Threads Analyzer

A Python-driven Comparative Study

SWE 486 – Cloud Computing and Big Data

Done By:

Renad Nigr Alsubaie

Mais Ashraf Alkhatib

Jouri Mufadhi Alanazi

Monerah Faris Alsubaie

Abeer Sami Alshaya

Supervised By: Dr. Afshan Jafri

Content

- Introduction
- Data exploration
- Data preprocessing
- Model planning and building
- Communicate Results
- Conclusion

Introduction

- Threads is an online social media service operated by Meta Platforms. It allows users to post and share text, images, and videos, and interact with other users' posts. The app is closely linked to Instagram and requires users to have an Instagram account. Threads gained over 100 million users in its first five days, making it the fastest-growing consumer software application in history. It initially launched as a separate app with messaging and video chat features but was discontinued in December 2021. The app aims to provide a microblogging experience similar to Twitter but lacks certain features like hashtags and direct messaging. In this project, we will analyze user reviews of Threads on Google Play Store and the App Store and conduct sentiment analysis.

project goals

Evaluate the level of user satisfaction with the Threads application

Extract valuable information about user preferences and opinions regarding Threads.

Use user outcome of this analysis to improve the overall experience of Threads application users.

Initial Hypothesis

Our hypotheses for analyzing the Threads app reviews dataset are:

H1: Compare user satisfaction levels between Twitter and Threads.
H2: Measure and evaluate user satisfaction with the Threads app.
H3: Compare user engagement, preferences, content relevance, user experience, and performance between Twitter and Threads.

Data Exploration

After collecting the reviews data about threads platform from google play and app store in the previous phase, we started exploring and understanding our data, the following shows several information about it

the following is what we found about our data

Total number of reviews in the dataset:

32910

Names of columns in the dataset:

Source
Review _description
Rating
Review _date
Dtype

Types of data in each column it the dataset:

Source; object
Review _description: object
Rating :int64
Review _date: object
Dtype : object

Data Preprocessing

Upon exploring and reviewing the collected data, we were able to identify several issues that our dataset may have.

Issues That we Addressed

19925	Google Play	Great App!
19926	Google Play	Great App!
19927	Google Play	Great App!
19928	Google Play	Great app!
19929	Google Play	Great app!
19930	Google Play	Nice mark!
19931	Google Play	Great app!
19932	Google Play	Great app!

Remove Duplicate Reviews

Google Play	💪💪💪💪🤌🤌🤌🤌❤️❤️❤️♥♥♥♥🔥🔥🔥🔥
Google Play	Superb ❤️😘😋😘🥰
Google Play	amazing app ❤️
Google Play	Bhut achha hai ye app ❤️❤️❤️
Google Play	Best experience ❤️
Google Play	Nice 🤌 Awesome 😎
Google Play	Simply 😎 Awesome❤️❤️🔥🔥
Google Play	Mr. Copycat😂😂
Google Play	Hello, Threads! ❤️😊

Unrelated Characters – emojis, special character, URLs

Issues That we Addressed

Google Play	സ്വപ്നങ്ങളാണ് വിജയത്തിലേയ്ക്കയുള്ള ആദ്യ ചവിട്ടു പാടി. സ്വപ്നങ്ങൾ കാണുക.....
Google Play	Yahan bhi ladkiyan bhaw nahi deti... Worst app 🙄🙄
Google Play	pode melhorar, mas tá muito bom pra uma rede social nova
Google Play	Ye aap bhi bekar hai bhai koi ladki reply nahi karti hai😒
Google Play	Bekaar app hai yaar, larrkiyon ko msg b nae kr sakte
Google Play	Solo pudiera competir con Twitter, si no tuviera tantas restricciones como facebook.
Google Play	Que dura el app creo que mejor que Twitter 😅
Google Play	Elon, ketar ketir wkwk
Google Play	Kitna ghatiya app hai🤣
Google Play	Login vayena mero ma 😜
Google Play	বাপের পকেটে টাকা ফাক করার জন্য নতুন পদ্ধতি মার্ক এর😑

Non-English Reviews

Google Play	Use of threads after one day, Shuru mazboori me kiye the ab mza aa rha hai
Google Play	Breathtaking freshness!!
Google Play	Useless -_-

