# Data Mining and Analysis Project

POSITIVE NEUTRAL NEGATIVE

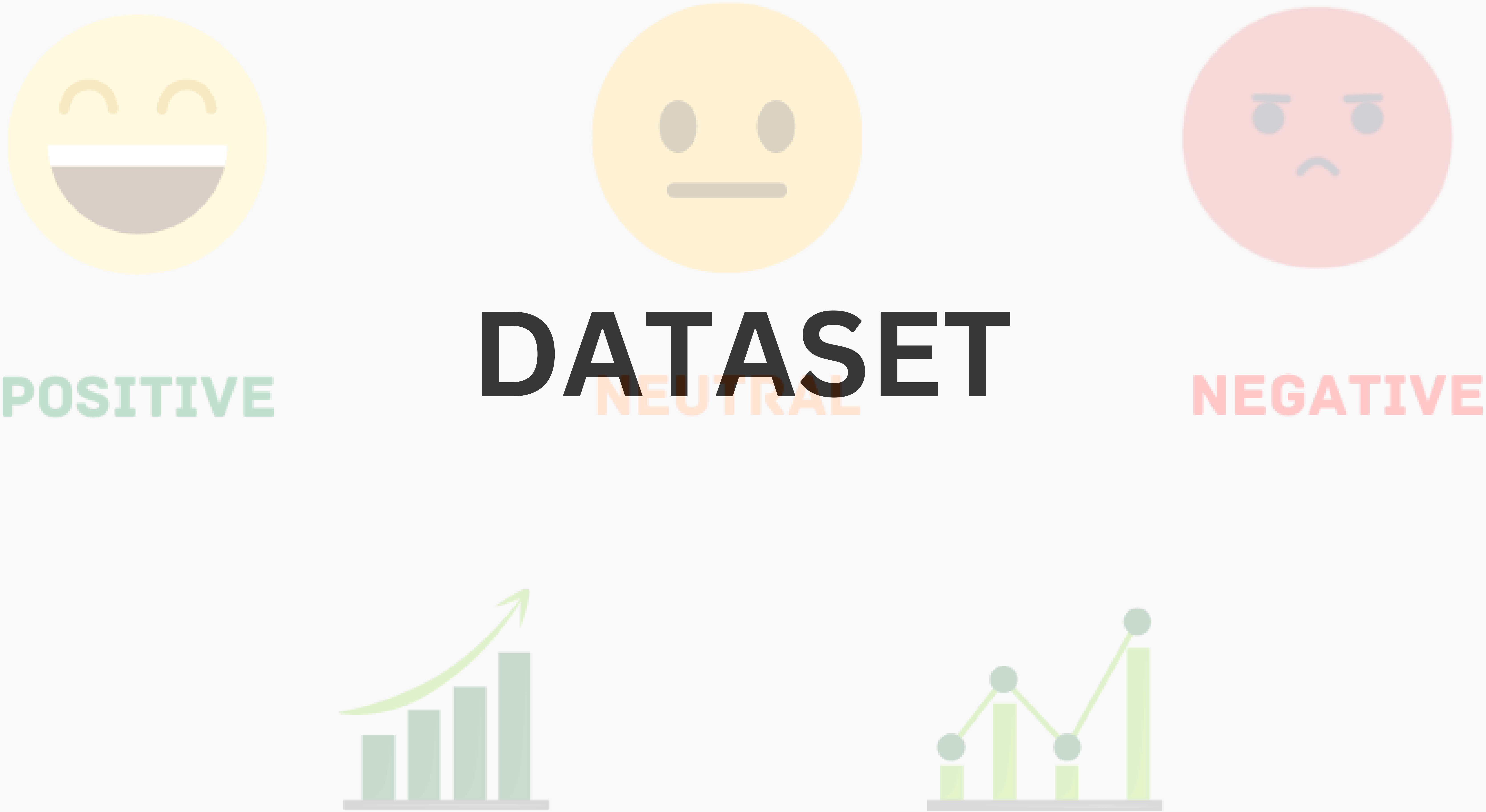
# Sentiment Analysis

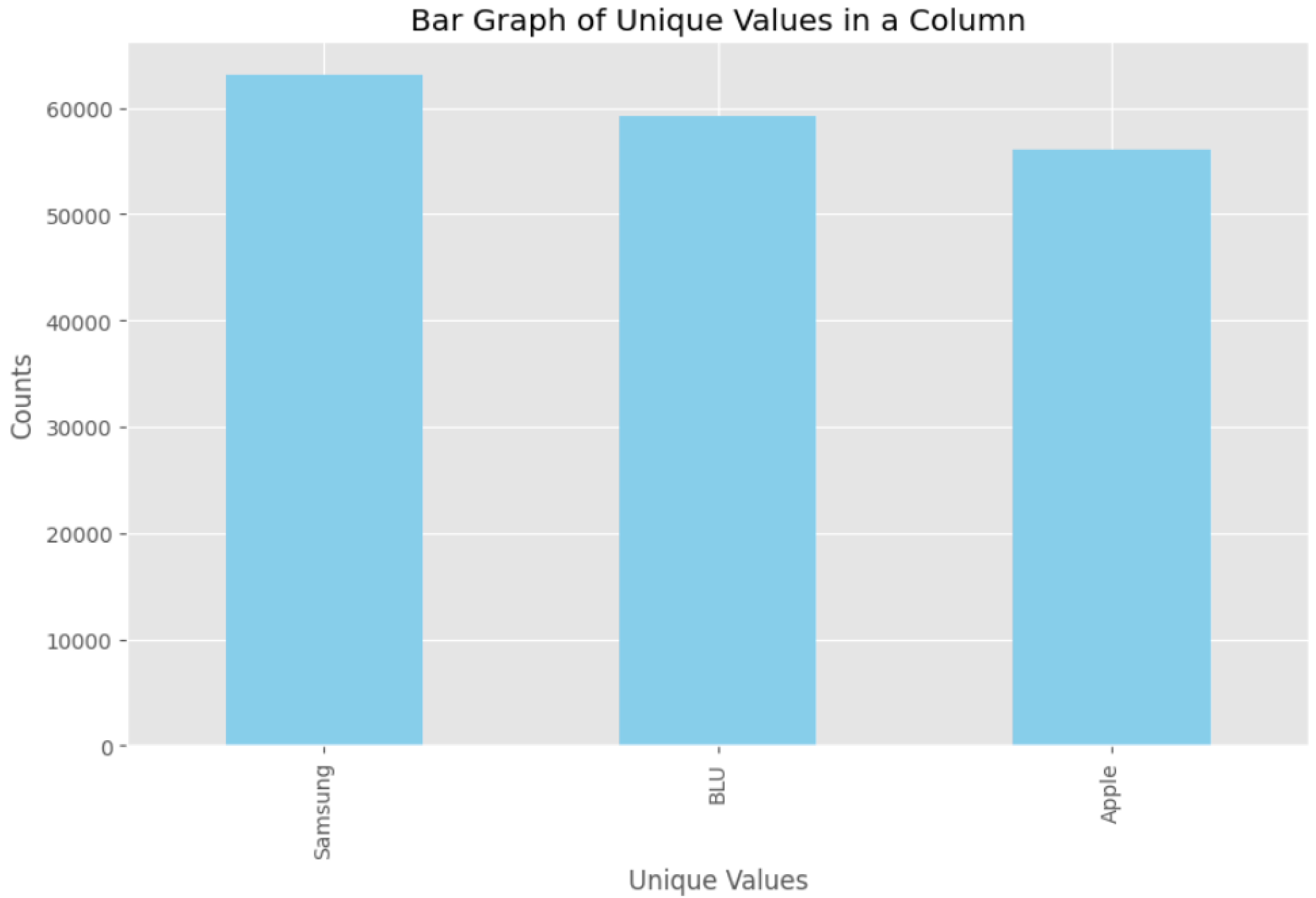
 

# INTRODUCTION

The idea was to implement ***<u>Sentiment Analysis on Amazon reviews involves using natural language processing</u>*** to assess and understand the emotional tone and opinions expressed