Yuwen Sang

COMP 3705: Text Mining – Assignment 2

# Sentiment Analysis Report of The Women’s E-Commerce Clothing Reviews

## *Introduction*

This report is an analysis of the women e-commerce clothing reviews. There are two text files (*input1.txt* and *input2.txt*) including the high rating reviews and low rating reviews respectively. This report will explore what the customers like or dislike of the products.

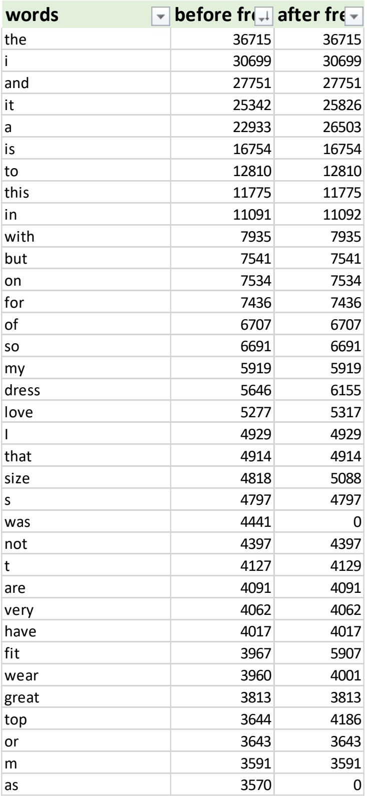
This report only shows the most valuable result of the experiment, and the word list from the two text files are lemmatized in almost steps, since it can reflect more accurate and more valuable analysis results. The analysis highly rely on counting the word frequency in the different ways.

Finally, since the two input files has different size, the data comparison between the two files will be converted into percentage format

## *Data and Analysis*

### *The Effect of Lemmatization*

As shown in figure 1, which shows the top 35 frequently appeared words based on before lemmatization counting. However, we can see that after lemmatization, the word such as “was”, its frequency become zero, while the words such as “dress,” “wear,” their frequency are increased.  This can make the analysis more accurate. Thus, the further analysis of the words are all lemmatized.



| words | before fr | after fre |
|---|---|---|
| the | 36715 | 36715 |
| i | 30699 | 30699 |
| and | 27751 | 27751 |
| it | 25342 | 25826 |
| a | 22933 | 26503 |
| is | 16754 | 16754 |
| to | 12810 | 12810 |
| this | 11775 | 11775 |
| in | 11091 | 11092 |
| with | 7935 | 7935 |
| but | 7541 | 7541 |
| on | 7534 | 7534 |
| for | 7436 | 7436 |
| of | 6707 | 6707 |
| so | 6691 | 6691 |
| my | 5919 | 5919 |
| dress | 5646 | 6155 |
| love | 5277 | 5317 |
| I | 4929 | 4929 |
| that | 4914 | 4914 |
| size | 4818 | 5088 |
| s | 4797 | 4797 |
| was | 4441 | 0 |
| not | 4397 | 4397 |
| t | 4127 | 4129 |
| are | 4091 | 4091 |
| very | 4062 | 4062 |
| have | 4017 | 4017 |
| fit | 3967 | 5907 |
| wear | 3960 | 4001 |
| great | 3813 | 3813 |
| top | 3644 | 4186 |
| or | 3643 | 3643 |
| m | 3591 | 3591 |
| as | 3570 | 0 |

*Figure 1: Word Counting from input1.txt*

### *The Word Counting Based on Word Tags*

Nouns

As showing in figure 2, the most frequent nouns in both files are “dress,” “fabric,” “color,” “sweater,” “jean,” “size,” “waist,” etc., which implies the topics that the customers are most interested in, and the kind of the products that are the most popular in the online store.

From the noun’s frequency chart, we cannot have more clues on the review’s sentiment.

