

IMPERIAL

App Store Applications Reviews Analysis

How Reviews Help Developers Improve Their Products

Baoqi Zhang
April 26, 2025

Contents

Population of Interest:

App users who leave reviews.

App developers who regularly check reviews.

Quantities of Interest:

- App rating and Categories Predictions
- Common topics in user reviews

01 Data Introduction

02 Star Rating Model and Prediction

03 Transfer learning

04 Multinomial Model Category Analysis

05 Topic Model Analysis

Data Introduction

- Data source website: AppFollow
- We explore 10 different categories of apps: Finance, Games, Food Delivery, Dating, Entertainment, Shopping, Productivity, Social Media, Travel, and Music
- Each category contains information on 5–10 different apps
- Each document has about 30,000–50,000 data points, forming a powerful dataset for analysis

The screenshot shows the AppFollow interface. On the left, there's a sidebar with various analytical tools: Home hub, Reviews (selected), Automation hub, Tags, Templates, Monitor dashboards, Rating analysis, Reviews analysis, Tags analysis, Agent performance, Automation performance, Custom reporting, Get insights, Exec report™, Semantic analysis (with AI icon), and Phrase analysis.

The main area is titled "Reviews feed" and displays data for the "Monzo - Banking made..." app. It shows 227 reviews, an average rating of 3.23, and 12% replies. There are tabs for "Featured reviews" (382), "All reviews" (227), "No replies" (227), and "Pending approval". A prominent feature is an AI-driven automation tool that suggests saving time by letting AI handle replies. Below this, a specific review from "LEANNEDY92" dated Feb 22, 2025, at 2:04 AM is shown, rated 5 stars. The review text is: "VERY GOOD DOWNLOAD MONZO. VERY EASY TO USE QUICK AND LOVE THE DAY EARLY PAYDAY. DEFINITELY HAVE AND WILL HIGHLY RECOMMEND YOU". The review includes fields for Country (United Kingdom), Review language (English), and App version (6.34.0). At the bottom, there are options to "Generate AI reply" (with suggestions like "Motivation", "Satisfied user", "Feature request", and "Add tag"), "Thank you", "General positive", and a "Personalize AI replies with Custom Instructions" button.

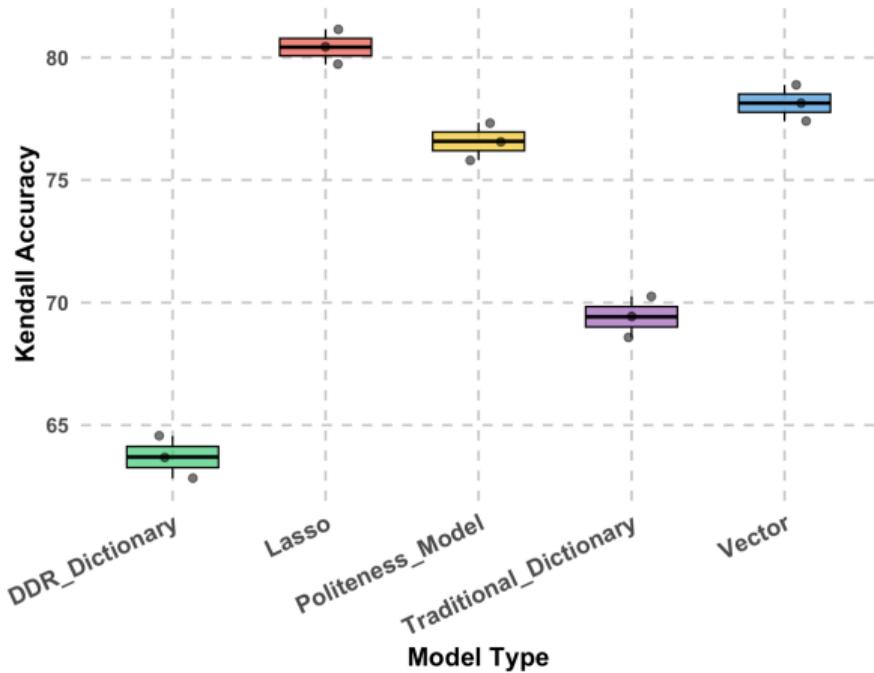
Lasso and Benchmarks Comparison

Financial Apps Dataset Only

Key Insights:

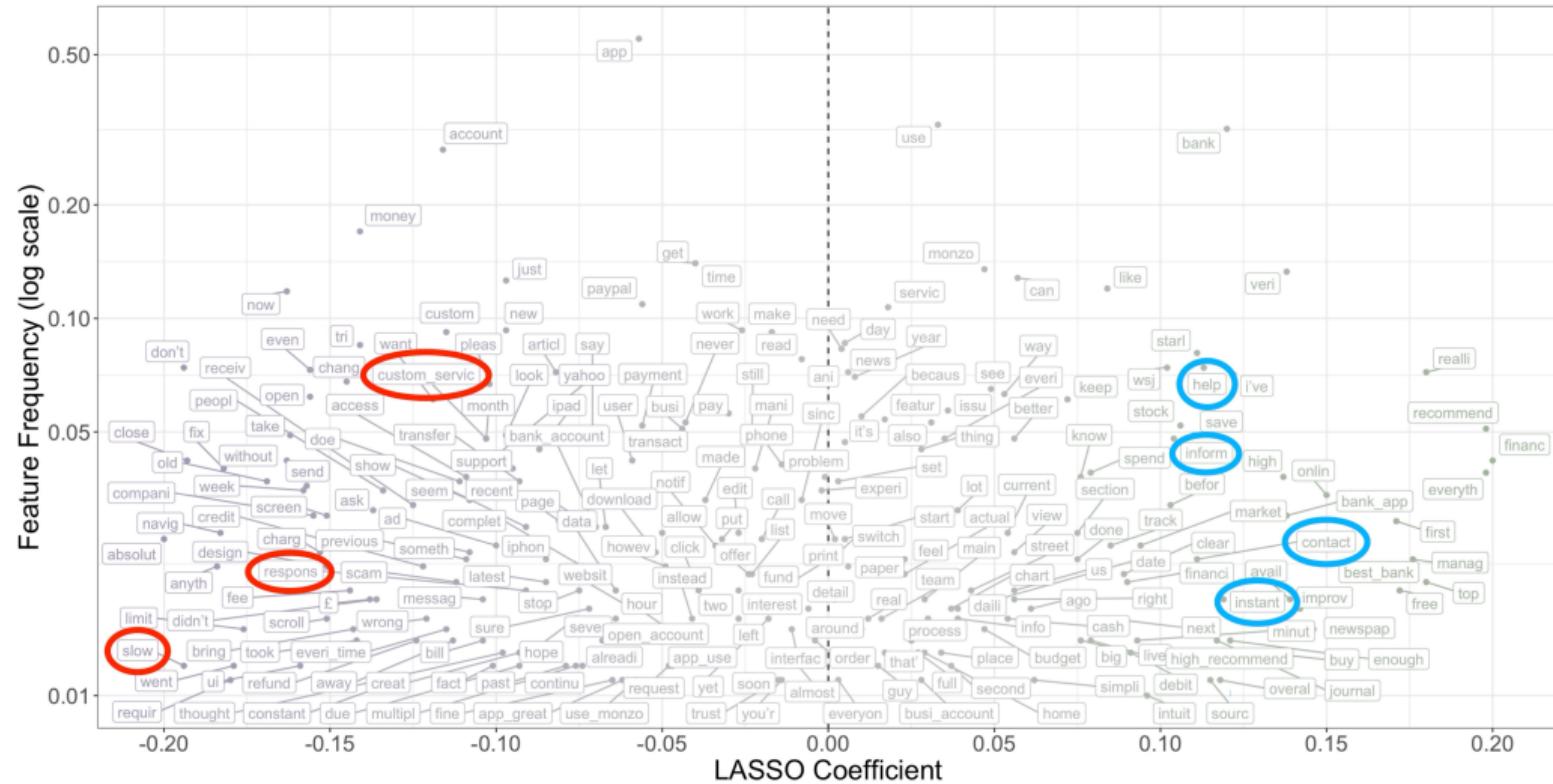
- Use LASSO regression as the main model and compare it to four benchmark models
- LASSO regression achieves the best performance
- LASSO-based sentiment analysis will be solid in our future research

