Content

[Introduction 3](#_Toc193016946)

[Data Collection 4](#_Toc193016947)

[1. Data richness 4](#_Toc193016948)

[2. The data set meets the requirements of research objectives 4](#_Toc193016949)

[3. Well-structured data 4](#_Toc193016950)

[Data processing 5](#_Toc193016951)

[1. Removing Redundancies, Merging Text, and Handling Missing Data 6](#_Toc193016952)

[2. Constructing a Data Cleaning Function for Text Standardization 6](#_Toc193016953)

[3. Stop Words Removal and Text Tokenization 7](#_Toc193016954)

[4. Lemmatization 8](#_Toc193016955)

[5. Cloud visualization 9](#_Toc193016956)

[Sentiment Analysis 10](#_Toc193016957)

[VADER 10](#_Toc193016958)

[1.1. Coding of VADER 10](#_Toc193016959)

[1.2. Visualization of VADER 12](#_Toc193016960)

[Logistic Regression 13](#_Toc193016961)

[1.1. Coding of Logistic Regression 13](#_Toc193016962)

[1.2. Visulazation of Logistic Regression 15](#_Toc193016963)

[Comparing Logistic Regression and VADER 15](#_Toc193016964)

[Limitation 16](#_Toc193016965)

[LDA 17](#_Toc193016966)

[Finding the optimal number of topics 17](#_Toc193016967)

[Identifying Key words of Topics 18](#_Toc193016968)

[Limitation 18](#_Toc193016969)

[Business Insights and Recommendations 19](#_Toc193016970)

[Conclusion 19](#_Toc193016971)

Introduction

In the current digital age, smartphones have become an important aspect of modern life, transforming communication, entertainment and productivity. Carbonell, Oberst, and Beranuy (2013) emphasized that cell phones have become an integral part of an individual's identity and daily life, and that people form a strong emotional bond with their devices. This has made people increasingly inseparable from their cell phones, which also indicates a strong demand for cell phone products and the market for the cell phone industry has a good scope for development. Storbacka, Strandvik, and Grönroos (1994) proposed a service quality model that underscores the critical link between service quality and customer satisfaction. They further emphasized that customer satisfaction plays a pivotal role in fostering customer loyalty, which, in turn, enhances the profitability of companies. Given this relationship, customer feedback serves as a crucial determinant of success or failure of products, as it directly influences service improvements and business performance. Understanding consumer sentiment through customer reviews can help a cell phone company identify key issues. Based on these issues the company can understand the direction of improvement and thus provide better decisions and products for the customers. This data-driven decision-making approach not only enhances brand competitiveness but also promotes market share growth and product sales (Kim, Lee & Ahn, 2006). When user experience is enhanced, and user needs are met, ultimately the company can achieve increased market competitiveness and profitability.

This study utilizes a dataset obtained from Kaggle that consists of Amazon customer reviews of smartphones. This study intends to discuss concern of consumer and potential demand for the mobile phone industry by combining sentiment analysis and topic modeling methods: sentiment analysis will quantify attitude of user toward products, while topic modeling will identify high-frequency discussion topics through text clustering.

Data Collection

The dataset, from Kaggle, contains smartphone user reviews from the Amazon platform. Several reasons led me to choose this dataset:

1. Data richness

The dataset contains more than 60,000 user reviews, covering several well-known brands, and is both data-rich and representative. This enables researchers to extract comprehensive consumer behavior information from diverse data. In addition, the data comes from Amazon, one of the world's largest e-commerce platforms (Moriset, 2018), whose review system is an important channel for consumers to express their purchase experience and product opinions and occupies an important position in the mobile phone consumption industry. This ensures a wide range and reliability of data sources.

2. The data set meets the requirements of research objectives

The dataset contains detailed text comments and user ratings to support sentiment analysis and topic modeling. User ratings provide labeling basis for emotion classification, and comment text provides sufficient material for mining consumers' concerns.

