College of Business
香港城市大學
City University of Hong Kong

Sentiment Analytics for Streaming Platform User Experience

Uncovering User Sentiment Patterns in Disney+ Application Reviews

Group 10086

Xinran GAO	58670679	xinrangao3-c@my.cityu.edu.hk
Qinshu DU	58819313	qinshudu2-c@my.city.edu.hk
Yihui JIANG	58878190	yihujiang4-c@my.cityu.edu.hk
Wei ZHAO	58903615	wzhao67-c@my.cityu.edu.hk
Yiyang ZHENG	58838131	yiyzheng5-c@my.cityu.edu.hk

IS 6941 Machn Lrn & Soc Medi Analytics

Final Report

CITY UNIVERSITY OF HONG KONG

April 2024

Executive Summary

This project aims to analyze user sentiment and emotions towards Disney+ through a comprehensive review analysis. Disney+, launched in 2019, has a rich content library but faces increasing competition. Understanding user feedback is crucial for enhancing user satisfaction and loyalty.

We collected 4,608 user reviews from Google Play using a Python crawler. We performed sentiment and emotion analysis using tools such as VADER, NRC lexicon, and machine learning algorithms, and performed KNN and cluster analysis.

The analysis revealed that Disney+ ratings declined from 2.5 stars at launch to approximately 1.5 stars by 2022–2023, with no version reversing this trend. Technical issues, such as app unresponsiveness and streaming problems, as well as intrusive advertising, were the main sources of dissatisfaction. Neural Networks achieved over 86% accuracy in predicting user satisfaction. Negative emotions, such as anger and disgust, correlated with lower ratings, while positive emotions, such as joy and anticipation, correlated with higher ratings.

Based on these findings, we recommend prioritizing the resolution of technical issues, especially device compatibility problems, revising the advertising strategy to be less intrusive, enhancing offline functionality, and improving customer support with faster resolution times, focusing on technical and content accessibility issues. Future research directions include conducting competitive benchmarking with other streaming services to identify best practices, evaluating the impact of specific features on user satisfaction through controlled studies, and analyzing the correlation between negative sentiment and subscription cancellations to identify key churn drivers.

CONTENTS

1 Introduction

1.1 Business Background

In today's digital age, streaming services have become an important channel for people to obtain entertainment content. Disney+ is a streaming platform launched by Disney in 2019. With its rich film and television resources and strong brand effect, it has quickly attracted a large number of users. The platform not only brings together popular content such as classic Disney animations, Pixar movies, Marvel series and Star Wars, but also continuously launches original programs to meet the needs of different user groups.

The success of Disney+ lies in its strong brand influence and extensive content library. However, with the intensification of competition, it is particularly important to understand user feedback. By analyzing user reviews, companies can gain in-depth understanding of user satisfaction, preferences and potential problems, and then formulate corresponding improvement measures. Therefore, through data-driven insights, Disney+ can better meet user needs and enhance user stickiness and brand loyalty.

1.2 Project Objectives

The goal of this project is to conduct a comprehensive sentiment and emotion analysis of the user experience of the Disney+ streaming platform. We use python crawlers to collect user reviews and feedback from online social media platforms, such as Google Play, to create a robust dataset for analysis. By summarizing the data using descriptive analysis, we assess the overall user perception of Disney+. We will use sentiment and emotion analysis techniques (including methods such as VADER sentiment analysis, NRC lexicon, TextBlob, etc.) to identify specific emotional relationships expressed in the reviews like joy, disappointment, and anger. We will also explore methods such as Latent Dirichlet Allocation (LDA) and K-Nearest Neighbor (KNN) algorithms in our analysis to identify patterns to classify emotions among different user groups. We will then utilize data visualization techniques to present the analysis results in a thoughtful way so that stakeholders can easily grasp the user feedback and sentiment trends. Finally, based on our findings, we will propose preliminary recommendations to enhance the product design and user satisfaction of Disney+, thereby enriching the overall user experience.

1.3 Business Process

As part of the enterprise's business processes, this social media analytics methodology can be adopted to conduct data analysis in the following scenarios:

(1) Identification

- Determine the analysis objectives, such as understanding user satisfaction and sentiment towards products/services
- Identify data sources, such as social media platforms and app stores for user reviews

(2) Collection

- Use web crawlers and other tools to collect relevant user review data from various channels
- Clean the collected data by removing irrelevant information

(3) Cleaning

- Perform data cleansing and pre-processing to remove noise and irrelevant content
- Ensure data quality to provide a reliable foundation for subsequent analysis

(4) Analyzing

- Apply sentiment analysis, topic analysis, and other techniques to uncover user sentiment and preferences
- Utilize machine learning models to discover user group characteristics and content preference patterns

(5) Visualization

- Use charts, dashboards, and other visual tools to present the analysis results
- Facilitate management's understanding of user feedback and support decision-making

(6) Interpretation

- Based on the analysis results, formulate content optimization and marketing strategies
- Continuously monitor the effectiveness and optimize the analysis process and business decisions

Through this series of processes, enterprises can gain a deeper understanding of user needs, improve product/service quality, enhance user loyalty, and increase their market competitiveness.

2 Problem Definition

This study investigates user experience with the Disney+ streaming service through analysis of app reviews. Four key research questions guide our investigation:

Q1: How has user sentiment toward Disney+ evolved over time and across different versions?

Hypothesis (H1): User satisfaction with Disney+ has improved since its launch, with each major update showing progressively higher ratings.

Q2: What main factors drive negative reviews of the Disney+ app?

Hypothesis (H2): Technical issues (app functionality, streaming quality) and advertising complaints are the primary sources of dissatisfaction, with varying importance across versions.

Q3: Can machine learning effectively predict user satisfaction from review content?

Hypothesis (H3): Neural Networks can predict ratings with over 80% accuracy, with emotional language and technical terms being the strongest predictors.

Q4: How do different emotions in reviews relate to overall ratings?

Hypothesis (H4): Negative emotions (anger, disgust, fear) strongly correlate with lower ratings, while positive emotions (joy, anticipation) correlate with higher ratings.

3 Data Description

3.1 Data Sources

The data was collected through a custom web scraping process targeting the Disney+ application's review pages on Google Play. The scraping was performed using Python-based tools that collected each review's rating, text content, submission date, app version, and reply to information.

3.2 Data Preprocessing

- **Column Removal**: Eliminated irrelevant columns), including: reviewId, userName, userImage, reviewCreatedVersion, game_id.
- **Null Value Handling**: Removed 392 rows with null values in the appVersion column.
- **Emoji Removal**: Eliminated emoji characters from all review content to prevent encoding issues and ensure text consistency.
- **Text Cleaning**: Removed special characters from content and replyContent fields, preserving only basic punctuation.
- **Text Standardization**: Converted all text to lowercase for consistent analysis.
- **Feature Engineering**: Created metrics including review length and sentiment scores to enhance analytical capabilities.
- **Format Standardization**: Normalized date formats and version numbers to enable temporal and version-based analysis.

3.3 Data Overview

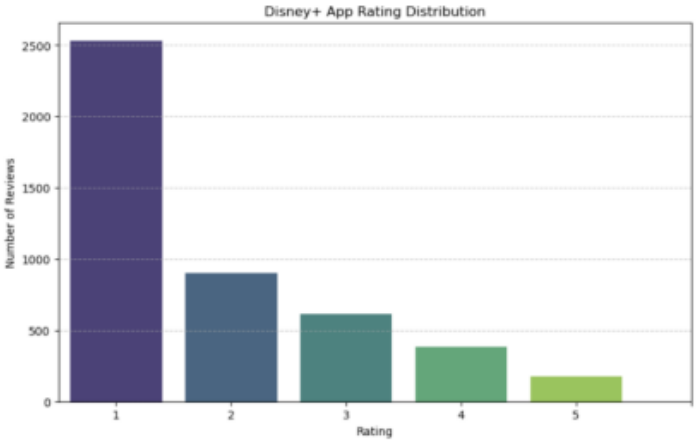
The cleaned dataset comprises 4,608 Disney+ app reviews with 14 variables capturing comprehensive user feedback. It includes both original and cleaned review text, numeric ratings, thumbs-up counts, and temporal data. The data also documents developer responses with corresponding timestamps and tracks app version information, along with review length.

