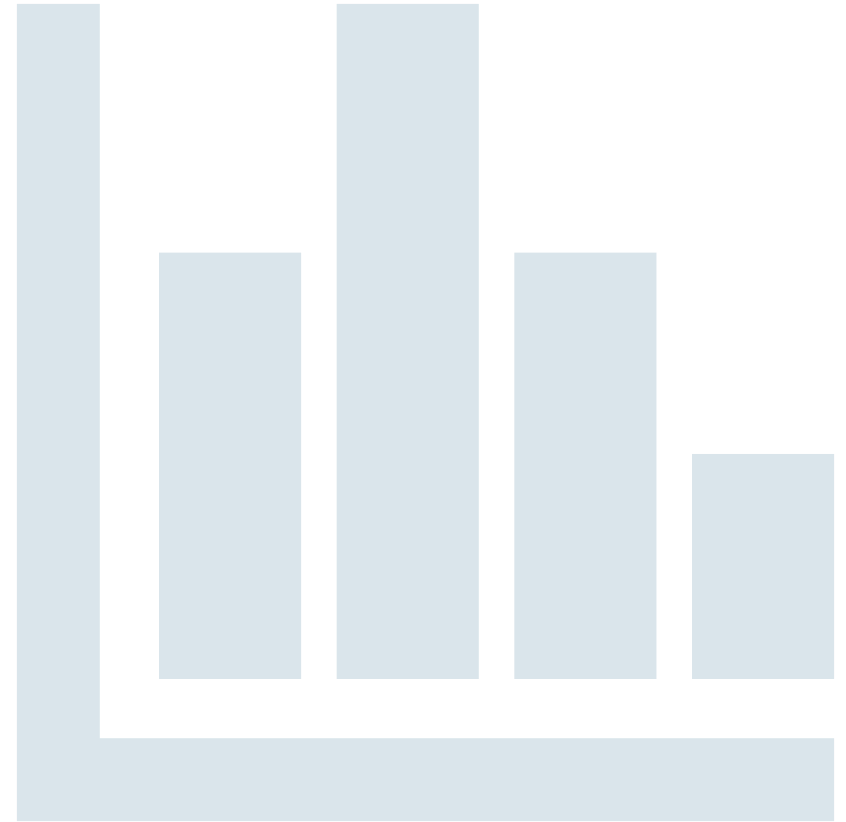


Big Data Analytics Project

Amazon Reviews Analysis:

iOS vs Android

By: Mohamad Alloush & Sam EL Saati

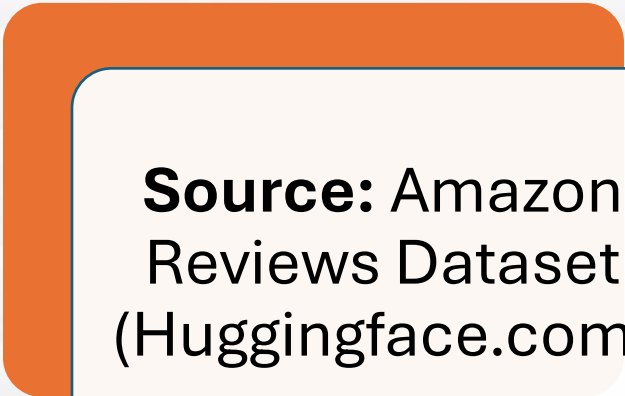




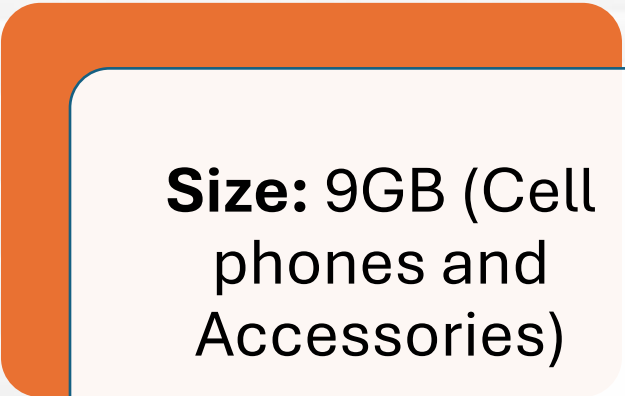
Introduction

Objective: Analyzing Amazon reviews for Android and iOS users; using big data techniques.