3. Well-structured data

The dataset contains structured variables such as review text, rating, product brand and review time. According to research reports, structured data sets contribute to information extraction and business analysis, improve data processing efficiency and enhance the reliability of analysis results (Chen, Chiang & Storey, 2012). A well-structured data set can help researchers conduct in-depth analysis of consumer thinking and consumption trends. As shown in Figure 1:图形用户界面, 文本, 应用程序

AI 生成的内容可能不正确。

Figure 1: Table of Mobile Product Reviews

The dataset contains the following variables:

图形用户界面, 文本, 应用程序, 电子邮件

AI 生成的内容可能不正确。

Data processing

In the data cleaning section, I followed the following steps to complete my data preprocessing:

1. Removing Redundancies, Merging Text, and Handling Missing Data

I removed unnecessary fields such as ***asin, name, date, verified***, and ***helpfulVotes*** to simplify the dataset. Additionally, I removed rows where ***body*** was empty to address potential missing data issues in subsequent analysis.

Some reviews contained only a ***title*** or only the ***body***, which did not form a complete review text. To resolve this, I merged title and body, creating a new column, ***full\_review***, to store the complete review content.

I also used ***fillna("")*** method to handle any missing values in the title column. The results are shown in the following figure.

**图形用户界面, 文本, 应用程序, 电子邮件

AI 生成的内容可能不正确。**

Figure 2：The result of removing redundant columns and missing values

2. Constructing a Data Cleaning Function for Text Standardization

To ensure the usability of the text, we designed the ***clean\_text*** function for preprocessing. The function included the following steps: converting text to lowercase for uniform formatting, removing URLs, mentions, and hashtags to eliminate irrelevant information, and removing punctuation and special characters to retain only letters and numbers. Additionally, spaces were normalized to ensure the text remains well-structured. The process is illustrated in the figure below:

文本

AI 生成的内容可能不正确。

Figure 3: Details of the ***clean\_text*** Function

The result is a new column, ***cleaned\_text***, that contains only letters and numbers:

图形用户界面, 文本, 应用程序

AI 生成的内容可能不正确。

Figure 4: Results of Applying the ***clean\_text*** Function

3. Stop Words Removal and Text Tokenization

Stop words are high-frequency words that carry little to no meaningful information, such as "is," "the," and "and." According to the study by Ghag and Shah (2015), removing stop words helps reduce irrelevant information, which improves machine training and classification efficiency. Therefore, I used the built-in stop words list in NLTK library and a customs stop words set to clean the text. The customized stop words come from testing for the result, obtained by running the code repeatedly many times.

The detailed implementation is shown below:

文本, 散点图

AI 生成的内容可能不正确。

Figure 5: Details of Stop word Removal and Tokenization Function

Finally, the ***tokenized\_text*** column contains the results of text tokenization:

图形用户界面, 文本

AI 生成的内容可能不正确。

Figure 6: Tokenization Results

4. Lemmatization

Lemmatization is an essential step in NLP for reducing words to their base forms. Unlike stemming, lemmatization relies on dictionaries and linguistic rules, making it more effective in preserving semantic meaning and providing higher accuracy (Pramana et al., 2022). In this study, I used ***WordNetLemmatizer*** for lemmatization, with the specific implementation as follows: 图形用户界面, 文本, 应用程序

AI 生成的内容可能不正确。

Figure 7: Lemmatization Implementation Process

A new column, **lemmatized\_review**, is generated as follows:

图片包含 图形用户界面

AI 生成的内容可能不正确。  
Figure 8: Lemmatization Results

5. Cloud visualization

Finally, I used a word cloud to get an overall view of the most frequently mentioned words. The visualization result is shown below:

报纸上的字

AI 生成的内容可能不正确。

Figure 9: Word Cloud Result

From the content in the image, overall evaluation of users of mobile phones appears to be positive. Words like "great", "five star", and "good" are prominent, indicating that many reviews express positive sentiments. However, words like "problem" and "issue" also stand out, suggesting that some users have mentioned issues or drawbacks.