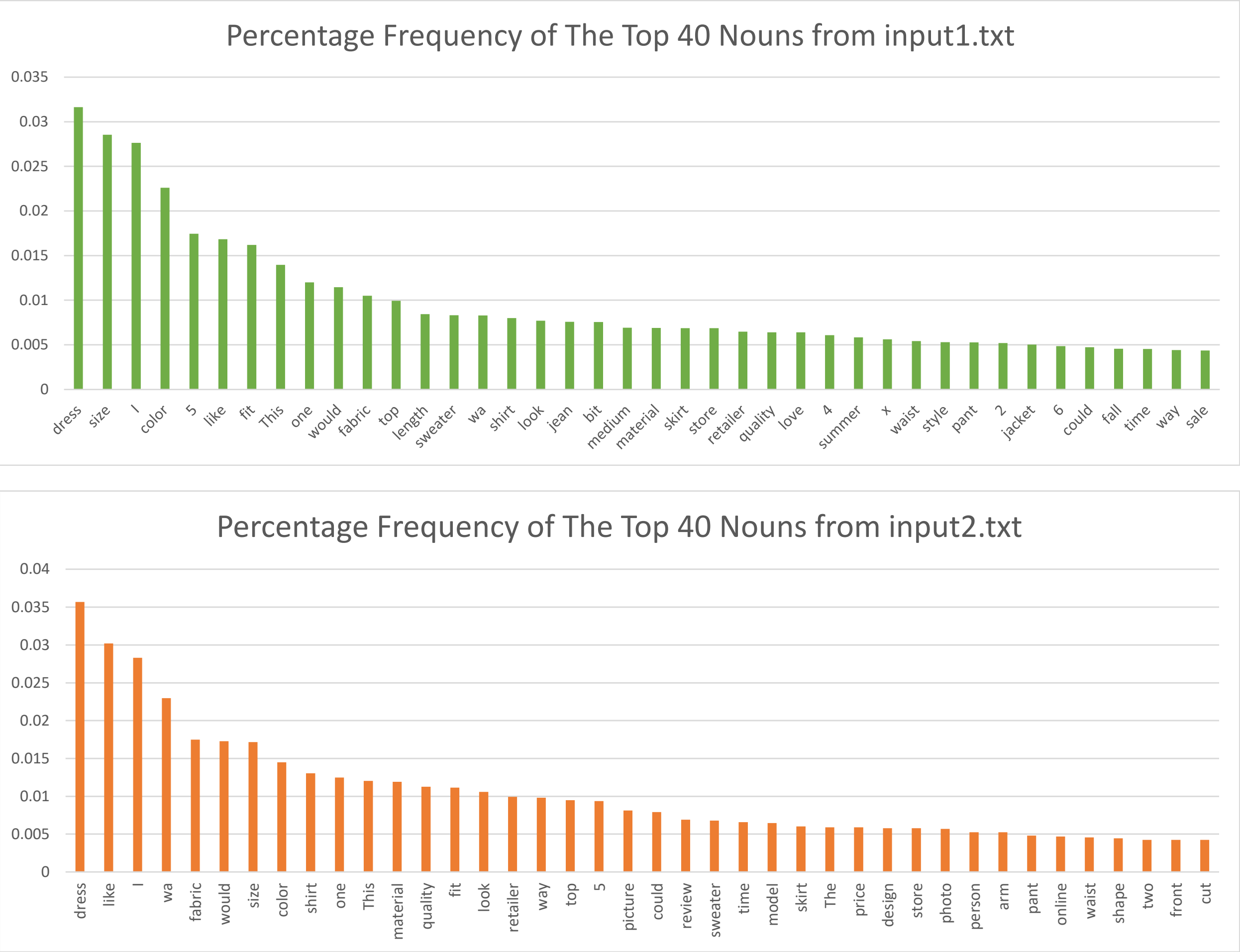


*Figure 2: Nouns Frequency*

## Verbs

Figure 3 show the verbs frequency charts. We can see that, the most frequent verbs that have strong sentiments from input1 file are “love,” “flattering,” and “recommend.” These words strongly implied the consumers love the products they bought. The verbs from input2 including “love,” and “returned,” which are have more negative meaning.