Punctuation & Stop Words

Sample of our Data Set after Cleaning

Google Play	Pretty good first launch easy use self explanatory say algorithm good well great potential things need improvement ability use hashtags would make easier find t
Google Play	brand new app well optimized However missing quite features apps like Twitter way timeline show threads re following would also nice able switch accounts qui
Google Play	Great app lot potential However lot needs fixed example option mute accounts lagging delayed feature definitely needed stage part next problem Another issue
Google Play	app good needs lot functionality example searching topic find anything related topic meaning anything comes via main page incase lowers lot outreach possibility
Google Play	Currently challenging use dark mode want change brighten apparently change Instagram Dark mode needs improved tough eyes UI visually unbearable Twitter r
Google Play	still want see content people specifically follow unless explicitly search search want see option complete searches user searches bare bones poor clone Twitter
Google Play	Could great pages loaded clicked Sometimes Posts usually load biggest thing follow someone Threads automatically follow Instagram well Either re two separa
Google Play	liking concept room improvement though Everytime try attach photo thread app crashes able add photo also seem find way view list people follow always folks r
Google Play	bad first launch still room improvements would like see trending page longer videos edit button threads posts uploaded user Improve loading time Add pause but
Google Play	UI app good Using easy visually clean Unfortunately fails functionality department home feed 99 random people accounts actually follow defeats purpose followir
Google Play	Nice crashed access photo folders crashed twice today far middle browsing Also reason click file button attach photo let browse individual folders gallery shows
Google Play	Pointless following anyone feed algo show content actually follow want see find opening app wanting like quickly get annoyed seeing randomness like re invited
Google Play	quite ready prime time Needs feed specifically accounts follow one Instagram account Threads allow easy switching accounts without involving logging out loggir
Google Play	frustrating experience Feed full users follow care content Switching company private accounts forces complete sign reverify Picture upload quality compressed T
Google Play	like simplicity app definitely bugs need get worked choosing sync IG accounts follow weeded unfollowed reappeared ve done 3x closing app refresh keeps hap
Google Play	needs lot work constantly crashes point barely use auto following everyone follow insta search terms lead topic interested layout feels quite lacking needs sort v
Google Play	first thoughts Since web version yet difficult using upright tablet going try phones today sure work fine web version comes really useful expected leaks app first a

Mode Planning and Building

1 Sentiment Analysis

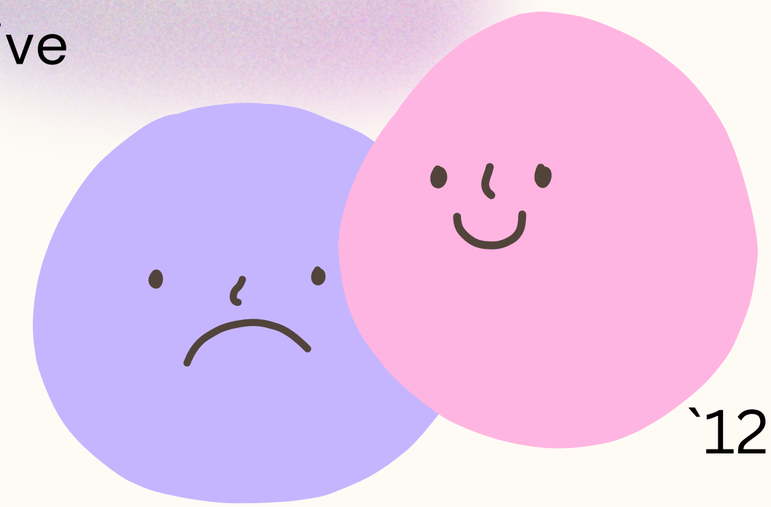
2 Descriptive Analysis

3 Predictive Analysis

Sentiment Analysis

Why Sentiment analysis?

- To understand Public Opinion and to Develop Feedback Analysis which aligns with our hypothesis
- Our sentiment analysis is based on the “Rating” column:
  - Rating from 1-2: Negative
  - Rating 3: Neutral
  - Rating 4-5: Positive



Descriptive Analysis

What is Descriptive Analysis ?

- an approach to analyzing data to provide meaningful descriptions, displays, or summaries of data points.
- Its purpose is to show patterns that align with all data conditions

Tools used

- we used the Describe() method to find the Mean - Median - Standard Deviation and variance
- we used the info() method to get general information about the data.
- we used the value_count() method to calculate the number of positive, neutral, and negative reviews.

Descriptive Analysis

Tools Used (Cont.)

- we used the CountVectorizer method to calculate word occuronces

```
dfidf.sort_values(by = ['idf_weights']).head(15)
```

	idf_weights
app	2.011888
twitter	2.531545
instagram	3.074609
like	3.127682
threads	3.142041
good	3.184175
account	3.682801
better	3.716224
see	3.772938
people	3.774538
use	3.787431
follow	3.932031
please	3.950953
great	4.038746
new	4.084621

Most frequent

```
[ ]  dfidf.sort_values(by = ['idf_weights']).tail(15)
```

	idf_weights
independly	10.21149
independence	10.21149
indecency	10.21149
ind	10.21149
incrementally	10.21149
incremental	10.21149
increased	10.21149
increadible	10.21149
incorporation	10.21149
incorporating	10.21149
incorporates	10.21149
inconsistencies	10.21149
inconsiderate	10.21149
incompetent	10.21149
kneecap	10.21149

Least frequent

Predictive Analysis

Predictive analytics is the process of using data to forecast future outcomes. We have selected two models for our Predictive Analysis: Naïve Bayes and Logistic Regression.