Sentiment Analysis

Before conducting sentiment analysis, I utilized the rating column in the dataset to label each text entry. Ratings of 1-2 were categorized as negative, a rating of 3 was categorized as neutral, and ratings of 4-5 were categorized as positive. These labels were stored in the ***rating\_sentiment*** column and later used as ground truth sentiment labels. The code of process is shown below:

表格

AI 生成的内容可能不正确。

Figure 10: Creation of Ground Truth Sentiment Labels

VADER

* 1. Coding of VADER

In traditional sentiment analysis methods, VADER is an effective tool that can quickly assign sentiment labels. Research has shown that VADER performs well in analyzing social media text and excels in multi-class sentiment classification (Elbagir & Yang, 2019). Therefore, I applied this method to label the sentiment of the textual content in this dataset. The code is shown in the following figure:

图形用户界面, 文本, 应用程序

AI 生成的内容可能不正确。

Figure 11: VADER Sentiment Analysis Method

First, I imported ***SentimentIntensityAnalyzer*** from the ***nltk.sentiment*** module. To process textual data, I applied the apply method, which iterates through each row in the dataset and applies the sentiment analysis function. Then, the computed sentiment scores were stored in the ***VADER\_sentiment\_score*** column. Finally, I set sentiment scores greater than 0.05 were labeled as positive, scores less than -0.05 were labeled as negative, and scores in between were classified as neutral.

The results are displayed in the figure below:

日历

AI 生成的内容可能不正确。

Figure 12: VADER Sentiment Analysis Results

* 1. Visualization of VADER

For visualization, I used bar charts and histograms to display the sentiment distribution and sentiment score distribution. The results are shown below:

图表, 直方图

AI 生成的内容可能不正确。

Figure 13: Distribution of Sentiment and Sentiment Scores

For the first bar chart, as seen in the chart, the "positive" category significantly outweighs both "neutral" and "negative" categories, indicating that most reviews or textual data express positive emotions. Negative sentiment is the least frequent, suggesting that in 2019, people's overall sentiment toward mobile phones was predominantly positive.

For the second chart, from the distribution, we observe multiple peaks, with scores mainly clustering around 0.0 and 0.75 to 1.0. This suggests that most of the text exhibits either neutral or highly positive sentiment. The distribution near -1.0 is relatively sparse, indicating that extreme negative sentiment data is rare. This pattern aligns with the trend in the first chart, confirming that most text entries display strong positive sentiment, while the proportion of negative sentiment is relatively low. Additionally, the presence of multiple minor peaks between 0.25 and 0.75 indicates that a considerable number of reviews fall within the moderate positive sentiment range, rather than being strongly positive.

Logistic Regression

* 1. Coding of Logistic Regression

To compare with VADER, I used Logistic Regression for binary sentiment analysis. First, for computation, I removed the "neutral" label and converted the text ***lemmatized\_review*** into numerical vectors. I set ***lemmatized\_review*** as variable x and ***rating\_sentiment*** as variable y. Then, I splited the dataset using ***train\_test\_split***, with 80% for training and 20% for testing. I extracted TF-IDF features using ***TfidfVectorizer***, fitting it only on the training set. After training the ***LogisticRegression*** model, I predicted sentiment labels on the test set and calculated both training and test accuracy. Next, I evaluated the model’s performance using ***classification\_report***, which includes precision, recall, and F1-score. Finally, I applied the model to predict sentiment labels for the entire dataset.

The following figures present the code of mode and performance results:

文本

AI 生成的内容可能不正确。

Figure 13: Logistic Regression

图形用户界面, 应用程序

AI 生成的内容可能不正确。  
Figure 14: Precision, Recall, and F1-score

From the results, this classification report indicates that the Logistic Regression model performs well in the sentiment analysis task. It has high accuracy and stability. The training accuracy is 94.83%, and the test accuracy is 93.34%, showing that the model does not have significant overfitting issues. In terms of classification performance, the model recognizes positive reviews better than negative reviews, with precision, recall, and F1-score for positive reviews being 0.95, 0.96, and 0.96, respectively, while the corresponding values for negative reviews are 0.89, 0.85, and 0.87. Overall, the weighted average F1-score is 0.93, indicating that the model performs stably in the classification task. However, since the recall for negative reviews 0.85 is lower than for positive reviews 0.96, some negative reviews are misclassified as positive. The possible reason for this phenomenon could be the imbalance in the number of reviews. In general, this model can accurately distinguish between positive and negative sentiments.