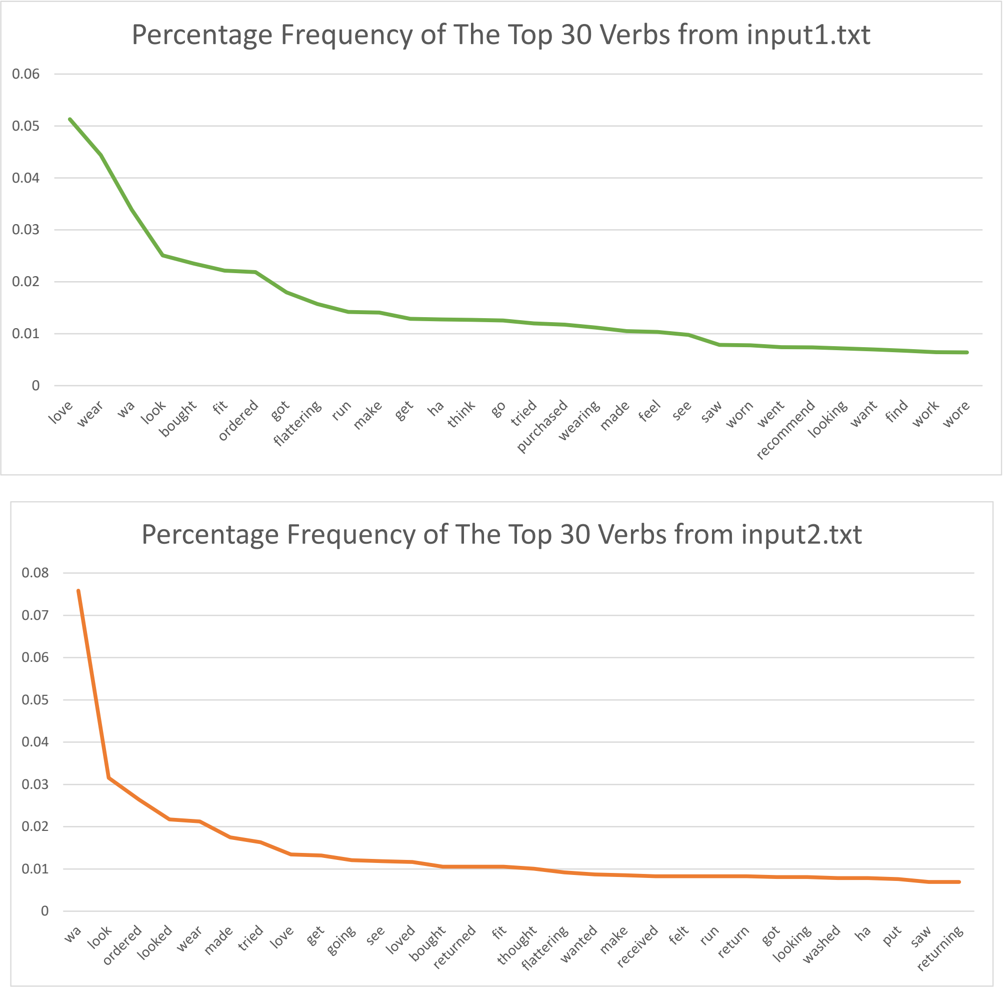


*Figure 3: Verbs Frequency*

## Adjectives

Figure 4's charts show the adjective frequencies. From adjectives, we can see more sentimental words, such as "great," "soft," "perfect," and "comfortable," from the positive reviews, and the words "small," "thin," "disappointed," etc., from the negative reviews. However, one interesting thing is that some positive words also have high frequency in the negative reviews, and some negative words also have high frequency in the high rating reviews.

For example, the word "small" also shown in the positive reviews, while the words "great," "beautiful" are also shown in the negative reviews. This might imply that the design of the clothes are nice, while the size is too small for many customers.

