tests. The positive percentages highlighted in red indicate that it has increased compared to the previous year, and p values highlighted in yellow indicate that the result is statistically significant to reject the null hypothesis. We conclude that the positive review percentages for 2006, 2016, 2017 and 2018 are higher than their previous years.

year	review_count	positive_count	positive_pct	p_value
2002	5	4	0.800000	nan
2003	19	16	0.842000	0.411100
2004	52	36	0.692000	0.896600
2005	154	95	0.617000	0.835800
2006	326	249	0.764000	0.000400
2007	1473	1123	0.762000	0.521700
2008	1716	1319	0.769000	0.338700
2009	1956	1506	0.770000	0.463100
2010	2083	1581	0.759000	0.793400
2011	3019	2230	0.739000	0.949800
2012	7547	5578	0.739000	0.481200
2013	25118	18051	0.719000	0.999800
2014	86723	59439	0.685000	1.000000
2015	195397	132669	0.679000	0.999600
2016	306711	210145	0.685000	0.000000
2017	165984	117416	0.707000	0.000000
2018	54306	39212	0.722000	0.000000

Table 1. Two Samples Proportion Z-Test Results for 2003 - 2018

Platforms like Amazon Fashion constantly make updates on their products, features and services, which makes the root cause for the changes in the positive review percentage multifaceted and hard to determine. While finding the underlying cause is beyond the scope of this project, it is still beneficial to understand if significant changes exist for related studies on customer sentiment.

Aspect Extraction

Aspect Extraction

Typical ABSA requires labeled data containing aspect terms and aspect categories (aspects) for each statement along with its sentiment score. Since no labeled data is available for the fashion industry, we use an unsupervised approach to extract aspects from each review.

We first provide a list of six aspects the fashion industry is commonly interested in: *color*, *design*, *material*, *price*, *quality*, and *sizing*. The ability to understand the customers' sentiment towards these specific aspects of their product can help fashion brands gain deeper insights and take more targeted actions.

To extract aspect terms and their corresponding aspects from the reviews, we follow below steps:

- Use the part-of-speech tagging technique to extract nouns and adjectives from the text corpus. This is done as part of the natural language processing.
- Embed each extracted word and each aspect from the given list into a vector format using the pre-trained word2vec model, compute and store the semantic similarity (cosine similarity) between each word-aspect pair.
- Examine the word-aspect pairs with the highest semantic similarities, and select the similarity threshold for the aspect extraction step of the ABSA pipeline.
- Extract word-aspect pairs with semantic similarities higher than the threshold, words in these pairs are our aspect terms.

For the word embedding step, another approach besides using the general pre-trained model is to train a domain specific model. For comparison, we train a domain-specific fasttext model to embed the extracted words and aspects, and examine the word-aspect pairs with high semantic similarities. The domain specific model is able to capture some meaningful word-aspect pairs such as ('teal', 'color') and ('stretchy', 'material'), but does not outperform the pre-trained word2vec model.

Top Aspect Term Visualization

Aspect terms are closely relevant to the given aspect or might be the aspect itself. We consider the top 15 most frequent aspect terms for each aspect, and visualize them using PCA as the dimensionality reduction technique.

From the scatter plot below (Figure 5), we can see that top aspect terms for the same aspect are located in close proximity to each other in the word embedding space.