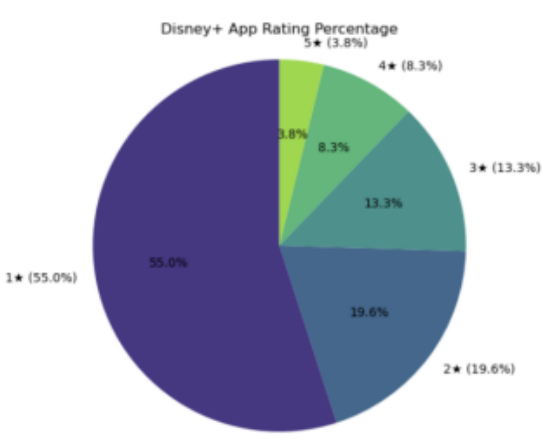
Table 1: Dataset Description: disney_plus_reviews

Col.	Column Name	Type	Description
1	content	Nominal	Text content of each review (emoji removed)
2	content_cleaned	Nominal	Modified "content" (emoji removed & special characters removed)
3	score	Interval	Score of each review, integer value within 1-5
4	thumbsUpCount	Interval	Number of thumb-up's received by the review
5	timeCreated	DateTime	The date and time when the review was created (including the Year-Month-Day and exact hour:minute:second of the day)
6	year	Interval	"Year" extracted from "timeCreated"
7	month	Interval	"Month" extracted from "timeCreated"
8	day	Interval	"Day" extracted from "timeCreated"
9	replyContent	Nominal	Text content of app developer's reply to the review
10	replyContent_cleaned	Nominal	Modified "replyContent" (special characters removed)
11	timeReplied	DateTime	The date and time when the app developer replied to the review (including the Year-Month-Day and exact hour:minute:second of the day)
12	appVersion	Nominal	The corresponding version of the Disney+ app
13	version_main	Nominal	The main version number extracted from "appVersion" (e.g., 2.10.7)
14	review_length	Interval	Length of the review content

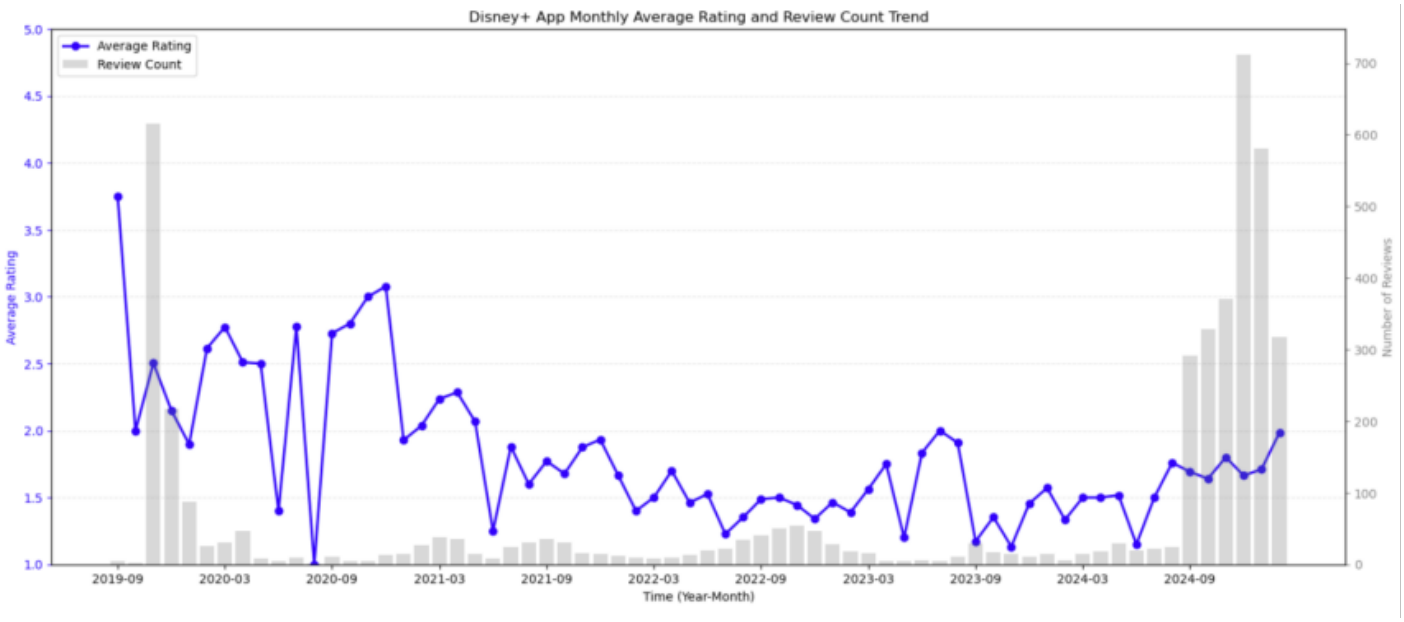
3.4 Descriptive Analysis

Rating Distribution: User ratings for the Disney+ app show overwhelming dissatisfaction. More than half (55%) of users gave the lowest possible 1-star rating. Only 12% of users expressed satisfaction with 4-5 star ratings. This distribution clearly demonstrates widespread user disappointment with the application.

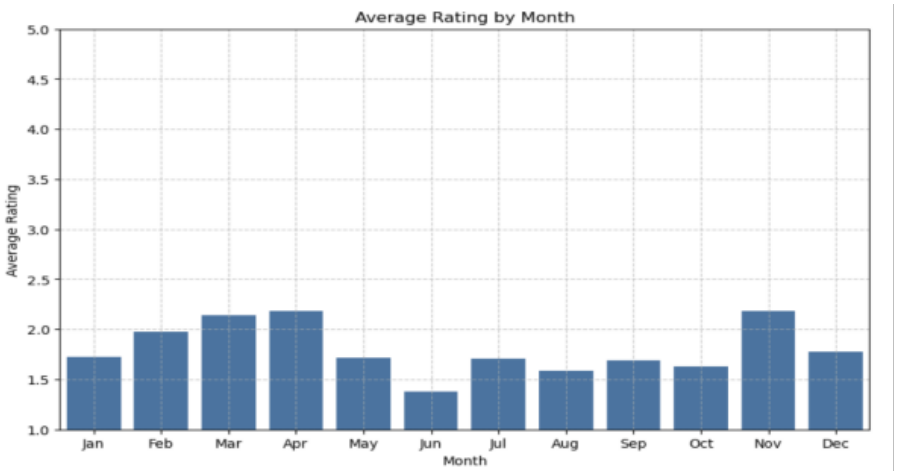




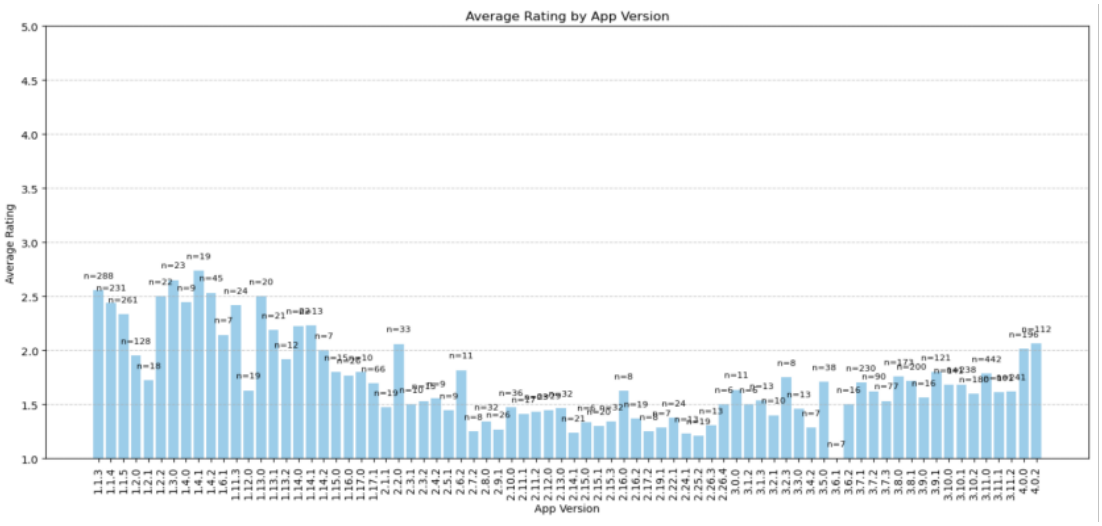
Rating Trends Over Time: Disney+ ratings have steadily declined from 3.8 stars at launch in 2019 to approximately 1.5 stars in 2022-2023. Despite five years of development, ratings consistently remained below 3 stars, with a sharp increase in negative reviews in late 2024 indicating intensifying user dissatisfaction. This persistent downward trend suggests fundamental issues remain unresolved.



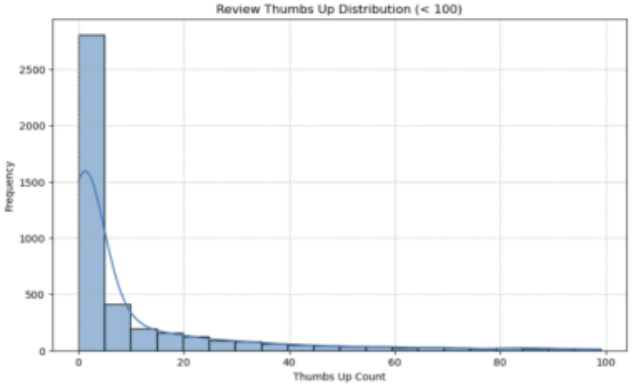
Seasonality Analysis: Spring months (March-April) and autumn months (November) have higher ratings. Summer months (particularly June) have comparatively lower user ratings. This reflects a seasonality trend in the performance or content of the app.

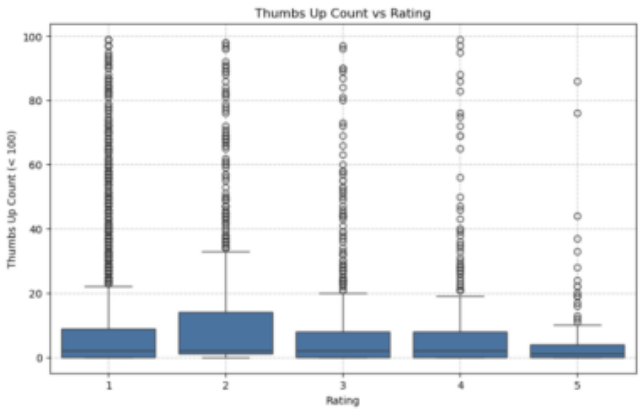


App Version Analysis: Disney+ app ratings declined significantly across versions, with early 1.x releases averaging around 2.5 stars before dropping sharply to approximately 1.5 stars during the 2.x series. More recent 3.x and 4.x versions failed to reverse this trend, consistently remaining below 2.0 stars despite continued development.

