Using Sentiment Dictionaries on Customer Reviews Dataset

Created	@July 20, 2023 7:57 PM
Tags	Personal Project
Project Type	Text Analysis

About the Project

In this project, I will be exploring sentiment dictionaries on a customer review dataset titled, “Women’s E-commerce Clothing Review”. This project can be divided into three parts: The first part will use the built-in Quanteda dictionary called “DictionaryGI” to predict the number of negative and positive terms in each text in the form of a document term matrix. I improve on this dtm by adding on the length of each text which would then contextualise the prediction further. Additionally, to make the dtm more understandable to the general audience, I created three types of universal sentiment scores to illustrate whether the text has more positive or negative sentiments. Secondly, I will explore how my dictionary can be improved by analysing whether the model has predicted terms into negative or positive correctly. Lastly, I will use the corpustools() library to conduct data visualisation on our dataset.

This project illustrates my confidence and experience in text analysis, working with a large and real world dataset, in the fields of marketing.

Retrieving the Data

The Women’s E-commerce Clothing Review dataset was retrieved from kaggle. This is a real world dataset which has been anonymised by the source. As stated in the website, the dataset includes 23,486 rows and 10 feature variables. Each row corresponds to a customer review, and includes the following variables:

- **Clothing ID:** Integer Categorical variable that refers to the specific piece being reviewed.
- **Age:** Positive Integer variable of the reviewers age.
- **Title:** String variable for the title of the review.
- **Review Text:** String variable for the review body.
- **Rating:** Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst, to 5 Best.
- **Recommended IND:** Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.
- **Positive Feedback Count:** Positive Integer documenting the number of other customers who found this review positive.

- **Division Name:** Categorical name of the product high level division.
- **Department Name:** Categorical name of the product department name.
- **Class Name:** Categorical name of the product class name.

Analysis

You can access the full R Markdown file below:

RPubs - Using Sentiment Dictionaries on Customer Reviews Dataset
https://rpubs.com/annisaptr/customer_reviews

Part 1: Basic Sentiment Analysis

In this first part, I have created a dictionary using quanteda's "DictionariGI" and applied it to my dataframe. I have also converted my dataframe into a tibble to visualise it in a more convenient manner. As a result, we have the following first 10 rows of the dataset, showing the doc_id, the amount of negative words and positive words in each corresponding doc_id.

Table 1: DTM of Negative and Positive Words

Document ID	Frequence of Negative Words	Frequency of Positive Words
0	0	2
1	0	6
2	2	4
3	3	6
4	2	4
5	1	5
6	1	2
7	4	3
8	1	2
9	3	4

We can further improve the table above by creating a new column which counts the total amount of words there are in each document ID. As a result, this will give us more context on our documents. For example, in Document ID 1 we see that there are 6 positive terms and 0 negative terms. At first hand, we may deduce that this document is in overall, an entirely positive document. However, when we consider the length of the document of 70 words, we may think instead that the text is not as positive as it may seem.

Table 2: DTM with the Length Column

Document ID	Frequence of Negative Words	Frequency of Positive Words	Length
0	0	2	8
1	0	6	70
2	2	4	111

3	3	6	30
4	2	4	41
5	1	5	107
6	1	2	121
7	4	3	117
8	1	2	37
9	3	4	88

To make this document term matrix more understandable by the general audience, we can create a universal sentiment score, with the scoring range of -1 to 1 to show whether a document is more or less positive. There are different ways in creating a universal sentiment score - I have presented three options:

1. Sentiment 1 gives results ranging from -1 to ., -1 showing it is negative and 1 showing it is more positive. It is calculated in the following way:

   ```
   sentiment1=(positive - negative) / (positive + negative)
   ```

2. Sentiment 2 also has a scoring range of -1 and 1, but a score of 1 means that all the terms in the document is positive and vice versa. It is calculated in the following way:

   ```
   sentiment2=(positive - negative) / length
   ```

3. Subjectivity is calculated in the following way:

   ```
   subjectivity=(positive + negative) / length)
   ```

As a result, we have the following document term matrix:

Table 3: DTM with Sentiment Scores

Document ID	Frequence of Negative Words	Frequency of Positive Words	Length	Sentiment 1	Sentiment 2	Subjectivity
0	0	2	8	1.00	0.25	0.25
1	0	6	70	1.00	0.09	0.09
2	2	4	111	0.33	0.02	0.05
3	3	6	30	0.33	0.10	0.30
4	2	4	41	0.33	0.05	0.15
5	1	5	107	0.67	0.04	0.06
6	1	2	121	0.33	0.01	0.02
7	4	3	117	-0.14	-0.01	0.06
8	1	2	37	0.33	0.03	0.08
9	3	4	88	0.14	0.01	0.08

Part 2: Improving my Dictionary

Sometimes, our model wrongly predicts terms into positive or negative words. We can see this in the following way. When examining positive words, I was curious how ther term “just” was classified into this sentiment group. That said, I used key word in context, kwic() function, with a window of 4 words to see how the term is used.

