Problem Chosen	2020	Team Control Number
C	**MCM/ICM** **Summary Sheet**	**2010617**

From data to wealth: textual information extraction and numeric processing

Summary

Sunshine Company wants to get advice on the sale strategies on their three products based on data from Amazon.com, especially the time-based patterns. In doing so, we did the following work and provided substantial and detailed conclusions, and in this process we built three main models:

In the first part, we mined as many measures as we could from the given data sets, mainly using TF-IDF moedel, semantic lexicon and Spectral Co-Clustering algorithm. Based on the mathematically processed and newly extracted measures, we constructed the review helpfulness evaluation model using the comprehensive weighing method of EWM and TOPSIS, and calculate an significant measure 'helpfulness' that indicates how helpful a review is.

In the second part, we focused on the solutions to specific problems:

- Aggregate three factors that respectively represent the basic, 'official' and 'unofficial' rating of a review. Then a classified-aggregate regression model is built to find the most informative measures to track. The result shows the 'unofficial' rating matters most, which means measures extracted from consumers not the Amazon.com are vital.
- Use the three factors in the previous step to do fitting based on Ordinary Least Squares, and obtain three possible reputation curves. We found the latter two of them can be used to detect trends of reputation.
- Detect the inflection points (symbolize the potential success or failure of products) based on MannKenddall method. Then another regression model was built to find the influential measure combinations.
- Build a fluctuation correlation analysis model that can examine the sequence and causality between two series. We used it to explore relationship between ratings and review types, and drawed conclusions like high star ratings can incite more positive reviews.
- Find out the relationship between ratings and review descriptors on the basis of quantification. And the results vary from descriptors.

In the last part, by using TF-IDF together with Co-Clustering, we generated many word clouds that can visually reflect the characteristics of a particular product, which provides specific reference for the improvement of the products.

Contents

Letter to the Marketing Director

To: Marketing Director, Sunshine Company

From: MCM Team 2010617

Subject: Strategies to win the success of new products

Date: March 12, 2020

Dear Marketing Director,

Textual data and time-based pattern are What's so special about these data sets. Our team put emphasis on these two parts and draw substantial and novel conclusions. Furthermore, our study is able to help you not only with the conclusions, but also with the available models.

In doing so, our team dug firstly deep into the ratings and reviews of these three products on Amazon using various scientific methods, such as TF-IDF moedel, Dhillons Spectral Co-Clustering algorithm and the comprehensive weighing method of EWM and TOPSIS.

Then, three models were built during the process to draw useful conclusions from known data:

- the first model evaluates how helpful a review is, which can help you save a lot of energy by selecting the most useful of the many reviews available for analysis after the new products are launched;
- the second model decides the most informative measures to track for company.
- the last one can be used to judge the sequence and causality between two sets of numbers, which is helpful for business analysis.

In terms of conclusions, we found that:

- The 'unofficial' rating of products is the most informative measures to track. Companies should pay special attention to the quality of customer reviews content, which contributes the most for the sales. To be more specific, to detect the measures such as helpful votes, ratio of helpful, complexity of reviews and so on.
- star rating weighted by helpfulness can best detect the trends of reputation.

- The measure combinations that can be used to detect the potential success or failure of products are review's two sides attribute (the review contains both negative and positive content, E. G. ′...is good, but....′), degree of positive and complexity.

- High star ratings can incite more positive reviews. So it's a good idea to find ways to get the high star reviews seen by more customers.

- Some descriptors, such as great, is generally strongly associated with rating levels, and the higher the level, the more it appears. However,according to our sentiment measure, although great appears in 5 star review, it doesnt necessarily mean this review is positive. But the conclusions vary from different descriptors, as is shown in appendix.

- We have also extracted features for different products, and visualize them using world clouds. You can intuitively see the keywords of user reviews for each specific product, so as to obtain the potentially important design features that caters to the market. For example, for the hair dryer 423960 ($product\_parent$) with high sales, the feature words include reliable, works well and fancy.

Overall product sales on Amazon are increasing year by year, and there are opportunities for companies to bring new products into the store. In our analysis, there are some products where the latter are ahead. We hope our team analysis and results can help you in your online sales strategy and product design.

Yours sincerely,

MCM Team 2010617

1 Introduction

1. Problem Summary

   On Amazon, customers can rate and review purchases through the star ratings reviews helpfulness rating mechanism. Sunshine Company wants to get advice on the sales strategy of its three new products based on existing data, which involves information under Amazon's rating system over the time periods.

2. Our model

   The main technical route is as follows:

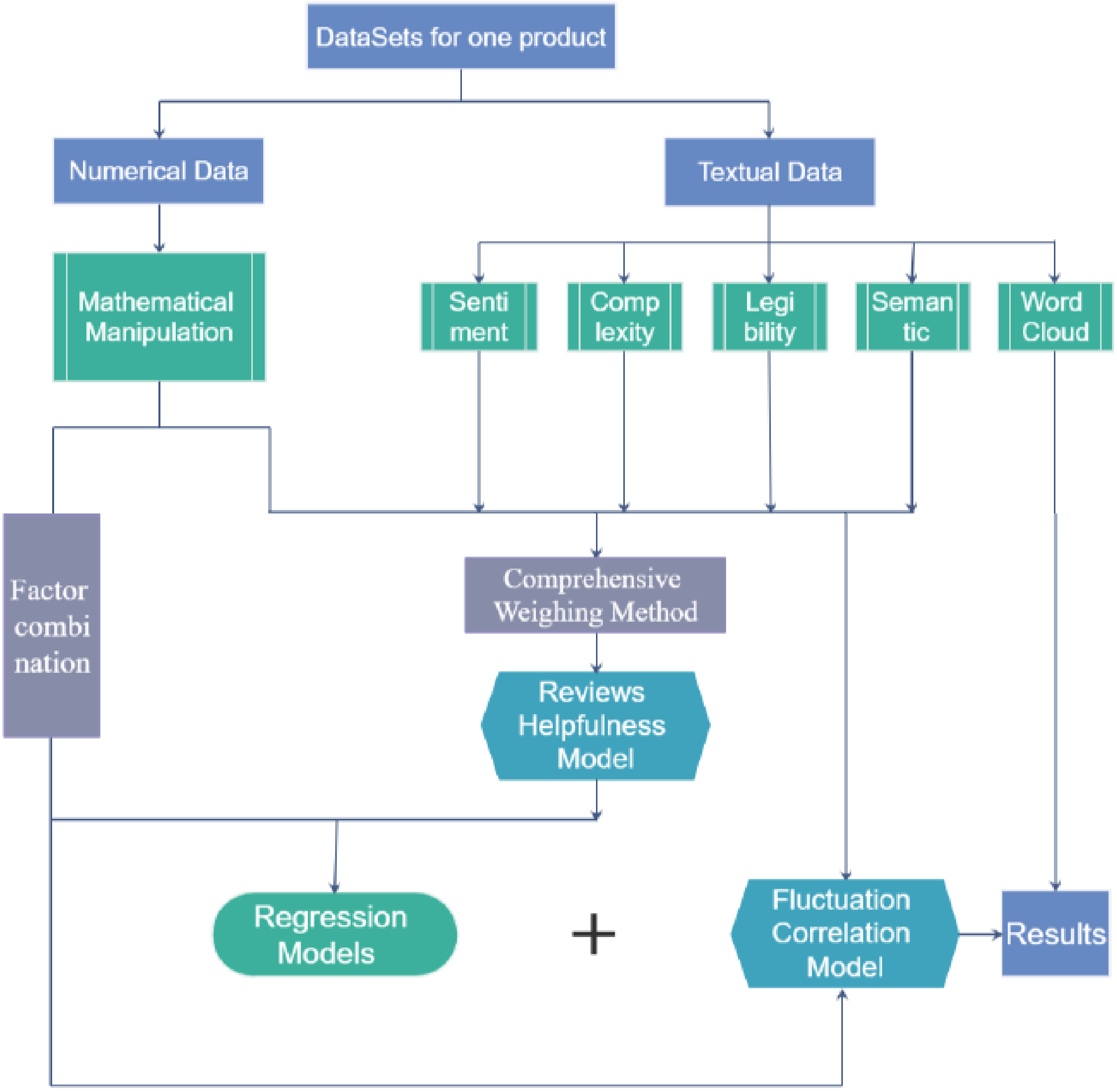


Figure 1: Technical Route

2 Background

Analysis of online reviews has an important enlightening effect on sales performance and business decision-making[4]. Lots of researches have been conducted in this area, and here are two relevant aspects.

1. Reviews Helpfulness

   Due to the large amount of reviews data, "what factors make a review helpful for potential customers to make purchasing decisions" has become an important research topic. According to a number of studies, consumers tend to judge review helpfulness based on two major features : the extent to which the review familiarizes consumers with the product and the perceived credibility [5]. To be more specific, such as:

   - length : reviews with medium length is considered more helpful
   - legibility : reviews that are easy to read and understand
   - sentiment : positive, neutral, negative or mixed ones
   - professionalism : a professional consumer′s review is more helpful

2. Natural Language Processing

   Since reviews are natural language, to extract as much information as possible from them, it is natural to use Natural Language Processing (NLP) technologies. When apply to the analysis of online reviews, NLP is often used to do sentiment identification[6,7], semantic analysis[8] and other information retrieval[9,10].

3 Assumptions

- Assuming everyone voted in a fair manner on whether others' reviews are helpful, so the voting results are objective and accurate.
- Assuming that the reviews written by Customers who were authorized as Amazon Vine Voices are accurate and can be over-weighted.
- Although someone who bought a product may not review online, there is still a positive correlation between the sales of products and number of reviews, so we can use it as an approximate substitute for sales of particular products.

4 Part I: Mining for the given data sets

4.1 Data Preprocessing

4.1.1 Missing values processing

A total of 4 values were found to be missing, two for ′review_headline′ and two for ′review_body′. Considering that the number of missing is small, and there are lots of examples in other reviews have the same contents in ′review_headline′ and ′review_body′ or ′review_headline′ and ′star_rating′ ,for instance, "5" in ′star_rating′ and "Five Stars" in ′review_headline′, we use those two methods to fill missings. The results are shown as follows:

Table 1: Fill in missing values

review_id	star_rating	review_headline	review_body
R74VTHE48J4IQ	3	Smaller	Smaller
R3BVF5UJ5TMXHK	5	Five Stars	Five Stars
R2TW4FSXQ60M75	5	Five Stars	Used once ...
R24Y12M6JKTTQM	4	Four Stars	It is very nice ...

4.1.2 Duplicate values processing

We found 4 and 12 duplicate values in each of the two datasets, microwave ovens and baby pacifiers. While it's reasonable for parents to buy a pacifier product several times at the same date and write same reviews, it seems that the duplicated reviews of microwave oven, written by a person for 4 times at a single day, are just repeated complaints about his awful experience. Also given that all the duplicated pacifier reviews were purchase-verified which were not for ovens′, we keep all the pacifier reviews but only one for those 4 duplicated microwave oven reviews.

4.1.3 Generate new features

We combine some ratio-like features. For example, we divide the $total\_votes$ by $helpful\_votes$ to get $helpful\_ratio$.

4.1.4 Data Normalization

Normalization can provide an approach for comparison of different numerical features and reflect the combined results of different factors.

We use min-max normalization, doing linear transformation of the raw data, mapping data set $x = \{x_1, x_2, \ldots, x_n\}$ into $[0, 1]$. The normalized value is

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{1}$$

4.2 Review Information Extraction

4.2.1 Extract structured word vector

Algorithms often cannot directly process text data so we need to extract some features out of the texts and convert them into numeric features.

We focus on our attention on using TF-IDF model to remove stop words and extract key features hiden in the 'review_body'. In a given document, term frequency (TF) refers to how often a given term appears in the document. If a word appears very frequently in a particular document and very low in other documents, then the word is likely to be a word unique to the document and can describe the document well. Word d_i's frequency TF_{ij} in document D_j is:

$$TF_{ij} = \frac{n_{ij}}{n_{dj}} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{2}$$

with n_{ij} is the number of occurrences of the keyword d_i in the document D_j, n_{dj} is the total number of words in the document D_j. And inverse document frequency is defined as below:

$$IDF_i = \log\frac{N}{d} = \log\frac{N}{\{j : d_i \in D_j\}} \tag{3}$$

with N is the total number of documents in the data set, and d is the number of documents containing the keyword d_i. And finally we get TF-IDF:

$$TF - IDF_i = TF_{ij} \cdot IDF_i \tag{4}$$

4.2.2 Sentiment Extraction

The TF-IDF vectorized posts form a word frequency matrix, which is then biclustered using Dhillons Spectral Co-Clustering algorithm.

Clusters rows and columns of an array X to solve the relaxed normalized cut of the bipartite graph created from X as follows: the edge between row vertex i and column vertex j has weight $X[i, j]$.

For a few of the best biclusters, its most important words get extracted. Then look each word up in a senti-dictionary, called $vader\_lexicon$, and compute its tendency to three attitude levels: $positive$, $negtive$ and $neutral$. We calculate three scores for each review accordiing to those three attitude level. Then we do normalization to make three scores sum up to 1 and generate the sentiment vector.

4.2.3 Legibility Extraction

A large number of studies have shown that lexical and syntactic difficulties have a high predictive power for text legibility. For General Purposes, formulas containing these two variables can make fairly accurate and direct predictions of text legibility[11]. In this paper, we make use of the vocabulary angle, based on the study of vocabulary in Brown Corpus to compile the common vocabulary list, and according to the coverage of words in these lists to the text corpus, we can determine the legibility of the text. And from that, we've got the new measure $understandable$ which can range from 0 to 1.