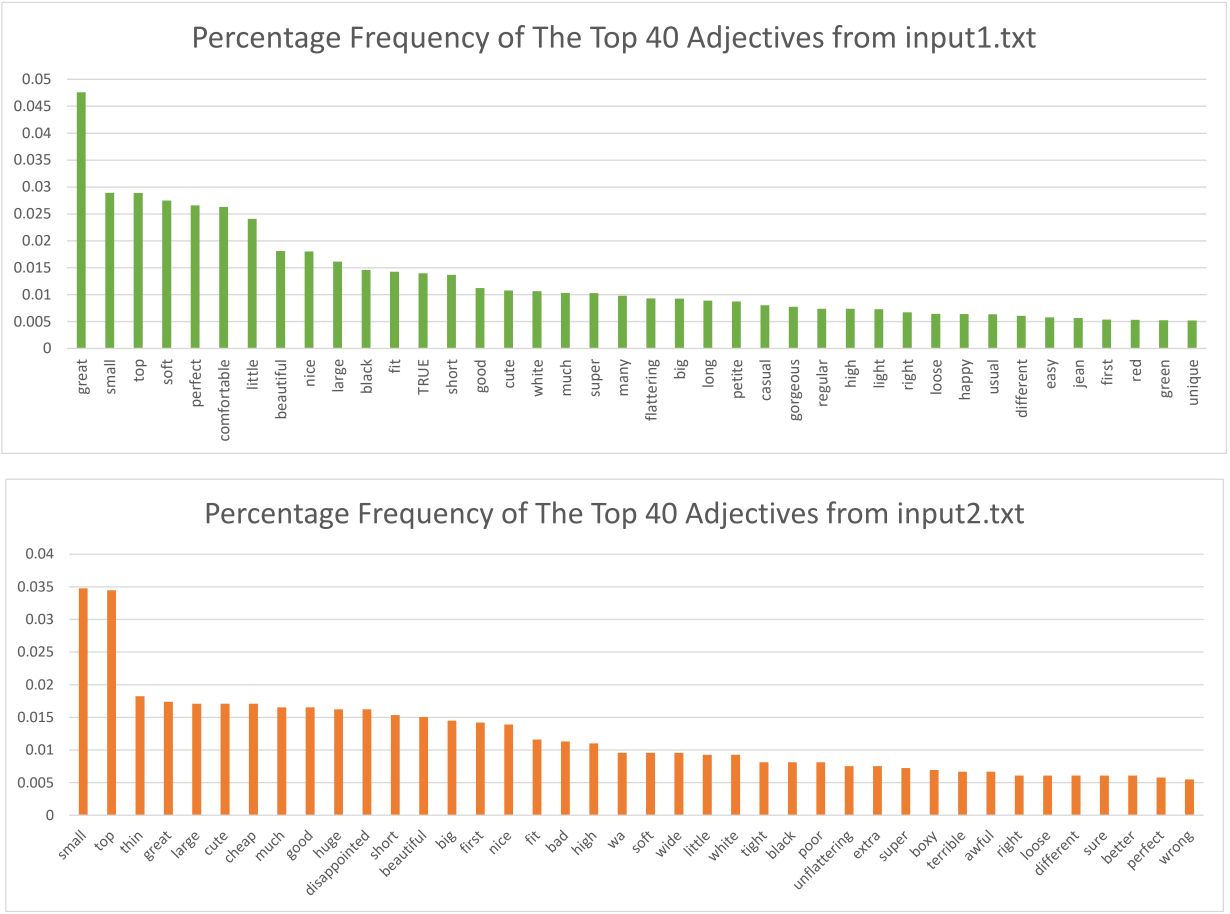
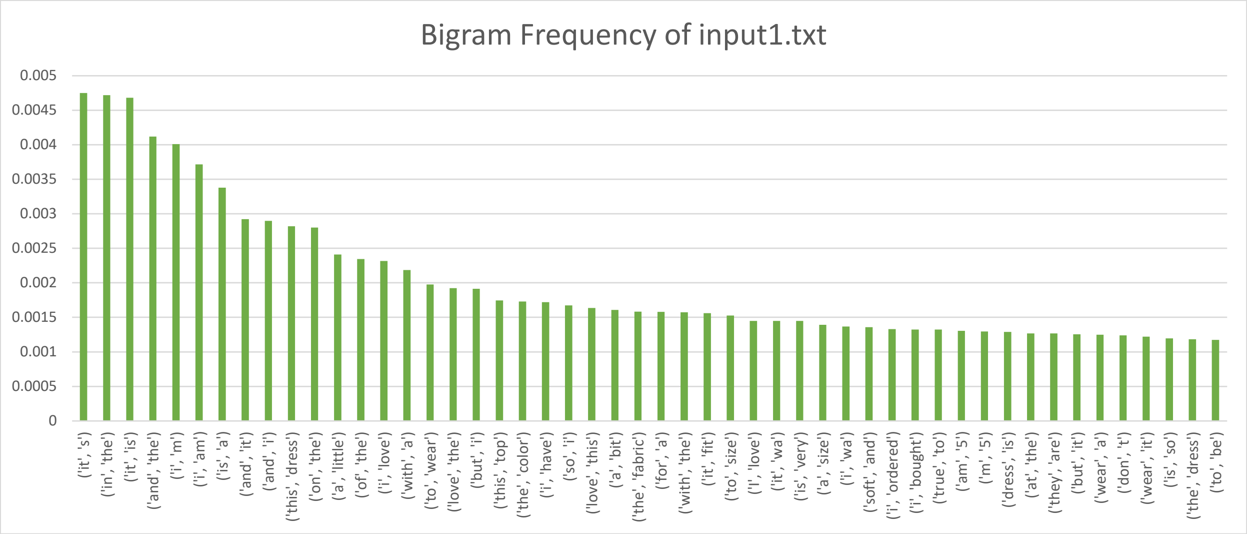