* 1. Visulazation of Logistic Regression

For the visualization of the results, I also used a confusion matrix to gain deeper insights into the performance of Logistic Regression. The confusion matrix is shown below:

**图表, 树状图

AI 生成的内容可能不正确。**

Figure 15: Confusion Matrix for Logistic Regression  
This confusion matrix indicates that Logistic Regression performs well in the sentiment analysis task but still has some misclassification cases. The model correctly classifies 2861 negative reviews and 8939 positive reviews, but 493 negative reviews are misclassified as positive, and 349 positive reviews are misclassified as negative. From the misclassification pattern, the model tends to misclassify negative reviews more frequently, suggesting a bias toward predicting positive sentiment.

Comparing Logistic Regression and VADER

As shown in the figure below, since the Logistic Regression model does not have a neutral category, I only compared the positive and negative cases. I used ***accuracy\_score*** to calculate the accuracy of the VADER method.

**图表

AI 生成的内容可能不正确。**

Figure 16: Comparison of Logistic Regression and VADER

VADER predicts a higher number of positive reviews than Logistic Regression, which indicates that VADER tends to classify sentiment as positive more frequently. In contrast, Logistic Regression predicts more negative reviews, suggesting that it applies stricter criteria for identifying negative sentiment. Additionally, Logistic Regression achieves an accuracy of 0.9334, which is higher than the accuracy of VADER (0.8664), indicating that Logistic Regression may be more accurate than traditional sentiment analysis methods on this dataset.

Limitation

In summary, for the data set, since there are far more positive than negative sentiments in the dataset, the model may have a classification bias. It is worth noting that I removed neutral sentiment in the logistic regression analysis and when comparing the two methods. This may have resulted in the sentiment trends across the text appearing more extreme, indeed authenticating some of the data.

For the analysis method, VADER favors positive sentiment and categorized more data into positive labels. But logistic regression is more sensitive in its performance for negative sentiment judgments, and it considers more negative sentiment labels.

LDA

Finding the optimal number of topics

For LDA, first, I needed to determine the optimal number of topics for topic modeling. For simple calculation, I randomly sampled 2,000 entries from the dataset and performed tokenization and dictionary construction. Then, I defined a function ***compute\_coherence\_values*** to train LDA models across a range of topic numbers. The function calculated the coherence score for each model to evaluate topic quality. It iterated multiple topics numbers, records the coherence score. And then I plotted a line chart showing the variation of coherence score with the number of topics. Finally, the topic number with the highest coherence score was selected as the optimal number of topics and printed.

文本

AI 生成的内容可能不正确。

Figure 17: Determine the Optimal Number of Topics

As shown in Figure below, the optimal number of topics in this dataset is 4:

**图表, 折线图

AI 生成的内容可能不正确。**

Figure 18: Result of Optimal Number of Topics

Identifying Key words of Topics

This code uses LDA topic modeling to analyze text data and identifies four main topics. Topic 1 focuses on phone issues, such as phone, issue, problem, reflecting user feedback on device malfunctions. Topic 2 mainly concerns user experience, such as great, good, love, reflecting positive reviews. Topic 3 discusses phone features, such as sim, unlocked, brand, involving unlocking status and brands. Topic 4 focuses on technology and hardware, such as android, fingerprint, video, covering Android systems, fingerprint recognition, and video quality.