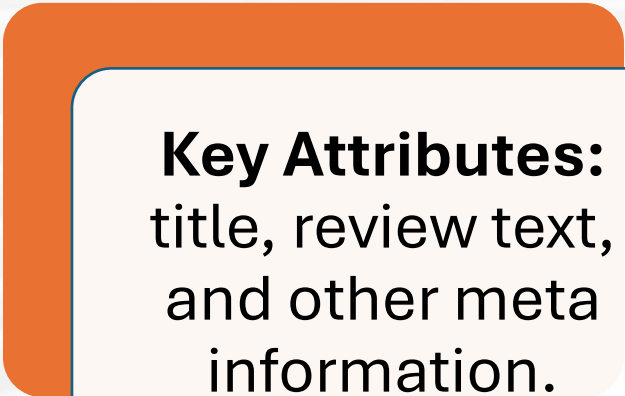
Dataset Overview



Source: Amazon
Reviews Dataset
(Huggingface.com)



Size: 9GB (Cell
phones and
Accessories)



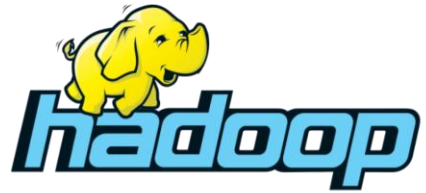
Key Attributes:
title, review text,
and other meta
information.

Sample of the Raw Dataset



	asin	rating	text	timestamp	verified_purchase	title
0	B08L6L3X1S	4.0	I bought this bc I thought it had the nice white background. Turns out it's clear & since my phone is blue it doesn't look anything like this. If I had known that I would have purchased something else. It works ok.	1612044451196	true	No white background! It's clear!
1	B0798PGF6C	5.0	Perfect. How pissed am I that I recently paid \$20 for 1 Fitbit cable and promptly lost the damned thing? Extremely pissed! I keep the spare in my medicine bag so hopefully I won't lose it and my grandson can't get to it and try to use it as a belt or a dog leash or any of the other nutty things he's been using the other one for.	1534443517349	true	Awesome! Great price! Works well!

