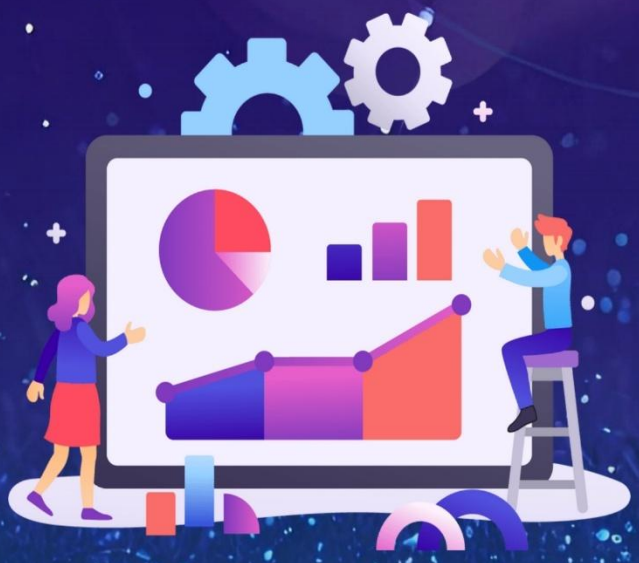
Prodigy InfoTech Internship



Data Science

Note:
It's not compulsory to use the provided dataset

Track: DS

Name: Suriya. B
Linkdin: www.linkedin.com/in/suriya0210
Email: bsuriya223@gmail.com

PRODIGY INFOTECH TASK-4

Analyse and visualize sentiment patterns in social media data to understand public opinion and attitudes towards specific topics or brands

Abstract

This project aims to analyse and visualize sentiment patterns in social media data to understand public opinion and attitudes toward specific topics and brands. By leveraging tweets from a publicly available dataset, we performed comprehensive data cleaning, exploratory data analysis (EDA), and advanced visualization techniques. The project highlights the proportion of positive, negative, neutral, and irrelevant sentiments and investigates the most common keywords associated with each sentiment. Using visualizations like bar plots, stacked charts, and word clouds, businesses and researchers can gain actionable insights into public perceptions, enabling better decision-making and brand management.

Introduction

Explain the context and relevance of sentiment analysis.

- **What is Sentiment Analysis?**
    - Process of identifying emotions and opinions in text data.
    - Widely used for brand monitoring, marketing, and customer experience improvement.
- **Why Social Media?**
    - Social media platforms generate massive volumes of user opinions.
    - Useful for real-time tracking of public attitudes.

- **Project Goal:**
    - To analyse Twitter data and determine public sentiment towards different topics/brands.

Objectives

Clearly list the project objectives:

1. To clean and preprocess social media text data.
2. To perform exploratory data analysis (EDA) for understanding patterns.
3. To visualize sentiment distribution across topics and brands.
4. To identify the most frequently used positive, negative, and neutral keywords.
5. To generate insights for brand perception and decision-making.

Literature Review

Provide a short background of previous works:

- Studies showing how companies like Amazon, Flipkart, or Netflix use sentiment analysis for customer feedback.
- Example: Research papers on **Natural Language Processing (NLP)** applications in marketing and social media analytics.

Dataset Description

Describe your dataset thoroughly.

- **Dataset Name**: Twitter Training Data (twitter_training.csv)
- **Source**: Open-source dataset (e.g., Kaggle or other repositories).
- **Size**: 74,681 records and 4 columns.

Column Name	Description
tweet_id	Unique ID for each tweet
topic	Brand or topic mentioned in the tweet
sentiment	Label sentiment (Positive/Negative/Neutral/Irrelevant)
tweet_text	The actual tweet text

Methodology

Step 1: Importing Libraries and Data

- Python libraries used:
  pandas, matplotlib, seaborn, word cloud, re, collections.

Step 2: Data Cleaning

- Handle missing values.
- Remove:
  - URLs
  - Hashtags
  - Mentions
  - Non-alphabetic characters
- Create a cleaned_text column for processed tweets.

Step 3: Exploratory Data Analysis (EDA)

EDA Tasks:

1. **Sentiment Distribution**

- Visualize counts of Positive, Negative, Neutral, and Irrelevant tweets.

2. **Tweet Length Analysis**
    - Identify whether long or short tweets are more common.
3. **Sentiment by Topic/Brand**
    - Use stacked bar charts to compare sentiments across different topics.
4. **Word Clouds**
    - Visualize common words for each sentiment type.

Step 4: Sentiment Pattern Analysis

Goal:
Analyse public opinion about specific topics and brands.

Techniques:

1. **Word Clouds for Each Sentiment**
    - Highlight the most discussed terms in positive, negative, and neutral tweets.
2. **Top Keywords**
    - Use frequency counts to identify top 10 keywords per sentiment.
3. **Sentiment Ratio by Brand**
    - Show which brands are viewed positively or negatively.

Results and Visualizations

Include key graphs and explain them:

1. **Sentiment Distribution**

- *Finding:* Negative tweets slightly dominate the dataset.

2. **Stacked Bar by Topic**
    - *Example:*
        - Brand A: 70% Positive, 20% Neutral, 10% Negative.
        - Brand B: Higher negative sentiment → potential issue.
3. **Word Clouds**
    - Positive words: "love", "great", "amazing".
    - Negative words: "hate", "bad", "worst".
4. **Top Keywords**
    - Useful for identifying trending concerns or praises.

Insights

Business-level interpretation of results:

1. Negative sentiment is higher for certain brands, indicating dissatisfaction.
2. Positive sentiment keywords focus on product quality and user experience.
3. Word clouds reveal trending topics, helping brands tailor campaigns.
4. Topics with balanced sentiment indicate mixed customer feedback.

Conclusion

- The project successfully identified patterns in social media sentiments.
- Businesses can leverage these insights to:

- Improve customer engagement.
- Address negative feedback proactively.
- Track changes in brand perception over time.

Future Scope

1. Implement **machine learning models** to predict sentiment automatically.
2. Integrate **real-time Twitter API** for live data monitoring.
3. Use **deep learning models** like BERT for advanced NLP tasks.
4. Expand the dataset to multiple languages and regions.

References

- Research papers and articles on sentiment analysis.
- Python documentation for Pandas, Seaborn, and WordCloud.
- Kaggle datasets related to Twitter sentiment analysis.

Code and results:

Import libraries and load data:

```
# ============================================================
# Sentiment Analysis of Social Media Data
# ============================================================

# ========== Step 1. Import Libraries and Load Data ==========
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud, STOPWORDS
import re
from collections import Counter

# Load the dataset
file_path ="C:\\Users\\SURIYA\\Downloads\\twitter_training.csv"  # Make sure CSV is in same folder
df = pd.read_csv(file_path)

# Rename columns for clarity
df.columns = ["tweet_id", "topic", "sentiment", "tweet_text"]

print("===== Dataset Loaded Successfully =====")
print("Shape of dataset:", df.shape)
print("\nFirst 5 rows of the dataset:\n", df.head())
```

df	DataFrame	[73995, 6]	Column names: tweet_id, topic, sentiment, tweet_text, cleaned_text, tw ...