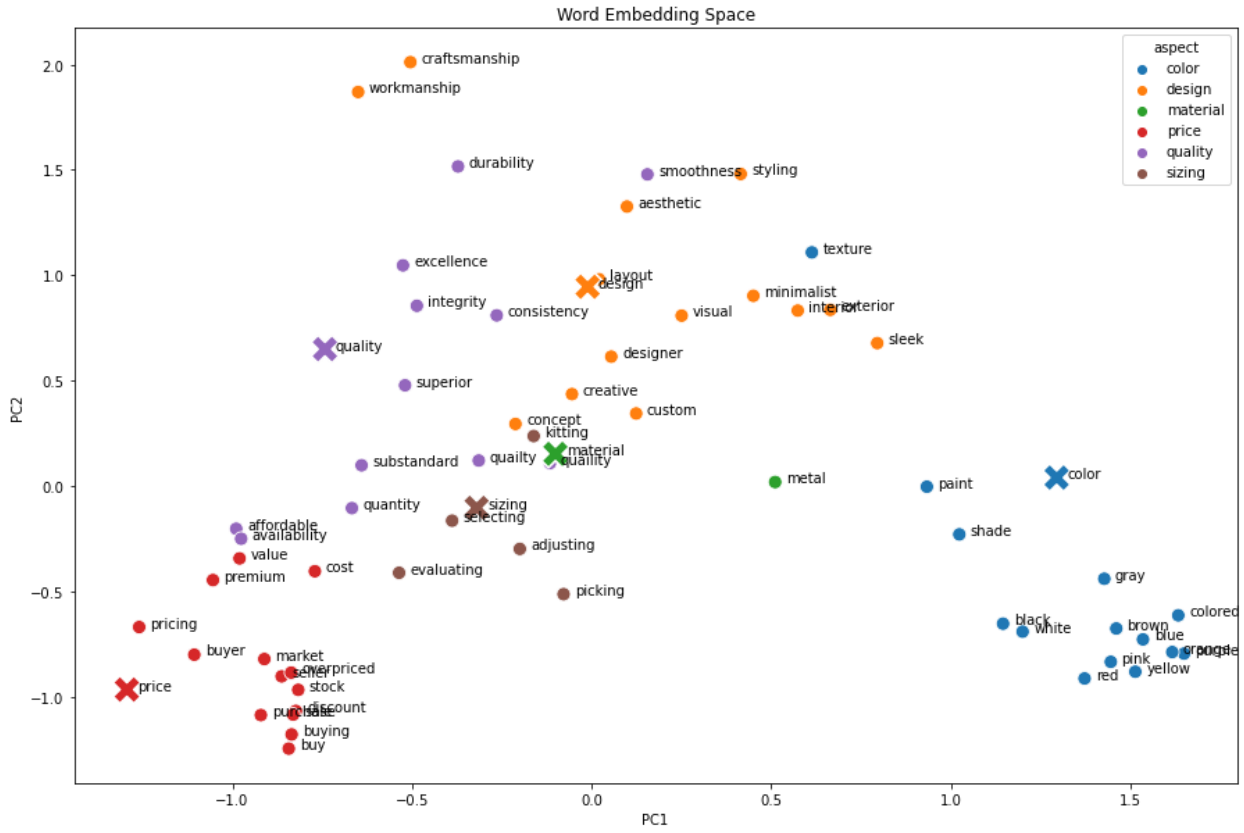

Figure 5. Scatter Plot for Top Aspect Terms

Sentiment Analysis

The Label

Sentiment prediction is a classification problem and the label is the underlying sentiment of a review. We map reviews with ratings of 5 or 4 as Positive, 3 as Neutral, 2 or 1 as Negative. We observe that the dataset is imbalanced with the majority of the reviews labeled as Positive:

- 69.28% of reviews are labeled as Positive
- 11.03% of reviews are labeled as Neutral
- 19.69% of reviews are labeled as Negative

This indicates that a baseline model always predicting Positive as the label for any review has an accuracy of 69%, and our model should achieve an accuracy higher than the baseline model to provide meaningful results.

Feature Engineering

To extract features from textual data, we embed tokens as vectors using below common techniques depending on the characteristics of the algorithms:

- Bag of Words (BOW) with unigram and bigram features
- Term Frequency-Inverse Document Frequency (TF-IDF) with unigram and bigram features
- Word2Vec using the pre-trained model on Google News dataset

Model Training

We evaluate below classical machine learning algorithms trained on BOW and TF-IDF features:

- Logistic Regression
- Multinomial Naive Bayes
- Random Forest

Neural network algorithms are computationally extensive, we utilize the Google Colab GPU runtime to accelerate the model training process, but are also constrained by the limited RAM. We evaluate below neural network algorithms trained on Word2Vec features:

- LSTM using the Keras api
    - LSTM
    - Bidirectional LSTM
    - LSTM with attention layer
- Transformer (BERT model) using the Simple Transformers library

Model Evaluation

Model Performance Comparison

When evaluating models, we need to take into consideration that the dataset is imbalanced with the majority of the reviews being positive and minority being neutral. For this specific project, reviews with neutral sentiment do not provide as much valuable information as positive and negative reviews, so we focus the model evaluation on the weighted average accuracy and F1 metrics.

Negative reviews are important in a sense that fashion brands can extract actionable insights from them immediately. For social listening applications, swift responses on negative reviews are crucial since they spread fast and generate undesired publicity. For this reason, we pay close attention to the F1 score for the negative sentiment.

F1 score is the harmonic mean of the precision and recall and punishes extreme values. When we have an extremely low precision, we are predicting reviews to be negative while a lot of them are not, so companies waste resource analyzing and responding to

these reviews. When we have an extremely low recall, we are only capturing very few out of the many negative reviews, so companies miss out on opportunities to improve their customer experience.