**图形用户界面, 文本, 应用程序

AI 生成的内容可能不正确。**

Figure 19: Topic Keywords and Document Distribution

Limitation

There are some limitations in the topic division of LDA topic modeling. The keywords of some themes are vague, resulting in a lack of precision in the interpretation of the model. For example, the keywords of theme 3 include “brand”, “sim” and “unlocked”, which are less relevant to each other. Theme 4 keywords such as “android”, “fingerprint” and “video” cover both operating system, fingerprint recognition and video quality. This leads to a lack of clarity in the theme boundaries. In summary, these facts suggest that one needs more debugging and changes to adapt to LDA.

Business Insights and Recommendations

Conclusion

**References**

Carbonell, X., Oberst, U. and Beranuy, M. (2013). *The Cell Phone in the Twenty-First Century*. *Principles of Addiction*, pp. 901–909.

Chen, H., Chiang, R. H. L. & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact, *MIS Quarterly*, 36(4), pp. 1165-1188. Available at: <https://www.jstor.org/stable/41703503>

Elbagir, S. and Yang, J., 2019. *Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment.* Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS 2019), Hong Kong, 13-15 March, pp. 12-16. Available at: <https://www.iaeng.org/publication/IMECS2019/IMECS2019_pp12-16.pdf>

Ghag, K.V. and Shah, K. (2015). Comparative analysis of effect of stopwords removal on sentiment classification. *2015 International Conference on Computer, Communication and Control (IC4)*. Available at: <https://ieeexplore.ieee.org/abstract/document/7375527>

Kim, J., Lee, D.-J., & Ahn, J. (2006) A dynamic competition analysis on the Korean mobile phone market using competitive diffusion model. *Computers & Industrial Engineering*, 51(1), pp. 174–182. Available at: <https://www.sciencedirect.com/science/article/pii/S0167923611001229?casa_token=sI1edxfgM0MAAAAA:ydc2fK7chtBWBaUQL585AbDUIGWYRUO0SNfTwYys7i3wK_MK-hrdwy5-dnQr0glTqaTNGM3ByT8>

Moriset, B. (2018). *e-Business and e-Commerce*, *HAL Open Science*. Available at: <https://shs.hal.science/halshs-01764594>

Rio Pramana, Debora, N., Jonathan Jansen Subroto, Agung, A. and None Anderies (2022). Systematic Literature Review of Stemming and Lemmatization Performance for Sentence Similarity. *2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*. Available at: <https://ieeexplore.ieee.org/abstract/document/9971451?casa_token=eGuMUXFs200AAAAA:flYRlxpGs2g63yXgMLM4AZmCwXoGORXxJVoDiBl_gzlUYg8rJ3_Yi51lUHyqvsLH0aG2OUdX4bA>

Rio Pramana, Debora, N., Jonathan Jansen Subroto, Agung, A. and None Anderies (2022). Systematic Literature Review of Stemming and Lemmatization Performance for Sentence Similarity. *2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*. Available at: <https://ieeexplore.ieee.org/abstract/document/9971451?casa_token=yslIPfrF6-sAAAAA:OjopdXIrGK3sphogvRx1k6P0qM3prfEHPC7VnOqzRPkbsqMUpAIys94PBqaZDwEW_En8t-Fwq_8>

Storbacka, K., Strandvik, T. & Grönroos, C. (1994). Managing customer relationships for profit: The dynamics of relationship quality, *International Journal of Service Industry Management*, 5(5), pp. 2. Available at: <https://www.emerald.com/insight/content/doi/10.1108/09564239410074358/full/html?casa_token=UzfLI83Z1ScAAAAA:hJMneCYR4ZkNsYHGvUFtLKQ13SF15LLWErz5qMIsp4EzJWmGf6mZowpYuwL7zddxfkwZNznsZVUHPmeb5P_TpZDxOmhz4wzaBaB_3re9-dZNbihIrFc3>

**Appendix**

1. **Kaggle (2019) *Amazon Cell Phones Reviews*. Available at:** [**https://www.kaggle.com/datasets/grikomsn/amazon-cell-phones-reviews**](https://www.kaggle.com/datasets/grikomsn/amazon-cell-phones-reviews)**-reviews**