Data Cleaning and Understanding:

```python
# ========== Step 2. Data Cleaning and Understanding ==========
# Check data info
print("\n===== Dataset Information =====")
print(df.info())

# Check for missing values
print("\nMissing values in each column:\n", df.isnull().sum())

# Drop rows where tweet_text is missing
df.dropna(subset=['tweet_text'], inplace=True)

# Clean tweet text: remove URLs, hashtags, mentions, and non-alphabetic characters
def clean_text(text):
    text = re.sub(r"http\S+|www\S+|https\S+", '', text)  # Remove URLs
    text = re.sub(r"@\w+|#\w+", '', text)                # Remove mentions & hashtags
    text = re.sub(r"[^A-Za-z\s]", '', text)              # Keep only letters
    text = text.lower().strip()
    return text

df['cleaned_text'] = df['tweet_text'].apply(clean_text)

# Verify cleaning
print("\n===== Cleaned Data Sample =====")
print(df[['tweet_text', 'cleaned_text']].head())

# Check for missing values
print("\nMissing values in each column:\n", df.isnull().sum())
```

```
===== Dataset Information =====
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 74681 entries, 0 to 74680
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   tweet_id    74681 non-null  int64
 1   topic       74681 non-null  object
 2   sentiment   74681 non-null  object
 3   tweet_text  73995 non-null  object
dtypes: int64(1), object(3)
memory usage: 2.3+ MB
None
```

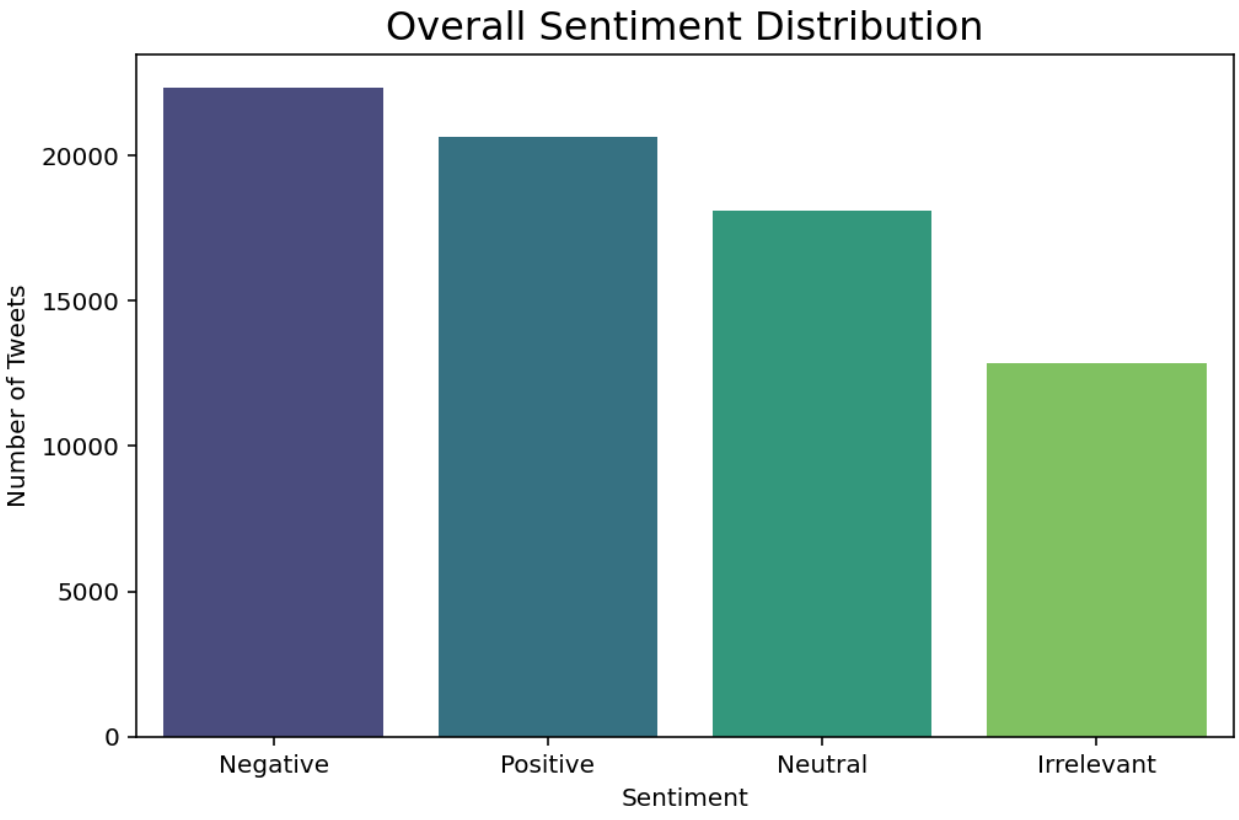
```
In [6]:
   ...: print("\nMissing values in each column:\n", df.isnull().sum())
   ...:

Missing values in each column:
 tweet_id        0
topic            0
sentiment        0
tweet_text     686
dtype: int64
```

```
In [9]:
   ...: print("\n===== Cleaned Data Sample =====")
   ...: print(df[['tweet_text', 'cleaned_text']].head())

===== Cleaned Data Sample =====
                                          tweet_text                                       cleaned_text
0  I am coming to the borders and I will kill you...  i am coming to the borders and i will kill you...
1  im getting on borderlands and i will kill you ...  im getting on borderlands and i will kill you all
2  im coming on borderlands and i will murder you...  im coming on borderlands and i will murder you...
3  im getting on borderlands 2 and i will murder ...  im getting on borderlands  and i will murder y...
4  im getting into borderlands and i can murder y...  im getting into borderlands and i can murder y...
```

```
In [10]:
    ...: print("\nMissing values in each column:\n", df.isnull().sum())

Missing values in each column:
 tweet_id        0
topic            0
sentiment        0
tweet_text       0
cleaned_text     0
dtype: int64
```

Full EDA and Visualizations:

Sentiment distribution



```
    ...:
    ...: sentiment_counts = df['sentiment'].value_counts()
    ...: print("\n===== Sentiment Distribution =====")
    ...: print(sentiment_counts)
    ...:
    ...: plt.figure(figsize=(8, 5))
    ...: sns.barplot(x=sentiment_counts.index, y=sentiment_counts.values, palette="viridis")
    ...: plt.title("Overall Sentiment Distribution", fontsize=16)
    ...: plt.xlabel("Sentiment")
    ...: plt.ylabel("Number of Tweets")
    ...: plt.show()

===== Sentiment Distribution =====
sentiment
Negative      22358
Positive      20654
Neutral       18108
Irrelevant    12875
Name: count, dtype: int64
```