From the model performance comparison table (Table 2) below, we can see that the transformer model has the highest weighted accuracy and F1, as well as F1 for negative reviews. The logistic regression model trained on TF-IDF features also has decent performance, and the prediction time on the test dataset is significantly faster than the transformer model.

Model	Feature	Accuracy	F1	F1 Positive	F1 Neutral	F1 Negative	Predict Time (seconds)
Logistic Regression	BOW	0.85	0.84	0.93	0.36	0.79	
	TF-IDF	0.86	0.84	0.93	0.34	0.80	0.125
Multinomial Naive Bayes	BOW	0.84	0.83	0.92	0.31	0.78	
	TF-IDF	0.77	0.70	0.86	0.00	0.53	
Random Forest	BOW	0.82	0.78	0.90	0.11	0.74	
	TF-IDF	0.82	0.78	0.90	0.09	0.74	
LSTM	Word2Vec	0.82	0.80	0.91	0.23	0.73	
Bi-LSTM	Word2Vec	0.83	0.81	0.91	0.27	0.73	
Bi-LSTM + Attention	Word2Vec	0.83	0.81	0.91	0.30	0.74	
Transformer	Built-in	0.87	0.87	0.95	0.42	0.82	512.725

Table 2. Model Performance Comparison Table

LSTM Model Learning Curve

We also observe that the logistic regression models and the multinomial naive bayes model trained on BOW features all outperform the LSTM models by a decent margin. We are interested in diving deeper into the LSTM models to diagnose if there are any problems with the training process.

Visualizing the training loss vs. validation loss over a number of epochs is a good way to evaluate the training process of a LSTM model. We want to make sure that the model is not undertrained with room left for improvement, or overtrained such that it starts memorizing the training data and losing the ability to generalize on the test data.

Due to the RAM constraints on Google Colab, we have to train the LSTM models on a subset of the entire dataset, and we evaluated two different feature engineering approaches.

For the first approach, we truncate/pad each review input data to 16 words and use all reviews from the dataset to train the model. Figure 6 below shows the training and validation learning curves of the bi-directional LSTM with attention layer model.

Figure 6. Training vs Validation Loss

For the second approach, we truncate/pad each review input data to the 90% percentile of the length of all reviews, 74 words, and use a 25% random sample of the entire dataset to train the model. Figure 7 below shows the training and validation learning curves of the bi-directional LSTM with attention layer model.

Figure 7. Training vs Validation Loss

We can see from Figure 6 and Figure 7 that both approaches suffer from overfitting. After around 5 epochs, the model starts to become more and more specialized to the training data, and the validation loss reaches a minimum and begins increasing. This indicates that the model has too much flexibility for our data and one way to mitigate it is to train the model on more data. As we add more data, the model becomes unable to overfit all the samples, and is forced to generalize to make progress.

Transformer Model Intuition

The transformer neural network architecture was introduced by Google in 2018, which turned out to be a groundbreaking milestone in NLP. BERT (Bidirectional Encoder Representations from Transformers) is a popular transformer-based model that is pre-trained on a large corpus of unlabeled text and can be fine-tuned for NLP tasks.

For this project, we use a pre-trained BERT model as a starting point, then further train the model on our relatively smaller Amazon Fashion review dataset.

The reason that the transformer model can achieve better results for the sentiment prediction task is that, unlike the LSTM model that processes an input sequence token by token, it takes the entire input sequence as a whole. As a result, the transformer model is able to remember important information from the distant past, while the LSTM model needs to backpropagate the error and eventually forgets some of the information from before in long reviews.

Recommendations

Our recommendations are twofold for this ABSA project and fashion brands should adopt the machine learning model that best suits their needs.

For social listening purposes, companies need a model that can detect any negative sentiment quickly so that they can respond before the word-of-mouth starts damaging their brands. We recommend the logistic regression model as it offers a good speed-accuracy trade-off. The logistic regression model provides decent prediction results, and is significantly faster than the transformer model in making the predictions.

For other common purposes like customer experience management and product development, companies need a model that can detect review aspects and sentiment accurately because they aim to extract actionable insights for long-term planning. We

recommend the transformer model as it provides better accuracy and F1 scores even though it takes longer time and more computational power to train.

ABSA Pipeline

We create a novel ABSA pipeline for the fashion industry. The pipeline preprocesses the textual data, extracts the underlying aspect(s), and predicts the sentiment for the extracted aspect(s) in a streamlined process.