by customers in their product reviews on the Amazon platform. It is performed on the dataset **“amazon mobile phone reviews”** from kaggle and two models namely **VADER and roBERTA** are used.

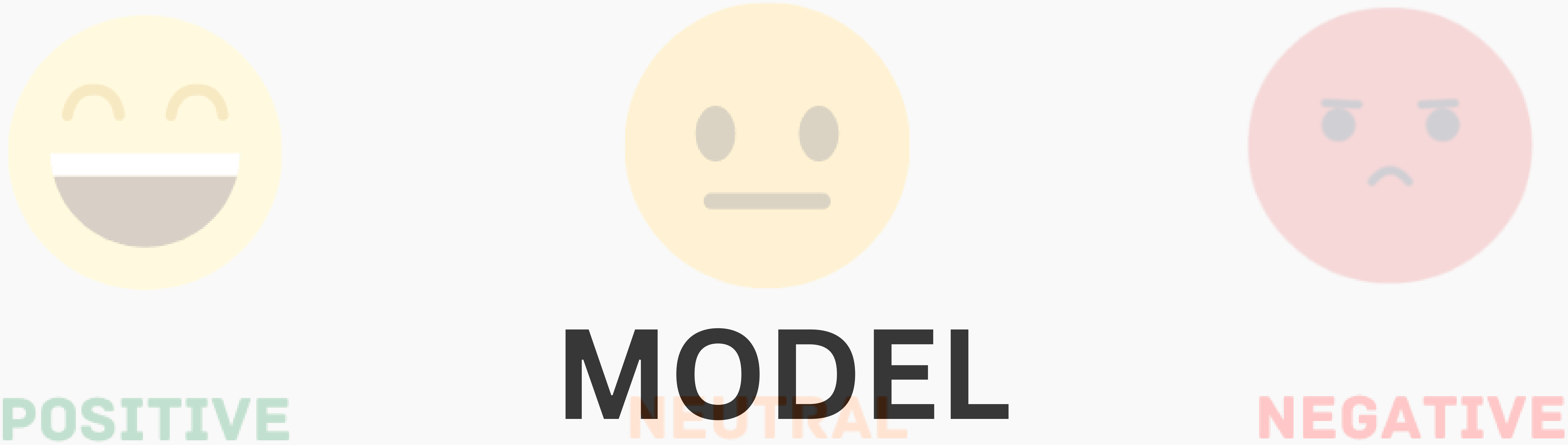
# DATASET




Bar Graph of Unique Values in a Column
Counts
60000
50000
40000
30000
20000
10000
0
Samsung
BLU
Apple
Unique Values

