Social Media Sentiment Analysis of Netflix

Sneha Jain 58648566

Aloiso Marques De Oliveira Junio 48215888

Xiaoshan Zhou 31740706

1. Introduction

Project Goal: To **analyze public sentiment** about **Netflix using data from Reddit.**

Business Value: This project will help Netflix **understand their users better** and help improve their marketing strategies.

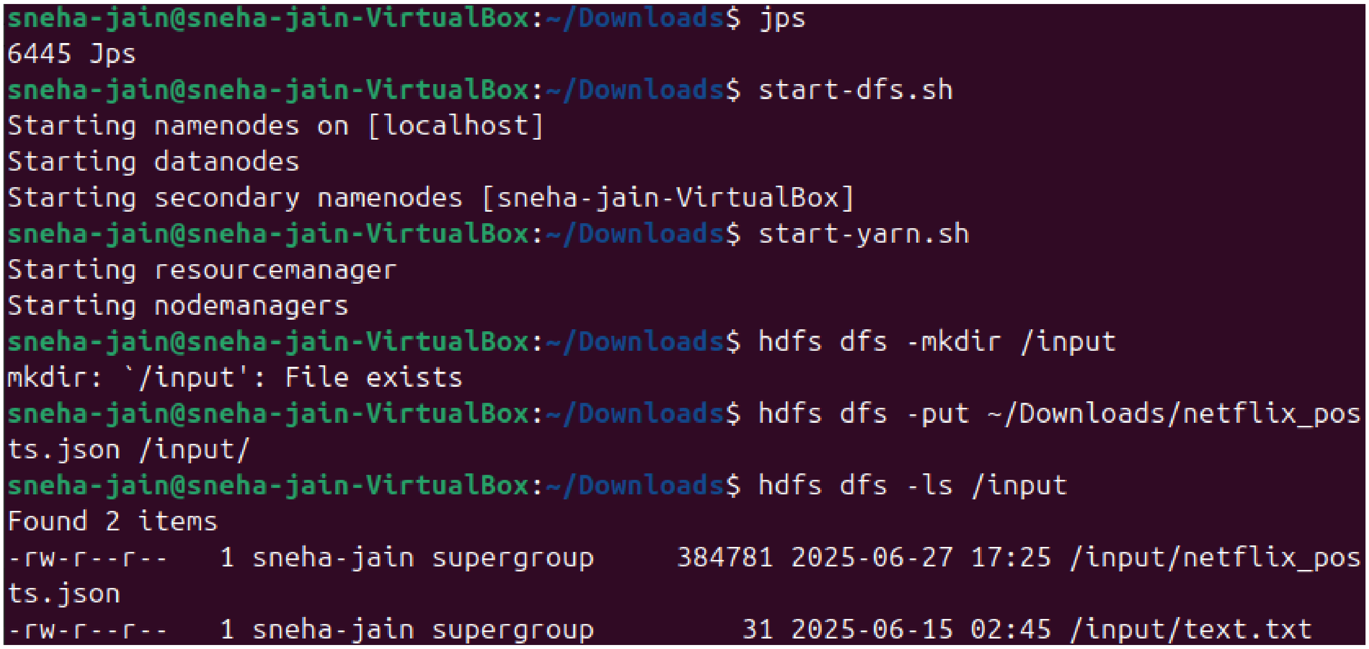
Methodology: The project involves **data collection via Reddit API**, storing the **data in Hadoop**, **cleaning** and processing using **Python** and MapReduce, **analyzing sentiment** with the **TextBlob** python library, and **visualizing** the results **using Tableau**.

2. Data Collection

APIs Used: Reddit API via the `praw` Python library.

Tools: Python, JSON files for storage, Reddit Developer App for authentication, and Hadoop HDFS for large-scale storage.

Data Acquisition Process: Posts containing the keyword 'Netflix' were collected from Reddit.



```
sneha-jain@sneha-jain-VirtualBox:~/Downloads$ jps
6445 Jps
sneha-jain@sneha-jain-VirtualBox:~/Downloads$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [sneha-jain-VirtualBox]
sneha-jain@sneha-jain-VirtualBox:~/Downloads$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
sneha-jain@sneha-jain-VirtualBox:~/Downloads$ hdfs dfs -mkdir /input
mkdir: `/input': File exists
sneha-jain@sneha-jain-VirtualBox:~/Downloads$ hdfs dfs -put ~/Downloads/netflix_pos
ts.json /input/
sneha-jain@sneha-jain-VirtualBox:~/Downloads$ hdfs dfs -ls /input
Found 2 items
-rw-r--r--   1 sneha-jain supergroup     384781 2025-06-27 17:25 /input/netflix_pos
ts.json
-rw-r--r--   1 sneha-jain supergroup         31 2025-06-15 02:45 /input/text.txt
```

*Data was collected and stored in Hadoop as shown above.

3. Data Cleaning

Cleaning Steps:

1. Removed Noise (URLs, stop words, emojis etc.)
2. Converted text to lowercase
3. Normalized white spaces

Challenges:

- There were words beginning with Hadoop, which might have malfunctioned during the data cleaning process.