Naive Bayes

a supervised machine learning algorithm commonly used for solving classification problems. It is based on ' Bayes' theorem, which calculates the probability of a certain event occurring given prior knowledge.

Logistic Regression

a supervised machinelearning algorithm used primarily for classification tasks
useful when the dependent variable is binary or categorical

Predictive Analysis Steps

1. Import Libraries.
2. Remove missing data (NAN) and neutral sentiments.
3. Convert the 'sentiment' column to numerical (0 for Negative, 1 for Positive).
4. Split the dataset into features (X) and targets (Y).
5. Vectorize Textual Data.
6. Balance Dataset.
7. Prepare datasets for training by transforming textual data.
8. Train Naive Bayes and Logistic Regression Models.
9. Evaluate Model Performance:
    - accuracy, confusion matrix, and ROC.

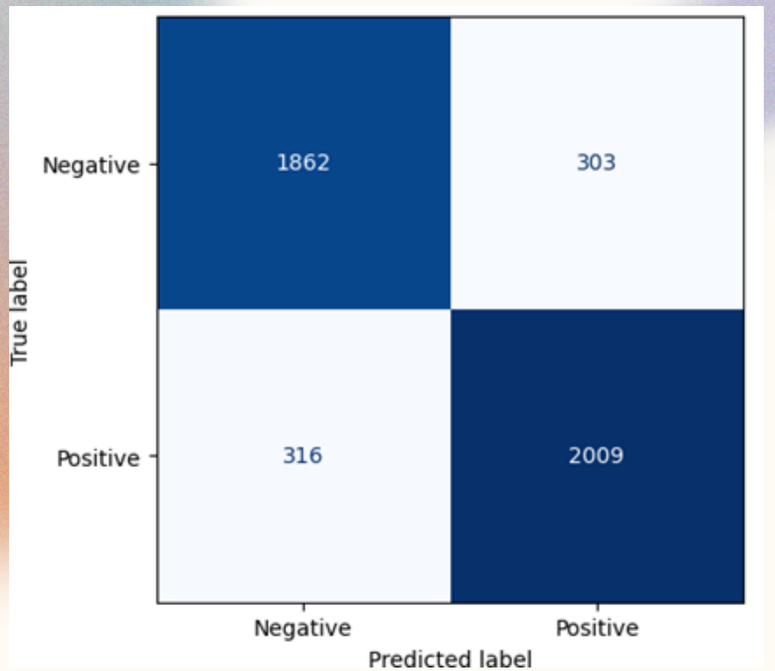
Naive Bayes

Accuracy

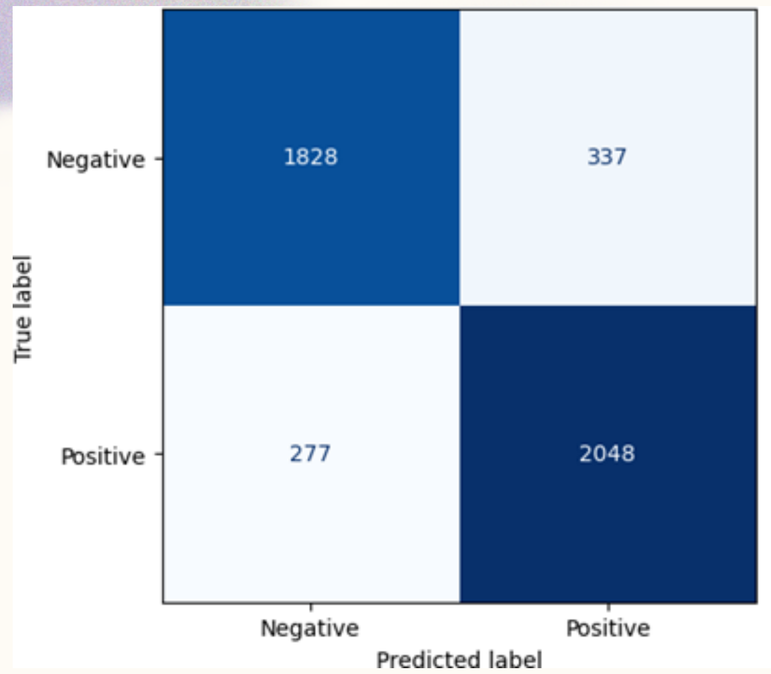
```
Accuracy Report
Model Accuracy: 0.90
10-Fold Cross Validation: 0.85
Training Score: 0.90
Testing Score: 0.86
```