| product_n... | brand_name | price | rating | reviews | review_vot... |
|---|---|---|---|---|---|
| "CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D700*FRONT CAMERA*ANDROID *SLIDER*QWERTY KEYBOARD*TOUCH S... | Samsung | 199.99 | 5 | I feel so LUCKY to have found this used (phone to us & not used hard at all), phone on line from som... | 1 |
| "CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D700*FRONT CAMERA*ANDROID *SLIDER*QWERTY KEYBOARD*TOUCH S... | Samsung | 199.99 | 4 | nice phone, nice up grade from my pantach revue. Very clean set up and easy set up. never had an and... | 0 |
| "CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D700*FRONT | Samsung | 199.99 | 5 | Very pleased | 0 |

# MODEL



[17]:

```python
from nltk.sentiment import SentimentIntensityAnalyzer
from tqdm.notebook import tqdm

sia = SentimentIntensityAnalyzer()
```

```
/opt/conda/lib/python3.10/site-packages/nltk/twitter/__init__.py:20: UserWarning: The twython library has not been installed. Some functionality from the tw
itter package will not be available.
  warnings.warn("The twython library has not been installed. "
```

[18]:

```python
sia.polarity_scores('I am so happy!')
```

[18]: {'neg': 0.0, 'neu': 0.318, 'pos': 0.682, 'compound': 0.6468}

```python
sia.polarity_scores('This is the worst thing ever.')
```

[19]: {'neg': 0.451, 'neu': 0.549, 'pos': 0.0, 'compound': -0.6249}

+ Code  + Markdown

[20]:

```python
sia.polarity_scores(example)
```

[20]: {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

```python
from transformers import AutoTokenizer
from transformers import AutoModelForSequenceClassification
from scipy.special import softmax
```

```python
MODEL = f"cardiffnlp/twitter-roberta-base-sentiment"
tokenizer = AutoTokenizer.from_pretrained(MODEL)
model = AutoModelForSequenceClassification.from_pretrained(MODEL)
```

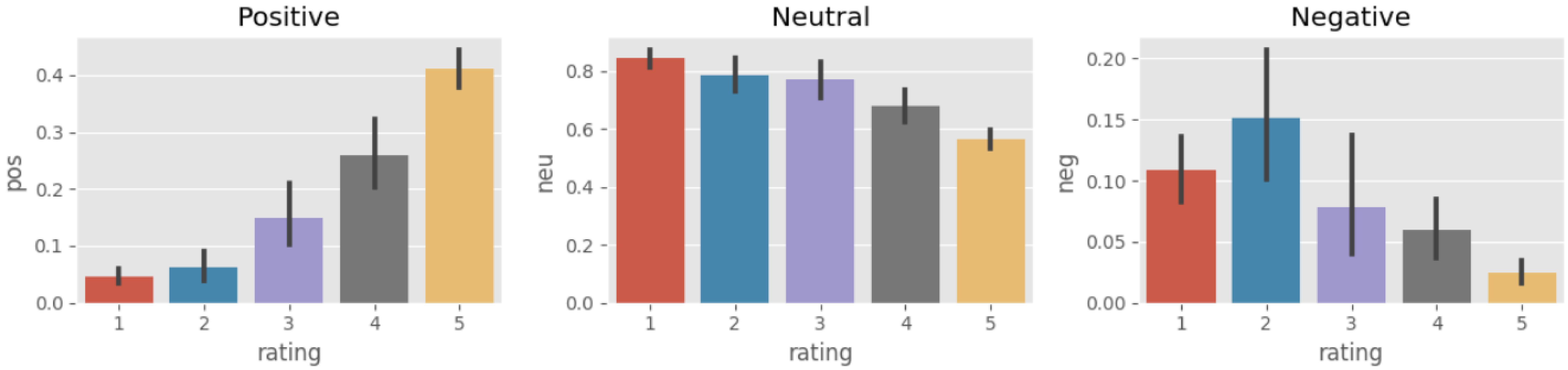
```python
res = {}
for i, row in tqdm(df.iterrows(), total=len(df)):
    try:
        text = row['reviews']
        myid = row['id']
        vader_result = sia.polarity_scores(text)
        vader_result_rename = {}
        for key, value in vader_result.items():
            vader_result_rename[f"vader_{key}"] = value
        roberta_result = polarity_scores_roberta(text)
        both = {**vader_result_rename, **roberta_result}
        res[myid] = both
    except RuntimeError:
        print(f'Broke for id {myid}')
```