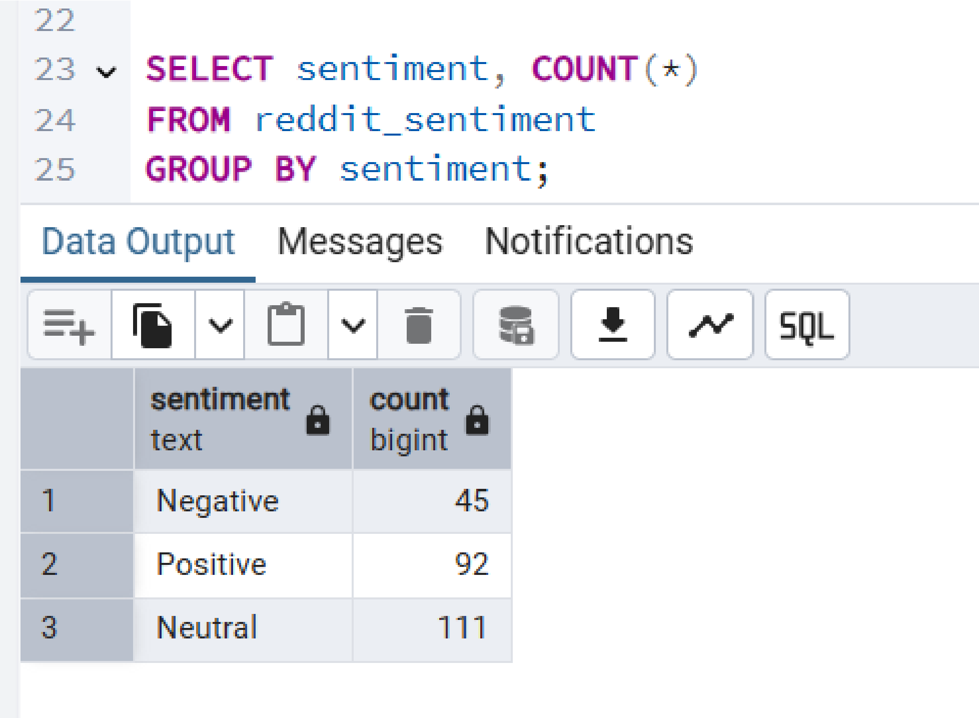
4. Data Analysis

1. **Transfer Data to PostgreSQL**
   The cleaned sentiment data, **processed using Python (Netflix_sentiment.py)**, was exported to a CSV file. This file was then **imported into a PostgreSQL** database using the pgAdmin 4. A table named **reddit_sentiment** was created with appropriate columns (**date, text, score, sentiment**) and the data was inserted successfully.

   This allowed SQL-based querying and enabled further analysis through structured queries**.**

2. **SQL Queries:**

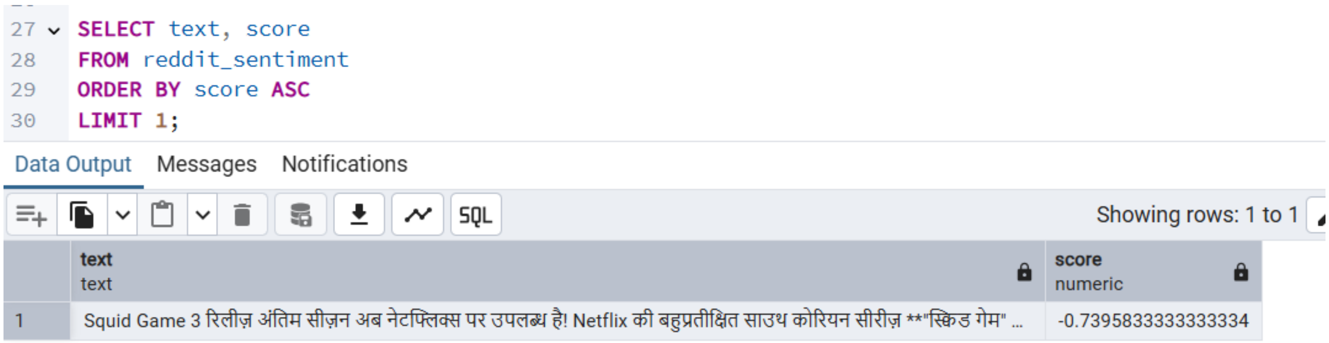
- Count of posts grouped by sentiment



```sql
SELECT sentiment, COUNT(*)
FROM reddit_sentiment
GROUP BY sentiment;
```