*Figure 4: Adjectives Frequency*

### *Bigrams and Trigrams Frequency*

The bigram and trigram frequency chart are shown in figure 5 and figure 6, respectively. We can see that the high frequency bigram and trigrams are always related with the key words "dress," "color," "size," "fabric," and "material," while the trigram frequency tables do not have very clear sentiments. The trigrams are more corresponding to the sentiments, such as ("look", "great", "with") and ("can", "t", "wait") from the highest rating reviews. On the other hand, the low rating reviews have ("i", "don", "t"), ("not", "at", "all"), ("i", "can", "t"), etc., which shows a more negative sentiment.



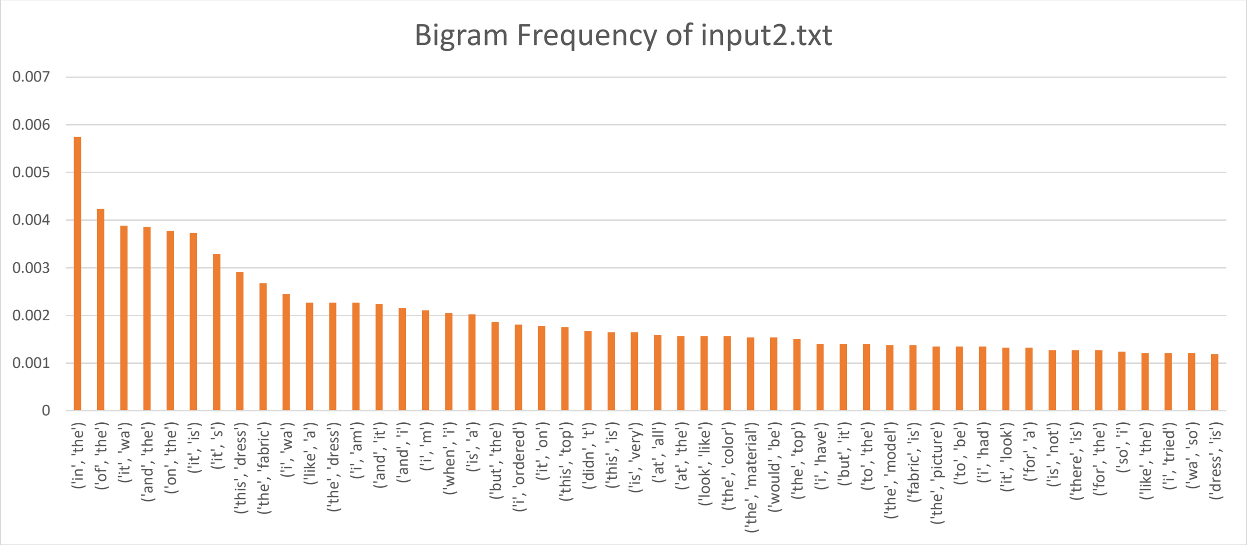
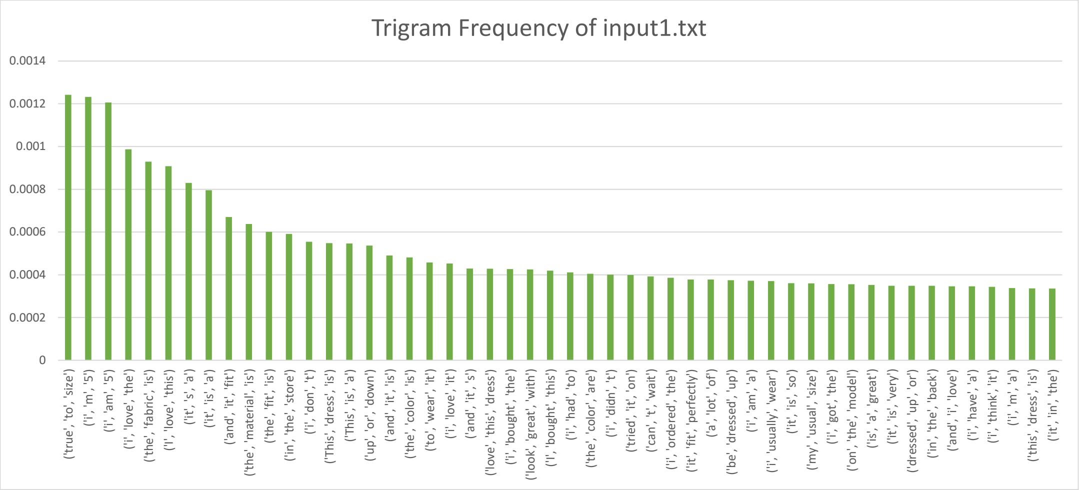