Interpretation:

1. Negative tweets are highest (\~22,000), showing a strong presence of dissatisfaction or criticism.

2. Positive tweets are slightly lower (\~20,000), indicating a good amount of appreciation and support.

3. Negative sentiment dominates overall, highlighting potential issues to be addressed.

4. Neutral tweets are moderate (\~18,000), reflecting factual or emotionless discussions.

5. Irrelevant tweets are the least (\~13,000) and can be filtered for focused analysis.

6. The gap between negative and positive tweets is small but significant.

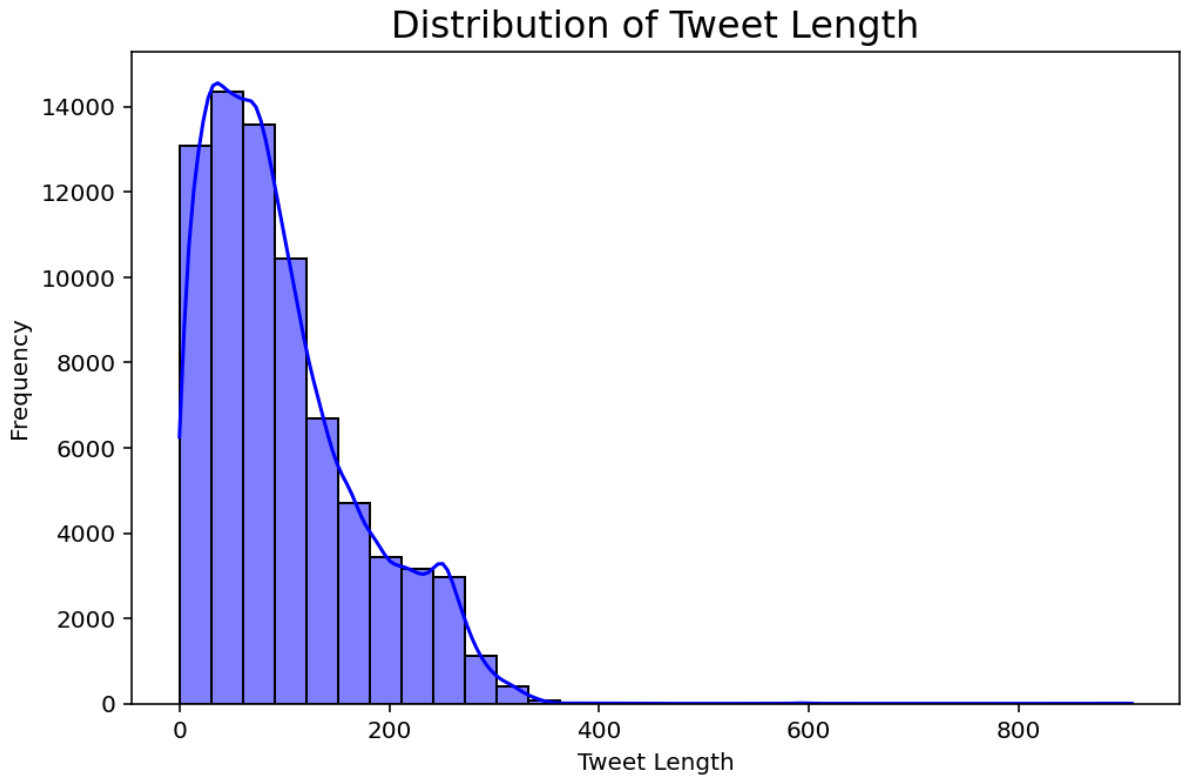
7. Negative feedback may indicate brand or product dissatisfaction.

8. Positive tweets show there is still a strong base of satisfied users.

9. Neutral tweets can be monitored to predict future sentiment shifts.

10. Companies should reduce negatives, boost positives, and track trends for better brand perception.

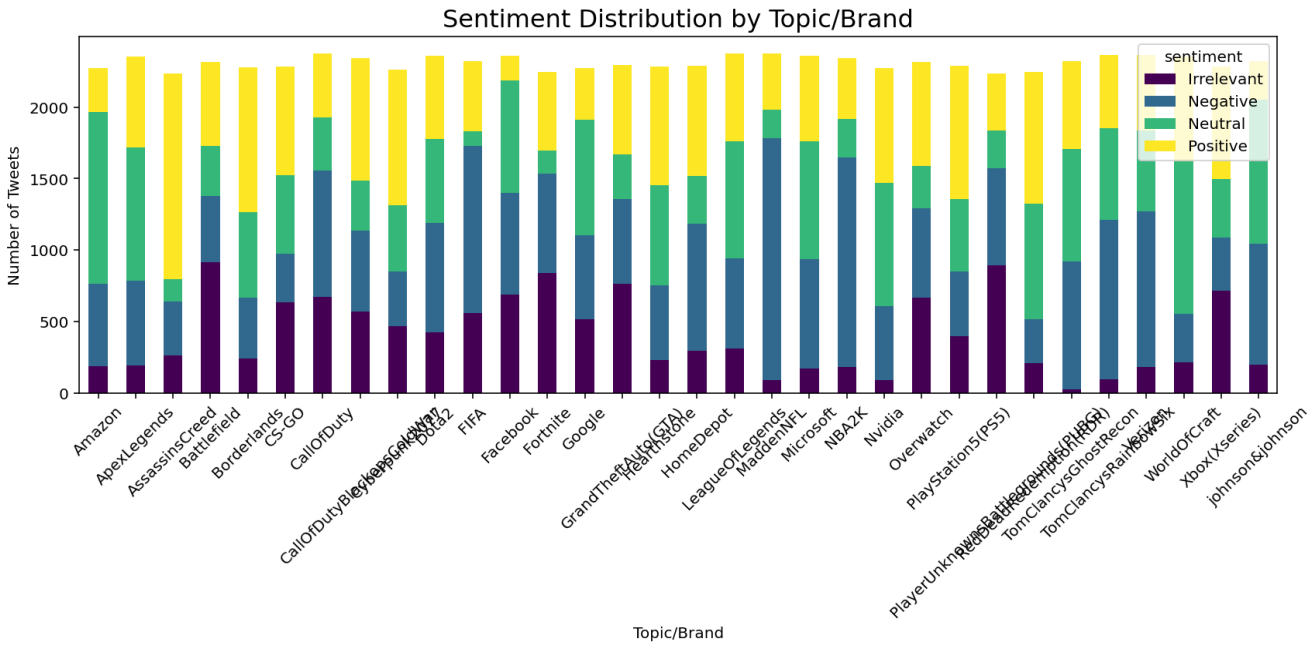
Tweet length analysis:



Interpretation:

1. Most tweets are short, with the highest frequency between 20–60 characters, showing concise user communication.
2. The distribution is right-skewed, meaning very few tweets are extremely long.
3. Tweets above 200 characters are rare, indicating users prefer brevity.
4. The peak suggests that short and impactful tweets dominate social media conversations.
5. Businesses should focus on clear and concise messaging to match common user behaviour.