Method

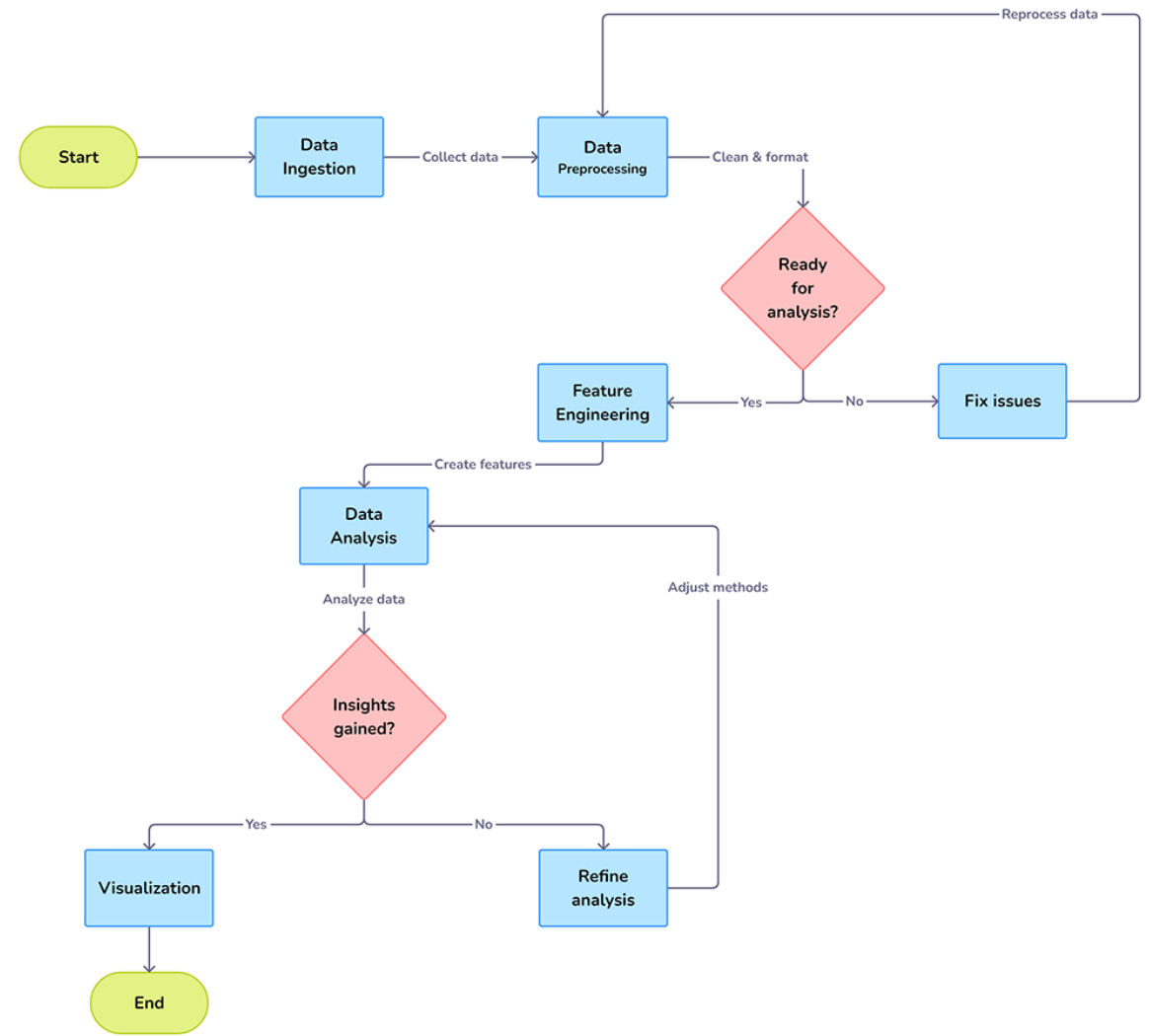


- Hadoop for storage and preprocessing.



- PySpark for analytics and visualization.

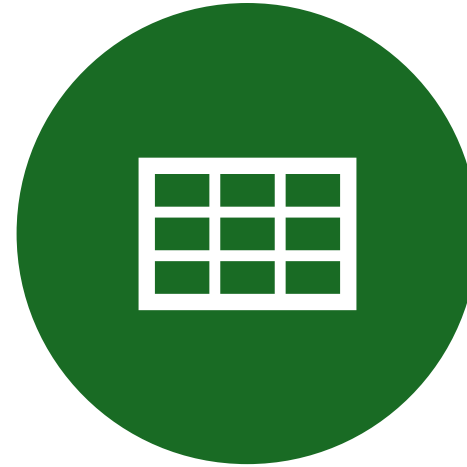
Workflow Diagram



Data Preprocessing



- CLEANING AND FEATURE ENGINEERING.



- TOKENIZING AND CATEGORIZING DATA FOR ANALYSIS.

Data Analysis

- User Behavior Insights
- Rating-based Sentiment Analysis
- Trend Analysis

Title	Text	Text (Cleaned)	Rating	Sentiment	Category
Worked but took a...	Overall very happ...	[overall, happy, ...]	5.0	Positive	iOS
Works Great with...	This item works g...	[item, works, gre...]	5.0	Positive	iOS
A bit complicated...	A bit complicated...	[bit, complicated...]	2.0	Negative	iOS
One Star	Fell apart right...	[fell, apart, rig...]	1.0	Negative	iOS

Results



Found & Analyzed the most used words at category level.