For the aspect extraction, the pipeline computes the semantic similarity between each extracted noun or adjective and each given aspect using the pre-trained word2vec model, and outputs the aspect when the similarity exceeds the 0.40 threshold. However, the semantic similarities between some common aspect terms and their corresponding aspects do not exceed the threshold. We add below common aspect terms (Table 3) into the aspect extraction step so that the pipeline outputs the corresponding aspects of a review when these aspect terms are present:

Aspect	Common Aspect Terms
material	'fabric', 'cotton', 'leather', 'suede', 'polyester', 'nylon', 'spandex', 'silk', 'wool', 'cashmere'
price	'cheap', 'expensive', 'deal', 'promotion', 'coupon'
sizing	'size', 'fit', 'small', 'medium', 'large', 'big', 'xs', 'xxs', 'xl', 'xxl', '2t', '3t', '4t', '5t'

Table 3. Aspects and Common Aspect Terms

Application Examples

We apply the ABSA pipeline on randomly selected Amazon Fashion reviews posted in 2019 or later. This is to make sure we do not accidentally use the training data to examine the pipeline and receive misleading results.

Review example 1 (Positive):

Warm, comfortable, and stylish

Reviewed in the United States on January 23, 2021

Size: 7 | Color: Black | **Verified Purchase**

I recently moved to a cold climate and needed shoes to protect my feet. These are definitely warm and comfortable. The build is a little bit stiffer than I expected and not as fluffy and soft on the inside but it definitely does what it says it will do. I found the fit to be comfortable and true to what I expected. I usually wear a 6.5 woman's but I ordered a 7 just in case and it has just enough room for thick socks. Walked around an outdoor mall for a few hours with no issues.

Pipeline output:

Review: i recently moved to a cold climate and needed shoes to protect my feet. these are definitely warm and comfortable. the build is a little bit stiffer than i expected and not as fluffy and soft on the inside but it definitely does what it says it will do.
Aspects: general
Sentiment: Positive

Review: i found the fit to be comfortable and true to what i expected.
Aspects: sizing
Sentiment: Positive

Review: i usually wear a 6.5 woman's but i ordered a 7 just in case and it has just enough room for thick socks. walked around an outdoor mall for a few hours with no issues.
Aspects: general
Sentiment: Positive

Review example 2 (Positive):

Colors are not correctly represented in the photos

Reviewed in the United States on September 24, 2020

Color: Matte Havana/Blue | Lens Width: 55 Millimeters | **Verified Purchase**

Just got the pair described as Matte Havana / Blue. The colors are not even close to the photos. In the photos on Amazon they look royal blue and orange. They are in fact slate blue, almost gray, on the outside of the ear pieces, and basically coffee or caramel colored on the inside. While the look is still very cool, it is not accurate on this site at all. I own another pair of these that I have enjoyed for many years, and I am going to keep these, but be aware that the colors are pretty hard to pick out on this site and they are WAY off, like, not even close.

Pipeline output:

Review: just got the pair described as matte havana / blue. the colors are not even close to the photos. in the photos on amazon they look royal blue and orange. they are in fact slate blue, almost gray, on the outside of the ear pieces, and basically coffee or caramel colored on the inside.
Aspects: color
Sentiment: Negative

Review: while the look is still very cool, it is not accurate on this site at all.
Aspects: general
Sentiment: Negative

Review: i own another pair of these that i have enjoyed for many years, and i am going to keep these, but be aware that the colors are pretty hard to pick out on this site and they are way off, like, not even close.
Aspects: color
Sentiment: Negative

Review example 3 (Neutral):

★★★☆☆ **Off center logo....**

Reviewed in the United States on June 20, 2020

Size: One Size | Color: Black/White

Kind of pissed. First one i got. Ended up being bleached from sweat which seems stupid as its an athletic hat and i only had it for two months in the winter.

Ordered another and the logo was off center.

Dont know if ill buy again if i cant even be guarenteed quality and cant speek to anyone about making sure the hats logo is centered. As it should be.