4.2.4 Complexity Extraction

Once the $understandable$ feature is done, we can easily define a new feature called 'complexity'. As reviews with strong legibility can convey their thoughts and feelings to other people immediately, a review with strong complexity may introduce more detailed informations. We'd like to find out whether detailed reviews will affect the sales of the product. The reviews' $complexity$ can be difined as follw:

$$complexity = e^{-understandable}, \quad understandable \in [0, 1] \tag{5}$$

with $complexity$ also range from 0 to 1 and is negatively related to $understandable$.

4.3 Evaluation Model for Reviews Helpfulness

In order to make the best use of the textual information of reviews, it is necessary to evaluate their helpfulness. On the basis of combing the previous conclusions in this field, we put forward an evaluation system model which is suitable for the existing data sets. On the basis of extracting the effective patterns and measures(hereinafter together referred to as the indexes), the comprehensive weighting method of Entropy-weight and TOPSIS is applied, and we successfully obtained the helpfulness degree of each comments.

4.3.1 Establishment of Index System

As for indexes, $helpful\_votes$ and $helpful\_ratio$ reflect two different mindsets among consumers: some are more focused on the quantity of reviews, while others prefer quality; vine can be used to determine whether a reviewer is "professional" and make whose review objective and clear. The indicators we have selected are shown in the table below.

Table 2: Indexes for Review Helpfulness Evaluation

Index	Type	Value	P/N^*
helpful_votes	Quantitative	Non-negtive Integer	P
helpful_ration	Quantitative	Decimal between 0∼1	P
understandable	Quantitative	Decimal between 0∼1	P
complexity	Quantitative	Decimal between 0∼1	N
vine	Qualitative	Boolean Y/N	P
verified_purchase	Qualitative	Boolean y/n	P
sides	Qualitative	One-sided or Two-sided	/

* P/N illustrates whether the effect on helpfulness is positive or negative

We also work with reviews' $sides$ when measuring review contents. People may often consider two-sided reviews to be more helpful than one-sided ones when the reviewers are experts[12]. $verified\_purchase$ is another important index which confirms the purchase. When a review is understandable, which means it doesn't use many obscure words or grammars. Higher readability makes the review become more helpful while complexity may introduce more details about the product which should be also taken into considerations[13].

4.3.2 Comprehensive Weighting Methodology

Now suppose there are p reviews, which poses the alternative set; n evaluation indexes, which poses the criteria set; then we can get the decision matrix $X = [x_{ij}]_{p\times n}$.

To obtain more reasonable weighting results, we adopted the combination of 'Entropy-weight' method and 'TOPSIS' method, whose principles are as follows:

(1) Entropy Weight Method

When applied to the entropy weight method, the higher the entropy value, the less effective information the index carries, and the more chaotic the system. Assuming that the data have been non-negative processed, the entropy of the jth evaluation index can be calculated:

$$H_j = -(\ln n)^{-1} \sum_{j=1}^{p} q_{ij} \ln q_{ij} \tag{6}$$

with $q_{ij} = x_{ij} / \sum_{i=1}^{p} x_{ij}$ represents the proportion of the ith review in the jth index. Further, the entropy weight of the jth criterion can be calculated, and then the entropy weight method score of each review can be calculated as below:

$$W_j = \frac{1 - H_j}{n - \sum_{j=1}^{n} H_j} \tag{7}$$

(2) TOPSIS Method

TOPSIS is another method based on the distance between the chosen alternative and the ideal solution. Here the optimal score of a certain attribute of review is the ideal solution of this attribute, while the negative ideal solution is the opposite, indicating the worst score of a certain attribute. We calculate the distance from the ith review to the positive ideal solution negative ideal solution by following formulas:

$$d_{1j} = \sqrt{\sum_{j=1}^{n} (x_{ij} - a_j)^2}, \quad d_{2j} = \sqrt{\sum_{j=1}^{n} (x_{ij} - b_j)^2} \tag{8}$$

Finally, we calculate the TOPSIS score of the ith review by those two distances above:

$$y_i = \frac{d_{2i}}{d_{1i} + d_{2i}} \tag{9}$$

4.3.3 Calculation and Results

We calculate the comprehensive weighting method based on a performance matrix composed of four numerical indexes, namely $helpful\_votes$, $helpful\_ratio$, $understandable$ and $complexity$. Since the four indexes were given by lots of consumers, which means it's "public and unofficial", unlike $vine$ and $verified\_purchase$, which are given directly by Amazon.com. So we call the results of this stage "Unofficial Helpfulness". Due to space constraints, table 2 only shows 7 reviews out of 32021. All the results are shown in appendix.

Table 3: Resluts of Helpfulness for Each Review

review_id	Entropy-Rank	Entropy-Score	TOPSIS-Rank	TOPSIS-Score
R35BHQJHXXJD59	16705.5	0.028909	16729.5	0.001014
R230LCPQDOFJJZ	24633.5	0.027926	24638.5	0.000733
R21NN9ONVZITI0	29482	0.001702	29482	0.000312
R18NK8BQ5LPMZZ	24633.5	0.027926	24638.5	0.000733
R2NEOR5Y0P7U8V	17808.5	0.027681	17828.5	0.000983
...	...	...	...	...
R3N0F2FKJOMGKK	2764.5	0.354135	1776.5	0.015989
R2WDYCZI91LL6K	27390	0.021149	27391	0.000558

Though the scores of Entropy-weight method and TOPSIS method are different in the overall size, but from the ranking point of view, the two results are correspondent which shows the comprehensive method of weight is very effective.

The second step is to apply Boolean indexes to score. When both conditions $vine$ is 'Y' and $sides$ is 'two-sided' are met, the score is rewarded; when $verified\_purchase$ is 'n', the score is reduced. The results obtained after this stage are comprehensive helpfulness:

$$\begin{aligned} ComprehensiveHelpfulness = {} & OfficialHelpfulness \\ & + UnofficialHelpfulness \end{aligned} \tag{10}$$

Now we have the helpfulness value for each review, which provides an important reference for exploring the role of review in business strategy.

5 Part II: Analysis and solution of specific problems

5.1 Most Informative Measures to Track

To find out the most informative measures, we construct a classified-aggregate regression model of three factors, constructs their regression relationship with sales. The dependent variable here is a $products' sales$, because it's the final transformation results and the most concern of business operators.

Among various extraction ways we have, $star\_rating$ is the most basic and intuitive feature to show a consumer's feeling, which is also the first layer of Amazon's mechanism. We can generate a scatter plot between the sales and $mean\_star\_rating$ of each hair dryer, aggregated by $product\_parent$. However, the plot turn out to be meaningless as all the points gathered together around a high $star\_rating$ while sales are very different. Therefore, review-based measures are essential and needed to be added.