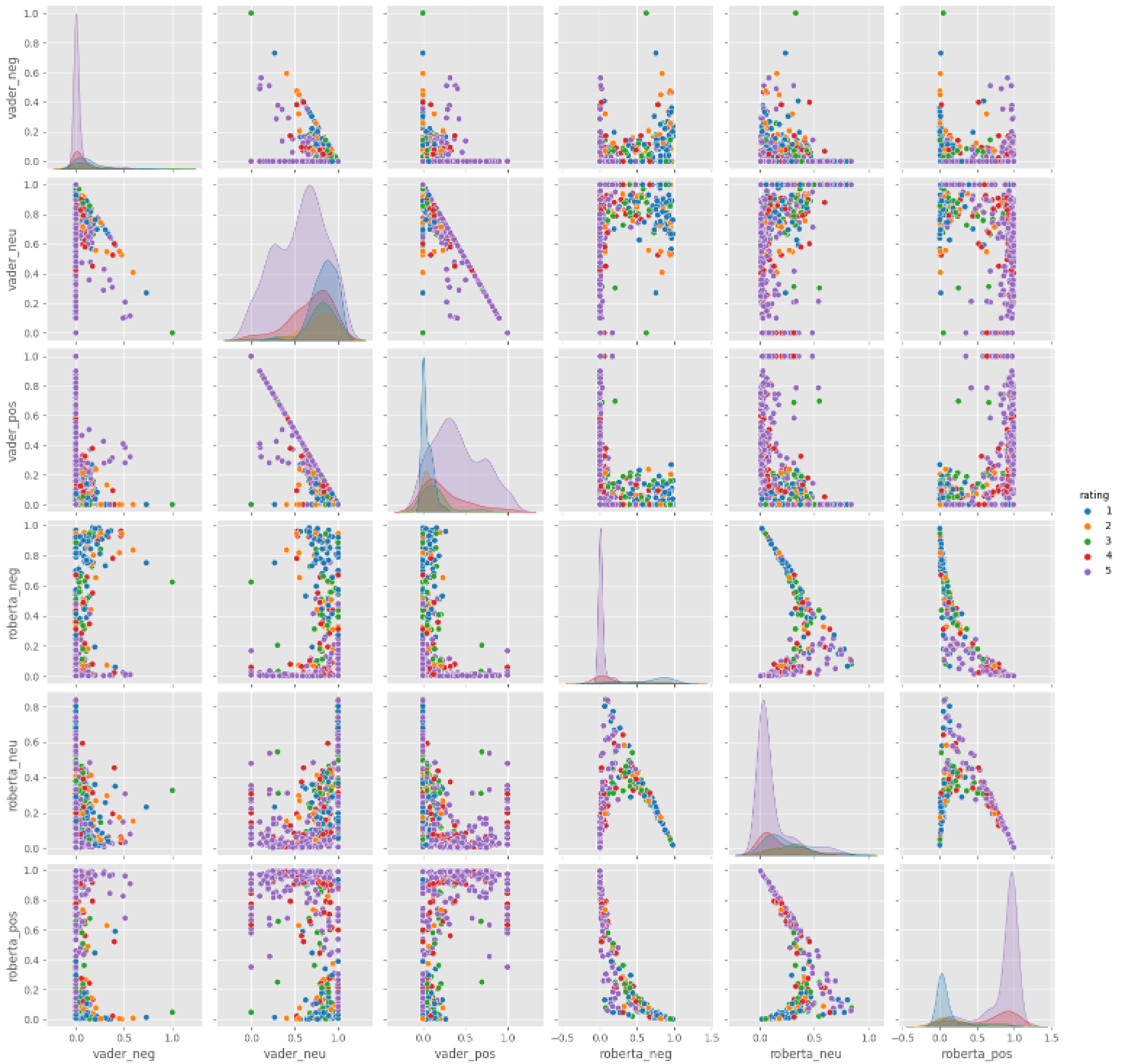
Loading widget...

```
Broke for id 77
Broke for id 121
Broke for id 430
```

# SAMSUNG RESULTS

Count of Reviews by Stars
300
250
200
150
100
50
0
1
2
3
4
5
rating


Positive
Neutral
Negative
pos
neu
neg
rating
0.0
0.1
0.2
0.3
0.4
0.6
0.8
0.00
0.05
0.10
0.15
0.20
1
2
3
4
5