Confusion Metrix

Balanced data



Unbalanced data



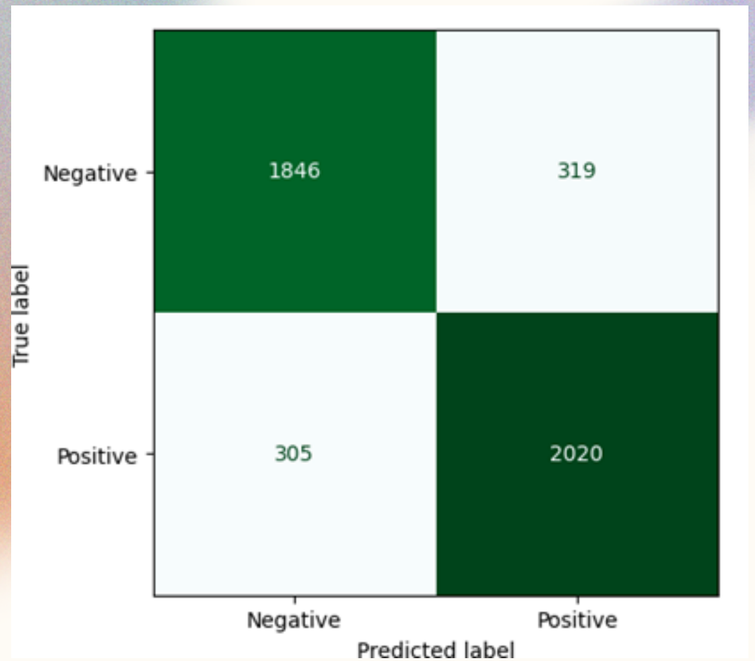
Logistic Regression

Accuracy

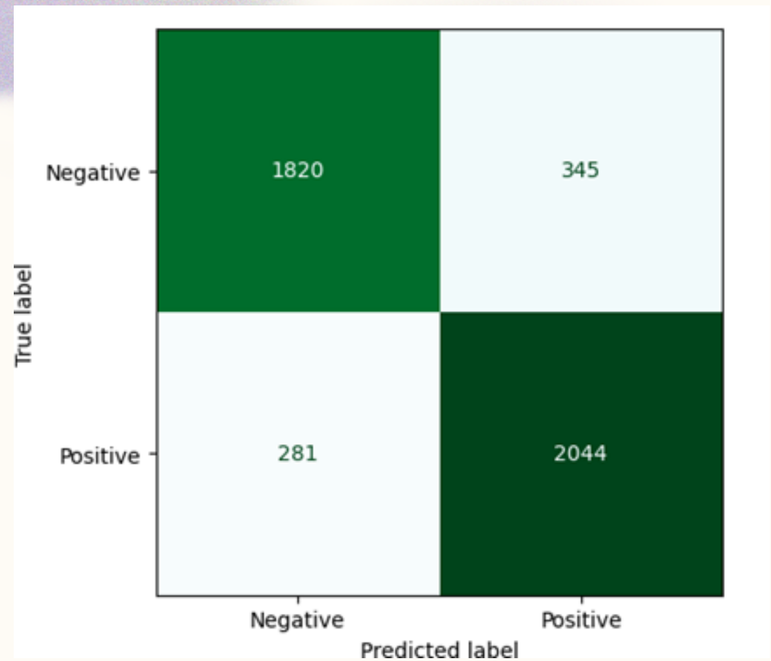
```
Accuracy Report
Model Accuracy: 0.90
10-Fold Cross Validation: 0.85
Training Score: 0.90
Testing Score: 0.86
```