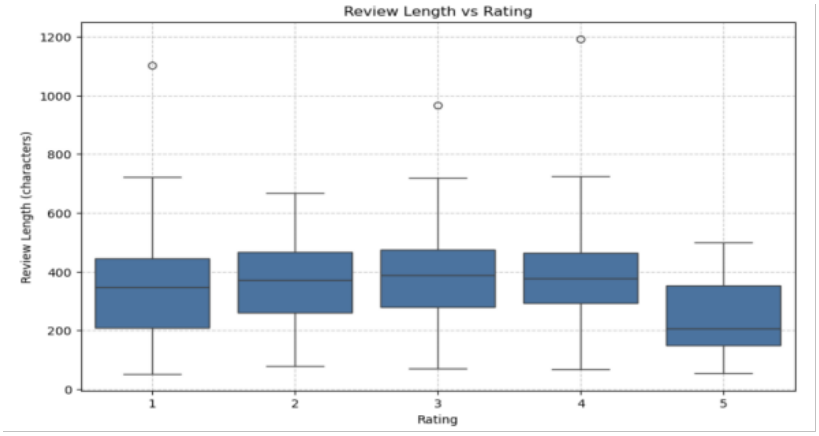


Thumbs Up Analysis: The left graph reveals that most reviews of the Disney+ app have extremely low "thumbs up" votes, with the vast majority having less than 10. The right graph demonstrates how ratings relate to thumbs up numbers. 2-star reviews had more thumbs up votes on average than the other ratings, which means these relatively moderate negative reviews resonate more with other users.

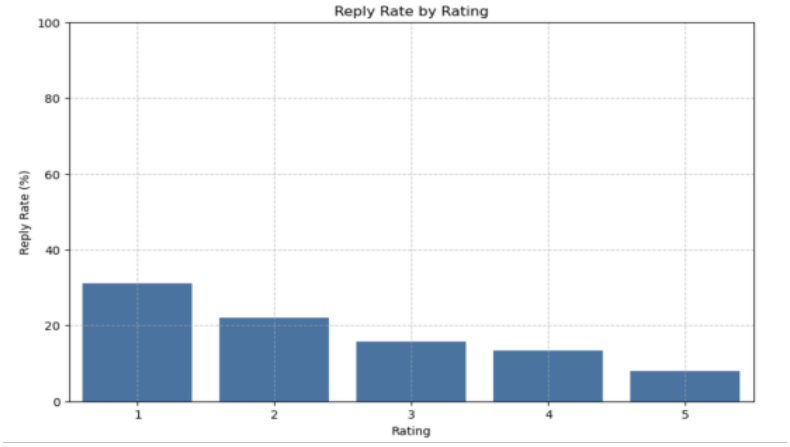




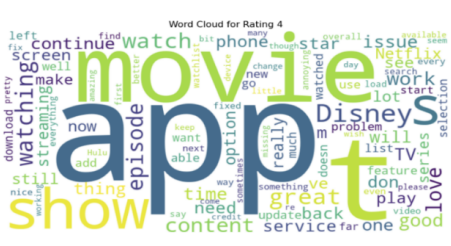
Review Length Analysis: Reviews reveal that dissatisfied users write significantly longer feedback (350-400 characters) than satisfied users.

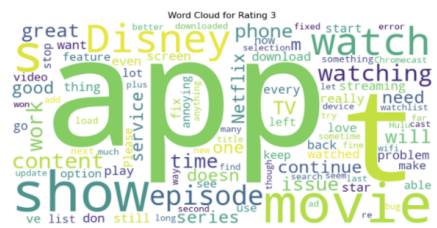


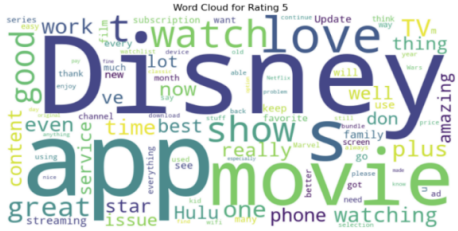
Reply Analysis: Disney+ demonstrates a strategic prioritization in customer service, responding to approximately 30% of 1-star reviews while addressing only about 7% of 5-star feedback. This pattern reveals the company prioritizes addressing negative feedback rather than acknowledging positive user experiences.



Review Content Text Analysis: Disney+ reviews show clear rating-based patterns: 1-star reviews primarily discuss technical failures with phrases like "doesn't work" and "issue"; 2-3 star reviews focus more on content concerns mentioning "show," "movie," and "episode"; while 4-5 star reviews use positive emotional language like "love" and "great" with minimal technical terminology.







4 Analysis and Findings

4.1 Sentiment analysis

In the modeling process, to ensure the accuracy of the analysis results, we used various tools and methods based on the characteristics of the dataset, such as processing special vocabulary, using VADER sentiment analyzer, calculating TF-IDF weights, combining user rating stars, etc., and evaluating the model at last.

In addition, we generated the top 10 high-frequency words with different ratings for different versions of the Disney+ app, and generated word clouds and line charts for further business analysis.

4.1.1 VADER Sentiment Analysis

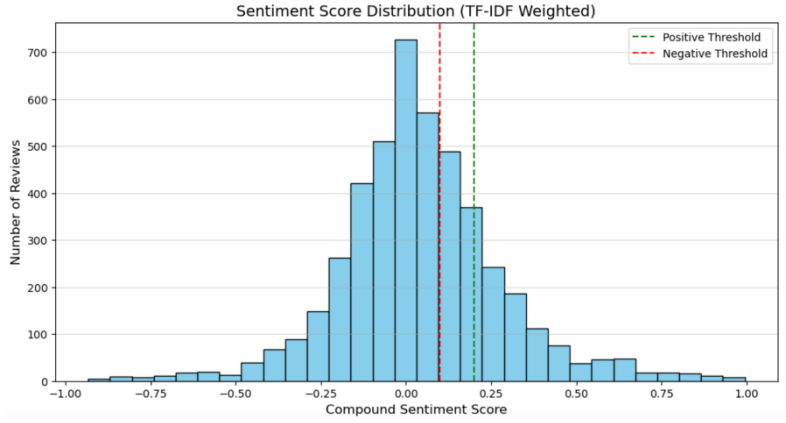
special words:

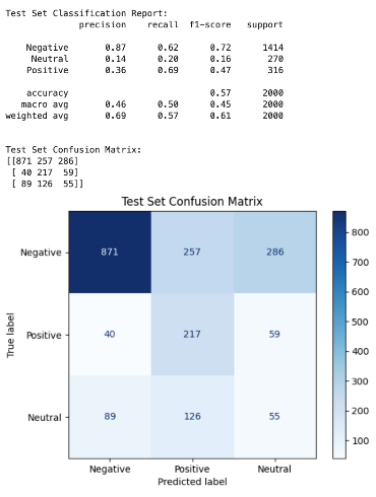
Considering the complexity of grammar leading misjudgments, such as negative/transitional words combined with positive words, which can have a negative effect, result in some comments being negative but the sentiment score being positive. Therefore, we set up special words to reduce the incorrect calculations.(Figure 4.1)

Sentiment calculation and model assessment:

At the beginning, we use the sent_tokenize function in the nltk library to split an entire comment into multiple parts with periods as the dividing points. The reason is to calculate the sentiment score of each sentence (column: sentence_dentiment_scores), and to integrate sentence_scores to obtain the sentiment score of the entire comment (column: overall_dentiment_scores), capturing more emotional differences in detail. (Figure 4.2)

During the calculation process, we use the VADER sentiment analyzer to calculate the sentiment scores, and calculate the TF-IDF weights using the TfidfVectorizer class in the scikit learn library. In addition, because user ratings can naturally serve as a reference for sentiment, we consider the weight of user star ratings, especially by strengthening the weight of negative reviews. And we choose to give lower weight to high-star reviews because they always contain constructive feedback, while the weight of 3-star reviews remains unchanged.





For defining the sentiment label, we used a dynamic neutrality threshold system because the number of negative comments is over 50% and lots of data in sample distribution are around 0. The original label “neutral” should have been located from -0.2 to 0.2 corresponding to 3-star reviews, but we ultimately adjusted it from 0.1 to 0.2. A sentiment score greater than 0.2 is considered as “positive” label, while a score less than 0.1 is considered as “negative” label.

Finally, we evaluated the model using a confusion matrix and found that the accuracy of negative was very high, at 87%. Neutral and positives are 14% and 36% respectively, although not high, in fact, after introducing special words and star weighting, they have actually increased by around 10% each.