Pipeline output:

Review: kind of pissed. first one i got. ended up being bleached from sweat which seems stupid as its an athletic hat and i only had it for two months in the winter. ordered another and the logo was off center.
Aspects: general
Sentiment: Negative

Review: do not know if ill buy again if i can not even be guarenteed quality and can not speek to anyone about making sure the hats logo is centered.
Aspects: quality
Sentiment: Negative

Review: as it should be.
Aspects: general
Sentiment: Positive

Review example 4 (Negative):

★☆☆☆☆ **Bas quality. Fake Product**

Reviewed in the United States on January 13, 2021

Color: 02blue | Size: Medium | **Verified Purchase**

Garbage quality product and nothing like the picture. Fabric looks thick and nice color on the picture. The real thing is dull and very cheap fabric that it looks secondhand. False advertisement

Pipeline output:

Review: garbage quality product and nothing like the picture.
Aspects: quality
Sentiment: Negative

Review: fabric looks thick and nice color on the picture.
Aspects: material, color
Sentiment: Positive

Review: the real thing is dull and very cheap fabric that it looks secondhand.
Aspects: price, material
Sentiment: Negative

Review: false advertisement
Aspects: general
Sentiment: Negative

Review example 5 (Negative):

I wanted to like this suit!

Reviewed in the United States on July 12, 2019

Color: Navy | Size: 16 | **Verified Purchase**

I really wanted to like this suit! The material was great and nicely made. I'm 5'9 with a long torso and 200 lbs. I do have a thicker butt and thighs. This was tight around my legs... like cut into my legs. Fit nicely in my mid section and in the chest. But because I'm so long waisted when i life my right arm my whole boob almost comes out the top of the suit! It was a no for me

Pipeline output:

Review: i really wanted to like this suit!
Aspects: general
Sentiment: Negative

Review: the material was great and nicely made.
Aspects: material
Sentiment: Positive

Review: i am 5'9 with a long torso and 200 lbs. i do have a thicker butt and thighs. this was tight around my legs... like cut into my legs. fit nicely in my mid section and in the chest. but because i am so long waisted when i life my right arm my whole boob almost comes out the top of the suit! it was a no for me
Aspects: general
Sentiment: Neutural

Pipeline Evaluation

From the examples above we can see that the ABSA pipeline is able to correctly extract the aspects of a review for the most part and are very reliable in predicting their sentiments. This is especially powerful when there are tens of thousands of reviews to go through.

Fashion brands can gain meaningful and actionable insights for improvement even from positive reviews. We can see from example 2 that the customer gave a 4 star rating for the product, which indicates that they were happy with the purchase overall. However, the *color* aspect shows up in the ABSA result twice and the sentiment is rather negative. This should immediately raise a red flag for the brand and trigger actions. They can quickly look into the review and find out that the actual color of the product is very different from the photos and updating the photos to be accurate will help improve the customer experience and retention of loyal customers.

We observe from example 1 and 5 that the pipeline fails to extract *sizing* as the aspect and outputs *general* when customers talk about sizes (*"..I usually wear a 6.5 woman's…"*) or their heights and weights (*"..I'm 5'9 with a long torso and 200lbs.."*). We keep the numbers in reviews during the preprocessing since they may provide additional information for sentiment analysis, but they can also be useful features in the aspect extraction, which can be explored in the future improvements of this project.

The unsupervised approach we take for the aspect extraction also inevitably has its drawbacks and limitations. In example 3, there are no nouns or adjectives in the last

sentence *"As it should be."*, as a result the pipeline cannot extract any aspects from it and outputs *general* as the aspect. In example 4, the adjective *cheap* in the sentence *"The real thing is dull and very cheap fabric that it looks secondhand."* is used to describe the material, but the pipeline outputs both *price* and *material* as the aspects because *cheap* is added as an aspect term for *price*.

Future Improvement

We are limited by certain constraints when working on this project, and below are future improvements we are interest in:

- Extract quantities, measurements and their units from numbers in the preprocessing step using packages such as quantulum, and utilize them to improve aspect extraction and sentiment analysis results.
- Train a model on annotated data for better aspect extraction. We did not have labeled data for supervision when building the aspect extraction step of the ABSA pipeline, so had to take the unsupervised approach. From the previous examples we can see that the aspect extraction step has some room for improvement, and the supervised approach may help.
- Include emoticons as features. As a common approach, we removed punctuations and special characters during the text preprocessing. However, emoticons composed of punctuation marks and special characters play an important role in the sentiment of texts nowadays. This is especially the case in colloquial writing such as product reviews and social media posts. We are interested in converting emoticons into tokens and taking them as features to help improve sentiment analysis performance.
- Tune hyperparameters for the transformer model. The Google Colab notebook can only stay connected for up to 24 hours, so we used the out-of-the-box transformer model without any hyperparameter tuning since it was very time consuming and computationally extensive.
- Train the LSTM model on the entire Amazon Fashion review dataset. Due to the RAM constraint on Google Colab, we were not able to train the LSTM models on the entire dataset. We can see from the learning curves that the models started to overfit in the early stage of the training process, and one way to mitigate overfitting is to train the model on more data.