Sentiment by Topic/Brand:



Interpretation:

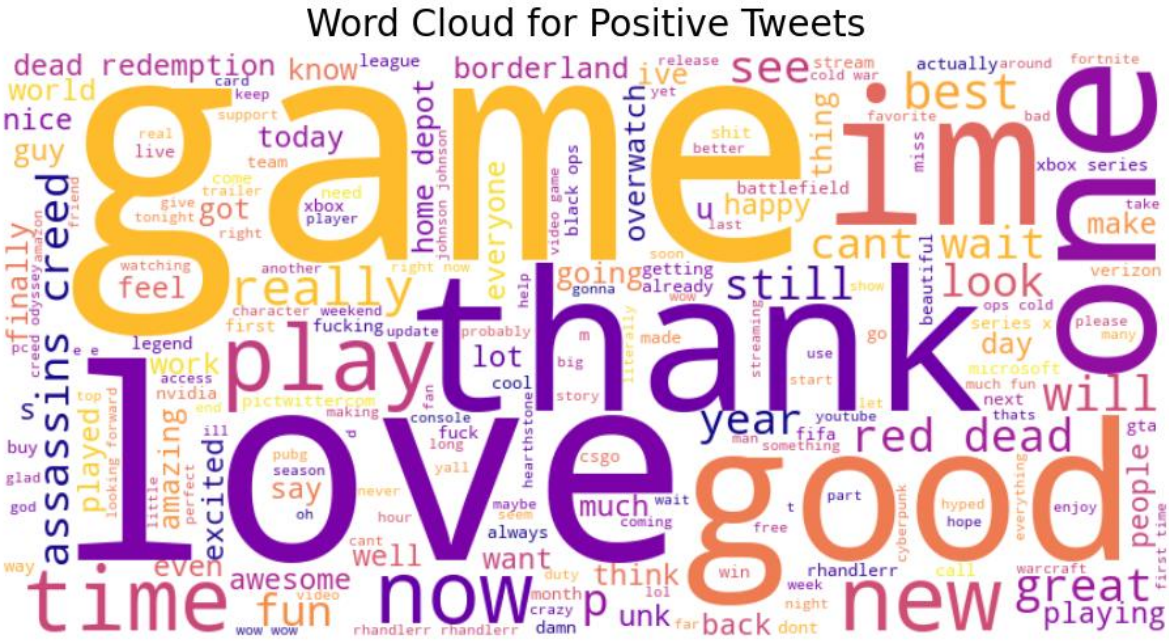
1. Most brands show a mix of positive, negative, neutral, and irrelevant sentiments, reflecting diverse public opinions.
2. Amazon, Apex Legends, and Assassin's Creed have higher positive sentiment, showing strong customer satisfaction, while Call of Duty, FIFA, and Fortnite face higher negative sentiment, indicating dissatisfaction.
3. Neutral tweets dominate for tech companies like Google and Microsoft, suggesting many factual or news-based discussions.
4. Gaming brands such as Overwatch, League of Legends, and NBA2K have significant negative sentiment, while some like Tom Clancy’s Ghost Recon show mostly neutral tweets.
5. These insights help businesses boost positive engagement, resolve negative feedback, and monitor brand reputation effectively.

Word Cloud for all tweets:

Interpretation:

1. The most prominent word is "game", indicating that gaming is a central theme of discussions in the dataset.
2. Positive words like "love", "thank", and "good" are highly frequent, reflecting appreciation and satisfaction among users.
3. Negative expressions such as "fuck", "shit", and "fucking" appear prominently, showing frustration or criticism.
4. Words like "play", "new", and "time" suggest active engagement and excitement for new releases or updates.
5. Brand mentions like "Amazon", "Xbox", "Facebook", and "Overwatch" highlight popular platforms or companies being discussed.

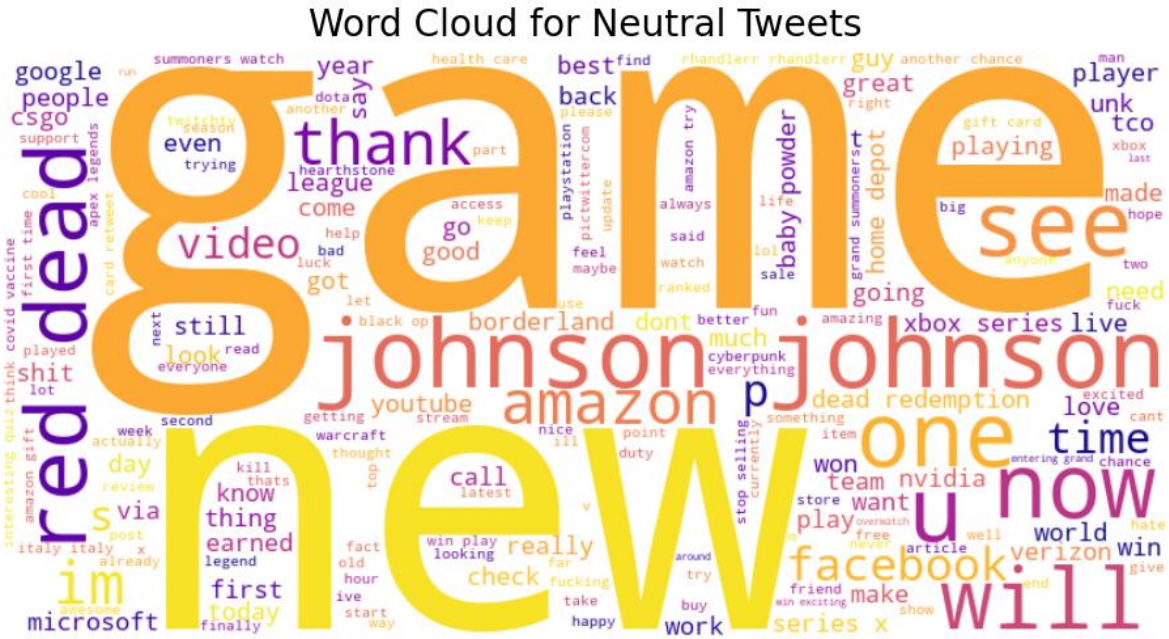
Word Cloud for Positive Tweets:



Interpretation:

1. The words "game," "love," "time," and "thank" are the most prominent, suggesting that the tweets are likely about a positive gaming experience.
2. Words like "now," "new," and "still" indicate a focus on current or ongoing events related to the games.
3. Specific game titles such as "red dead redemption," "overwatch," and "creed" are visible, pinpointing the games being discussed.
4. The presence of words like "good," "great," and "best" directly conveys positive sentiment.
5. Words related to social interaction, like "watching" and "people," suggest that gaming is also a social experience for these users.