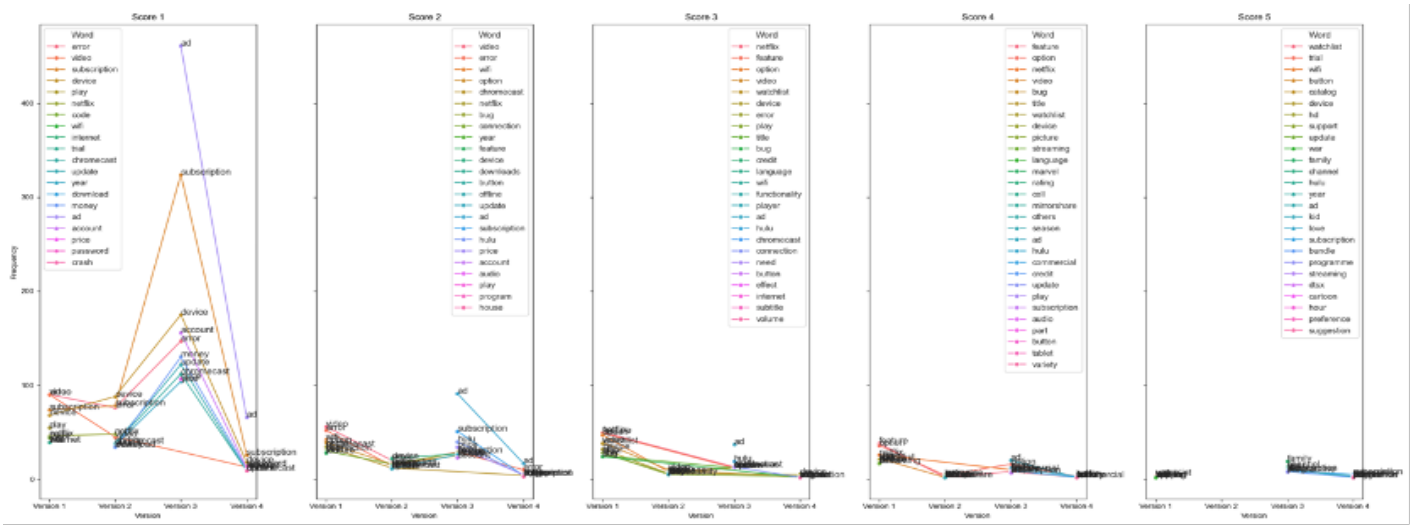
4.1.2 App Version Word Analysis

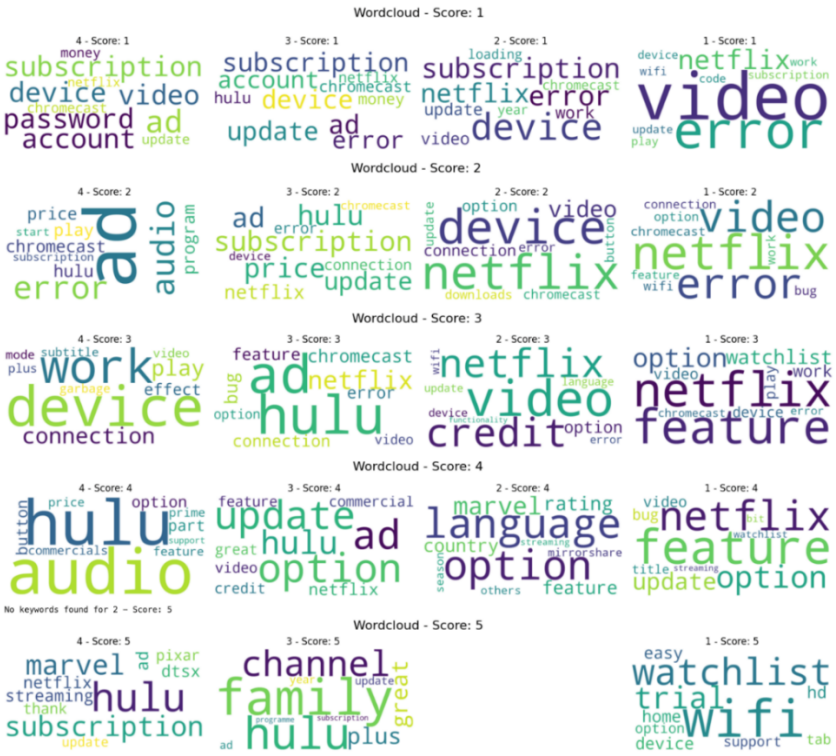
This analysis mainly focuses on optimizing the word cloud mentioned earlier. We divide the app into four versions based on the first digit of appVersion column, version 1 to version 4。 Then classify and output the top 10 words with different ratings for each version, excluding words with a frequency of 1. Due to the presence of high-frequency and non analytical words such as "Disney" that are common to all versions and score ranges, we added stop words multiple times based on the high-frequency words in the output, and finally obtained the top 10 vocabulary frequencies that can be used for analysis (Figure 4.3 & Figure 4.4).

Subsequently, we generated word clouds and data sequence line charts based on high-frequency vocabulary. From the above results, we can conclude that:

Low scoring (1-2 stars) high-frequency vocabulary is concentrated in "error", "subscription", "device", "video", "ad", etc. Low scoring users mainly focus on issues such as application errors, device playback problems, network problems, and ad placement. Through the version changing, people have gradually shifted from complaining about system errors to complaining about advertising placements. It can also be clearly seen from the line chart that the new version (Version 3-4) has a more serious issue with advertising placement.

High scoring (4-5 stars) high-frequency words are concentrated in "feature", "option", "family", "channel", "watchlist", "subscription", "streaming", Pay more attention to family viewing experience, film content, subscription services, streaming experience, etc. Through the version changing, people have gradually shifted from discussing detailed content such as watchlist and trailers to focusing on user experience and streaming experience.

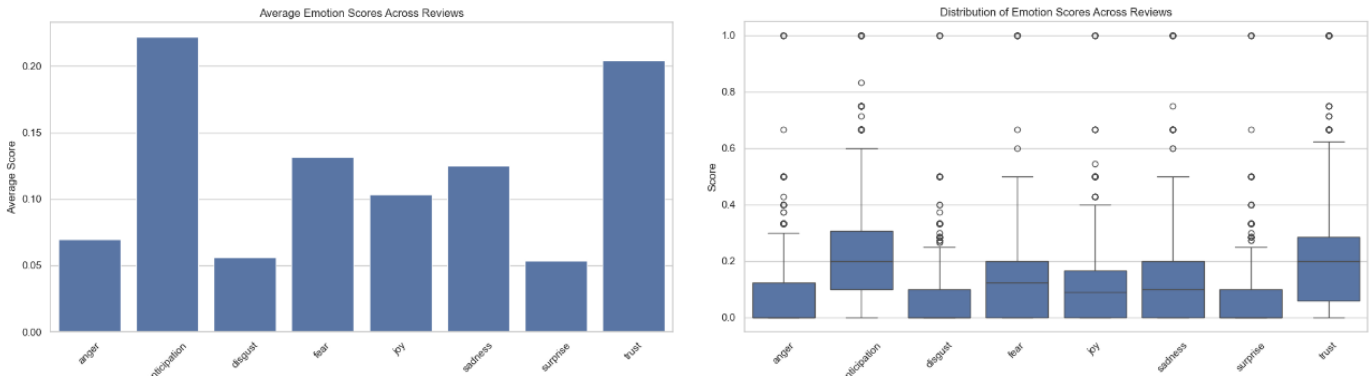




4.2 Emotion analysis

4.2.1 Distribution

We first dealt with the negation words as we did in sentiment analysis part, and then used NRC lexicon as our main method to calculate emotion scores of the reviews. The distributions of these 8 emotion scores are shown in the following bar chart and boxplot, and we found that *anticipation* and *trust* have the highest average scores, which may indicate potential loyalty and expectations of the app users.



4.2.2 Correlation

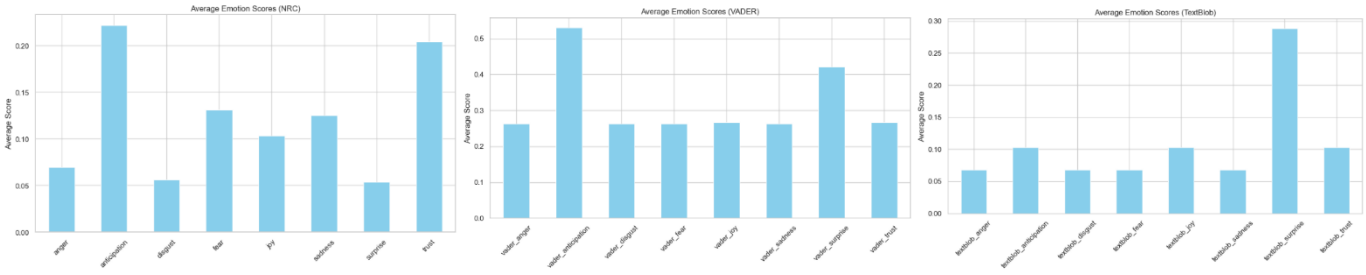
Next, we calculated the correlation between the emotion scores and the user ratings. We found that the emotions which often indicate positive feelings (anticipation, joy, surprise, and trust) are positively correlated with user ratings, while the emotions that often demonstrate negative feelings (anger, disgust, fear, and sadness) are negatively correlated with user ratings. This shows that the emotion score computed by NRC lexicon are consistent with the level of user satisfaction indicated by their ratings. At the same time, the negative emotions tend to be positively correlated with each other, implying the consistency among these emotional tendencies.