Table 4: Positive Terms

Term	Frequency	Rank	Document Frequency
love	8,940	24	7,415
fit	7,280	29	6,170
like	7,007	32	5,732
great	6,094	37	5,180
just	5,598	43	4,720
perfect	3,747	60	3,361
well	3,231	68	2,922
back	3,211	69	2,841
comfortable	3,046	74	2,941
cute	3,028	75	2,792

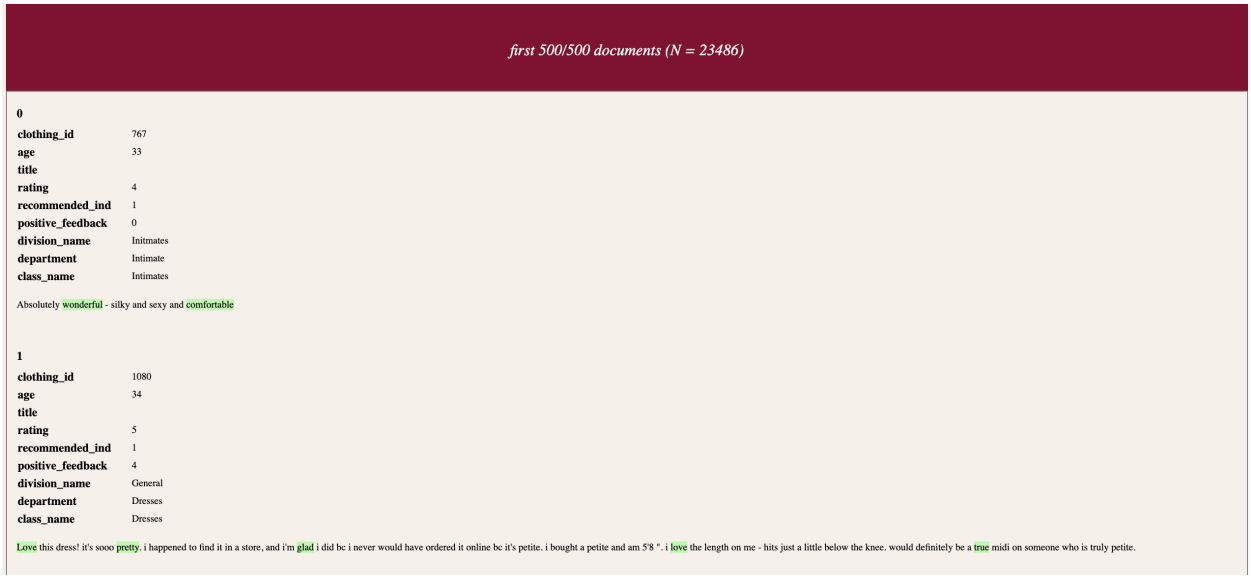
Table 5: Keyword In Context for “Just”

length on me- hits	**just**	a little below the
medium, which was	**just**	ok. overall,
petite. i am	**just**	under 5 feet tall
the style but it	**just**	did not work on
very cheap that even	**just**	pulling on it will
very form-fitting. falls	**just**	above the knee and

As can be seen, just can be used as either negative or positive or even neutral. For example, the second row uses just in an almost positive light with "just ok", whereas the third row uses the term to describe her height which makes this prediction innacurately used, while the fourth and fifth row used the term in a negative manner. That said, we can remove the term "just" from the list of positive words. The same can be done to false negative terms.

Part 3: Data Visualisation with Corpus Tools Package

Lastly, I experimented with the corpustools() package to create a tokenised list of each word in the dataset. I then used this list to create a full text browser which highlights words into positive with green and negative with red as seen in the screenshot below:



first 500/500 documents (N = 23486)

0

clothing_id	767
age	33
title	
rating	4
recommended_ind	1
positive_feedback	0
division_name	Initmates
department	Intimate
class_name	Intimates

Absolutely wonderful - silky and sexy and comfortable

1

clothing_id	1080
age	34
title	
rating	5
recommended_ind	1
positive_feedback	4
division_name	General
department	Dresses
class_name	Dresses

Love this dress! it's sooo pretty. i happened to find it in a store, and i'm glad i did bc i never would have ordered it online bc it's petite. i bought a petite and am 5'8 ". i love the length on me - hits just a little below the knee. would definitely be a true midi on someone who is truly petite.

10000

clothing_id	863
age	65
title	Don't recommend
rating	3
recommended_ind	0
positive_feedback	1
division_name	General Petite
department	Tops
class_name	Knits

This t is a beautiful color, and i wish it had worked. it is quite flimsy, however - very thin cotton fabric, without the nice drape that would have made this look good on. i also think it should have been cut a little longer; it flares out right at the hips, so instead of covering them, it highlights them. i'm disappointed, and am returning it.

10001

clothing_id	1110
age	35
title	Fun, feminine, versatile
rating	5
recommended_ind	1
positive_feedback	0
division_name	General
department	Dresses
class_name	Dresses

This dress is my new favorite! the style is trendy and feminine while forgiving in the tummy area without looking maternity or too baggy. it is so versatile because it can be dressed up or down and it's a great transitional dress b / c once it cools down, it will look great with a denim jacket, cardi, booties, etc. i normally wear a m, sometimes a l (i'm 5'7 ", 36dd) but the m fits great. the slip is a little snug across the chest but i definitely couldn't size up b / c then the dress would be too bi

Conclusion

In this project, I explored the use of sentiment dictionaries on a customer review dataset and divided it into three parts. In the first part, I used the Quanteda dictionary to predict the number of negative and positive terms in each text. I then contextualized the prediction further by adding the length of each text and created three types of universal sentiment scores to make the document term matrix more understandable to the general audience. In the second part, I analyzed whether my dictionary could be improved by examining whether the model had predicted terms into negative or positive correctly. Lastly, I used the corpustools() library to conduct data visualization on our dataset.

Overall, this project demonstrates the usefulness of sentiment dictionaries in analyzing customer review datasets and provides insights into how to improve them.