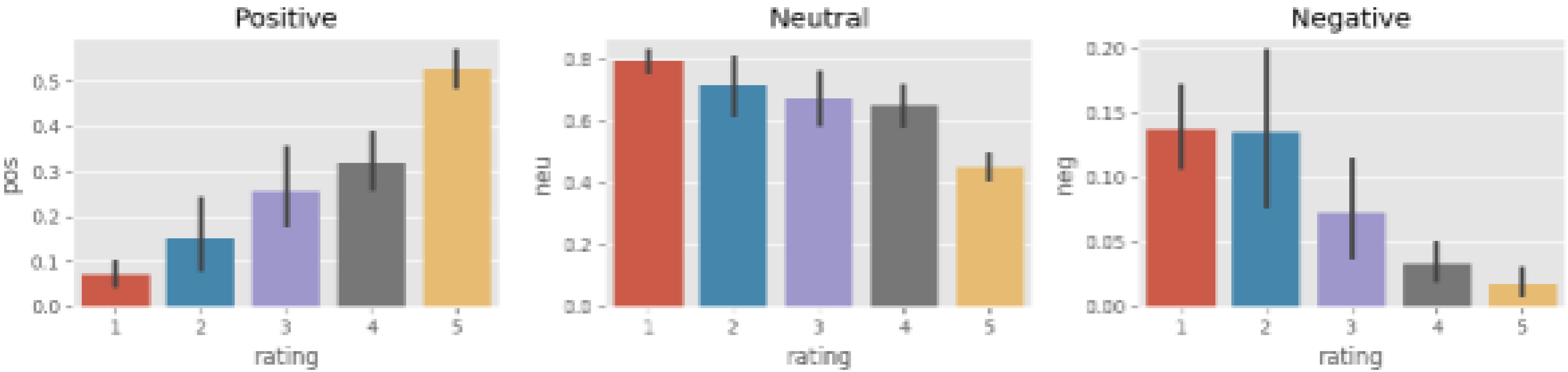


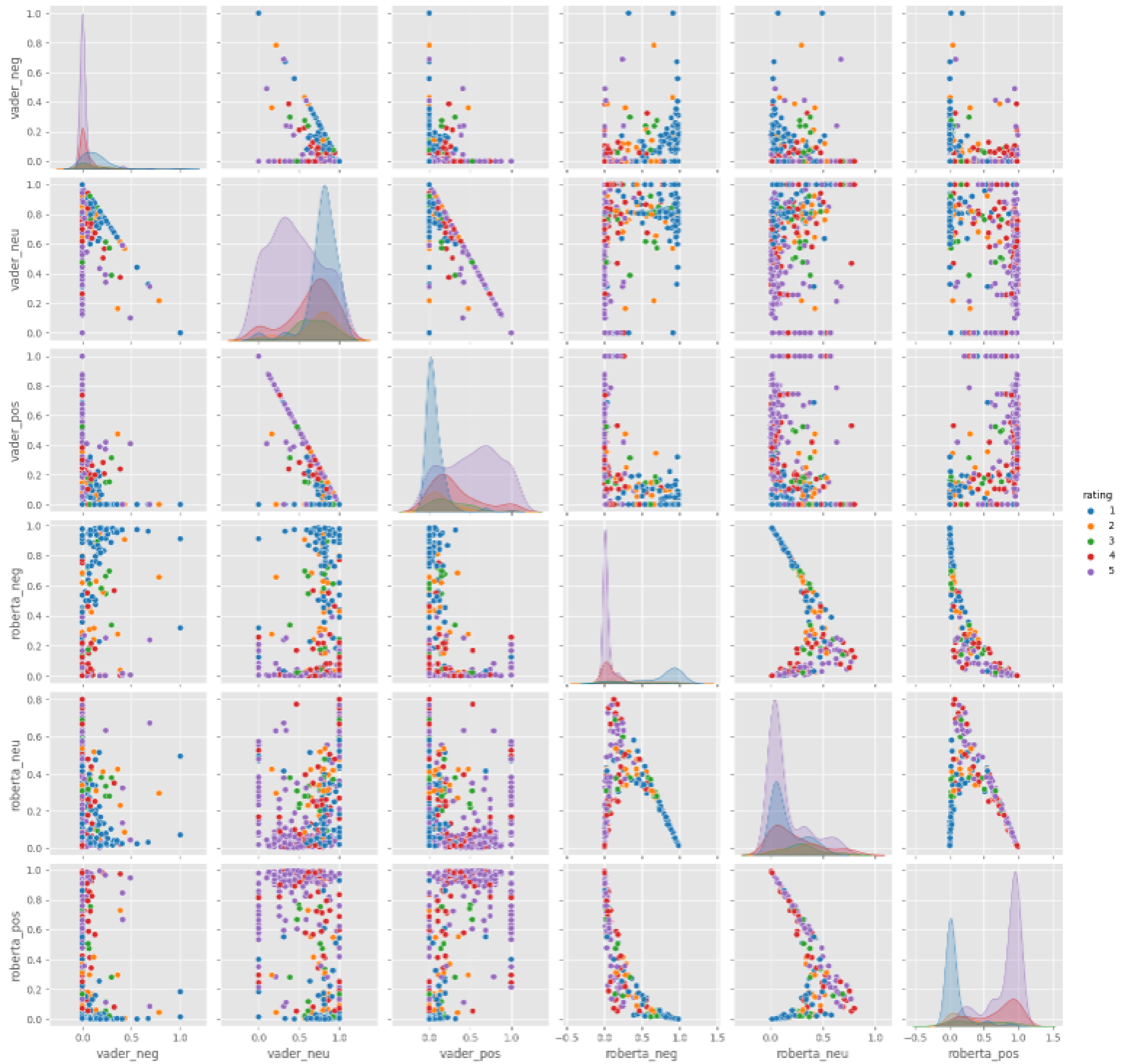

POSITIV

EGATIVE

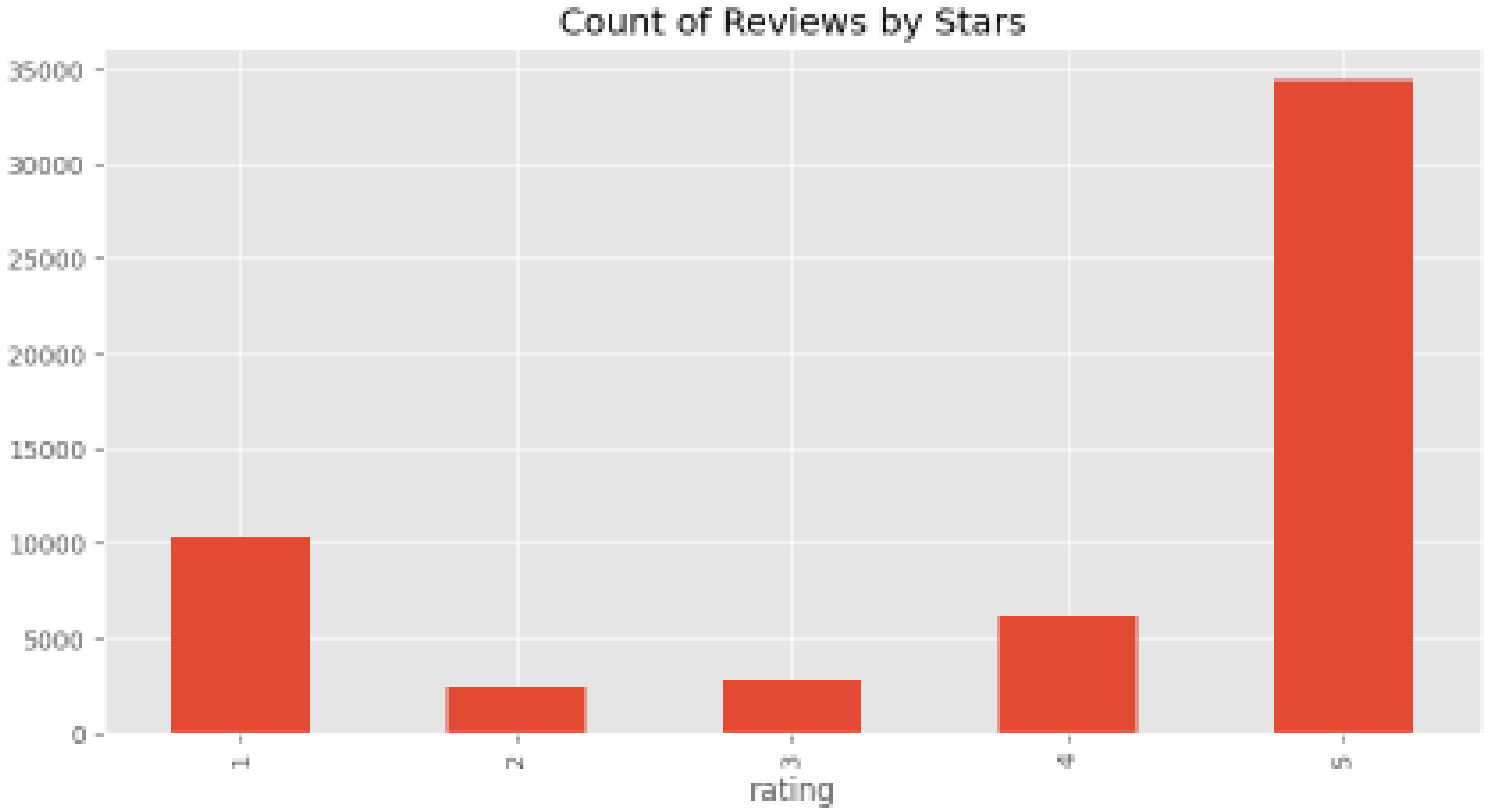
# BLU RESULTS

Count of Reviews by Stars
30000
25000
20000