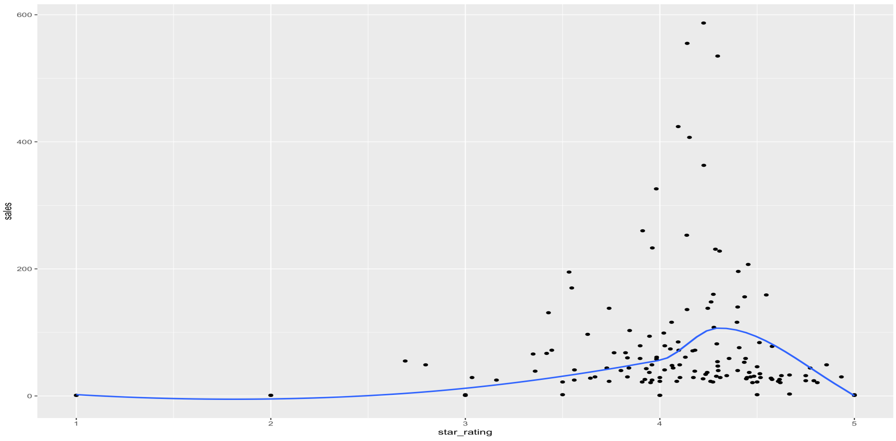
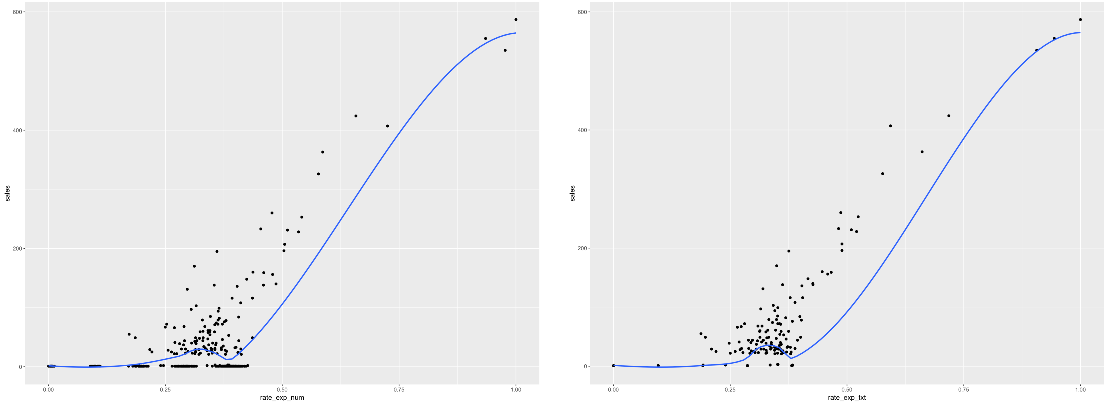


Figure 2: mean_star_rating for each dryer

When considering review-based measures, similar to part 4.3, we classify them into two types, Official and Unofficial. Actually, the measures are used to calculate the helpfulness of reviews, so we aggregate these measures by type and amplify the basic $star\_rating$:

- $x_1 : star\_rating = r$
- $x_2 : rate\_exp\_num = r \times e^{vine+verified\_purchase}$
- $x_3 : rate\_exp\_txt = r \times e^{helpfulness+two\_sides}$

Now we can generate dryers' sales scatter plots respectively based on $rate\_exp\_num$ and $rate\_exp\_txt$, which are significantly better than before:



(a) rate_exp_num (b) rate_exp_txt

Figure 3: Frames of the house and hotel data sets

Then we use linear regression model to fit products' sales and above three rating meatures. The results of three products are respectively shown in the table, which summarizes the importances of each measure variables:

Table 4: Importance given by Regression Model

measure variables	hair_dryer	microwave	pacifier
rate_exp_txt	0.7782913	0.4796037	0.5996864
rate_exp_num	0.1499111	0.4028003	0.3184868
star_rating	0.0717975	0.1175959	0.0818266

As is shown in the table, for all of the three products, the independent variable that has the greatest impact on sales is $rate\_exp\_txt$, which indicates that unofficial feelings about reviews have greater importance. Therefore, companies should pay special attention to the qualify of customer reviews content, which contributes the most for the sales. To be more specific, such as reviews' $helpful\_votes$, $helpful\_ratio$, $complexity$ and so on.

5.2 Measures that Suggests Reputation Trends

Considering where they came from, all of the above three variables $rate\_exp\_txt$, $rate\_exp\_num$ and $star\_rating$ can represent the reputation. The $star\_rating$ is the basic and intuitive symbol of reputation, while the other two are the scale-up of it based on different priorities.

In order to explore the trend of reputation change of each product, we fit the measures that can represent reputation in a chronological order, with specific steps as follows:

- Step 1 : Time aggregation

  Since the amount of data per day is too large, we first aggregate the data on a monthly basis, and the values of the three reputation variables are averaged. After aggregation, there are nine(3×3) reputation measures in time series.

- Step 2 : Draw the scatter plots

  According to the distribution of scattered points, the approximate and suitable curve type can be selected. We draw the nine scatter plots in this step. For example, the scatter plot of star_rating for hair dryer:

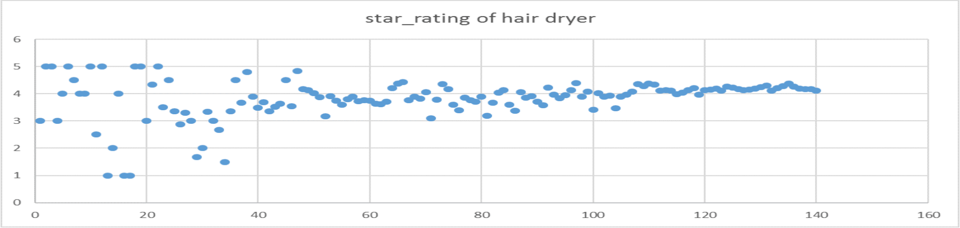


Figure 4: mean_star_rating for hair dryer

- Step 3 : Equation selection based on OLS

  Use OLS (ordinary least squares) method to find the best fit. Least squares are very common, and its principle is relatively simple, so we omit it here. Here we use the results of microwave ovens as an example:

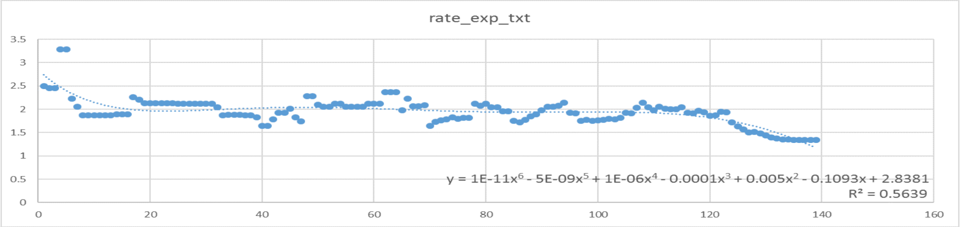


Figure 5: rate_exp_txt

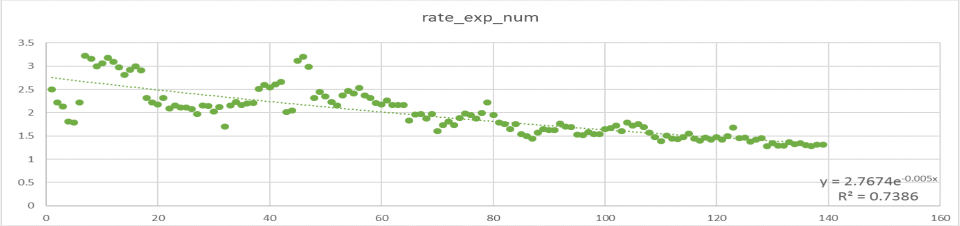


Figure 6: rate_exp_num

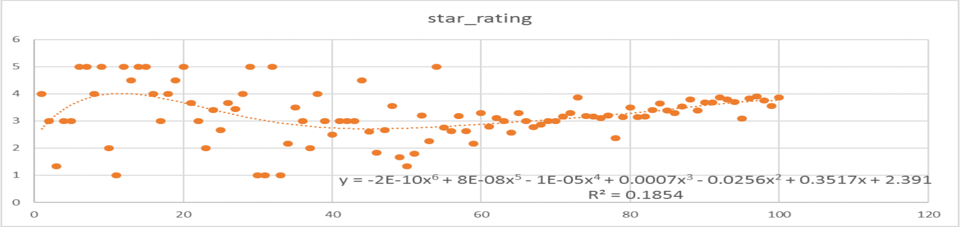


Figure 7: mean_star_rating

Table 5: Results of Reputation Fitting - MV

	R^2	Equation
rate_exp_txt	0.5639	$y = 0.0001x^3 + 0.005x^2 - 0.1093x + 2.8381$
rate_exp_num	0.7386	$y = 2.7674e^{0.005x}$
star_rating	0.1854	$y = 0.0007x^3 - 0.0256x^2 + 0.3517x + 2.391$

The fitting effects of three measures are different. As shown in the table 4, the fitting effect of $star\_rating$ is the worst, which means that if it is used to represent reputation, the trend over time will be difficult to detect, because it fluctuates much. From Figure 5-7, microwave ovens' reputation is on a downward trend according to $rate\_exp\_txt$ and $rate\_exp\_num$, however, according to star_rating, the trend change is not obvious. Actually, for baby pacifiers and hair dryers, the situations are similar.

5.3 Measure Combinations that Indicate Products' Success/ Failure

5.3.1 Aggregation Processing

For different products, the factors that affect their success or failure are different. we aggregated the product types by $product\_parent$ and gathered data on a monthly basis as the time factor is also indispensable for the prediction of the factors causing the success or failure.

5.3.2 Corner Detection

We view the inflection point in the product sales time series as a sign of potential success or failure, because it means that the product is starting to turn, and the company should recognize this and remedy it before the product actually fails. MannKenddall method is used to detect those inflection points, it is a non-parametric method, which means it does not need samples to follow a certain distribution, and is not disturbed by a few outliers. It is more suitable for sequential variables and easy to calculate:

For the time series X with n sample sizes, construct:

$$S_k = \sum_{i=1}^{k} r_i, \tag{11}$$

$$where \quad r_i = \begin{cases} 1 & x_i > x_j \\ 0 & else \end{cases} \quad j = 1, 2, ..., i \tag{12}$$

Under the assumption of random independence of time series:

$$UF_k = \frac{S_k - E(S_k)}{\sqrt{Var(S_k)}}, k = 1, 2, ..., n \tag{13}$$

where $UF_1 = 0$, mean $E(S_k) = \frac{n(n+1)}{4}$, and variance $Var(S_k) = \frac{n(n-1)(2n+5)}{72}$

In the case of a product with $product\_parent$ is 732252283, the result of detecting the inflection point is as followed:

As the UFK and UFK graphs show, when the value of UFK or UFK is greater than 0, the sequence shows an upward trend, while less than 0 indicates a downward trend. When they exceed the critical line, they indicate a significant upward or downward trend.The range beyond the critical line is defined as the time zone in which the inflection occurs. If two curves of UFk and UBk intersect and the intersection point is between the critical boundary, this point symbolizes the time when the inflection begins.

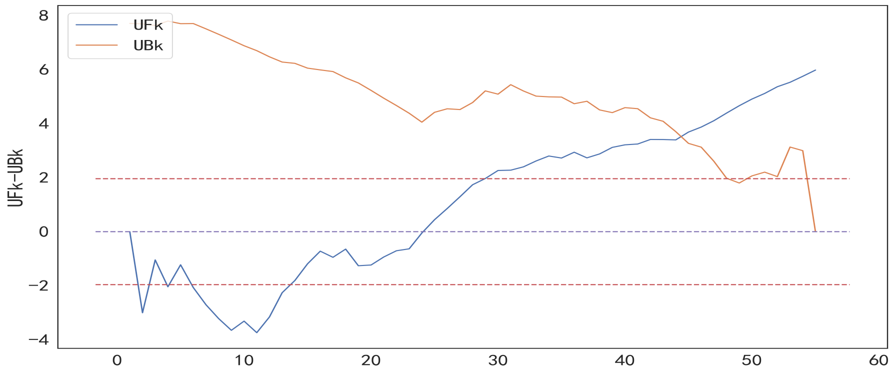


Figure 8: Results of Detection

5.3.3 Regression

To determine the combinations of text-based and ratings-based measures that best indicate a potentially success or failure, we construct another linear regression based on the data of the year before the inflection point.

Table 6: Importance given by Regression Model

variabels	importance	variabels	importance
two_sided	0.6430	understandable	0.0339
positive	0.1548	star_rating	0.0211
complexity	0.0457	att_pos	0.0140
negetive	0.0407	star_rating_helpful	0.0133

The results turned out the linear regression fits well, for 4.5820 for mean squared error and 0.9244 for R^2. From the result, the measure combinations that detect the potential success or failure of products are review's $two - sided$ attribute, $positive$ and $complexity$.

That means a product with more two-sided reviews and convey more positive, detailed information may lead to a success.

5.4 Relationship Between Ratings and Review Types

5.4.1 Fluctuation Correlation Model

We establish a model to test the fluctuation correlation between two groups of variables. Fluctuations are deviations from expected values. For each time series, extract out its fluctuation characteristics series(FCS)[14], then calculate the correlation between two FCS which indicates the sequence and causality between these two original series. The detailed fluctuation correlation modeling steps are as follows:

- Step 1 : Prediction method selection

  Actually we have tried many prediction methods, and use mean-square error(MSE) as the optimization criterion, and finally, exponential smoothing method is determined to be the prediction method. Since the principle of exponential smoothing is very basic, here we omit it.

- Step 2 : Extraction of the FCS

  The predicted series is subtracted from the original series to obtain FCS, as is shown in the figure below.