Model Performance on Finance Dataset



Lasso Model

Coefficient Plot



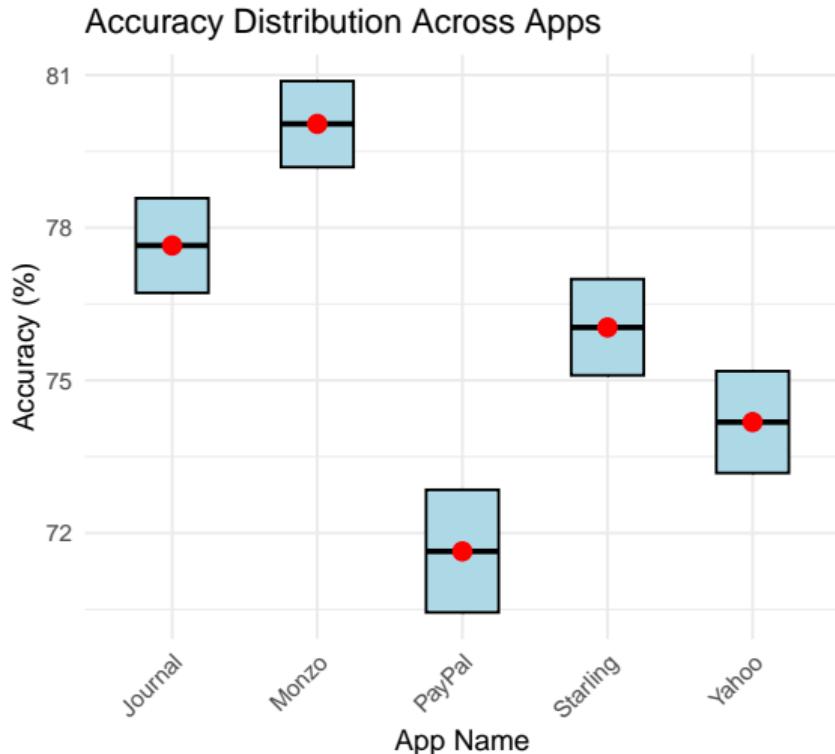
Transfer Learning: Rolling Predictions

Key Insights:

- **App-Level Predictions:** prediction accuracy differs across apps
- **Accuracy:** Generally high at the app level

Conclusion:

- Transfer learning helps predict app ratings
- Providing a solid logical foundation for further research



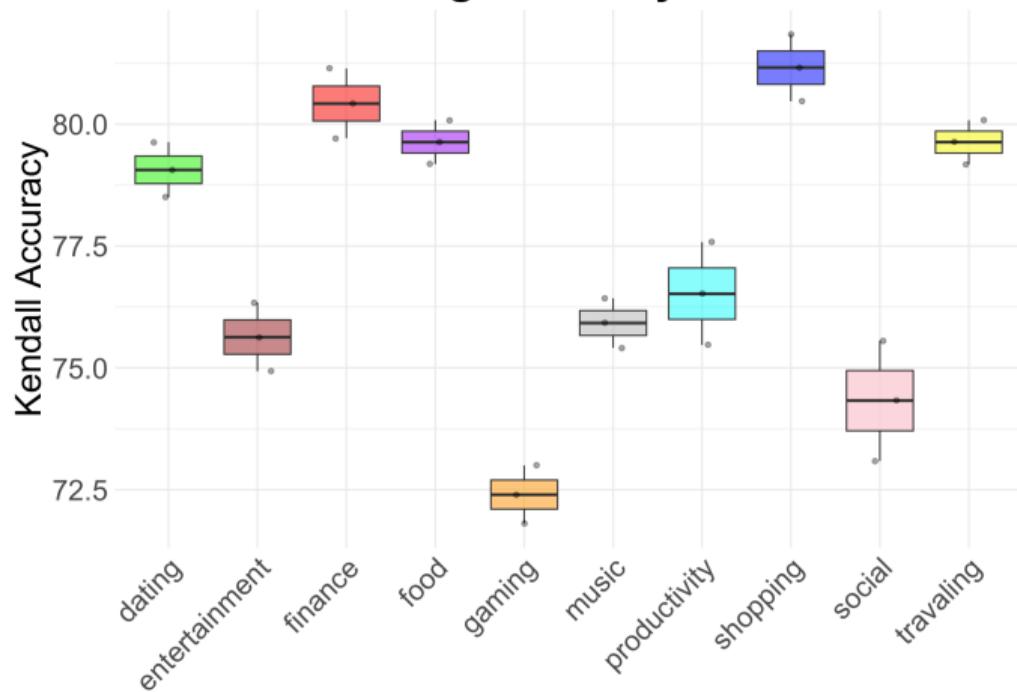
Transfer Learning:

Use financial Reviews to Predict
Other Categories

Key Insights

- **Categorical Level:** 10 categories analyzed
- **Target Categories** Dating, Entertainment, Finance, Food, Gaming, Music, Productivity, Shopping, Social, Traveling.
- **Accuracy:** Overall acceptable, but variations exist.