4.2.3 Time-series Analysis

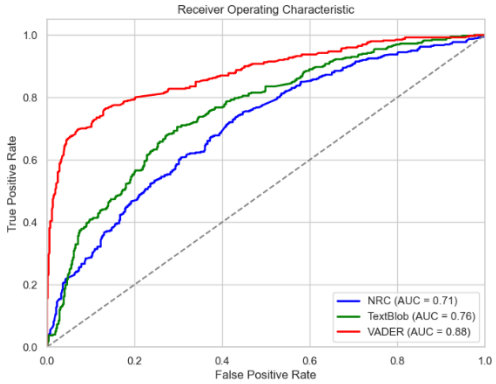
The following time-series plots show how the average emotion scores changed by month from 2020 to 2025. Some emotions changed with similar shapes, like anger and disgust, or joy and anticipation. Moreover, some emotion indicators show co-changing patterns around specific time points. For example, in the middle of 2023, there was a peak in *anticipation* with a sharp drop in *anger* and *disgust*; while in late 2023, there was a drop in *anticipation* with an increase in *anger* and *disgust*. Also, at the beginning od 2022, there was a peak followed by a sudden drop in *trust*, while the average scores of *surprise* and *sadness* moved in the opposite way: a decline followed by an increase. These changing patterns can assist us to understand what has happened to the app in different periods of time.

4.2.4 Other Methods

Apart from NRC, we tried other methods for emotion analysis, which are VADER and TextBlob. From the following figures, we found that the VADER lexicon gave us a more similar result with respect to average emotion scores, while the TextBlob result is totally different.



We tried to use logistic regression to see how the emotion scores calculated by these three methods can help in predicting the sentiment labels of each review (positive or negative). We used ROC curve to compare their model performances. As shown in the following figure, VADER had the best performance among the three, which might be because that VADER is specifically designed for analyzing short and informal texts derived from social medias, and it might have a better sense of the sentiment within the app reviews.



4.3 Clustering & KNN

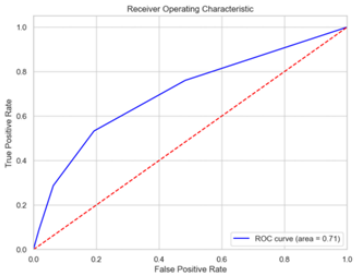
4.3.1 KNN

In this experiment, three ways were chosen to predict the dataset.

- KNN Algorithm:

X Test

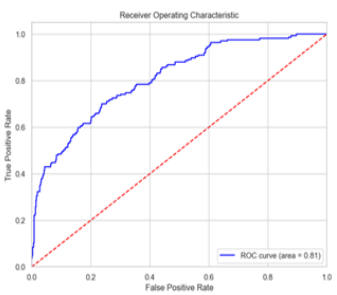
	precision	recall	f1-score	support
0	0.91	0.94	0.92	1210
1	0.39	0.29	0.33	167
accuracy			0.86	1377
macro avg	0.65	0.61	0.63	1377
weighted avg	0.84	0.86	0.85	1377



- NLP Algorithm:

X Test

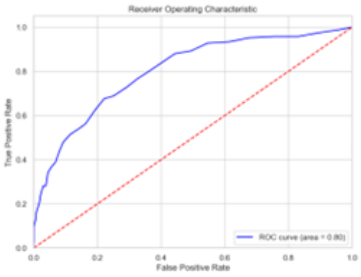
	precision	recall	f1-score	support
0	0.92	0.97	0.94	1210
1	0.62	0.38	0.47	167
accuracy			0.90	1377
macro avg	0.77	0.67	0.71	1377
weighted avg	0.88	0.90	0.88	1377



- Random Forest Algorithm:

X Test

	precision	recall	f1-score	support
0	0.88	1.00	0.94	1210
1	0.00	0.00	0.00	167
accuracy			0.88	1377
macro avg	0.44	0.50	0.47	1377
weighted avg	0.77	0.88	0.82	1377



According to the ROC graph, the ROC curve of the NLP algorithm is more willing to be diagonal and close to the upper left corner, from which we can see: the better the model classification performance of the NLP algorithm.

The random forest algorithm was further manipulated in the subsequent experiments. To enhance the performance of the model, we use two approaches:

1. Grid search tuning
2. Merge new features

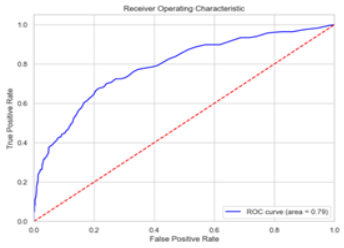
Due to the previous sentiment analysis, it was found that adverbs of degree, transitions, and negatives have a large impact on the correctness of the analysis, so five extended vocabularies were extracted (below)

```
NEGATIVE_WORDS = {'not', 'no', 'never', 'none', 'neither', 'nor', 'cannot', 'dont', 'havent', 'doesnt', 'couldnt', 'isnt', 'arent'}
BOOST_WORDS = {'very', 'extremely', 'absolutely', 'completely', 'really', 'totally', 'super'}
COMPARITIVE_WORDS = {'more', 'less', 'better', 'worse', 'compared', 'than'}
CONJUNCTIONS = {'but', 'however', 'although', 'though', 'yet', 'nevertheless', 'on the other hand'}
NEGATE = ["aint", "arent", "cannot", "cant", "couldnt", "darent", "didnt", "doesnt", "dont", "hadnt",
          "hasnt", "havent", "isnt", "mightnt", "mustnt", "neither", "don't", "hadn't", "hasn't", "haven't", "isn't",
          "mightn't", "mustn't", "neednt", "needn't", "never", "none", "nope", "nor", "not", "nothing", "nowhere",
          "oughtnt", "shant", "shouldnt", "uhuh", "wasnt", "werent", "without", "wont", "wouldnt", "won't", "wouldn't", "rarely",
          "seldom", "despite"]
```

- The optimised predictions are as follows:

X Test

	precision	recall	f1-score	support
0	0.88	1.00	0.94	1210
1	1.00	0.04	0.07	167
accuracy			0.88	1377
macro avg	0.94	0.52	0.50	1377
weighted avg	0.90	0.88	0.83	1377



Although the improved Random Forest predictions are not different from the original predictions, by looking at the detailed data, we can see that the direct use of the Random Forest is completely unable to identify the positive classes, and that the improved positive class recognition rate is very low even if there is an improvement. Based on the feature importance of the dataset generated by the model, we extracted the top ten important words and found that the influence of associative and negative words is very significant.

```
love: 0.01509422160071431
great: 0.00922183258677067
negative_count: 0.008081887483853047
amazing: 0.007430331930338453
movies: 0.007250717037292385
app: 0.006951365140979878
good: 0.006805895333061491
love app: 0.006505962916119423
disney: 0.0063043851180723505
conjunction_count: 0.006098089548243706
```

- The Word Significance：

Even if the influence of positive words such as: love, great, amazing, etc. is significant, the recognition rate of the positive class is almost 0. These indicate the predominance of negative comments in our dataset, which leads to an imbalance in training.

4.3.2 Clustering

We rated the dataset as 4 as the dividing line and classified the dataset into positive and negative (the main difference is that most of the comments with a score of 4 and above are giving advice or praise to the app, whereas most of the comments with a score of 3 and below are giving criticism or expressing dissatisfaction with the app).

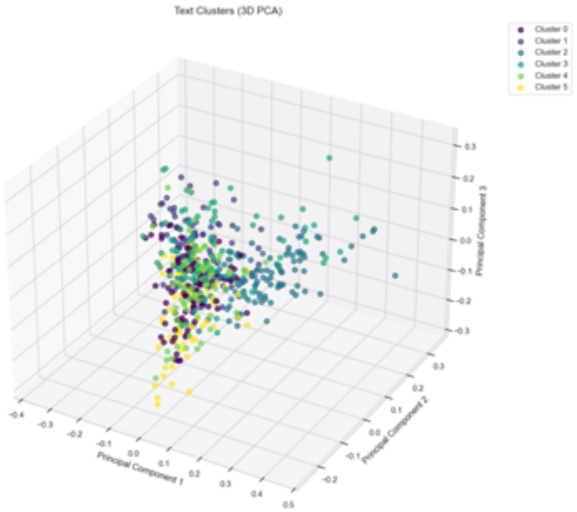
Choosing the right number of clusters is crucial in cluster analysis, and since simply using the elbow rule, profile coefficient, Davies-Bouldin Index results were not significant, so k-fold cross-validation was used to calculate and evaluate the average performance and select the appropriate number of clusters.