15000
10000
5000
0
1
2
3
4
5
rating

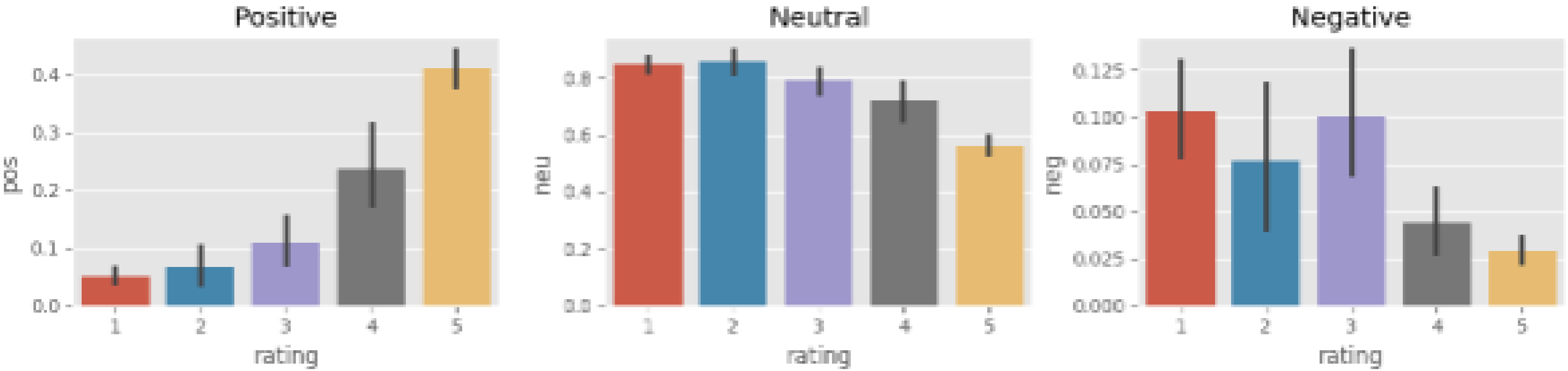

Positive
pos
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
rating
Neutral
neu
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