Confusion Metrix

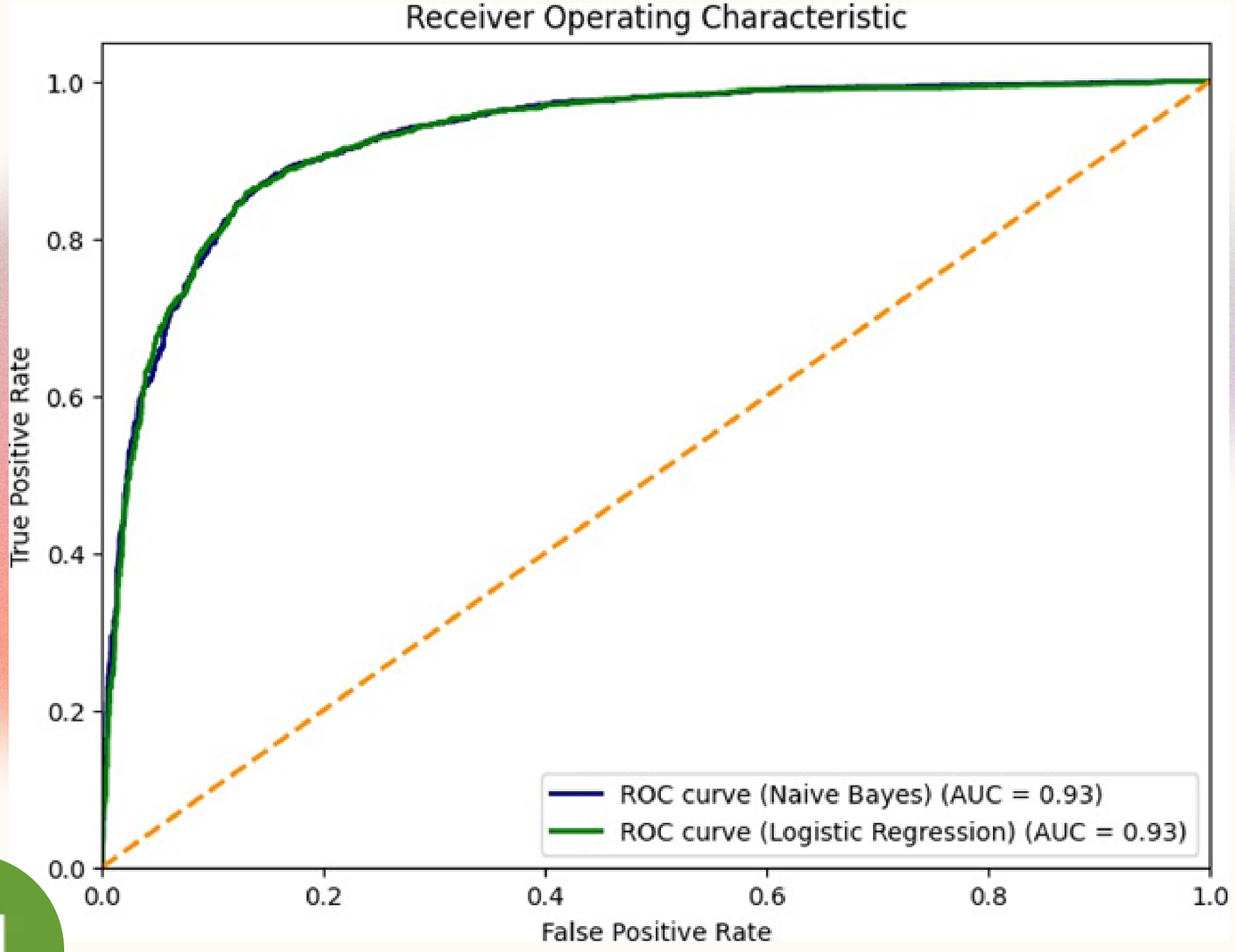
Balanced data



Unbalanced data



ROC Curve for Naive-Bayes and Logistic Regression



Communicate Results

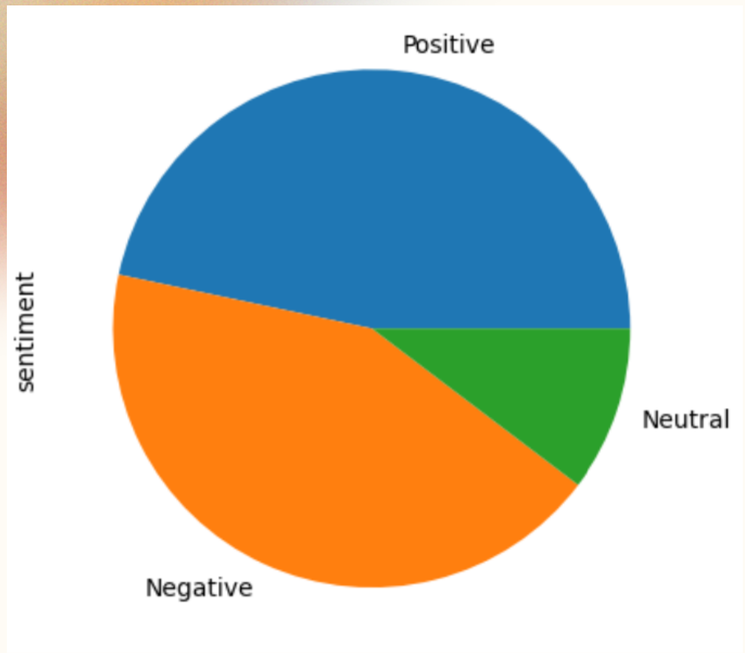
- Clear and effective communication of results is a crucial skill when dealing with different stakeholders.

- we will present the distribution of visualization reviews in terms of both the number of reviews and their proportions to support our initial hypothesis.