Transfer Learning Accuracy Across Datasets



Transfer Learning: Improving with Game Data

Key Insights

- Game accuracy particularly low
- Improve by add more reviews to training setSocial, Traveling.
- More reviews added, Higher the accuracy



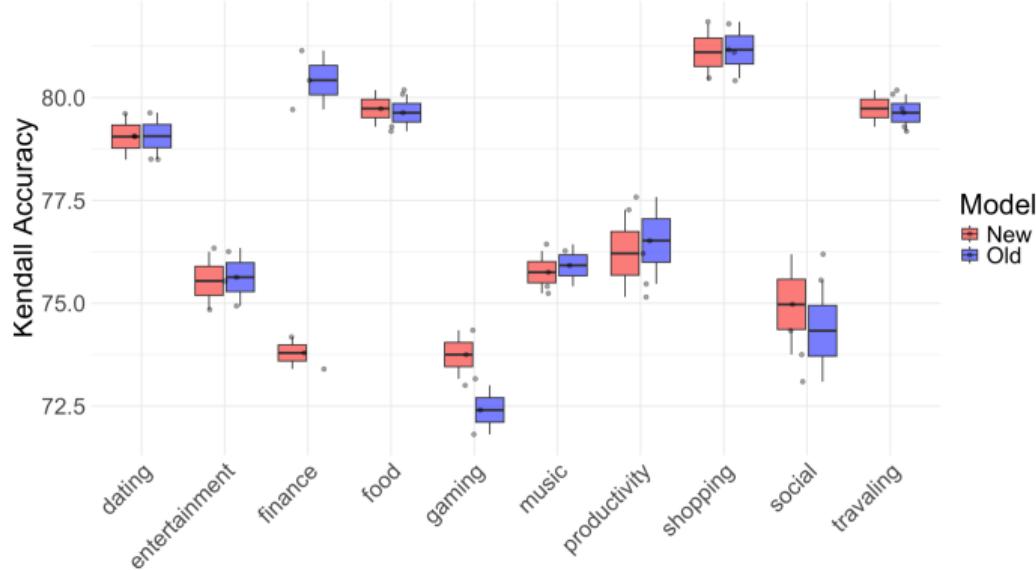
(Above) Topic Correlation with Star Ratings

Transfer Learning: Improved Model

Key Insights

- Trade-off in Transfer Learning
- Domain-Specific Knowledge Gets Weakened
- Structured Approach Required

Comparison of Old vs. New Model Accuracy Across Datasets



(Above) Topic Correlation with Star Ratings

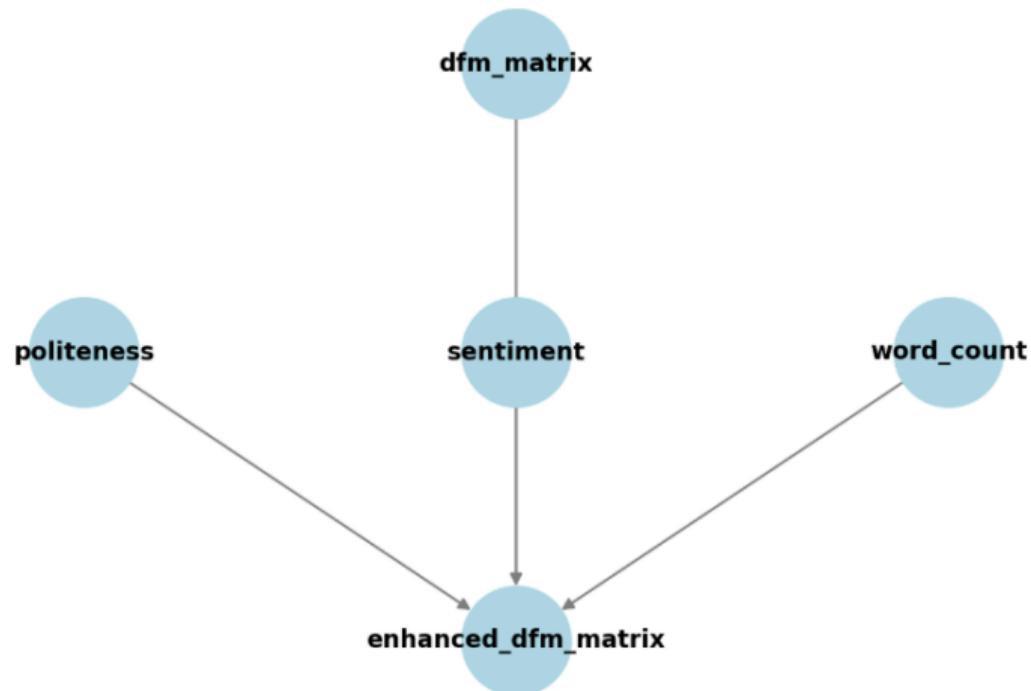
Pre-processing

Detailed Steps

Pre-processing Steps:

- Reviews often contain long sentences; we use trigrams, stopwords, tokenization, and word stemming.
- A minimum token proportion of 0.001 is set to avoid dropping too many words.
- For enhanced sentiment analysis, we apply negation review and lemmatisation (treating words following negations as having opposite sentiment).
- Additional features (dictionaries, politeness scores, and sentiment scores) are added using `cbind()` to boost model accuracy.