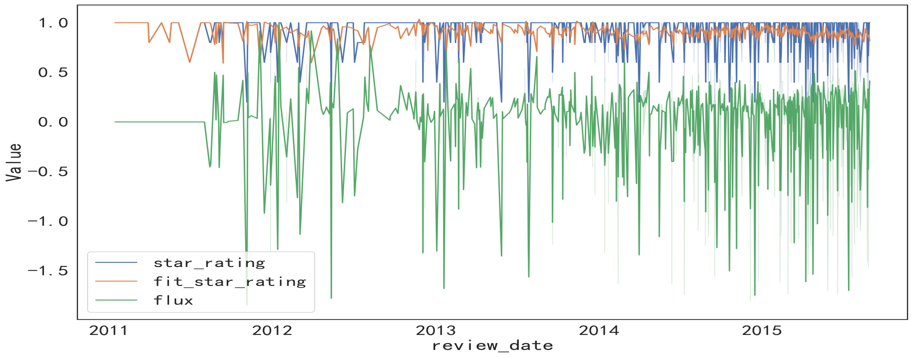


Figure 9: Step 2: Extraction of the FCS

- Step 3 : Correlation between FCS

  To calculate the correlation between FCS, and to get the final fluctuation correlation between the original series. There are many alternative methods of correlation analysis, such as J-measure, pearson correlation, spearman correlation, Granger and so on.

5.4.2 Quantification of Review Types

In Part I, we have obtained the attitude levels in part 4.2.2. Here we use $positive$,$negative$ and $neutral$ scores to symbolize the type of reviews. The larger the positive score is, the more positive the review.

5.4.3 Correlation Analysis and Results

Before correlation analysis, time series are shifted to judge whether start rating can cause the change of review types, since there is a certain time lag between when consumers see the star rating in the previous period and when they write down their own reviews. Here the series of positive scores is translated back one period (about one to sixty days according to the data).

The correlation analysis include two parts:

(1) Common Correlation Analysis

Analyze the correlation between the original time series of star_rating and positive degree. As is shown in the table (df is the degree of freedom), all the data do not conform to the normal distribution. Therefore, spearman correlation analysis is adopted to calculate the correlation.

Table 7: Test results of normality

	Kolmogorov-Smirnova			Shapiro-Wilk		
	coefficient	df	significance	coefficient	df	significance
f1	.255	833	.000	.768	833	.000
f2	.073	833	.000	.967	833	.000
star	.437	833	.000	.543	833	.000
pos	.152	833	.000	.867	833	.000

The result of spearman correlation analysis is 0.27, and P value = 0.000 < 0.01. So star rating and the positive degree of reviews are significantly positively correlated, while the degree of correlation is not high.

(2) Fluctuation Correlation Analysis

Extract FCS of these two original series (f1 and f2), and analyze the fluctuation correlation between the original time series by calculating the correlation coefficients of the two FCS.

f1 and f2 also do not conform to the normal distribution, and the result of spearman correlation is 0.224, with P value < 0.01. Therefore, we can assume that high star ratings can incite more positive reviews.

5.5 Relationship Between Ratings and Review Descriptors

5.5.1 Quantification of Review Descriptors

For each review descriptor, we give two quantitative measures. One is the simple counting, and the other is based on the sentiment extracted in part 4.2.2.

5.5.2 Results

Table 8: Distribution for ’great’ of Hair Dryer

star_rating	positive	neutral	negative	total
5	344	1608	2	1954
4	49	477	0	526

Table 8: Distribution for 'great' of Hair Dryer

star_rating	positive	neutral	negative	total
3	0	130	0	130
2	0	79	0	79
1	0	82	0	82

As shown in the table 5, the descriptor 'great' is generally strongly associated with rating levels, and the higher the level, the more it appears. However, according to our sentiment measure, although 'great' appears in 5 star review, it doesn't necessarily mean this review is positive. But the conclusions vary from different descriptors, as is shown in appendix.

6 Part III: Feature extraction of specific products

By using TF-IDF toghther with Spectral Co-Clustering algorithm described before, we can easily extract some best words by normalized cuts and comparing sums inside and outside the bicluster.

The extracted topic words can help with the analysis with the advantage and disadvantages about a product by combining the relationship between the contents of those keywords and the product's *star_rating*, including adjusted rating meatures by *num* and *txt* methods. Here are some of the keywords we extracted and we made them into several word clouds for a clearer presentation. You can view more example clouds in Appendix.

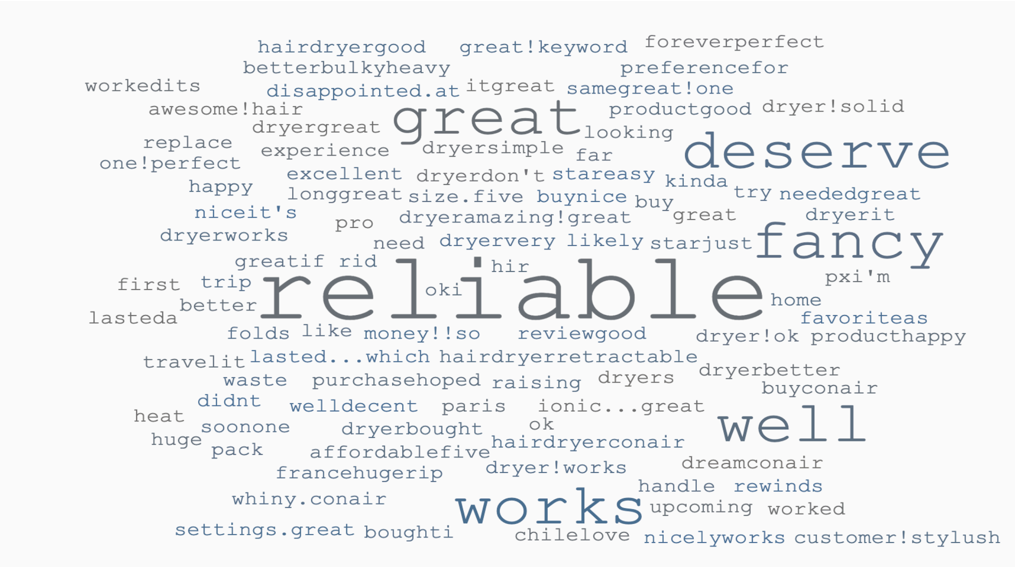


Figure 10: Key reviews of hair dryer 423960

7 Model Evaluation

7.1 Review Helpfulness Model

When using two or more methods to do evaluation, the consistency between the two evaluation methods' results should be tested, so it is necessary to carry out the correlation test for the two results in part 4.3.3. The normality test results of review helpfulness scores shows that, both of the results do not conform to the normal distribution. Therefore, spearman and kendall correlation analysis are adopted. Through calculation, the result is:

Table 9: Results of correlation analysis

	Spearman		Kendall	
	coefficient	significance	coefficient	significance
EWM v.s. TOPSIS	.974	.000	.916	.000

7.2 Regression Model

After several regression model, one thing that really matters is to do some diagnosis plots to see whether the model is well-fitted. Here we presents a example of hair dryer to illustrate the accuracy of our model.