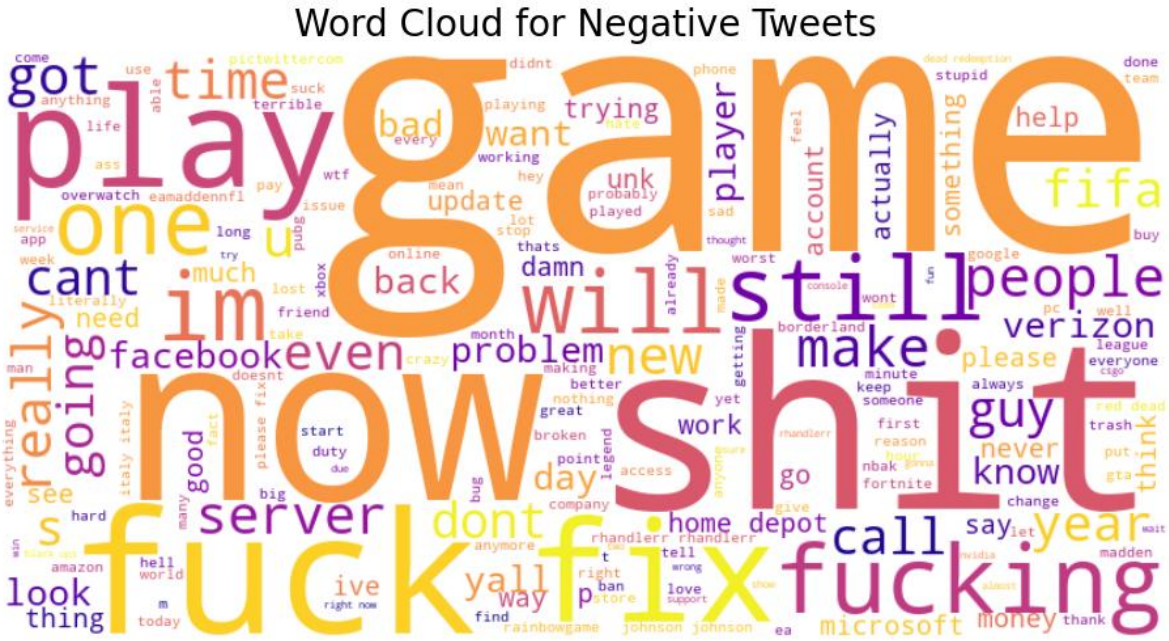
Word Cloud for Neutral Tweets:



Interpretation:

1. The most prominent words are **"game," "johnson,"** and **"new."** The large size of "game" suggests the tweets are likely about video games. The repeated "johnson" is a bit unusual, but could refer to a name, a specific product, or perhaps be an artifact of data processing. "New" indicates updates or releases are a common topic.
2. Other large words like **"see," "video,"** and **"thank"** suggest a focus on observation, media content, and possibly acknowledging something.
3. Words like **"try," "back," "one,"** and **"time"** are generic, and their neutrality is consistent with the overall tone.
4. Specific game titles such as **"dead redemption," "overwatch,"** and **"hearthstone"** are also present, indicating which games are being discussed.

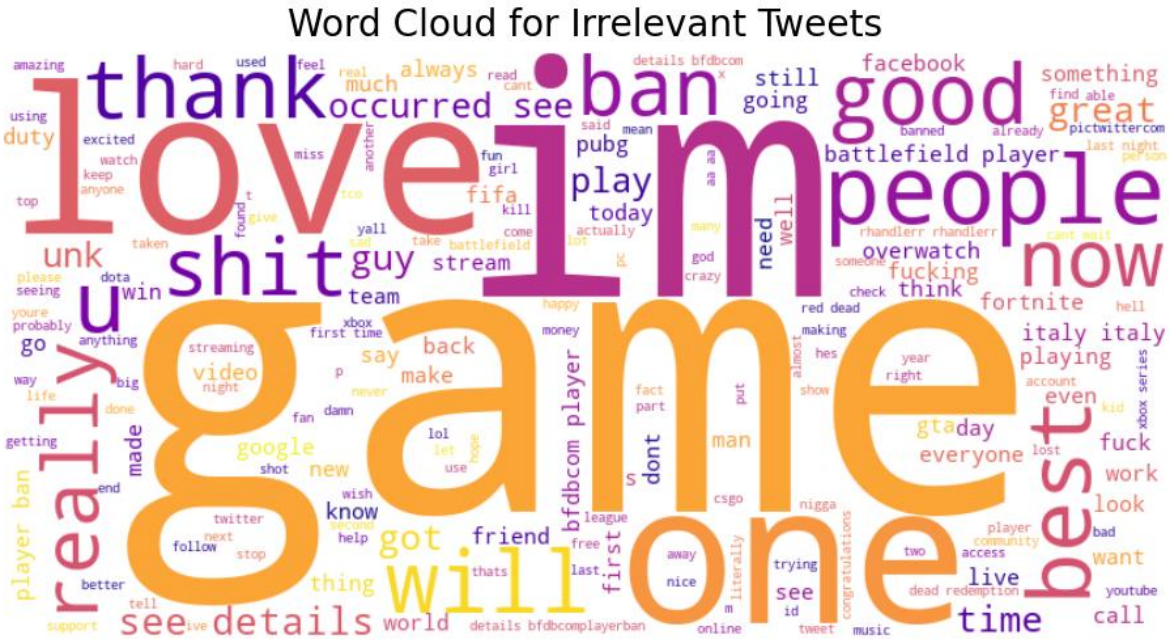
Word Cloud for Negative Tweets:



Interpretation:

1. The words "fuck," "shit," and "fucking" are the most prominent, immediately highlighting the strong negative sentiment.
2. "Game" is another large word, confirming that the negative sentiment is directed at a gaming experience.
3. Words like "server," "play," "player," and "back" point to specific issues, likely related to server stability, online gameplay, or a desire for a return to a previous, better state.
4. The presence of "call" and "people" suggests frustration with customer support or interactions with other players.
5. Other words like "problem," "stupid," and "don’t" further reinforce the negative and critical tone of the tweets.