Feature Addition using `cbind`



(Above) Feature Addition using `cbind()`

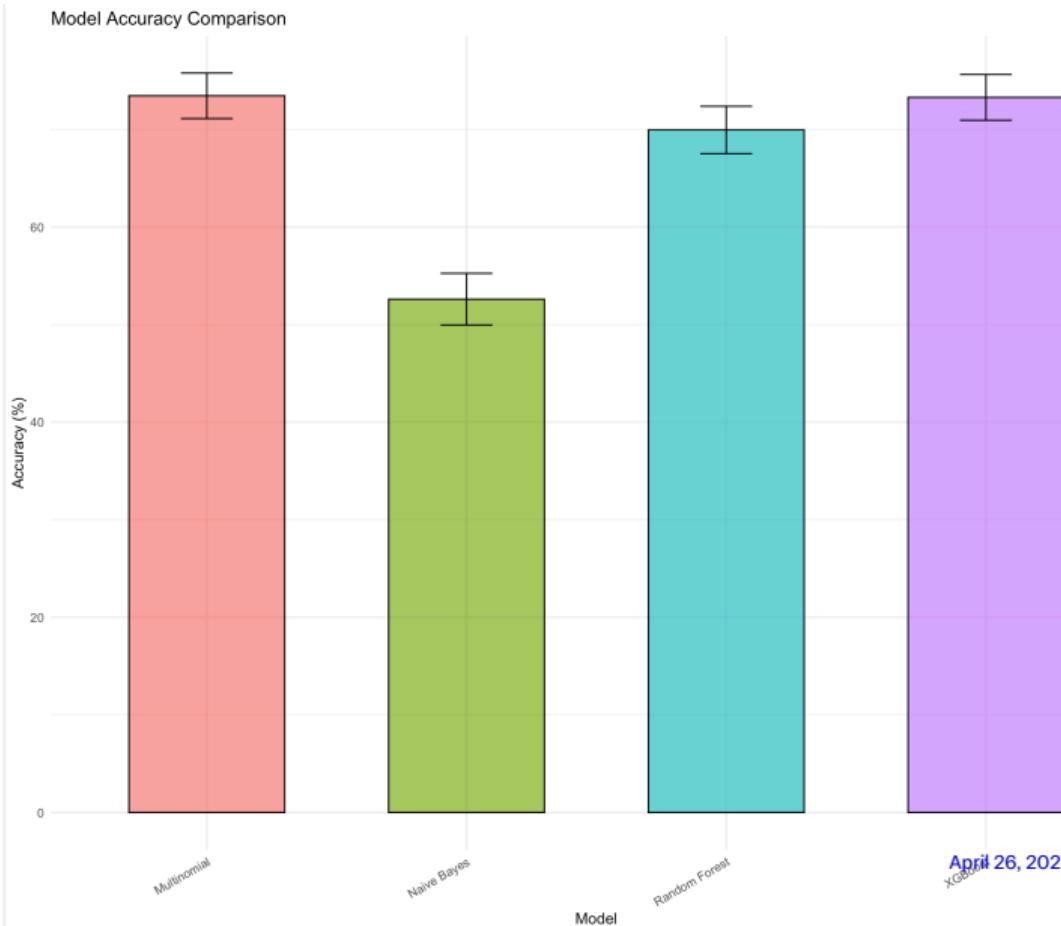
Multinomial Model

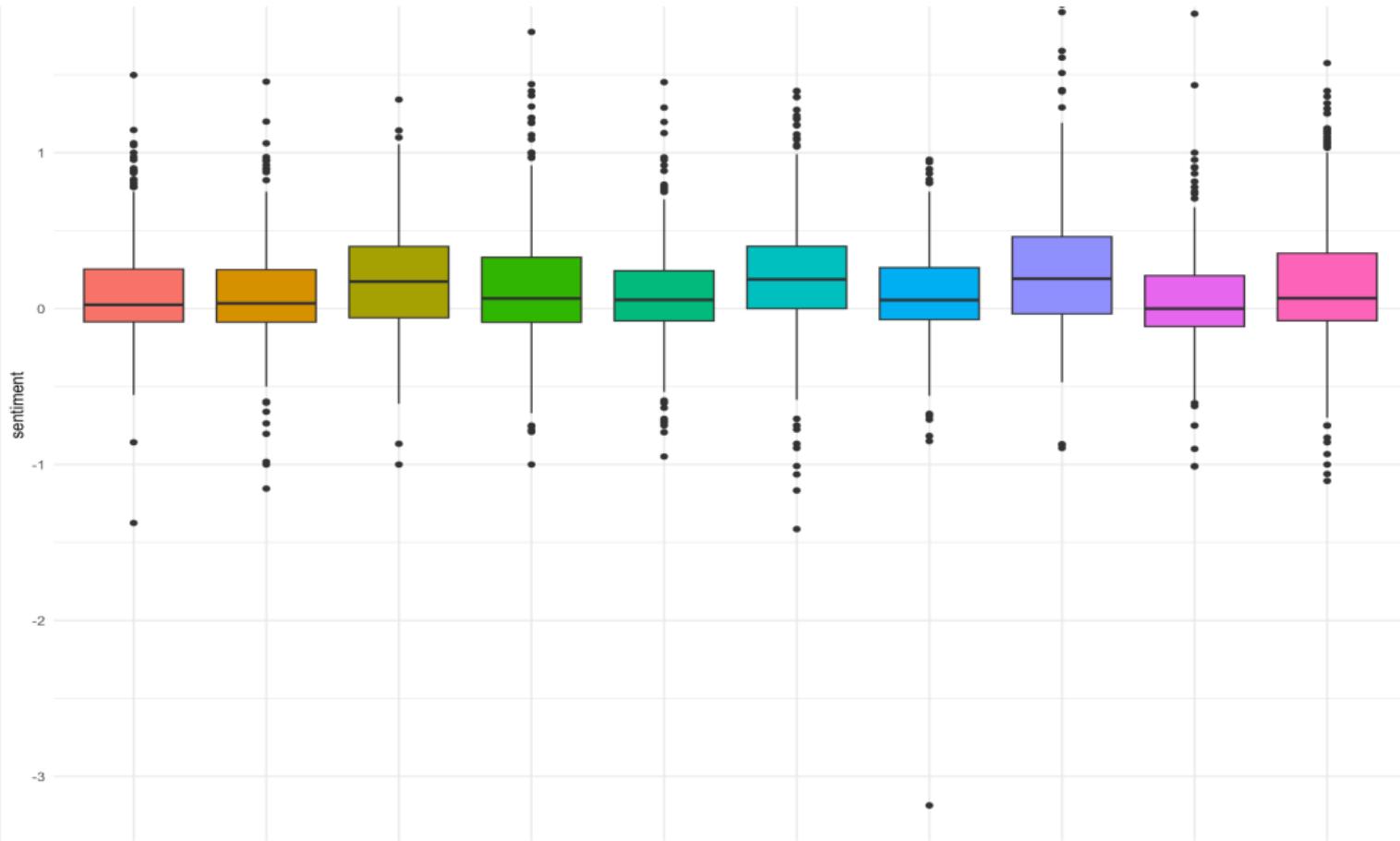
10 Categories Model

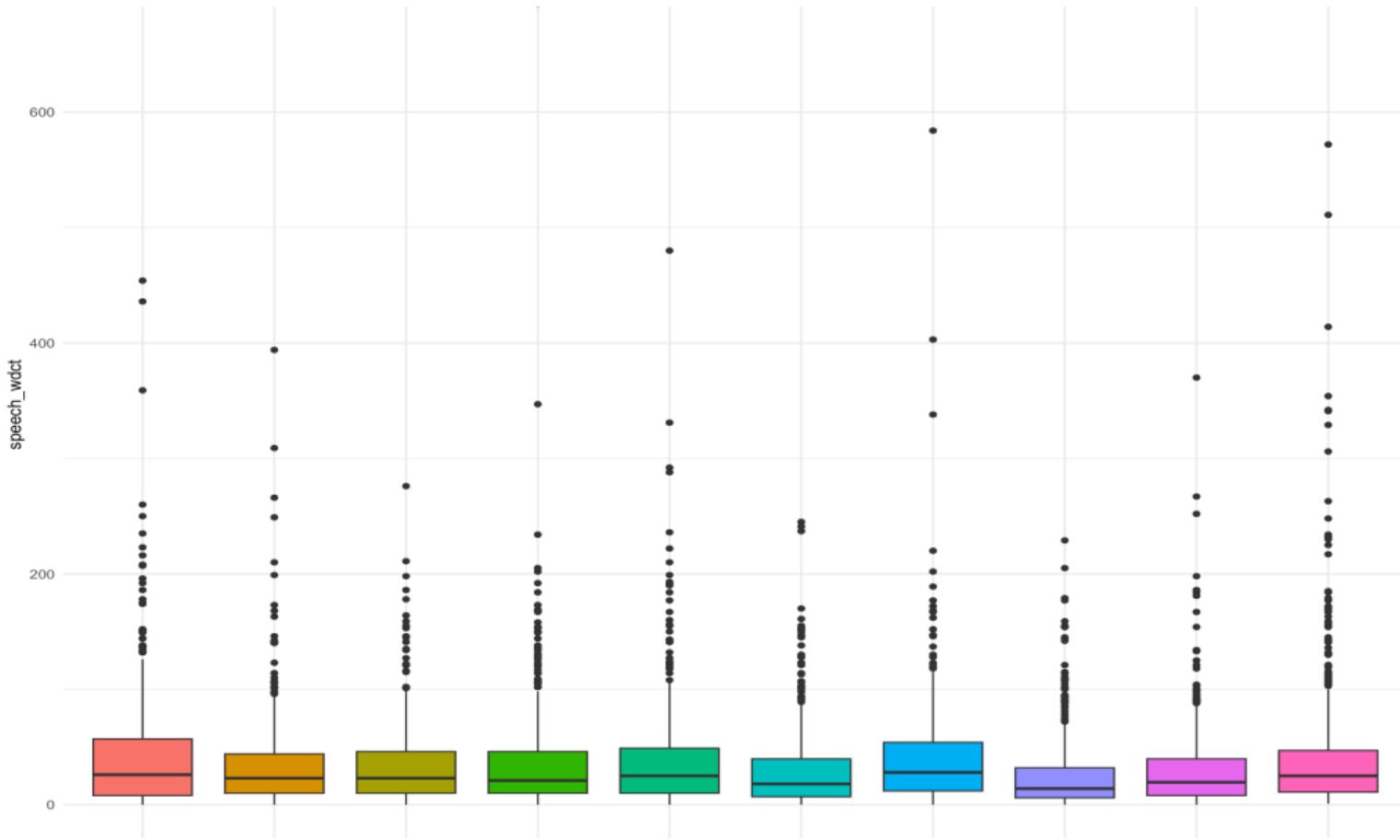
Model Comparison: Multinomial vs. Other Classifiers

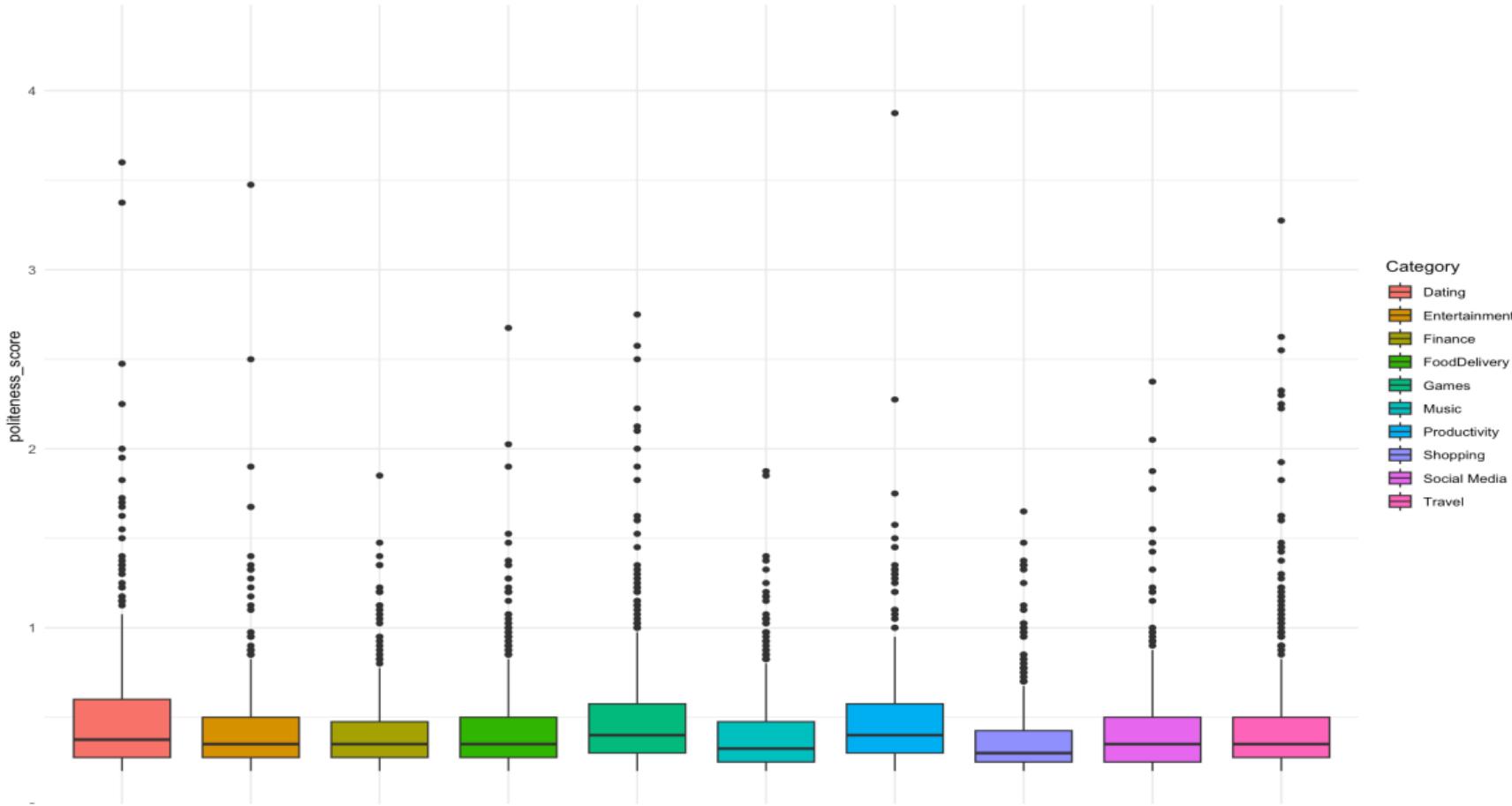
Key Findings:

- **Multinomial Model** achieved the highest accuracy, making it effective for review classification.
- **XGBoost** performed similarly well, benefiting from its ability to capture non-linear patterns.
- **Random Forest** showed moderate performance, outperforming Naïve Bayes.
- **Naïve Bayes** had the lowest accuracy due to its assumption of word independence.

