Adjective Counter: Women's E-Commerce Clothing Reviews

LING78100/73800 Methods in Computational Linguistics I

Leila Habibi

December 2024

1 Introduction

In this project, I analyzed online reviews of women's clothing by counting the frequency of specific adjectives used by customers. The dataset, sourced from Kaggle, was processed using a Python-based program developed for this purpose.

2 Data

The dataset consists of 23,486 reviews from female customers aged 18 to 99, focusing on various women's clothing items. After cleaning and organizing the data, I removed rows with missing values, using excel, reducing the dataset to 22,643 reviews. I then divided the data into eight groups, each representing a 10-year age interval.

3 Research Questions

The central question of this project is whether women from different age groups use distinct language to describe clothing items they purchased online. While this question opens the door to numerous considerations—such as whether the reviews are for the same item, the same seller, or whether the reviewers share similar economic backgrounds or household income levels, or even if the data is controlled for factors like weight—I focused on adjectives that are minimally influenced by these variables. The aim was to address a more specific question:

Do women of different age groups have different needs? In other words, do women of varying ages prioritize different characteristics in clothing? If so, can this be observed through a linguistic analysis of the data?

It is important to note that people describe things not only through adjectives but also using various linguistic tools, such as word order, descriptive clauses, slang, nouns, and so on. However, for the sake of practicality, this project focuses solely on the use of adjectives.

4 Program Design

To count the targeted adjectives, I first created a function to read a text file and return its content as a list of words. To ensure that both uppercase and lowercase instances of words were counted, I converted all text to lowercase. Next, I defined another function to calculate the frequency of each word and store the results in a dictionary, where words serve as keys and their frequencies as values. This approach allows for quick access to the count of any specific adjective by directly referencing it in the dictionary.

5 Results and Analysis

The results of the counts are provided in the following table:

Adjective	18-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
casual	122	293	274	177	120	18	7	-
versatile	70	182	155	75	59	4	1	1
soft	383	1037	879	522	390	50	27	3
lightweight	60	156	125	88	80	3	1	-
vibrant	32	95	91	49	31	9	1	-
simple	32	93	83	38	34	5	1	-
delicate	44	88	68	56	33	1	5	1
appropriate	26	71	37	22	26	2	4	-
showing	17	60	35	24	20	1	-	-
classy	23	58	47	18	20	2	2	-
chic	19	46	45	20	19	6	1	1
unflattering	44	110	97	48	22	47	1	2
comfortable	449	928	810	479	308	42	13	2
sexy	29	80	58	21	14	6	3	-
sturdy	12	35	14	8	4	1	-	-
timeless	11	17	18	16	4	4	1	-
vegan	2	7	7	4	4	-	1	-
trendy	4	16	15	8	3	3	-	-
functional	4	15	12	4	4	1	-	-
traditional	6	13	10	6	1	1	1	-

Table 1: Adjective count by Age Group

Table 1 shows the count numbers for each adjectives for every age groups available in the reviews. However, since the number of reviews have not been the same for each age group, the data was normalized by dividing the counts by the number of reviews for each age group. The result can be found here in Table 2:

Adjective	18-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
casual	0.0430	0.0399	0.0464	0.0462	0.0532	0.0507	0.0745	-
versatile	0.0247	0.0248	0.0263	0.0196	0.0262	0.0113	0.0106	0.0714
soft	0.1349	0.1412	0.1489	0.1362	0.1729	0.1408	0.2872	0.2143
lightweight	0.0211	0.0212	0.0212	0.0230	0.0355	0.0085	0.0106	-
vibrant	0.0113	0.0129	0.0145	0.0128	0.0137	0.0056	0.0106	-
simple	0.0113	0.0127	0.0141	0.0099	0.0151	0.0141	0.0106	-
delicate	0.0155	0.0120	0.0112	0.0146	0.0146	0.0028	0.0532	-
appropriate	0.0092	0.0097	0.0063	0.0057	0.0115	0.0056	0.0426	-
showing	0.0060	0.0082	0.0059	0.0063	0.0089	0.0028	-	-
classy	0.0080	0.0079	0.0075	0.0042	0.0115	0.0113	0.0213	-
chic	0.0067	0.0063	0.0076	0.0052	0.0084	0.0169	0.0106	0.0106
unflattering	0.0155	0.0150	0.0164	0.0125	0.0098	0.1324	0.0106	0.1429
comfortable	0.1582	0.1263	0.1372	0.1249	0.1365	0.1183	0.1383	0.1429
sexy	0.0102	0.0109	0.0098	0.0055	0.0062	0.0169	0.0319	-
sturdy	0.0044	0.0048	0.0024	0.0020	0.0027	0.0028	-	-
timeless	0.0039	0.0023	0.0030	0.0042	0.0042	0.0017	0.0106	-
vegan	0.0007	0.0010	0.0012	0.0010	0.0018	-	0.0106	-
trendy	0.0014	0.0022	0.0025	0.0021	0.0013	0.0028	-	-
functional	0.0014	0.0020	0.0020	0.0014	0.0018	0.0085	-	-
traditional	0.0021	0.0018	0.0017	0.0016	0.0044	0.0028	0.0106	-

Table 2: Normalized Adjective Use by Age Group (Rounded to 4 Decimal Places)

When comparing the results in Table 2, several interesting observations emerge: Softness and comfort appear to be the most sought-after features when women shop for clothing, with softness being particularly significant for women aged 80 to 99. The age group 80–89 also places far greater emphasis on the appropriateness of clothing compared to other age groups. Some unexpected patterns are also evident, such as "sexy" being most important for the 80–89 age group and "trendy" being most frequently mentioned by women aged 70–79 compared to other groups.

However, it is crucial to recognize that these results are influenced by numerous contextual factors, making it difficult to draw definitive conclusions from the counts alone. For instance, the sentiment of the sentence in which an adjective is used can significantly alter its meaning and the reviewer's intent. For example: "I love it, it is sexy and cute" conveys a positive sentiment, while "I don't like this outfit, it's too sexy" carries a negative tone. In general, this analysis holds potential sociolinguistic significance, among other areas of study, particularly if it is filtered and controlled across various dimensions such as polarity, mood, quotations, and more.

As for the program I developed, if I had the necessary knowledge, I would have designed it to eliminate the need for creating separate text files for the different age groups I defined. Instead, it would have been much more efficient and convenient to process a single text file

and handle the grouping dynamically within the same script. This approach would streamline the workflow, reduce manual steps, and make the program more adaptable for future use.

6   Final Thoughts

This project has the potential to serve as a valuable component of a larger initiative aimed at studying customer preferences and needs. By applying a similar analysis to a more extensive dataset encompassing a broader range of products, it could provide deeper insights into consumer behavior. Moreover, this approach could be leveraged as an inter-generational corpus analysis tool, examining the frequency and usage of specific words across different age groups and generations. Additionally, it offers an opportunity to explore semantic shifts in language over time, analyzing how the meaning and application of words evolve.