1-Distribution of classification visualization

1.1 Visualize the classification results using a Pie Chart.

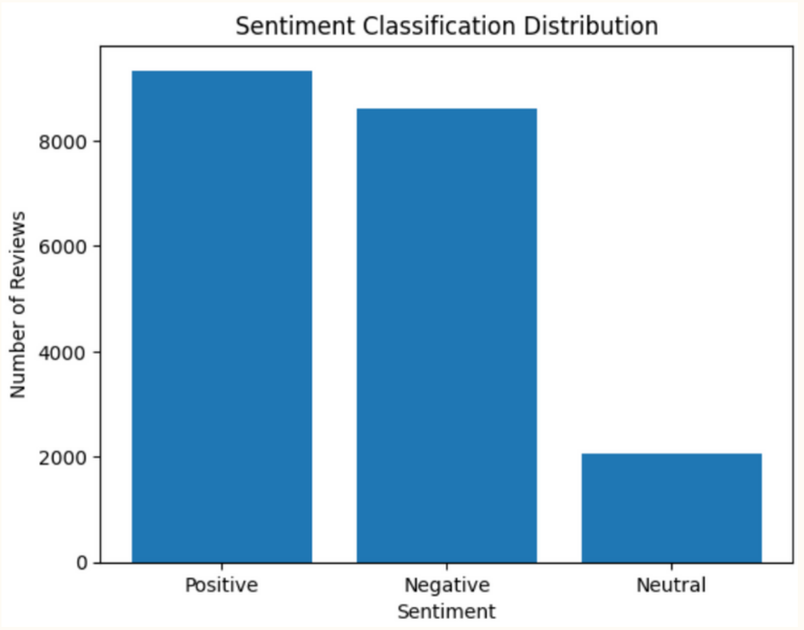
which effectively displays the proportion of each sentiment classification (neutral, positive, negative):



Pie Chart

1.2 Visualize the classification results using a bar chart.

which effectively displays the proportion of each sentiment classification (neutral, positive, negative):