- **Positive**

As can be seen from the figure below, the clustering performance is better when the number of positive comment clusters is 6.

```
Number of clusters 2 - Average profile coefficient: 0.0062249308622677425, Average Davies-Bouldin index: 7.790564261614658
Number of clusters 3 - Average profile coefficient: 0.004347189479950138, Average Davies-Bouldin index: 7.133502998695576
Number of clusters 4 - Average profile coefficient: 0.004607974730910796, Average Davies-Bouldin index: 6.297233326696242
Number of clusters 5 - Average profile coefficient: 0.0038144398268542975, Average Davies-Bouldin index: 5.698425414344588
Number of clusters 6 - Average profile coefficient: 0.0006206183485874541, Average Davies-Bouldin index: 5.339298398366387
Number of clusters 7 - Average profile coefficient: 0.0005104581065799991, Average Davies-Bouldin index: 4.964833581388389
Number of clusters 8 - Average profile coefficient: -0.0006225616456347055, Average Davies-Bouldin index: 4.468073612676823
Number of clusters 9 - Average profile coefficient: -0.0026010102850562944, Average Davies-Bouldin index: 4.28709828374128
Number of clusters 10 - Average profile coefficient: -0.003727048502294648, Average Davies-Bouldin index: 4.117199096108665
```

Since we are performing cluster analysis on the comments and the textual description has high dimensionality, we have used PCA (dimension=3) to visualize it in a dimensionality reduction and the results are shown in the figure below:



Visualisation Clustering：

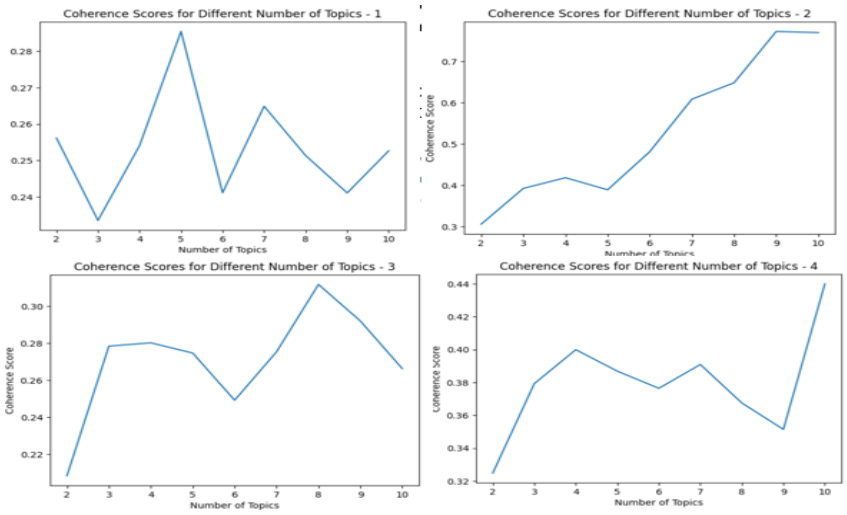
We later performed sentiment analysis on each cluster using a pre-trained large language model and found that in time the average score for each cluster was above 4, but only clusters 1 and 3 had positive sentiment.

```
Cluster 0:
  Positive: 0.38, Negative: 0.62, Neutral: 0.00
  Sentiment: Negative
Cluster 2:
  Positive: 0.20, Negative: 0.80, Neutral: 0.00
  Sentiment: Negative
Cluster 5:
  Positive: 0.42, Negative: 0.58, Neutral: 0.00
  Sentiment: Negative
Cluster 4:
  Positive: 0.34, Negative: 0.66, Neutral: 0.00
  Sentiment: Negative
Cluster 1:
  Positive: 0.55, Negative: 0.45, Neutral: 0.00
  Sentiment: Positive
Cluster 3:
  Positive: 0.58, Negative: 0.42, Neutral: 0.00
  Sentiment: Positive
```

After summarizing the sentences with the strongest emotions in each cluster, we can find the following two conclusions:

1. The app offers all the beloved Disney shows and movies, the content of which is great for families, and although some of the new seasons are slow to update, the overall experience feels good.
2. The app is missing the catalogue system, which alphabetically sorts the shows poorly, making for a smooth viewing experience; and updates tend to cause the app to fail to open.

Subsequently, to get a better understanding of the center of gravity of people's comments in the different versions, we used LDA to get the topic keywords and their coefficients for the different versions. Firstly, the optimal number of topics was analyzed by correlation and the results are as follows:



Then according to different number of topics, using LDA, topic feature words are extracted, and some of the results are as follows:

```
Cluster 1's best number of topics: 3
Cluster 1 - Topic 0: 0.027*"movie" + 0.019*"show" + 0.017*"watching" + 0.013*"love" + 0.011*"great" +
0.010*"continue" + 0.010*"episode" + 0.010*"watch" + 0.008*"need" + 0.008*"content"
Cluster 1 - Topic 1: 0.017*"show" + 0.014*"episode" + 0.011*"content" + 0.010*"great" + 0.009*"movie"
+ 0.009*"need" + 0.008*"issue" + 0.008*"watch" + 0.007*"ive" + 0.007*"watching"
Cluster 1 - Topic 2: 0.014*"movie" + 0.013*"show" + 0.013*"great" + 0.011*"issue" + 0.009*"content" +
0.009*"episode" + 0.009*"watch" + 0.008*"netflix" + 0.008*"watching" + 0.007*"fixed"
```

Based on these results, we summarize the strengths and areas for improvement of the four versions of Disney+:

1. Version 1: content requirements, viewing experience and potential problems.

2. Version 2: video playback smoothness, content usability and user experience.

3. Version 3: the emotional value embodied in videos such as films and programs and the services provided by the app.

4. Version 4: the discussion of film and television content, including the app's video recommendations, viewing experience and customer service quality.

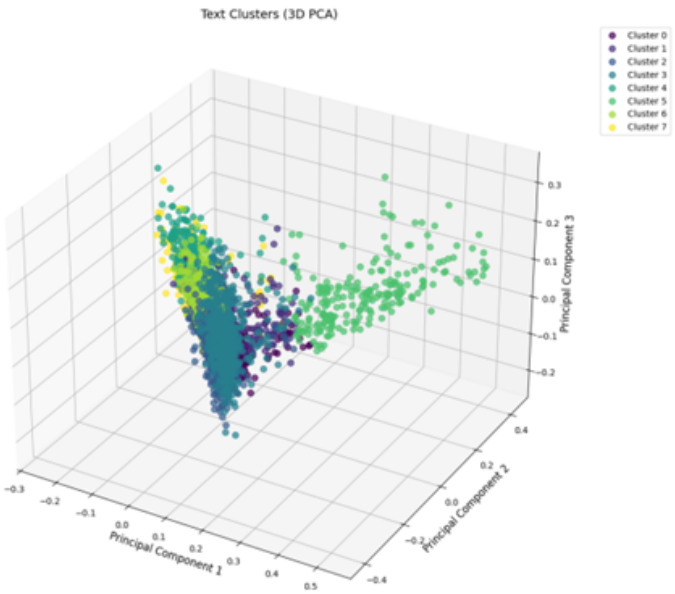
- **Negative**

For the negative dataset, we used the same method to derive that the optimal number of clusters is 8.

```
Number of clusters 2 - Average profile coefficient: 0.006128955036030641, Average Davies-Bouldin index: 9.271234370481544
Number of clusters 3 - Average profile coefficient: 0.00581080965481078, Average Davies-Bouldin index: 9.766720082848
Number of clusters 4 - Average profile coefficient: 0.005261317565499973, Average Davies-Bouldin index: 9.064545744669639
Number of clusters 5 - Average profile coefficient: 0.007135401926447038, Average Davies-Bouldin index: 8.851065169289068
Number of clusters 6 - Average profile coefficient: 0.007284977797580227, Average Davies-Bouldin index: 8.575719166693833
Number of clusters 7 - Average profile coefficient: 0.006595438300350127, Average Davies-Bouldin index: 8.157621903657821
Number of clusters 8 - Average profile coefficient: 0.006567764203786316, Average Davies-Bouldin index: 7.923770103624511
Number of clusters 9 - Average profile coefficient: 0.006589949791582649, Average Davies-Bouldin index: 7.7021718692405186
Number of clusters 10 - Average profile coefficient: 0.006479186540448991, Average Davies-Bouldin index: 7.673701925320418
```

The negative clustering visualization results are as follows and the sentiment analysis results for each cluster are negative. When we summarize the comments with the strongest sentiment for each cluster, the conclusions that can be drawn are as follows:

- Technical issues: the app is often unresponsive, black screens, and especially performs poorly on certain specific devices (e.g. Lenovo tablets and Samsung devices).
- Ad experience: Users expressed strong dissatisfaction with the frequent insertion of adverts and felt that the viewing experience was severely affected.
- Lack of offline functionality: Users were unable to use the app while offline or abroad and felt this was a major flaw.
- Inadequate customer support: Users have not resolved the issue after speaking with customer support.

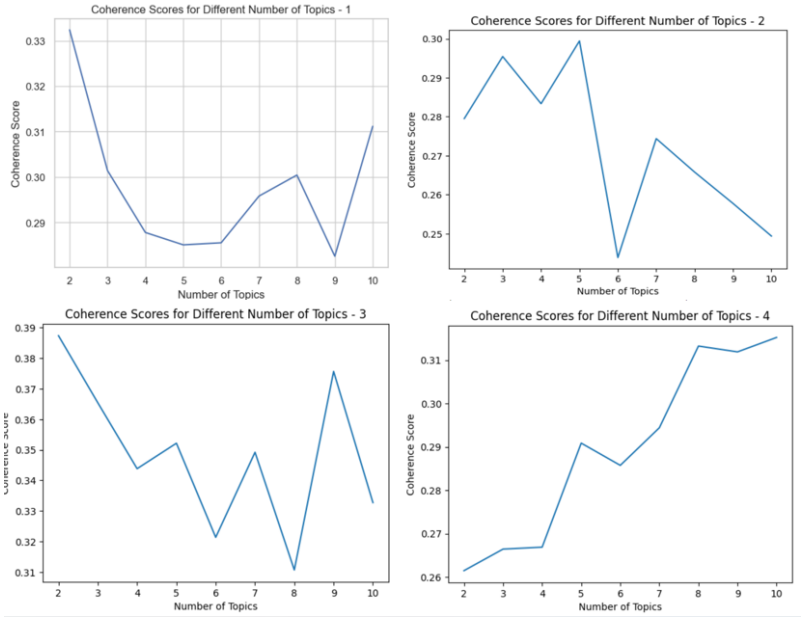


Visualisation Clustering:

We find that users are also more dissatisfied with paying for subscriptions, and the viewing experience of videos, especially paid films.