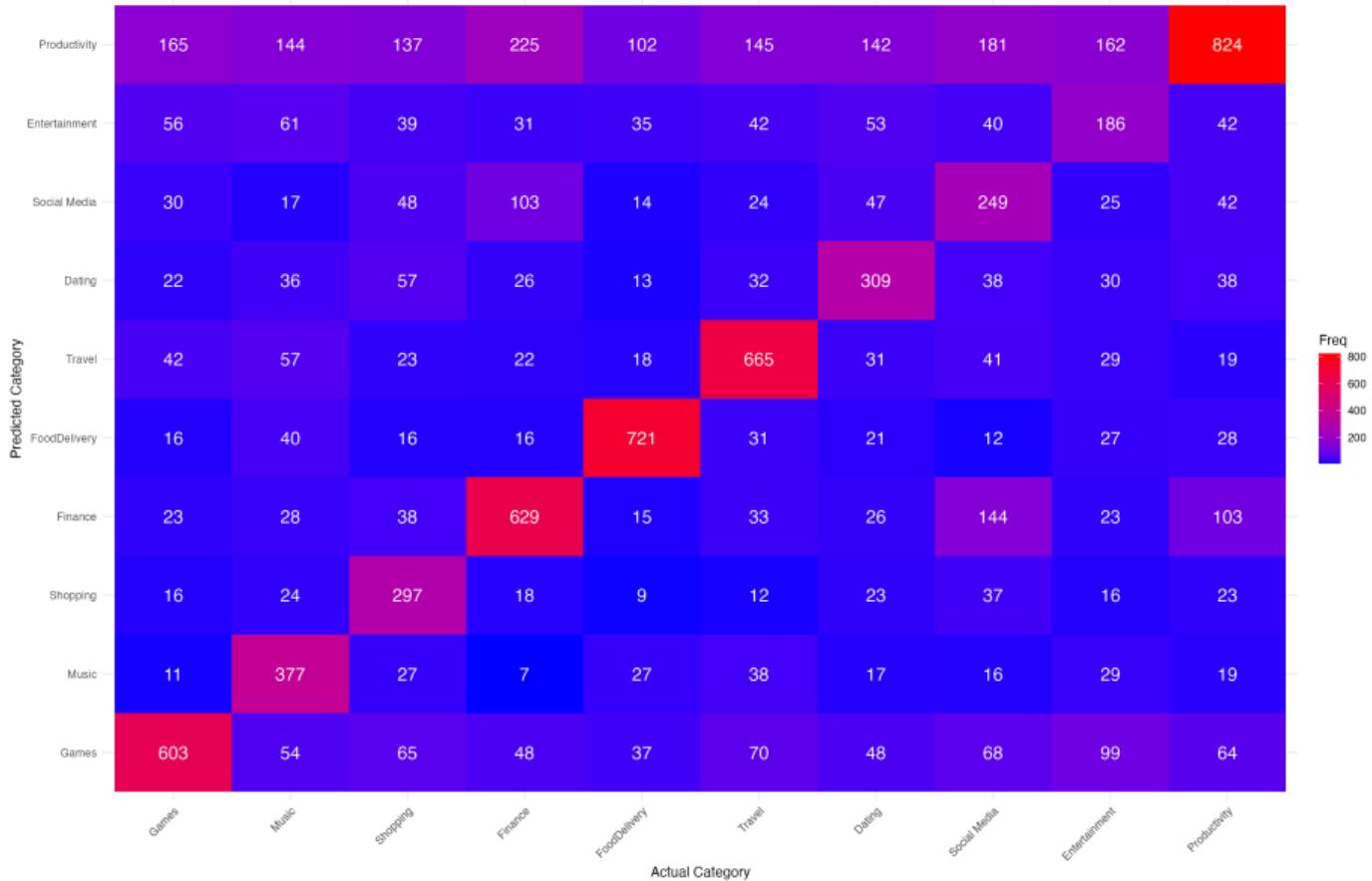








Confusion Matrix



Topic Modelling

Refining Topic Labels

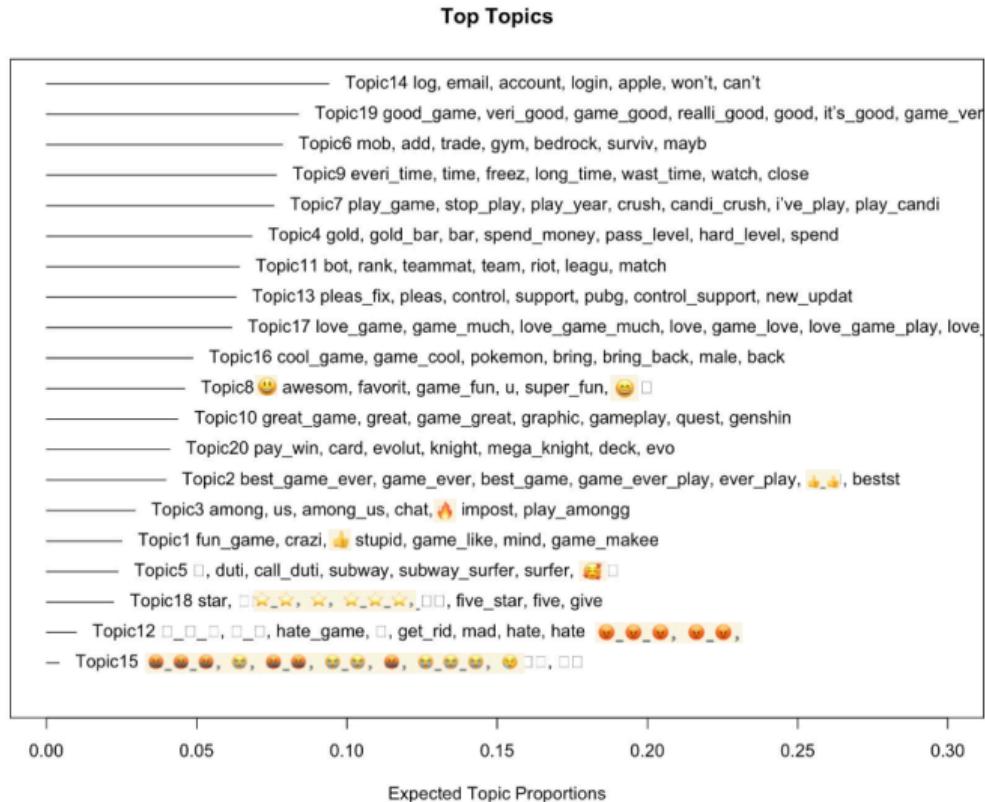
Key Insights

Topic Identification:

- Extracted key themes from app reviews.
- Topics highlight user experiences and concerns.