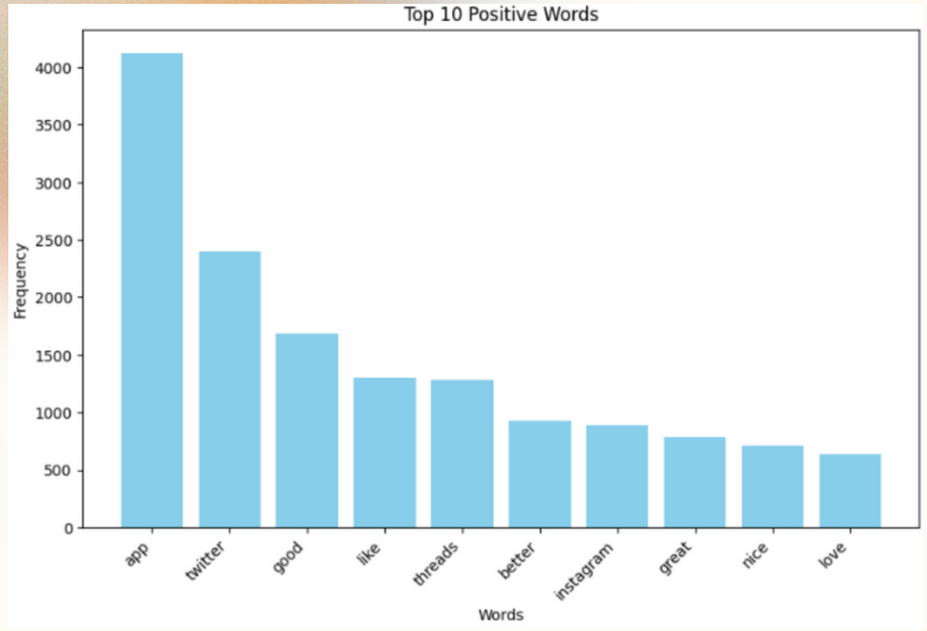


Bar Chart

2-Visualizing Common Positive and Negative Words

2.1 Visualizing Common Positive Words.

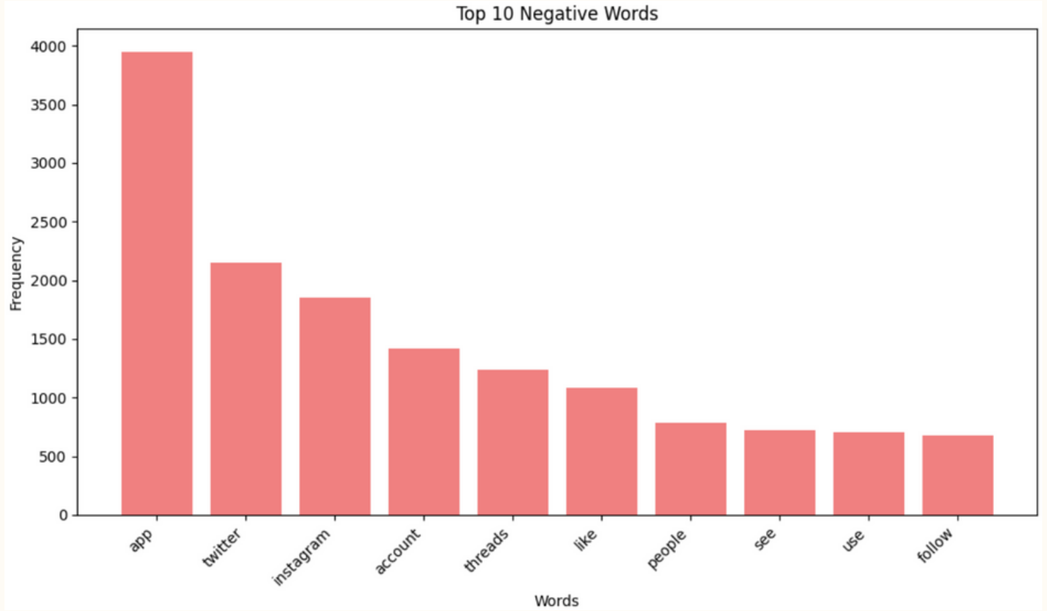
we visualize the top 10 positive words based on their frequency count.



most Common Positive Words bar chart

2.2 Visualizing Common Negative words.

we visualize the top 10 negative words based on their frequency count.

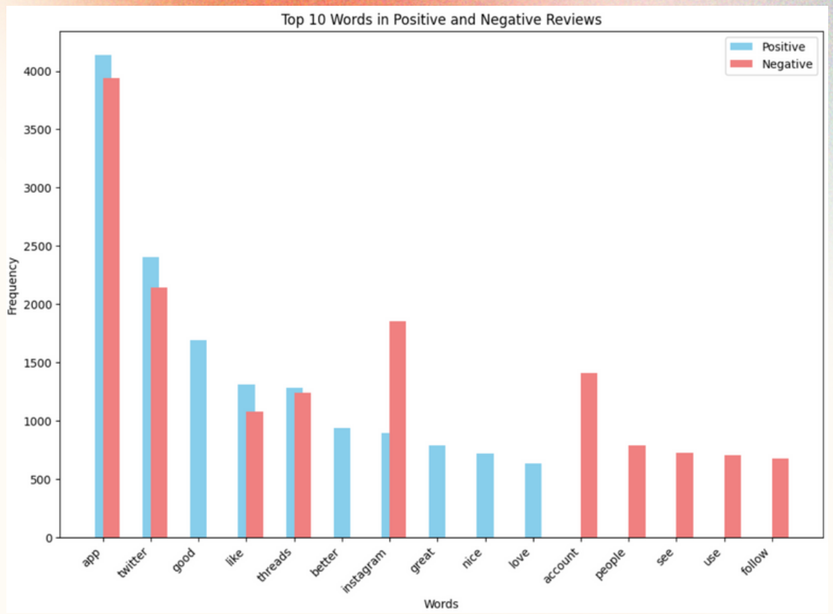


most Common Negative Words bar chart

3.Findings

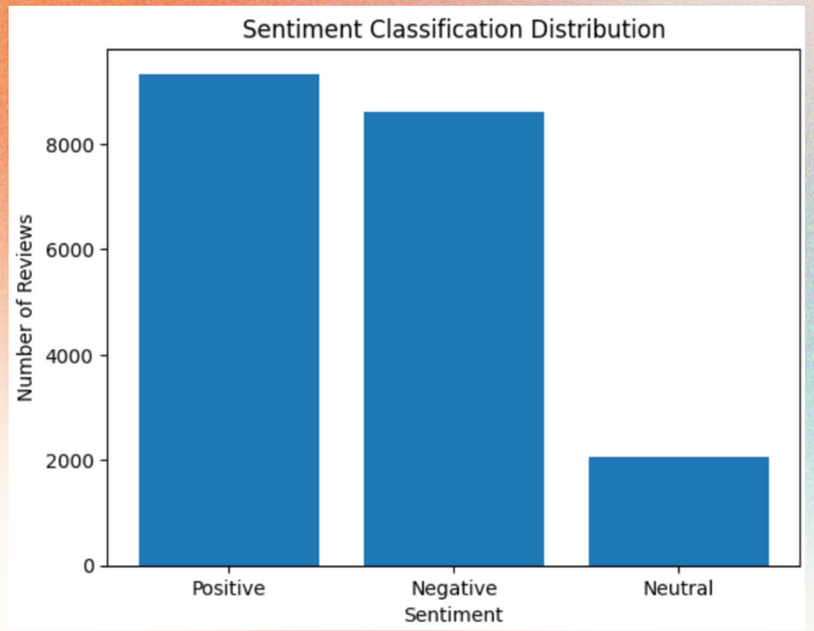
H1: Compare user satisfaction levels between Twitter and Threads.