So, we went on to explore the topic feature words in the reviews of the different versions of Disney+.

The number of themes and a graph of some of the results are shown below:




```
Cluster 1 's best number of topics: 2
Cluster 1 - Topic 0: 0.022*"show" + 0.021*"movie" + 0.019*"episode" + 0.018*"watching" + 0.012*"content" + 0.011*"watch" + 0.008*"great" + 0.008*"continue" + 0.008*"play" + 0.008*"scree
n"
Cluster 1 - Topic 1: 0.015*"phone" + 0.014*"watch" + 0.013*"movie" + 0.013*"issue" + 0.011*"tv" + 0.010*"content" + 0.009*"error" + 0.009*"play" + 0.009*"service" + 0.007*"work"
```

And comparing the feature words in the positive reviews, we found many who scored low were more likely to describe their requirements and preferences for personalized settings and that the placement of advertisements caused many users to be dissatisfied.

5 Conclusions

5.1 Key Findings

- **User Sentiment Evolution：** Contrary to H1, Disney+ ratings declined steadily from 2.5 stars at launch (2019) to approximately 1.5 stars (2022-2023). This downward trend persisted across all major version updates, with no version successfully reversing user dissatisfaction.
- **Negative Review Drivers：** Supporting H2, technical issues dominate negative feedback, including app unresponsiveness, streaming problems, and device compatibility issues. Additional key complaints include intrusive advertising, lack of offline functionality, poor customer support, and subscription payment concerns.
- **ML Prediction Effectiveness：** Confirming H3, our models predicted user satisfaction with high accuracy. Neural Networks achieved over 86% accuracy, with KNN and NLP approaches also performing well. The dataset's negative review imbalance created challenges for Random Forest models in identifying positive classes.
- **Emotions and Ratings Correlation：** Supporting H4, emotion analysis revealed strong correlations between emotional content and ratings. Negative emotions (anger, disgust, fear, sadness) consistently correspond with lower ratings, while positive emotions (joy, anticipation, trust) align with higher ratings. The sentiment score showed a 0.39 correlation with user ratings.

5.2 Business Recommendations

- **Prioritize Technical Stability**: Address the critical technical issues driving negative sentiment by establishing a dedicated team focused on resolving device compatibility problems, particularly for Samsung and Lenovo devices. Implement more rigorous quality assurance testing before releasing updates.
- **Revise Advertising Strategy**: Reconsider the current advertising approach that significantly impacts user satisfaction. Implement less intrusive ad placements and offer enhanced ad-free subscription options that better align with user expectations.
- **Enhance Offline Functionality**: Expand offline viewing capabilities to address a top user pain point identified in sentiment analysis.
- **Transform Customer Support**: Implement a more comprehensive response strategy with faster resolution times, focusing particularly on the technical issues and content accessibility concerns identified in the clustering analysis.

5.3 Future research directions

- **Competitive Benchmarking Analysis**: Compare user sentiment patterns with Netflix, HBO Max, and Amazon Prime to identify industry best practices that could improve Disney+'s performance and establish realistic targets for technical stability and customer satisfaction.
- **Feature Impact Evaluation**: Measure the specific impact of individual features (offline viewing, ad frequency, UI changes) on user satisfaction through controlled studies to prioritize development resources based on quantifiable user experience improvements.
- **Retention-Sentiment Correlation**: Analyze the relationship between negative sentiment patterns and subscription cancellations to identify which specific issues most directly contribute to churn versus temporary dissatisfaction.

6 References

Chen, W., Li, X. et al. A survey of sentiment analysis in social media. Knowl Inf Syst 60, 617–663 (2019). https://doi.org/10.1007/s10115-018-1236-4

Challa N P, Madhavi K R, Naseeba B, et al. Sentiment Analysis from TWITTER Using NLTK[C]//International Conference on Hybrid Intelligent Systems. Springer, Cham, 2023: 852-861.

Aljedaani W, Rustam F, Mkaouer M W, et al. Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry[J]. Knowledge-Based Systems, 2022, 255: 109780.

Hutto C, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text[C]//Proceedings of the international AAAI conference on web and social media. 2014, 8(1): 216-225.

7 Appendix

Figure 4.1. Special words

```
# Special words configuration
NEGATIVE_WORDS = {"aint", "arent", "cannot", "cant", "couldnt", "darent", "didnt", "doesnt",
     "ain't", "aren't", "can't", "couldn't", "daren't", "didn't", "doesn't",
     "dont", "hadnt", "hasnt", "havent", "isnt", "mightnt", "mustnt", "neither",
     "don't", "hadn't", "hasn't", "haven't", "isn't", "mightn't", "mustn't",
     "neednt", "needn't", "never", "none", "nope", "nor", "not", "nothing", "nowhere",
     "oughtnt", "shant", "shouldnt", "uhuh", "wasnt", "werent",
     "oughtn't", "shan't", "shouldn't", "uh-uh", "wasn't", "weren't",
     "without", "wont", "wouldnt", "won't", "wouldn't", "rarely", "seldom", "despite"}
BOOST_WORDS = {'very', 'extremely', 'absolutely', 'completely'}
COMPARITIVE_WORDS = {'more', 'less', 'better', 'worse', 'compared', 'than'}
CONJUNCTIONS = {'but', 'however', 'although', 'though', 'yet'}
```

Figure 4.2. New columns after sentiment calculation

	content	score	thumbsUpCount	timeCreated	version_category	overall_sentiment_score	sentence_sentiment_scores	sentiment_label
0	There are many issues that pop up from time to...	4	3	2025/2/13 6:00	4	0.1249	[0.0, 0.2732, 0.2263, 0.0]	Neutral
1	This app used to be so accessible and lovable....	2	68	2024/12/17 20:02	3	0.3474	[0.6444, 0.0, 0.6908, 0.4019, 0.0]	Positive
2	EDIT: Good job, Disney, fixing watchlist on ap...	5	536	2024/12/29 10:28	3	0.1713	[0.4926, 0.0, -0.3182, 0.5106, 0.0, 0.4927, 0....	Neutral
3	Absolutely loathe this app. It continuously lo...	2	46	2024/12/17 12:36	3	0.1306	[0.0, 0.0, 0.2263, 0.296]	Neutral
4	Update- App is worse. For a paid service, ther...	1	277	2024/12/22 15:15	3	-0.1220	[0.0, -0.4588, 0.0, -0.2732, 0.0, 0.0]	Negative

Figure 4.3. stop words of high-frequency vocabulary analysis

```
stop_words = set(stopwords.words('english') + [
    "aint", "arent", "cannot", "cant", "couldnt", "darent", "didnt", "doesnt",
    "ain't", "aren't", "can't", "couldn't", "daren't", "didn't", "doesn't",
    "dont", "hadnt", "hasnt", "havent", "isnt", "mightnt", "mustnt", "neither",
    "don't", "hadn't", "hasn't", "haven't", "isn't", "mightn't", "mustn't",
    "neednt", "needn't", "never", "none", "nope", "nor", "not", "nothing", "nowhere",
    "oughtnt", "shant", "shouldnt", "uhuh", "wasnt", "werent",
    "oughtn't", "shan't", "shouldn't", "uh-uh", "wasn't", "weren't",
    "without", "wont", "wouldnt", "won't", "wouldn't", "rarely", "seldom", "despite",
    "emoji", "example",
    "very", "really", "quite", "too", "just", "so", "such",
    "extremely", "highly", "fairly", "somewhat", "rather",
    "completely", "absolutely", "totally", "nearly", "almost",
    "amazingly", "awfully", "considerable", "considerably",
    "decidedly", "deeply", "effing", "enormous", "enormously",
    "entirely", "especially", "exceptional", "exceptionally",
    "extreme", "extremely",
    "fabulously", "flipping", "flippin", "frackin", "fracking",
    "fricking", "frickin", "frigging", "friggin", "fully",
    "fuckin", "fucking", "fuggin", "fugging",
    "greatly", "hella", "hugely",
    "incredible", "incredibly", "intensely",
    "major", "majorly", "more", "most", "particularly",
    "purely", "remarkably", "substantially",
    "thoroughly", "total", "tremendous", "tremendously",
    "uber", "unbelievably", "unusually", "utter", "utterly",
    "even", "app", "disney", "im", "one",
    "almost", "barely", "hardly", "just enough",
    "kind of", "kinda", "kindof", "kind-of",
    "less", "little", "marginal", "marginally",
    "occasional", "occasionally", "partly",
    "scarce", "scarcely", "slight", "slightly", "somewhat",
    "sort of", "sorta", "sortof", "sort-of",
    'like', 'get', 'go', 'make', 'see', 'know', 'say',
    'want', 'think', 'come', 'look', 'use', 'work',
    'back', 'also', 'since', 'still',
    'thing', 'things', 'one', 'two', 'three', 'time',
    'way', 'lot', 'people',
    'few', 'many', 'every', 'all', 'be', 'is', 'are', 'was', 'were', 'have', 'has', 'had',
    'do', 'does', 'did', 'can', 'could', 'will', 'would',
    'shall', 'should', 'may', 'might', 'must',
    "any", "some", "such", "all", "both", "each", "few", "many", "most", "several",
    "this", "that", "these", "those", "here", "there", "where", "when", "why", "how",
    "if", "though", "although", "while", "but", "and", "or", "nor", "for", "yet",
    "so", "then", "than", "as", "like", "just", "only", "rather", "still", "again",
    "another", "same", "different", "next", "last", "first", "many", "much", "more",
    "less", "either", "neither", "both", "each", "every", "few", "all", "none",
    "some", "any", "no", "not", "never", "always", "often", "sometimes", "rarely",
    'disney', 'app', 'apps', 'like', 'watch', 'the', 'this', 'that', 'thats', 'is', 'it', 'doesnt', 'dont', 'havent',
    'to', 'and', 'for', 'of', 'with', 'im', 'lot', 'one',
    'something', 'anything', 'everything', 'nothing', 'thing', 'way', 'please', 'fix', 'beginning', 'minute', 'reason', 'continue',
    'day', 'month', 'second', 'section', 'thought', 'quite', 'list', 'selection',
    'launch', 'experience', 'everything', 'data', 'quality', 'access',
    'issue', 'problem', 'stuff', 'star', 'phone', 'service', 'show', 'time', 'tv', 'movie', 'content', 'episode', 'series', 'film', 'screen',
    'movies'])
```