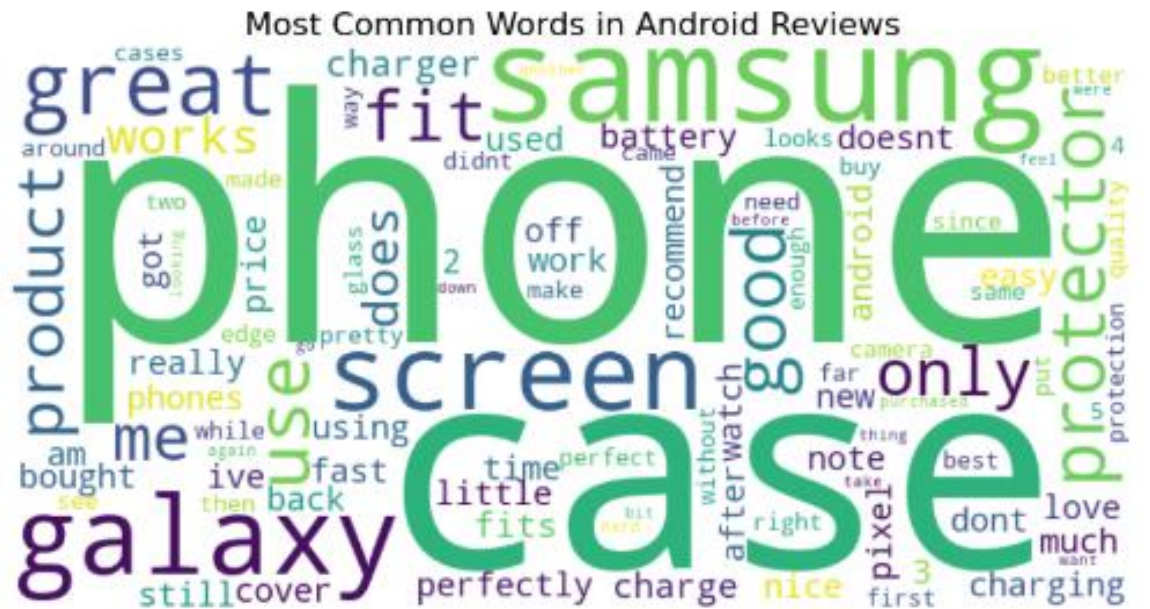


Studied the Distribution of Sentiments based on Ratings.



Explored the correlation between Ratings and time

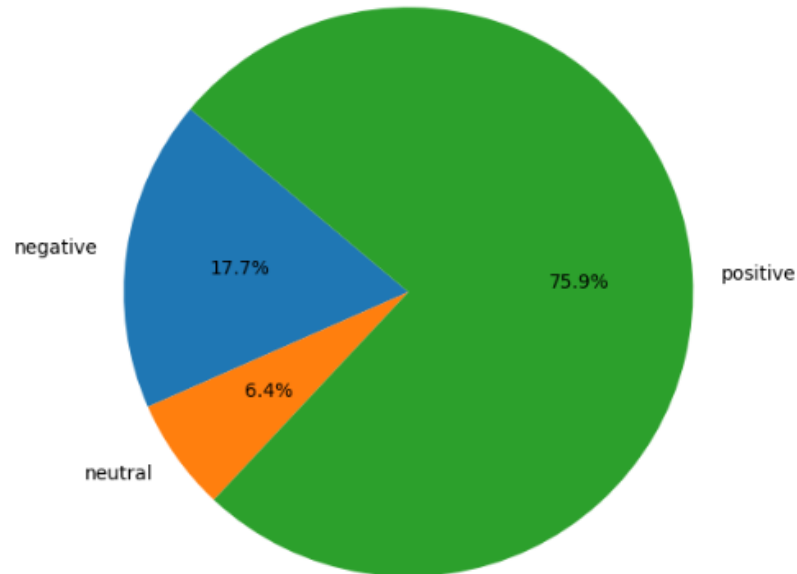
1



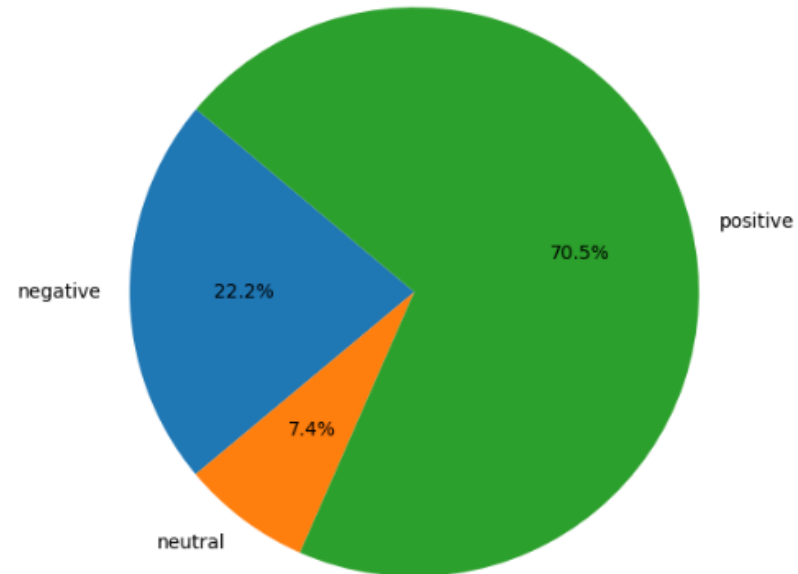
Visualization

- Rating-based sentiment analysis.