Word Cloud for Irrelevant Tweets:

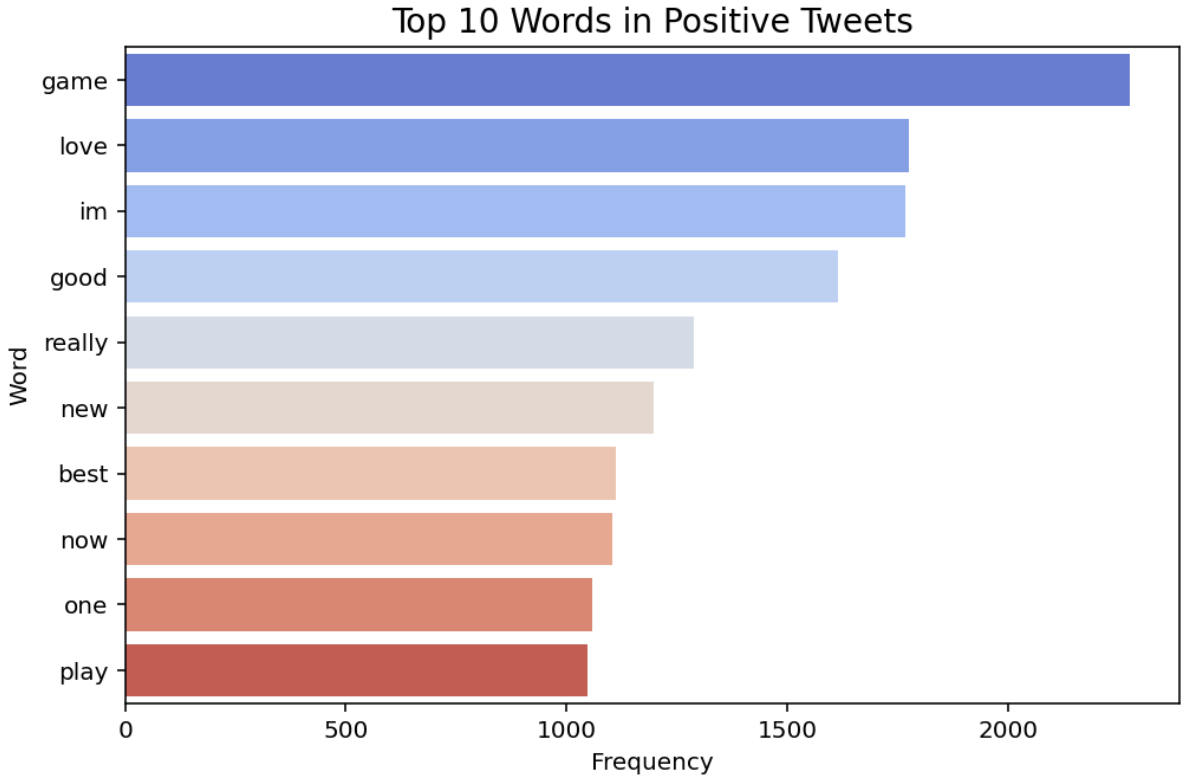


Interpretation:

1. The most prominent word is "game," which is not surprising, as it's a common term in gaming-related discussions. However, its large size indicates a high frequency of general game-related conversation.
2. Words like "thank," "love," and "good" are typically positive, while words like "shit" and "ban" are negative. The presence of both positive and negative terms could be why these tweets were classified as "irrelevant"—they might contain mixed emotions or sarcasm that is hard for an algorithm to interpret.
3. Words like "people," "now," "play," and "time" are generic and can be used in many contexts, making them less useful for sentiment analysis.
4. The presence of specific game titles such as "fortnite," "fifa," and "dota" anchors the discussion to a gaming context, but the overall mix of positive and negative words makes the sentiment ambiguous.

Top keywords for each sentiment:

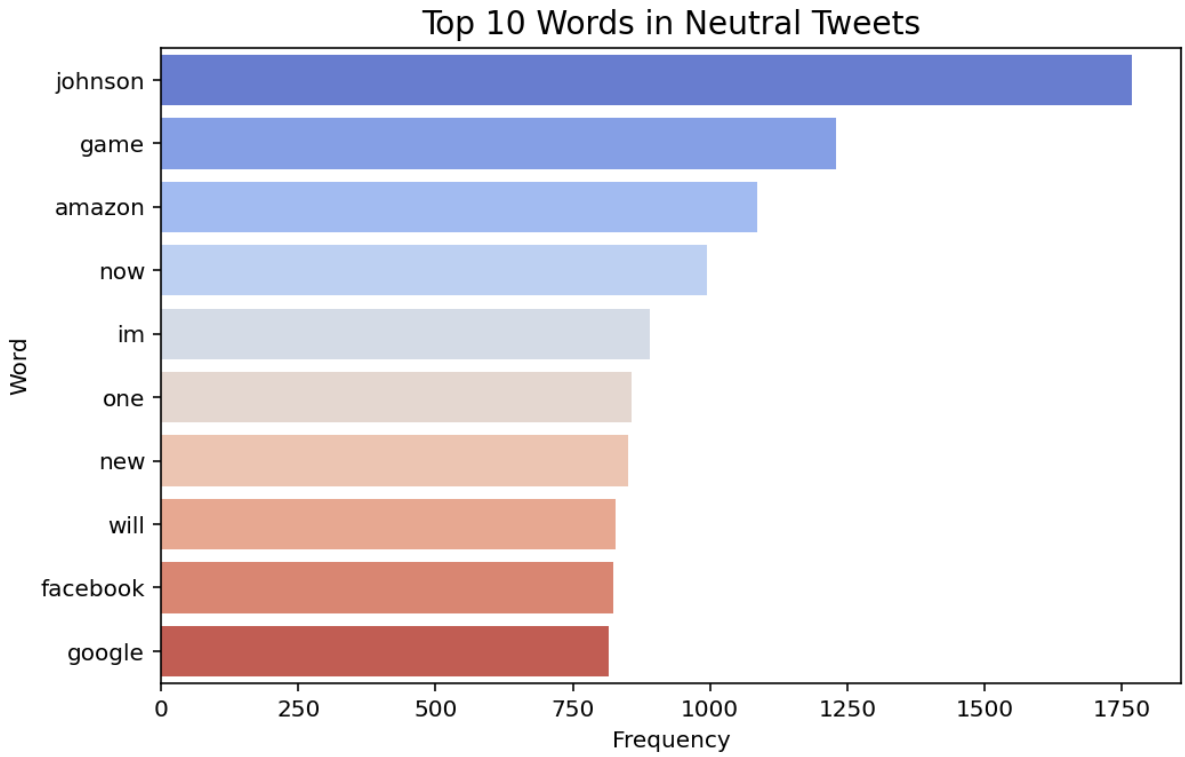
Top 10 Words in Positive Tweets:



Interpretation:

1. "Game" is the most frequent word, with a frequency of over 2,000, indicating that the tweets are predominantly about a gaming topic.
2. "Love" and "im" are the next most common words, with frequencies around 1,800. This directly reinforces the positive sentiment.
3. The words "good," "really," "new," and "best" all have high frequencies, ranging from just over 1,000 to around 1,600. These words are all positive or serve to intensify a positive statement.
4. The words "now," "one," and "play" also appear in the top 10, with frequencies just over 1,000. While not inherently positive, their presence in this context likely relates to an enjoyable, current, or recent gaming experience.