Redefining Topic Names:

- Original labels lacked clarity.
- Renamed topics based on keyword patterns.
- Some topics included **emoji artifacts**, Exaggerate the emotions of reviews.



wonder_game
love_subway gonna hope surfer game
never play_subway_surfers develop depress
gotta theme die feel amaze man polic game made
make impress track game_hope amaze_game real_life
teach better battl glad love_candi_fortin season dream because
wonder teach better battl love_candi just move
multiplay love_subway_surfers bye amazing jump WOW = boygamer
i'm action call duti_mobil
happi subway eat keep like
feel_like peace true i'm_happi
love_candi_crush lol thank real mobil
subway_surfers run son duti_mobil
call_duti

love_among game say
read person think think
mean gonna thank
cuz pls just US + seek get ask onli
map ban impost chat
thing talk whi word also tell wanna quick_chat give_us
make see use make word said everyone us firebean someth
parent use child crew mat room task let let report becaus
among_us know like lobbi
kill yeah
play_among_us typeinappropri play hack peopl
random i'm role play now don't
can hacker yall anyway ppl meet quick someon
stuffimpostor free chat instead one guy sorri
lie swear better free
play_among want name peopl_say

pokémon right_now
femal catch player year
start_right want natiaw away
rareyet feeltook actual
changecome back even
use just fairmake listen
servic bring_back niatic
shini everyth ball around bring
month Miss battlemale charact effect
complet ticket like playone
game cool antit
half also first get look girl
see sound came already onli
feel_like disappoint behind come
fan_servic turn still releas pokemon go got
back_game male communiti i'm
back_old fan dynamax gone
now go go_back cool
charact go_back
back

Topic Modelling

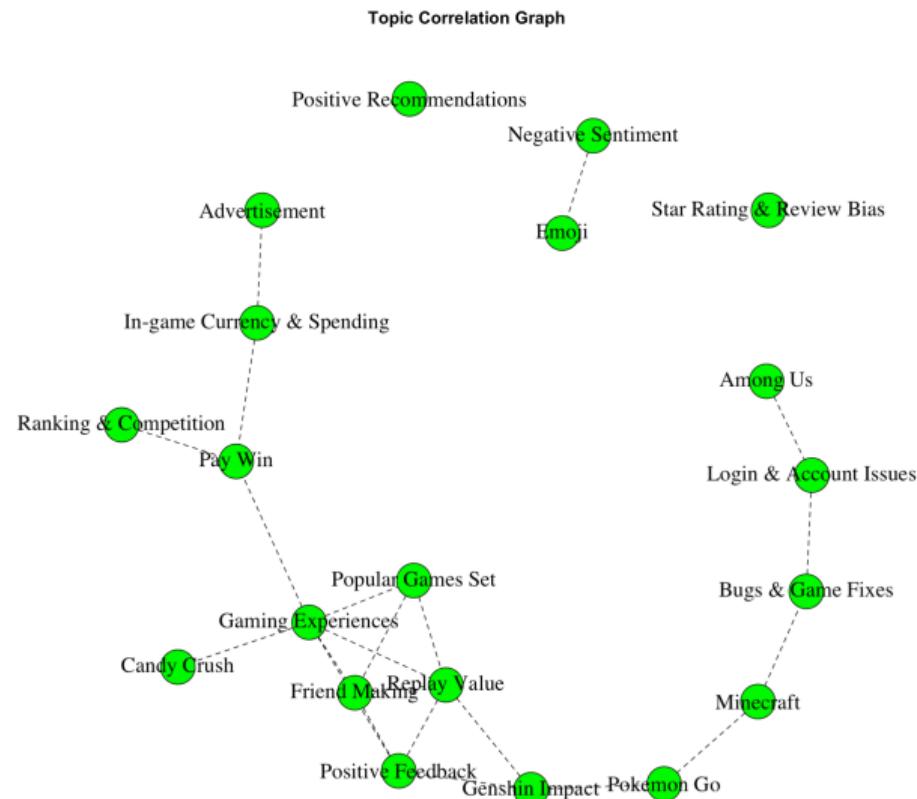
Topic Relationships and Correlations

Topic Clusters:

- **Game Reviews, Replay Value, Friend Making** are highly interconnected.
- **Pay-to-Win** links to **Ranking**, showing fairness concerns.

Isolated Topics:

- Positive Recommendation Picks and Star Ratings are standalone.
- Negative Sentiment interacts with Emoji.

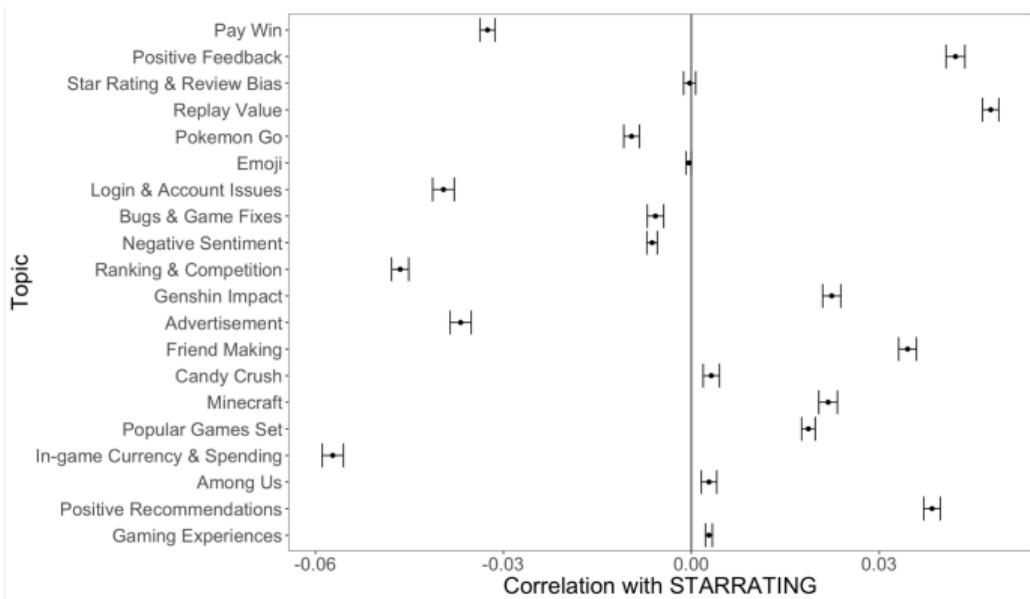


Topic Modelling

Topic Correlations with Ratings

Key Insights

- Some topics strongly correlate with **Star Ratings**.
- **Pay-to-Win, Good Game Reviews** show notable effects.
- **Negative impact:** Bug Fixes, Login Issues.
- **Positive impact:** Replay Value, Graphics.
- Some topics show weak influence on ratings.