Figure 4.4. result of top-10 high-frequency vocabulary

```
Version Version 1, Score 1 Top 10 Words:
['error (90)', 'video (90)', 'subscription (74)', 'device (68)', 'play (55)', 'netflix (46)', 'code (44)', 'wifi (40)', 'internet (39)', 'trial (39)']

Version Version 1, Score 2 Top 10 Words:
['video (56)', 'error (52)', 'wifi (42)', 'option (39)', 'chromecast (35)', 'netflix (35)', 'bug (32)', 'connection (30)', 'year (29)', 'feature (28)']

Version Version 1, Score 3 Top 10 Words:
['netflix (50)', 'feature (49)', 'option (47)', 'video (39)', 'watchlist (38)', 'device (32)', 'error (28)', 'play (28)', 'title (25)', 'bug (24)']

Version Version 1, Score 4 Top 10 Words:
['feature (38)', 'option (36)', 'netflix (26)', 'video (22)', 'bug (21)', 'title (21)', 'watchlist (21)', 'device (18)', 'picture (18)', 'streaming (17)']

Version Version 1, Score 5 Top 10 Words:
['watchlist (4)', 'trial (3)', 'wifi (3)', 'button (2)', 'catalog (2)', 'device (2)', 'hd (2)', 'support (2)', 'update (2)', 'war (2)']

Version Version 2, Score 1 Top 10 Words:
['device (88)', 'subscription (78)', 'error (76)', 'netflix (48)', 'video (45)', 'chromecast (38)', 'update (38)', 'year (36)', 'download (34)', 'money (34)']

Version Version 2, Score 2 Top 10 Words:
['device (22)', 'video (20)', 'connection (16)', 'option (16)', 'downloads (14)', 'error (14)', 'button (13)', 'chromecast (12)', 'offline (11)', 'update (11)']

Version Version 2, Score 3 Top 10 Words:
['video (10)', 'credit (8)', 'option (8)', 'language (7)', 'device (6)', 'error (6)', 'wifi (6)', 'functionality (5)', 'play (5)', 'player (5)']

Version Version 2, Score 4 Top 10 Words:
['language (4)', 'option (4)', 'marvel (3)', 'rating (3)', 'cell (2)', 'feature (2)', 'mirrorshare (2)', 'others (2)', 'season (2)', 'video (2)']

Version Version 2, Score 5 Top 10 Words:
[]

Version Version 3, Score 1 Top 10 Words:
['ad (461)', 'subscription (324)', 'device (175)', 'account (156)', 'error (147)', 'money (130)', 'update (122)', 'chromecast (112)', 'price (107)', 'year (104)']

Version Version 3, Score 2 Top 10 Words:
['ad (91)', 'subscription (51)', 'hulu (48)', 'price (34)', 'error (29)', 'connection (28)', 'device (27)', 'update (26)', 'option (25)', 'account (23)']

Version Version 3, Score 3 Top 10 Words:
['ad (37)', 'hulu (19)', 'chromecast (13)', 'connection (13)', 'need (13)', 'netflix (13)', 'error (12)', 'feature (12)', 'option (12)', 'title (12)']

Version Version 3, Score 4 Top 10 Words:
['ad (20)', 'option (16)', 'hulu (12)', 'netflix (10)', 'commercial (9)', 'credit (9)', 'update (9)', 'feature (8)', 'play (8)', 'subscription (7)']

Version Version 3, Score 5 Top 10 Words:
['family (19)', 'channel (13)', 'hulu (11)', 'year (11)', 'ad (10)', 'kid (9)', 'love (9)', 'subscription (9)', 'bundle (8)', 'programme (8)']

Version Version 4, Score 1 Top 10 Words:
['ad (66)', 'subscription (26)', 'device (18)', 'password (14)', 'video (13)', 'account (12)', 'money (11)', 'chromecast (9)', 'crash (9)', 'update (9)']

Version Version 4, Score 2 Top 10 Words:
['ad (17)', 'error (10)', 'price (5)', 'audio (4)', 'chromecast (4)', 'hulu (4)', 'play (4)', 'program (4)', 'subscription (4)', 'house (3)']

Version Version 4, Score 3 Top 10 Words:
['device (5)', 'play (3)', 'bug (2)', 'button (2)', 'connection (2)', 'effect (2)', 'internet (2)', 'subtitle (2)', 'video (2)', 'volume (2)']

Version Version 4, Score 4 Top 10 Words:
['audio (3)', 'hulu (3)', 'option (3)', 'part (3)', 'ad (2)', 'button (2)', 'commercial (2)', 'feature (2)', 'tablet (2)', 'variety (2)']

Version Version 4, Score 5 Top 10 Words:
['subscription (5)', 'streaming (4)', 'ad (3)', 'dtsx (3)', 'bundle (2)', 'cartoon (2)', 'hour (2)', 'hulu (2)', 'preference (2)', 'suggestion (2)']

Plotting data frame:
   Version  Score         Word  Frequency
0        1      1        error         90
1        1      1        video         90
2        1      1  subscription         74
3        1      1       device         68
4        1      1         play         55
```

Individual Self-reflection report

Student Name: DU Qinshu **Student No.:** 58819313

Part A	What are the main difficulties encountered in this analytics project?
	Data Collection: Though data scraping was quick, finalizing the dataset took weeks. Studied Kaggle's data structure, then decided to scrape app reviews due to website restrictions. Found usable code on GitHub. Initially scraped DeepSeek app reviews but switched due to a short time span. Tried Google Maps data but found too many professional terms. Finally chose Disney + reviews as they had the least "snarky" comments and were easier to analyze. **Sentiment Analysis**: The main issue was handling imbalanced data sources to improve model accuracy. The problem stemmed from data imbalance (over 50% one-star reviews) and the raw nature of directly scraped data, which wasn't as refined as Kaggle's processed data. Adopted strategies like special word handling, weighting different star reviews, and a dynamic neutrality threshold to enhance the model's accuracy, particularly for positive and neutral categories. **Clustering**: The key challenge was determining the optimal number of clusters. Used a line chart combining the elbow method and silhouette coefficient to assess different cluster numbers and ultimately determine the most suitable quantity.
Part B	How the concepts, methods, and techniques learnt in classes are applied to conduct the data analytics project, and how they solve the business problems?
	Concepts: Calculate sentiment scores using TF-IDF to evaluate the importance of words in reviews; Classify reviews into positive, neutral, or negative through sentiment analysis; Use clustering methods (e.g., K-means) to group data and identify patterns in reviews. **Methods**: Preprocess text with tokenization, stemming, and handle special linguistic phenomena (e.g., negations, adverbs); Score sentiments using VADER and adjust scores with TF-IDF weights; Cluster data using K-means and determine the optimal number of clusters with the elbow method. **Techniques**: Use the NLTK library for text processing and sentiment analysis; Use sklearn to compute TF-IDF vectors and evaluate classification performance; Visualize data with matplotlib and seaborn; Perform clustering analysis with K-means. **Application to Business Problem**: All above focuse on analyzing user reviews for the Disney+ app by examining their sentiment and extracting high-frequency words. It helps businesses quickly spot common issues and desired features mentioned by users. Clustering reviews into groups also reveals prevalent concerns or positive feedback, enabling companies to enhance the app and boost user satisfaction.

Individual Project Contribution Form

Student Name: DU Qinshu **Student ID:** 58819313

Project Title:

Sentiment Analytics for Streaming Platform User Experience: Uncovering User Sentiment Patterns in Disney+ Application Reviews

Project Group Name:

Group 10086

Contribution of Each Member in the Group:

Student ID	**Last Name**	**First Name**	**Contribution %**
58670679	GAO	Xinran	20%
58819313	DU	Qinshu	20%
58878190	JIANG	Yihui	20%
58903615	ZHAO	Wei	20%
58838131	ZHENG	Yiyang	20%
		Total	100%