Inferential Statistical Data Analysis Report: Sentiment Analysis of Amazon Product Reviews

Springboard Capstone Project 1

**Summary of Project**

In this capstone project I will use the amazon products dataset for three different product categories (baby products, grocery\_foods and cell phones) to determine the polarity of a review text based on ratings and reviews. The mostly used features of the dataset are review text, and rating (ranging from 1 to 5 star). From exploratory data analysis we found in all of the three product categories:

* Most of the reviews are 5 star
* Five and one star rated reviews are found to be shorter
* Reviews corresponding to 1-star and 5-star ratings have more uppercase letters and three star rated reviews have least upper case letters.
* The % of capitals show bimodal distribution, one peak centered at around 20% and the other peak centered at around 70% of capitals.
* 1 and 5 star reviews are more likely to contain >= 60% capital letters across product categories than other star ratings.

**Analysis of Review Text**

For the statistical data analysis all the ratings are divided into two groups which are 5-star and not\_5 star (1, 2, 3, and 4 star rating). The reviews associated with 5-star rating are compared with lower rated reviews (not\_5 star) based on the word counts and percent of uppercase letters present per review. Statistical comparison of these features between 5-star and other lower rated reviews are described below.

1. **Comparison of review length between 5-star and not\_5 star rating**

To compare the review length (measured by word count) for 5-star and not5-star rating, an independent two-sample t-test for unequal variance is used. Below are the hypotheses of interest.

**Null Hypothesis**: The review length for 5-star and other rated reviews are the same.

**Alternate Hypothesis**: The review length for 5-star and other rated reviews are different.

Baby Products: t = 41.398 , p = 0.000

Grocery\_Foods: t = 58.596, p = 0.000

Cell Phones: t = 23.899, p = 0.000

In all the three product categories, since p\_value is less than 0.05 we reject the null hypothesis and conclude that the word counts for 5-star rated reviews are significantly shorter than all 1-star through 4-star (not5) rated reviews.

1. **Comparison of %caps between 5-star and not\_5 star rating**

An independent two-sample t-test for unequal variance is used to compare the % caps between 5-star and other lower rated reviews.

**Null Hypothesis**: The % of capitals for 5-star rated reviews are the same as other rated reviews.

**Alternate Hypothesis**: The % of capitals for 5-star rated reviews are different from the other rated reviews.

Baby Products: t = 12.134, p = 0.000

Grocery\_Foods: t = 13.586, p = 0.000

Cell Phones: t = 12.302, p = 0.000

In all of the three product categories, since p\_value is less than 0.05 we reject the null hypothesis and conclude that the % of capitals for 5-star rated reviews are higher than the other rated reviews.

**Conclusion:**

Review length for 5-star rated reviews are shorter than other rated reviews and the % of capitals for 5-star rated reviews are higher than the other rated reviews.