rating
Negative
neg
0.00
0.05
0.10
0.15
0.20
1
2
3
4
5
rating

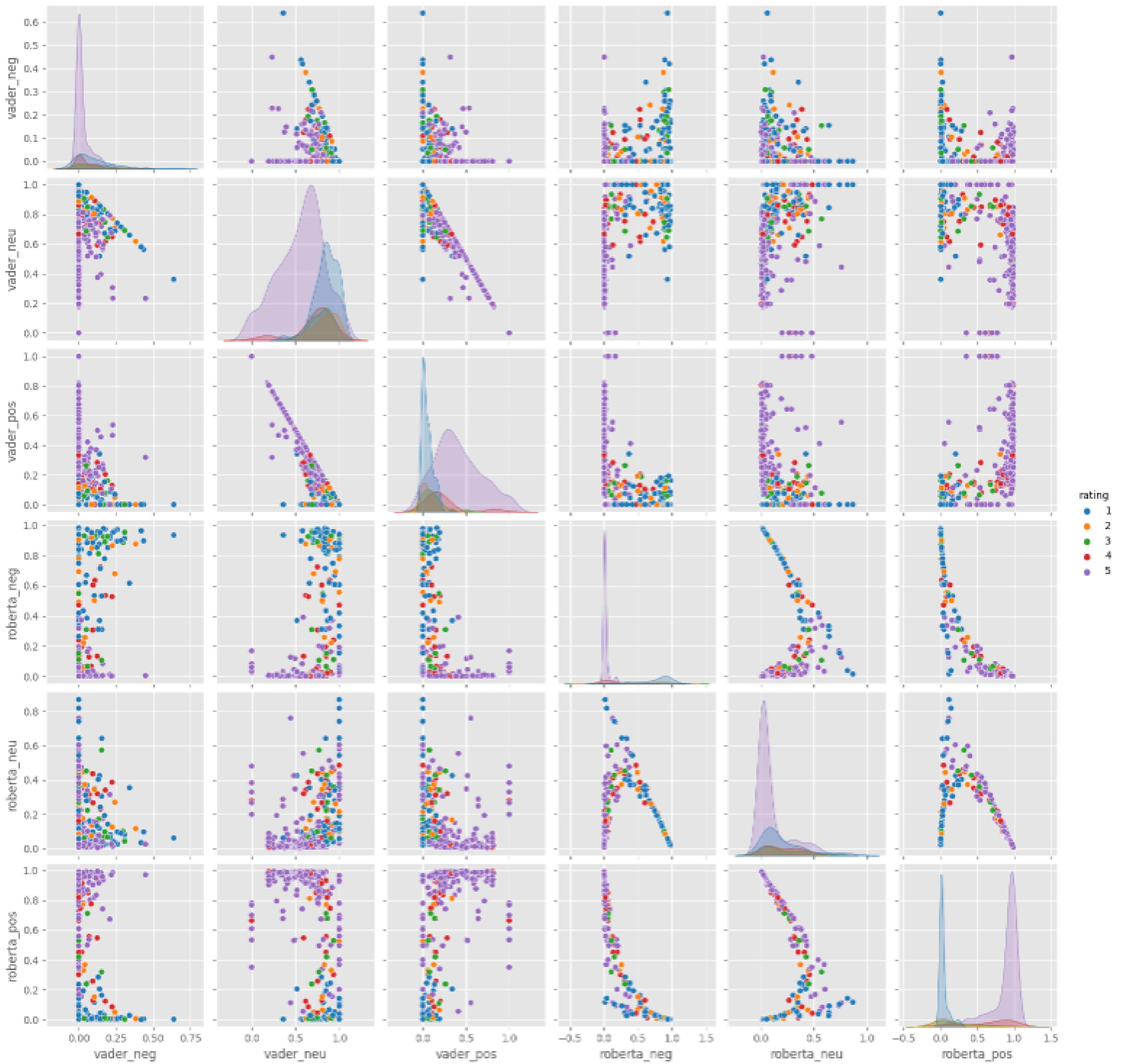

vader_neg
vader_neu
vader_pos
roberta_neg
roberta_neu
roberta_pos
rating
1
2
3
4
5