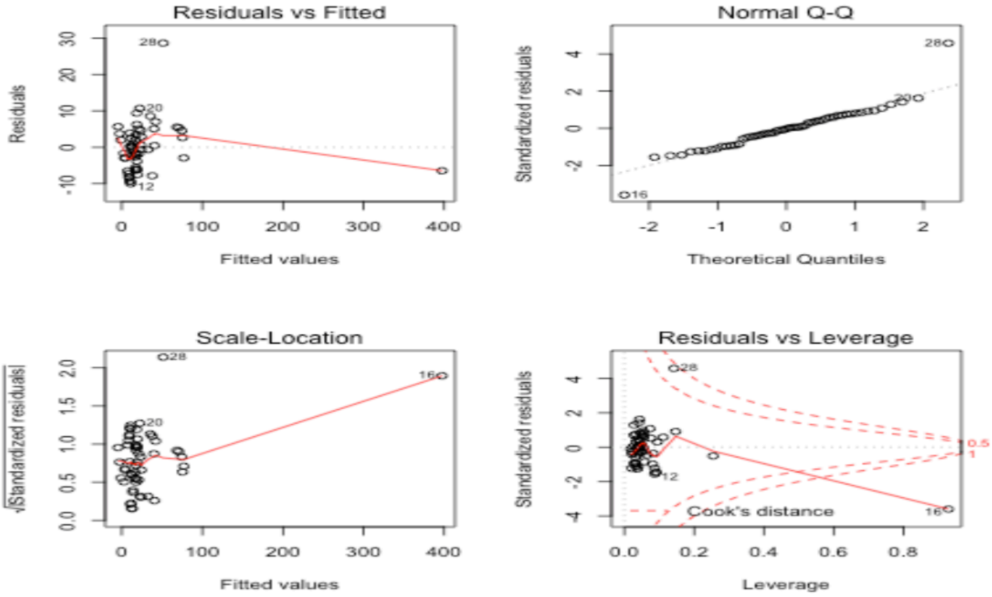


Figure 11: Regression diagnosis

These two sets of figures show that the polynomial regression fitting effect is ideal, and basically meets the linear assumption, residual normality test, and homoscedasticity. You can view more diagnosis plots in Appedix of other products.

8 Strengths and weaknesses

8.1 Strengths

- **Substantial and Detailed**
  Our research content is very rich, on the basis of deeply mining data sets and composing new measures, we also build more than three models to do analysis, and get a lot of considerable conclusions.
- **Solid theoretical support**
  All of our methods and models are supported by academic literature with available sources.

8.2 Weaknesses

Due to the limited content of the datasets, when building the review helpful evaluation model, the amount of indexes is far from enough. Therefore, the helpfulness score may be underestimated compared with the reality.

References

[1] Jia-Lang Seng, T.C. Chen, An analytic approach to select data mining for business decision, Expert Systems with Applications, Vol 37, pp. 8042-8057, 2010.

[2] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, AI Magazine (1996), pp. 3753.

[3] Indranil Bose, Radha K. Mahapatra, Business data mining a machine learning perspective, Information and Management, Vol 39, 2001, pp. 211-225,

[4] Shankhadeep Banerjee, Samadrita Bhattacharyya, Indranil Bose, Whose online reviews to trust? Understanding reviewer trustworthiness and its impact on business, Decision Support Systems, Vol 96, 2017, pp. 17-26,

[5] Zhou, S. S. and Guo, B. The order effect on online review helpfulness: A social influence perspective. Decision Support Systems 93, 2017, pp. 77-87.

[6] Hai, Z., et al. A statistical nlp approach for feature and sentiment identification from chinese reviews. in CIPS-SIGHAN Joint Conference on Chinese Language Processing. 2010.

[7] Ghorpade, T. and L. Ragha. Featured based sentiment classification for hotel reviews using NLP and Bayesian classification. 2012 International Conference on Communication, Information and Computing Technology. 2012.

[8] Flank, S. A layered approach to NLP-based information retrieval. 17th International Conference on Computational Linguistics, Volume 1. 1998.

[9] Li, H. and Z. Lu. Deep learning for information retrieval. in Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. 2016.

[10] Smeaton, A.F., Using NLP or NLP resources for information retrieval tasks, in Natural language information retrieval. 1999, Springer. p. 99-111

[11] Li, S. An overview of Legibility Research J. Journal of PLA Foreign Language Institute, 20004:1-5.

[12] Chen, M.-Y.J.O.I.R., Can two-sided messages increase the helpfulness of online reviews? 2016.

[13] Zhao, F. An analysis of the usefulness of online customer reviews in Amazon. com. 2018, Guangdong University of Foreign Studies

[14] Su, Y., et al. CoFlux: robustly correlating KPIs by fluctuations for service troubleshooting. in Proceedings of the International Symposium on Quality of Service. 2019.

Appendices

Appendix A

Table 10: Distribution for ′problem′ and ′problems′ of Microwave Ovens

star_rating	positive	neutral	negative	total
5	0	56	0	56
4	0	31	0	31
3	0	16	0	16
2	0	20	0	20
1	0	99	0	99

Table 11: Distribution for ′bad′ of pacifiers

star_rating	positive	neutral	negative	total
5	3	98	0	101
4	1	48	0	49
3	0	30	0	30
2	0	26	0	22
1	0	22	2	28