	sentiment text	count bigint
1	Negative	45
2	Positive	92
3	Neutral	111

- Most Negative Comment along with its score

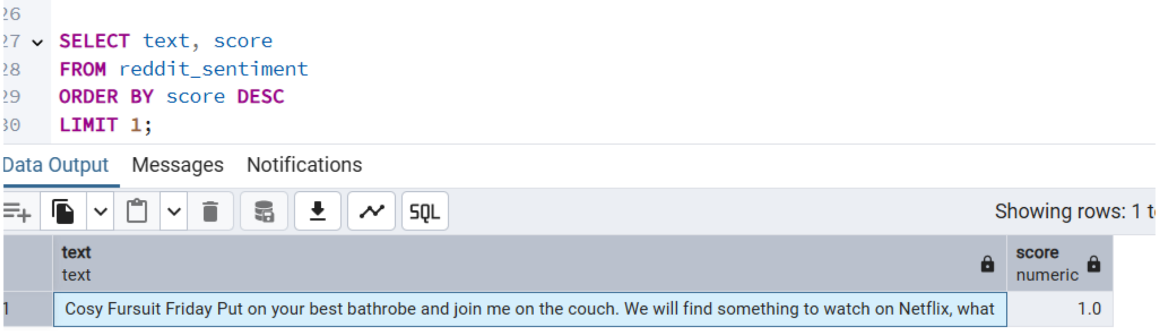


```
SELECT text, score
FROM reddit_sentiment
ORDER BY score ASC
LIMIT 1;
```

Showing rows: 1 to 1

	text text	score numeric
1	Squid Game 3 रिलीज़ अंतिम सीज़न अब नेटफ्लिक्स पर उपलब्ध है! Netflix की बहुप्रतीक्षित साउथ कोरियन सीरीज़ **"स्क्विड गेम" ...	-0.7395833333333334

- Most Positive Comment



```
SELECT text, score
FROM reddit_sentiment
ORDER BY score DESC
LIMIT 1;
```

Showing rows: 1 t

	text text	score numeric
1	Cosy Fursuit Friday Put on your best bathrobe and join me on the couch. We will find something to watch on Netflix, what	1.0

3. **Sentiment Analysis**:

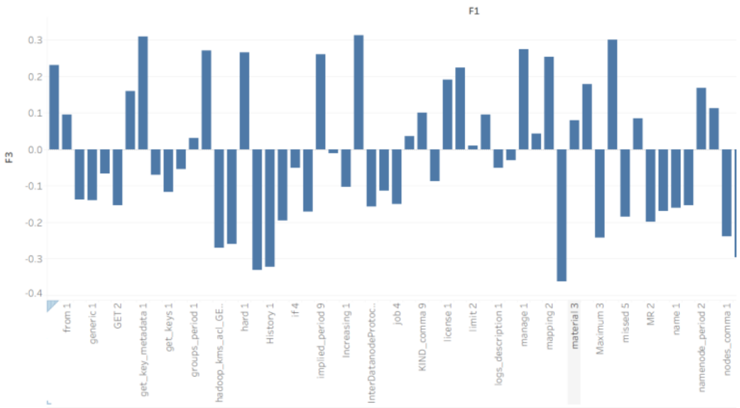
- TextBlob libraries was used to assign sentiment scores to each post. Posts were classified into three categories:
  - Positive (score > 0.1)
  - Neutral (between -0.1 and 0.1)
  - Negative (score < -0.1)

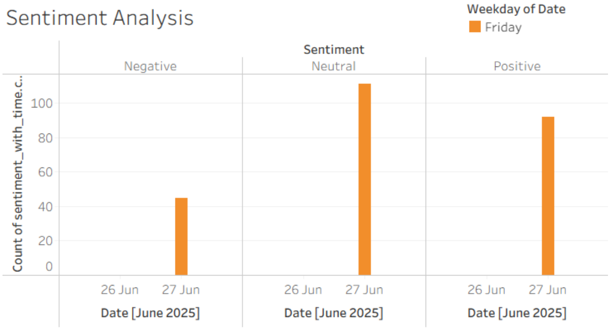
- Certain keywords were also extracted and they were given sentiment scores as well.

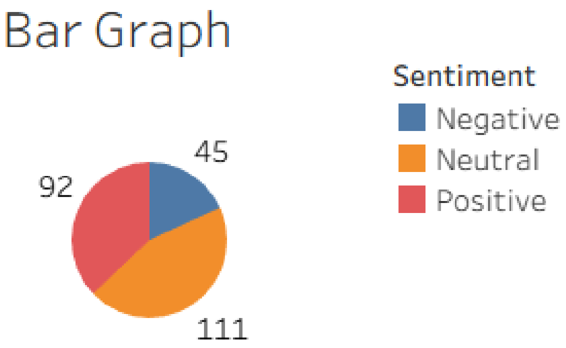
5. Visualizations

The final processed data was exported to a **CSV file**, which is a suitable format for **Tableau.**

Tableau Screenshots:







*Bar graph on the left shows words and their sentiment scores.

*Bar graph on the right shows the date and the number of neutral/positive/negative posts

*Pie Chart shows the number of posts grouped by sentiment.

6. Conclusion

Conclusion: From the pie chart and SQL query we can see that - most Reddit users had **neutral views** towards Netflix and the shows.

From the **word count and corresponding sentiment score**, we can observe:

Significant people showed excitement towards new seasons and some people also were disappointed by some new releases.

Recommendations:

The users don't seem to have any major complaints regrading the platform but to increase their users, Netflix can:

- Improve **Pricing strategies.**
- Get feedback from users.
- Improve **content quality.**