*Figure 5: Bigram Frequency*



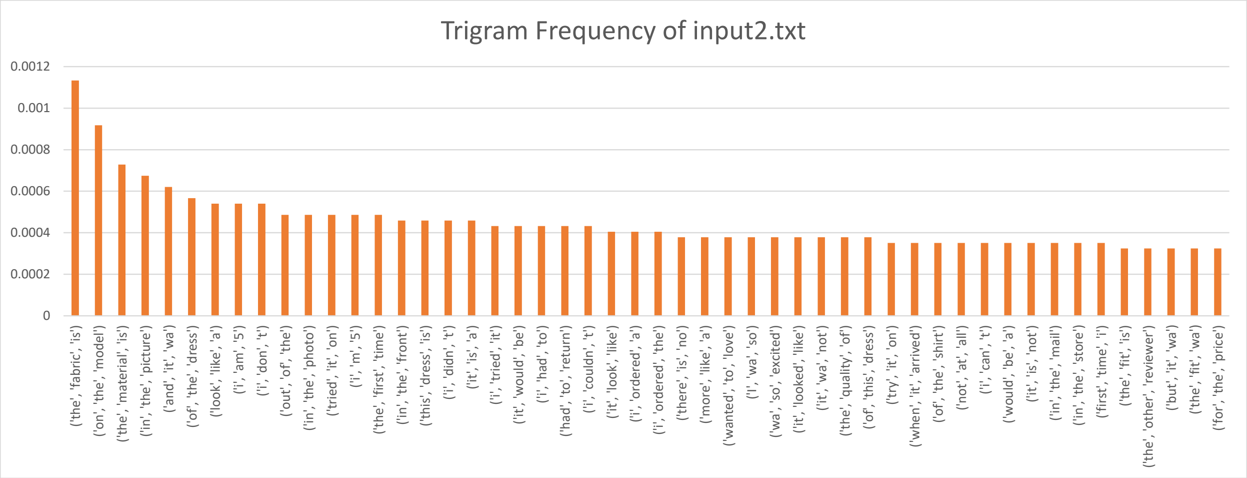


*Figure 6: Trigram Frequency*

## *Conclusions and Suggestions*

According to analyze the highest rating reviews and the lowest rating reviews, we can see that the topics that the customers are most concern about are "dress," "size," "material," and "fabric." This online store's dresses seemed too small for many of its customers. Even in the positive reviews, the word "size" is also a frequently mentioned problem. In other words, it seems that the customers that normally fit for some size cannot fit to this store's clothes with the same size. Therefore, the store should consider fixing the size problem to the customers.