Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2003717

Beat The Market: Comprehensive Exploration of Amazon Reviews and Ratings

Recently the Sunshine Company plans to launch three new products in the online marketplace, and our team is required to provide some insights of the given data, as well as giving strategic suggestions for improving the future sales and reputation of the products. Specific tasks are separated into two big questions.

As for question 1, we first apply data cleaning by removing unnecessary information, **tokenizing all texts and lemmatizing and stemming all words**. Then we apply **LDA topic model** to give an intuitive description of what reviews care about. Next we visualize the relationship of review amount, review length, star rating and the helpful votes by time and cross analysis. The results show that reviews with larger helpfulness ratings tend to accompany with high ratings and long review lengths. Moreover, **in the early stage the averaged number of helpful votes per review far outnumber that of the latter stage. In the mean time, the ratings, review lengths and other indexes dramatically fluctuate.** These all indicate the importance of keeping a good brand image in the cold start stage.

As to question 2(a), we propose three metrics for Sunshine to track: 1. **weighted rating ratio** which represents the occurring ratio of each rating weighted by number of helpful votes; 2. **weighted sentimental score for reviews**, where we apply **logistic regression** to calculate scores for each notional word and their weighted sums by helpfulness; 3.**preference vector**, where we sort out seven attributes for each product based on LDA's results , set up dictionaries which consist of related terms for each attribute and estimate people's preference ratio on these attributes based on weighted word frequency statistics with time decay. As a Result, Sunshine can allocate different efforts on improving different product features.

In question 2(b), we consider that the **reputation** of a product is related to the averaged star ratings, the authoritativeness of its reviews and the sales volume, among which we assume the sales volume is proportional to the review number in a fixed-length time window. Therefore the reputation over that fixed-length time window can be seen as the joint contribution of features concerning rate and reviews during that period, and after calculation, results show that for hair dryer and pacifier, their reputation scores increase in the early time and tend to be stable later, whereas those of microwave keep growing the whole time.

With regard to 2(c), we continue to use the time-varying **reputation** in 2(b) as a comprehensive indicator of both ratings and reviews and apply a **nested two-layer LSTM model** to predict its value for the review sequence, for considering that **this index takes nearly every informative given features into accounts (sales volume included)**.

In terms of 2(d), we consider the **ripple effect** of reviews. After analyzing the trend of the monthly averaged number of different ratings over time, we conclude that **reviews with rate 5 tend to incite more reviews**. Besides that, we unexpectedly observe that the reviews amount of rate 1 is significantly correlated with the length of reviews, and after **Granger Cause Test**, we find that **short reviews tend to lag within two months after a large ratio of low star ratings**.

Regarding 2(e), we sort out a dictionary for **affective words exclusively** and assign a score for each of them. Then we compute the new sentimental scores of all reviews and rank reviews into five ranks. After comparing the **confusion matrix** of star rating and review ranks, we find **a mapping asymmetricity** that some reviews with strongly affective words had mild ratings, and reviews with more mildly affective words has extreme ratings. We explain this phenomena by visualizing that for those reviews, greater chances are that **both positive words and negative affective words occur, or strongly affective words are replaced by words that describe product attributes.**

Last but not least, we summarize the pros and cons of our solutions and present our insights to Sunshine's market director for the purpose of helping Sunshine take a lead in online market.

Key Word: logistic regression; LSTM; LDA topic model; Granger Cause Test; sentiment analysis.