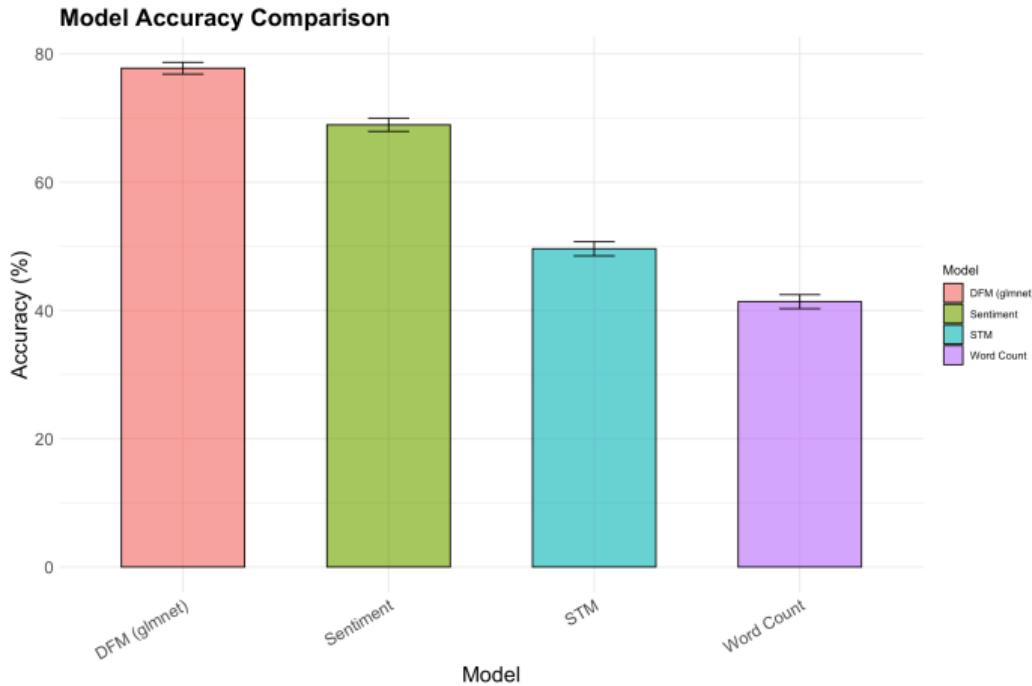
(Above) Topic Correlation with Star Ratings

Topic Modelling

Model Accuracy Comparison

Key Insights

- **DFM (glmnet)** achieves the highest accuracy.
- **Sentiment model** performs well but slightly lower.
- **STM shows moderate accuracy**, indicating topic structure alone is less predictive.
- **Word Count** is the weakest predictor.



(Above) Topic Correlation with Star Ratings

Model Limitations

Challenges and Constraints

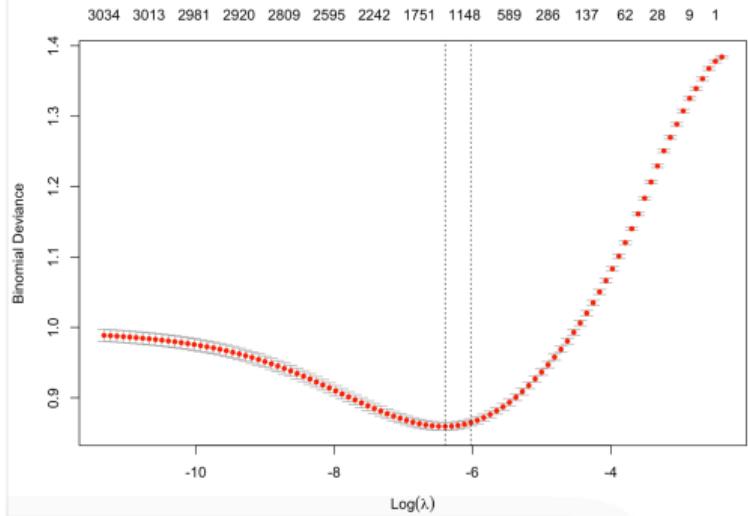
Models Limitations:

- **Lacks interpretability** – Lasso selects words but does not explain their impact.
- **Ignores context** – Lasso fails to capture negations or word dependencies.
- **Feature dropping** – Penalization removes useful correlated features.
- **Lower Predictive Accuracy** – Topic models prioritize theme extraction over rating prediction.

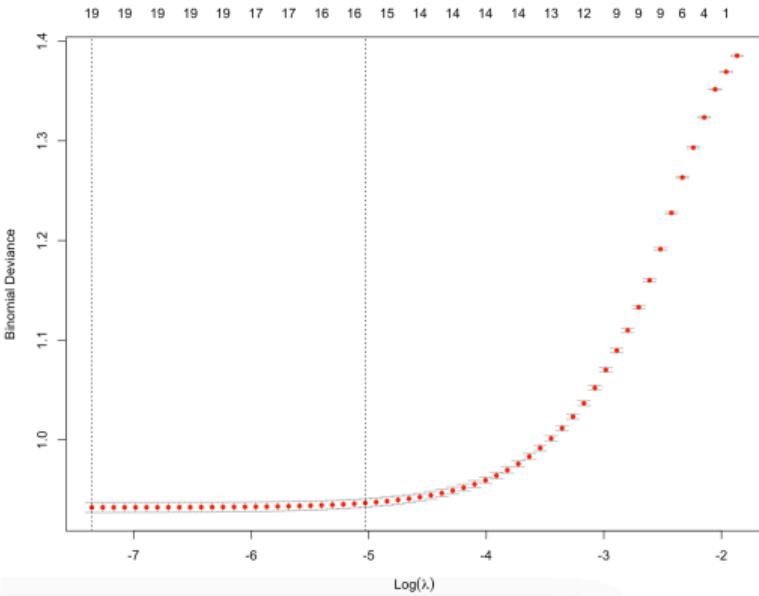
General Limitations:

- **Noisy data** – mixed positive and negative feedback in reviews.
- **Rating biases** – external factors (e.g., updates, promotions) influence scores.
- **Temporal shifts** – app changes over time but model relies on historical data.
- **Sentiment complexity** – sarcasm and nuanced expressions mislead models.
- **Data** – Only app store reviews may not reflect all aspects of users. Should combine with other platforms such as google store and etc.

Appendix



N-Grams Model



Topic Model

IMPERIAL

**Thank you.
Questions?**

**App Store Applications Reviews Analysis
April 26, 2025**