Sentiment Distribution - iOS

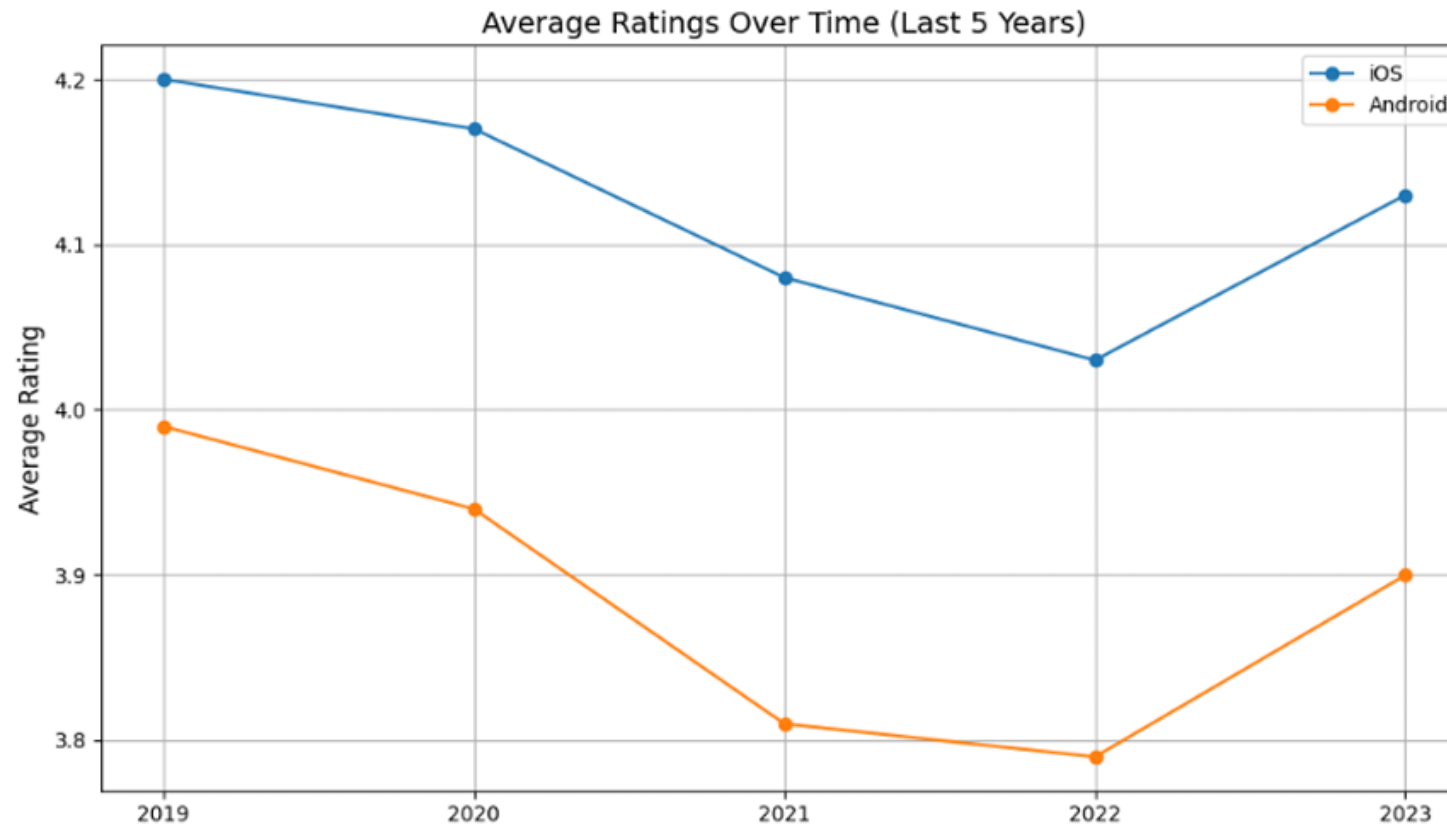


Sentiment Distribution - Android



Visualization

- Average Ratings Over Time.



Challenges



- Data volume and quality issues.



- System crashes and resource constraints.

Security Measures

Data access tracked using
HDFS audit logs.

```
if not defined HADOOP_SECURITY_LOGGER (  
    set HADOOP_SECURITY_LOGGER=INFO,RFAS  
)  
if not defined HDFS_AUDIT_LOGGER (  
    set HDFS_AUDIT_LOGGER=INFO,RFAAUDIT  
)
```

Conclusion

What went well?



Successfully analyzed a large dataset using big data tools (Hadoop and PySpark).



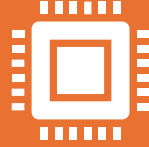
Produced insightful visualizations, such as word clouds and sentiment graphs.



Demonstrated clear distinctions between Android and iOS user priorities, supporting actionable insights for businesses.

Conclusion

What to improve?



Improve system resources to prevent crashes.



Use advanced sentiment tools like Vader or TextBlob with better resources.



Try new visualization methods for deeper insights and better audience engagement.

THE END