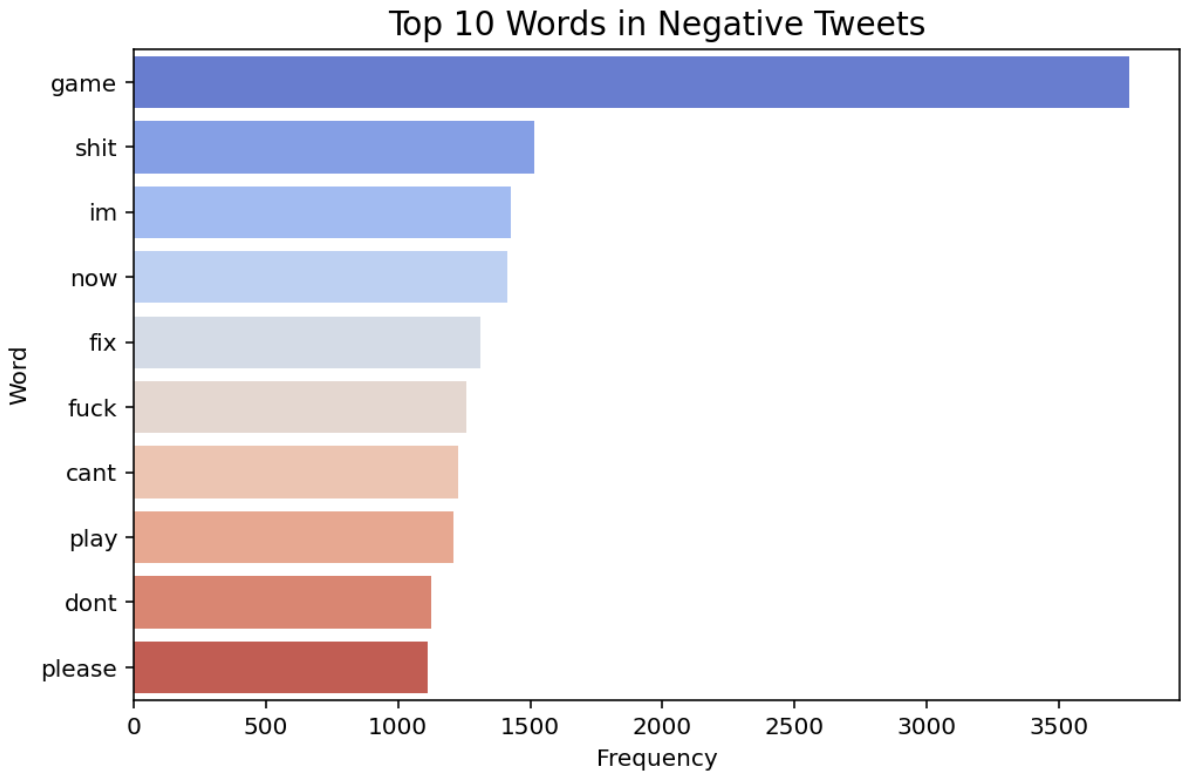
Top 10 Words in Neutral Tweets:



Interpretation:

1. "Johnson" is the most frequent word, with a frequency of over 1,750. This is unusual and suggests it could be a key entity or name that frequently appears in discussions, but without a clear positive or negative context. It may refer to a person, a product, or even an artifact of data processing.
2. "Game" is the second most frequent, with a frequency of around 1,250. Its high ranking confirms the topic is gaming-related, but its context within neutral tweets suggests it's mentioned in a descriptive, non-emotional way.
3. "Amazon," "now," and "im" are also common, with frequencies ranging from 900 to 1,100. These words are all generally neutral.
4. Words like "one," "new," "will," "facebook," and "google" also appear in the top 10. Their presence suggests these tweets often discuss new releases, future plans, or mention specific companies and platforms without expressing strong opinions.

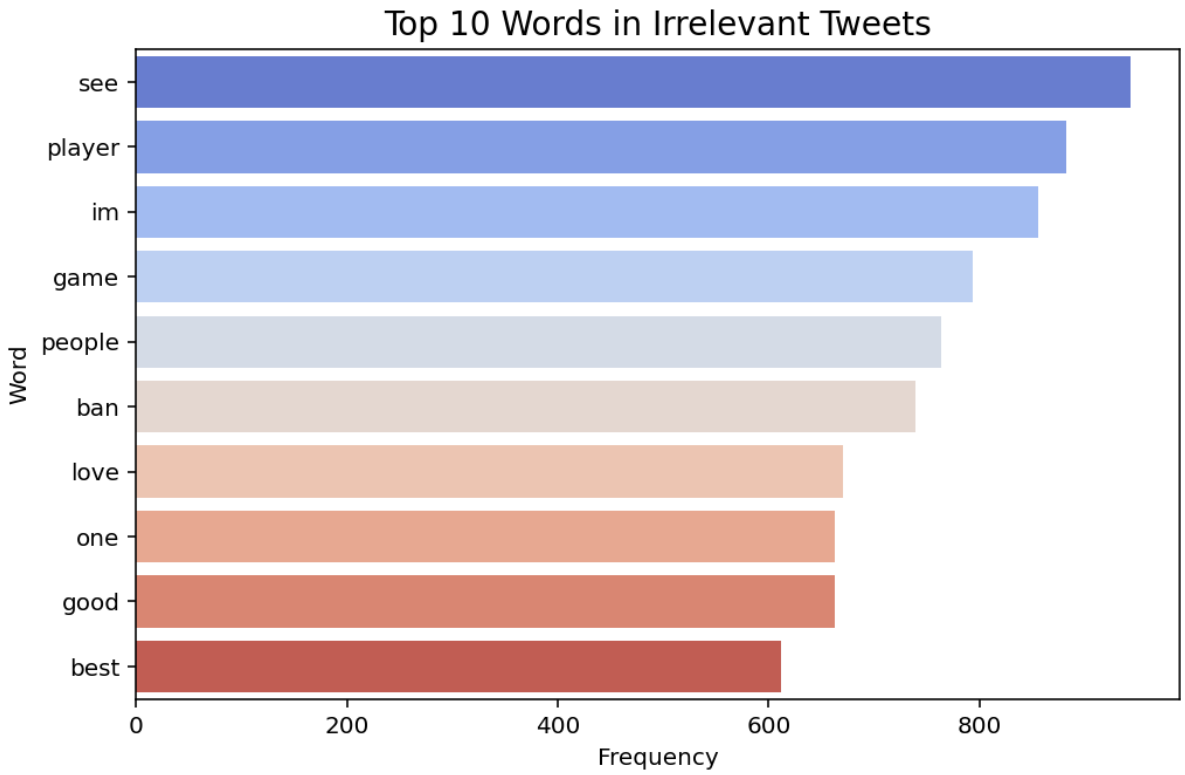
Top 10 Words in Negative Tweets:



Interpretation:

1. The word "game" is by far the most frequent, with a frequency of around 3,750, indicating that the negative sentiment is strongly tied to a gaming topic.
2. "Shit" is the second most frequent word, with a frequency of over 1,500, serving as a direct indicator of strong negative sentiment.
3. The words "im," "now," "fix," and "fuck" are next on the list. "Im" and "now" provide context, while "fix" and "fuck" clearly convey frustration and a desire for an issue to be resolved.
4. The words "can’t," "play," "don’t," and "please" are also in the top 10. "can’t" and "don’t" are explicit negative terms, while "play" likely refers to the inability to play a game due to a problem. The presence of "please" suggests a plea for help or a fix.

Top 10 Words in Irrelevant Tweets:

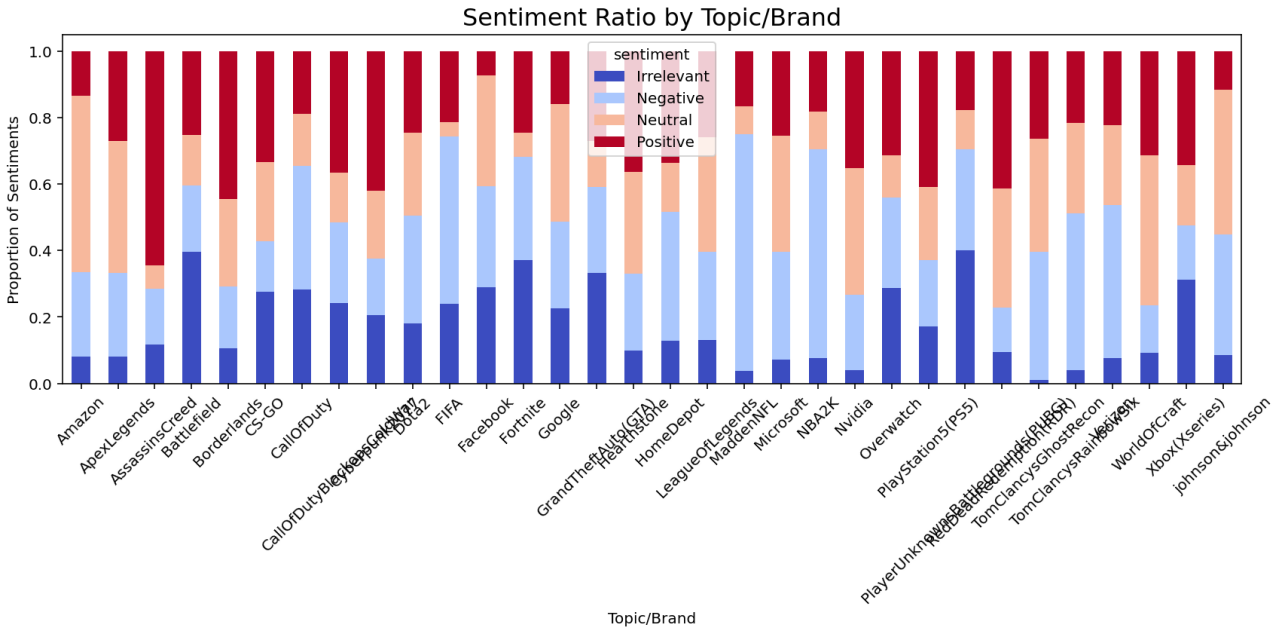


Interpretation:

1. The words "see", and "player" are the most frequent, both with frequencies above 850. These are general terms that could be used in a variety of contexts, making them less indicative of a specific sentiment.
2. The chart features a mix of seemingly positive and negative words. "Love," "good," and "best" are present, but so is "ban." This combination of conflicting sentiments is a key reason why these tweets might have been categorized as "irrelevant" by a sentiment analysis model.
3. The word "im" is also high on the list, a generic term. "Game" is present, but it's not the top word, as it was in the other charts. This suggests these tweets might be discussing topics related to gaming in a more tangential or ambiguous way.

4. The presence of general words like "people" and "one" further supports the idea that these tweets lack a strong, consistent sentiment, making them difficult to classify.

Sentiment Ratio by Topic/Brand:



Interpretation:

1. Dominance of Positive and Negative Sentiments: Across most topics, positive (red) and negative (light blue) tweets make up the largest proportions. This suggests that people are more likely to express strong opinions (either positive or negative) than neutral ones when discussing these topics.
2. Gaming and Entertainment Topics: Games like Fortnite, GrandTheftAuto, LeagueofLegends, and Battlefield have a mix of all four sentiment types, indicating a complex public response. Some topics, like Assassin'sCreed, ApexLegends, and Minecraft, show a high proportion of positive sentiment, suggesting they are generally well-received.
3. Hardware and Brands: Brands like Amazon, Google, Nvidia, and Microsoft also have varying sentiment distributions. Nvidia has a significant proportion of neutral tweets, perhaps related to technical specifications or product announcements.

4. Highly Negative Topics: The chart shows that some topics, particularly "Fifa" and certain gaming titles like "Call of Duty," have a very high proportion of negative tweets. This could indicate widespread player frustration with game mechanics, bugs, or other issues.
5. Neutral and Irrelevant Proportions: Topics like "Johnson" and "HomeDepot" have a high proportion of neutral or irrelevant tweets, suggesting that these are often mentioned in a non-emotional context. The high proportion of irrelevant tweets for some topics points to tweets that a sentiment model found difficult to classify.

Insights:

```
===== INSIGHTS =====
1. Negative tweets dominate the dataset, followed closely by positive tweets.
2. Sentiment distribution by brand highlights areas of satisfaction or dissatisfaction.
3. Word clouds reveal trending topics and emotions in tweets.
4. Businesses can leverage this data to understand customer feedback and improve strategies.
```

Interpretation:

1. Positive tweets are characterized by words like "love," "good," and "best," often in the context of games.
2. Negative tweets are dominated by strong negative language and frustration, with words like "fuck," "shit," and "fix" being prominent.
3. Neutral tweets feature more generic or specific, non-emotional terms like "Johnson" and "Amazon," suggesting factual or descriptive content.
4. Irrelevant tweets have a mix of positive and negative words, making their sentiment ambiguous and difficult to classify.
5. Overall, gaming topics show a high polarization of sentiment, with topics like Call of Duty and Fifa generating significant negativi