- mixed sentiment among users when comparing the two platforms. In both positive and negative reviews, the second most frequent word was "Twitter," suggesting a preference for Twitter over Threads. However, the term "Threads" also appeared frequently, indicating that there were users who preferred the Threads app. This finding suggests that user satisfaction levels varied between the two platforms



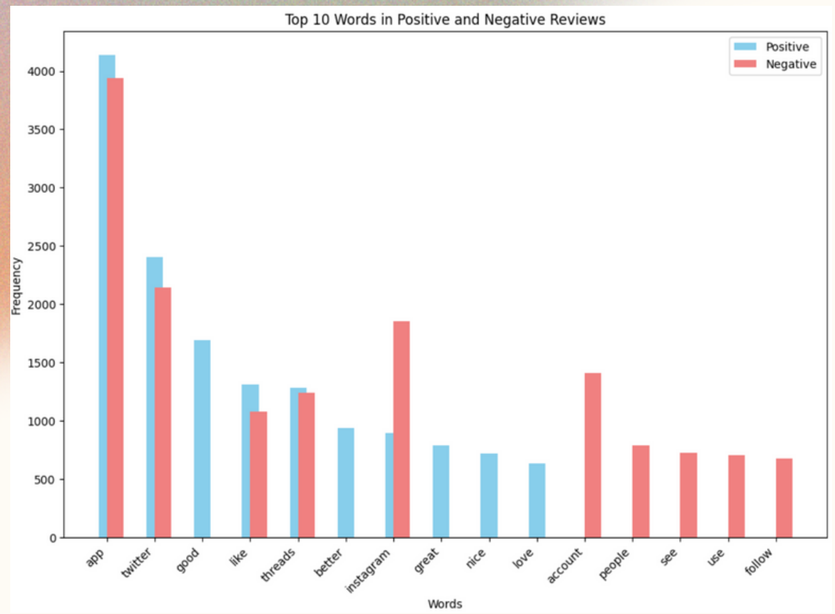
H2: Measure and evaluate user satisfaction with the Threads app.

- users have a positive perception of the Threads app based on the analyzed customer feedback. The higher proportion of positive sentiments indicates a higher level of user satisfaction.



H3: Compare user engagement, preferences, content relevance, user experience, and performance between Twitter and Threads.

- The analysis of positive and negative word visualization shows that "App" is the most frequent term in the positive chart, with "Twitter" and "Threads" ranking second and fifth respectively in both positive and negative charts. This suggests a balanced or neutral reviews between Twitter and the Threads app.



Conclusion

Performing data analysis is crucial for obtaining improved results and gaining insights into ongoing processes, enabling us to make informed decisions. However, when dealing with unclean data, it becomes challenging to make accurate decisions. Therefore, it is essential to prioritize data cleanliness and quality to enhance the effectiveness of our decision-making.

THANK YOU!

References

[1] Threads (social network) (2023) Wikipedia. Available at: (https://en.wikipedia.org/wiki/Threads_(social_network) (Accessed: 29 September 2023).

[2] The absolute basics for beginners# (no date) NumPy. Available at:https://numpy.org/doc/stable/user/absolute_beginners.html (Accessed: 30 September 2023).

[3] Visualization with python (no date) Matplotlib. Available at: https://matplotlib.org/ (Accessed: 30 September 2023).

[4] (No date) Pandas introduction. Available at: https://www.w3schools.com/python/pandas/pandas_intro.asp (Accessed: 30 September 2023).

[5] M, S. (2022) NLTK: A beginners hands-on guide to natural language processing, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/07/nltk- a-beginners- hands-on-guide-to-natural-language-processing/ (Accessed: 30 September 2023).

[6] Low-code data app development (no date) Plotly. Available at: https://plotly.com/ (Accessed: 30 September 2023).

[7] Google-play-scraper (no date) PyPI. Available at: https://pypi.org/project/google- play- scraper/ (Accessed: 30 September 2023).

[8] App-store-scraper (no date) PyPI. Available at: https://pypi.org/project/app-store- scraper/ (Accessed: 30 September 2023).

[9] Jhalani, S. (2023) Threads, an Instagram app reviews, Kaggle. Available at: https://www.kaggle.com/datasets/saloni1712/threads-an- instagram-app- reviews?select=threads_reviews.csv (Accessed: 29 September 2023).

[10] Google-play-scraper (no date) PyPI. Available at: https://pypi.org/project/google- play- scraper/#description (Accessed: 30 September 2023).

[11] Logistic regression (no date) Logistic Regression - an overview | ScienceDirect Topics. Available at: https://www.sciencedirect.com/topics/computer- science/logistic-regression (Accessed: 16 November 2023).

[12] naive Bayes (no date) scikit. Available at: https://scikit- learn.org/stable/modules/naive_bayes.html (Accessed: 16 November 2023).