Appendix B

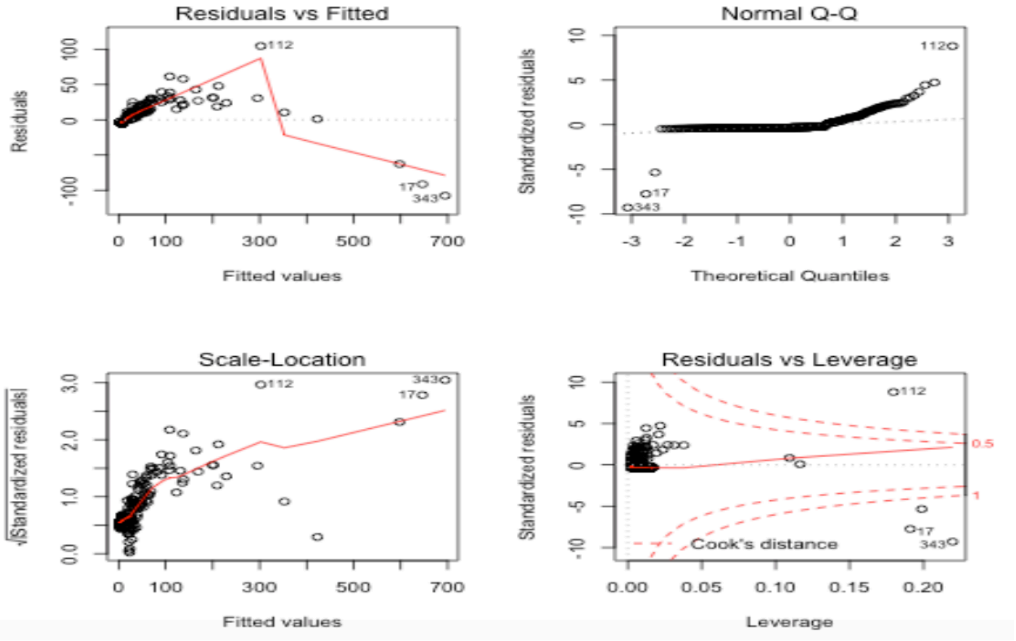


Figure 12: Regression diagnosis

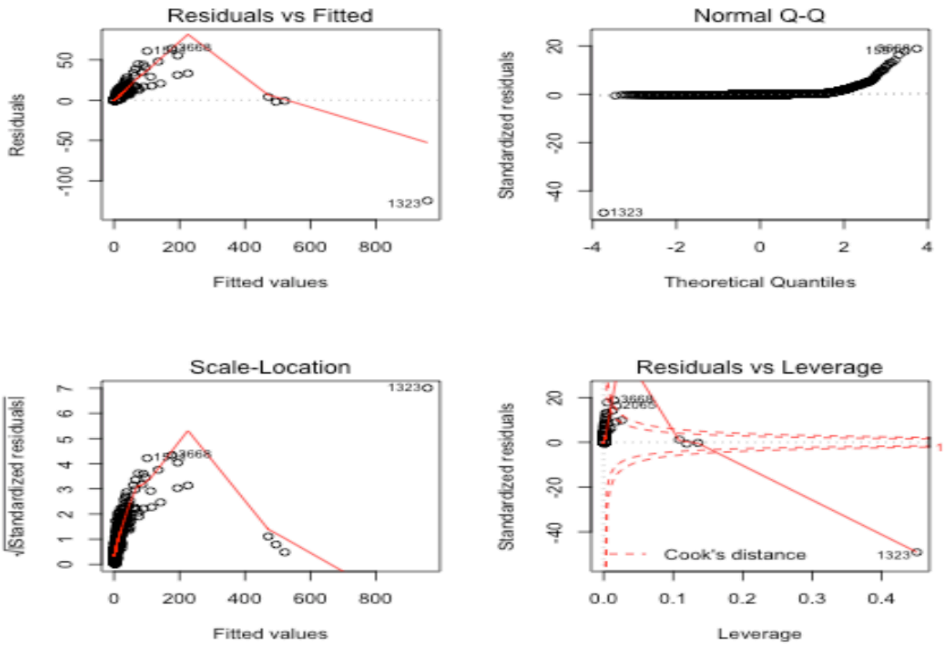


Figure 13: Regression diagnosis

Appendix C

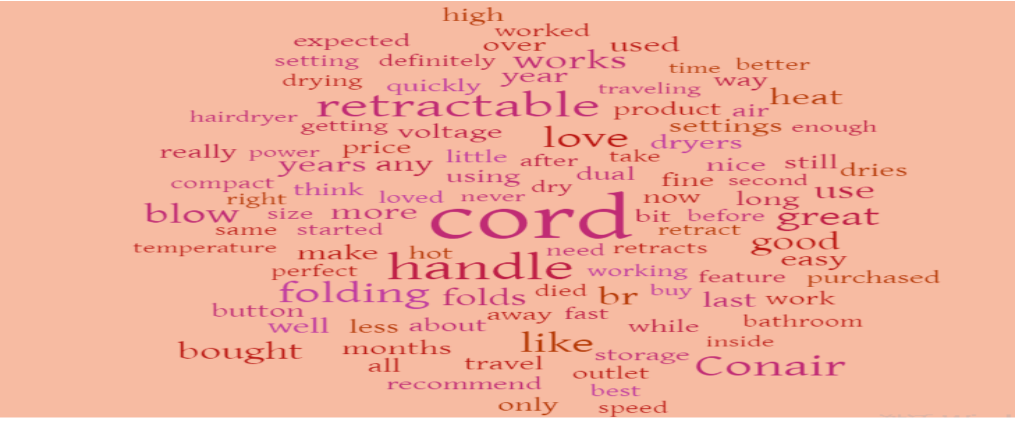


Figure 14: Key reviews of hair dryer 423960

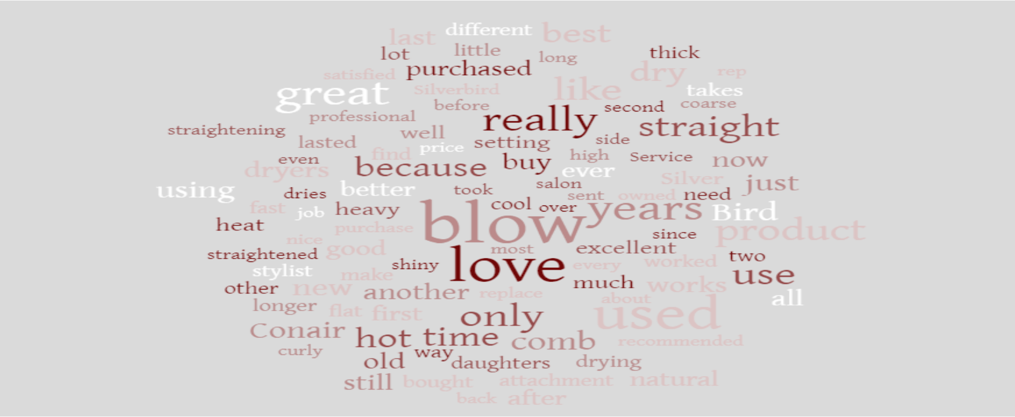


Figure 15: Key reviews of hair dryer 423960

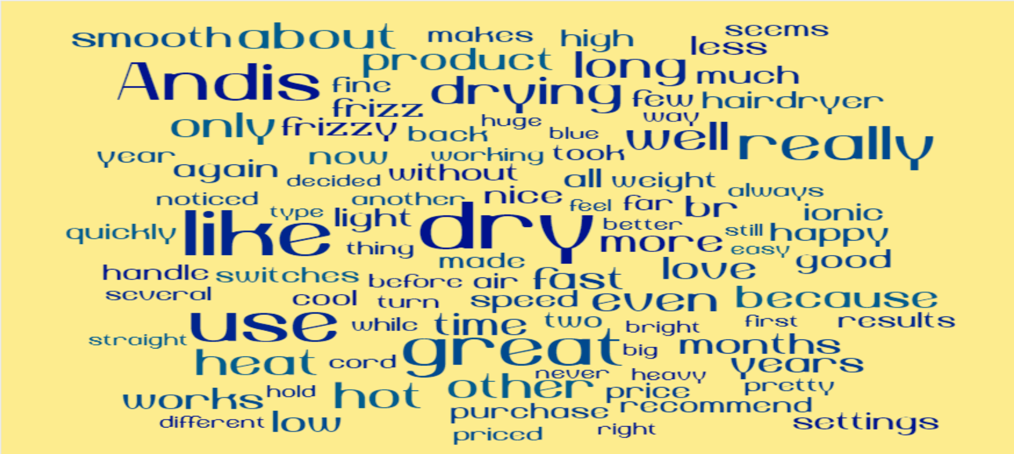


Figure 16: Key reviews of hair dryer 423960