# APPLE RESULTS



Count of Reviews by Stars
35000
30000
25000
20000
15000
10000
5000
0
1
2
3
4
5
rating


Positive
Neutral
Negative
pos
neu
neg
rating
rating
rating
0.0
0.1
0.2
0.3
0.4
0.0
0.2
0.4
0.6
0.8
0.000
0.025
0.050
0.075
0.100
0.125
1
2
3
4
5



vader_neg
vader_neu
vader_pos
roberta_neg
roberta_neu
roberta_pos
rating
1
2
3
4
5

# VADER

- Valence Aware Dictionary and Sentiment Reasoner.
- It is a lexicon-based model, which means that it uses a dictionary of words and their sentiment scores to calculate the sentiment of a text.
- VADER is relatively simple to use and does not require any training data.
- It is also very fast, which makes it ideal for real-time applications.
- However, VADER is not as accurate as transformer-based models, such as RoBERTa.

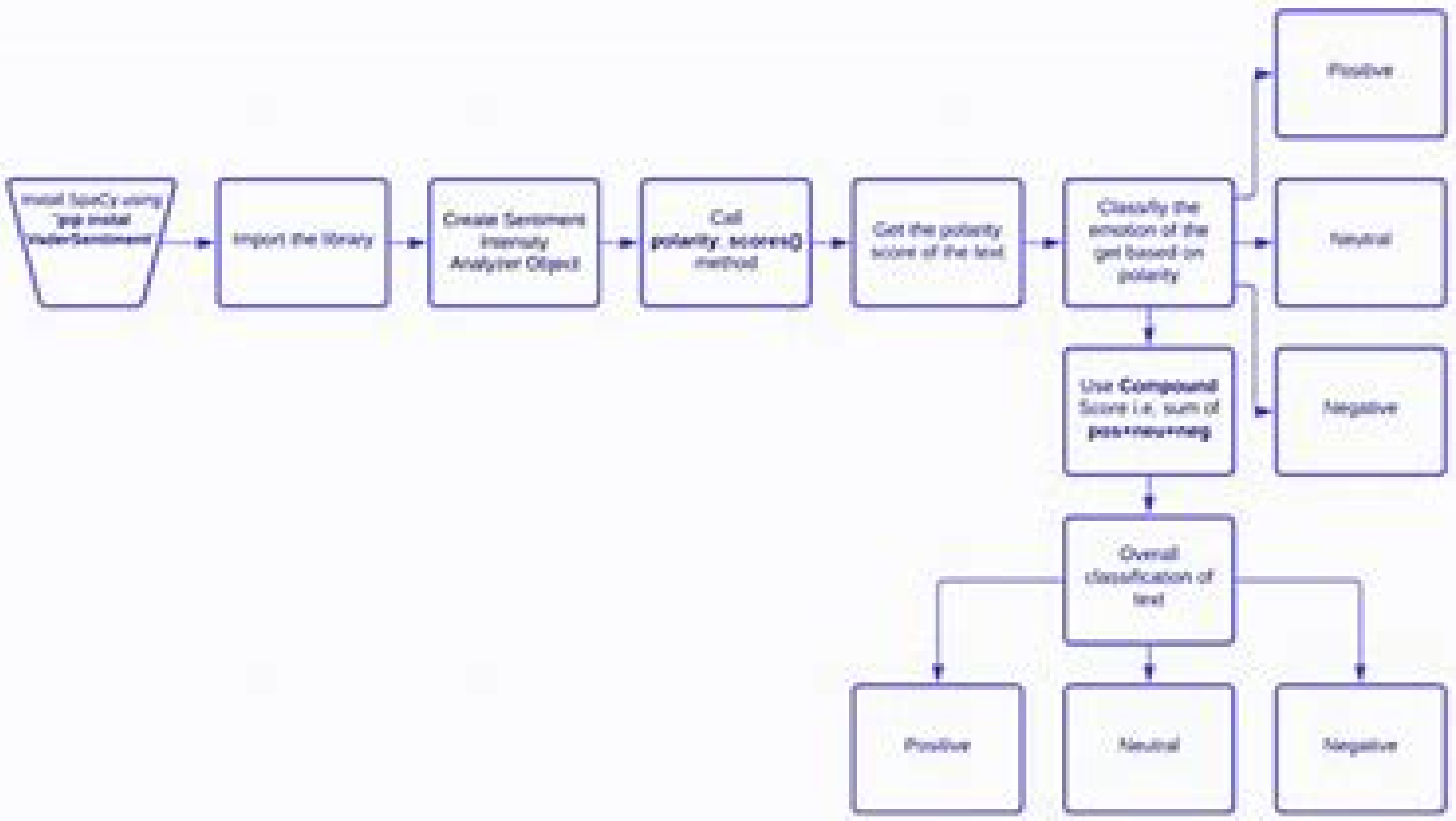
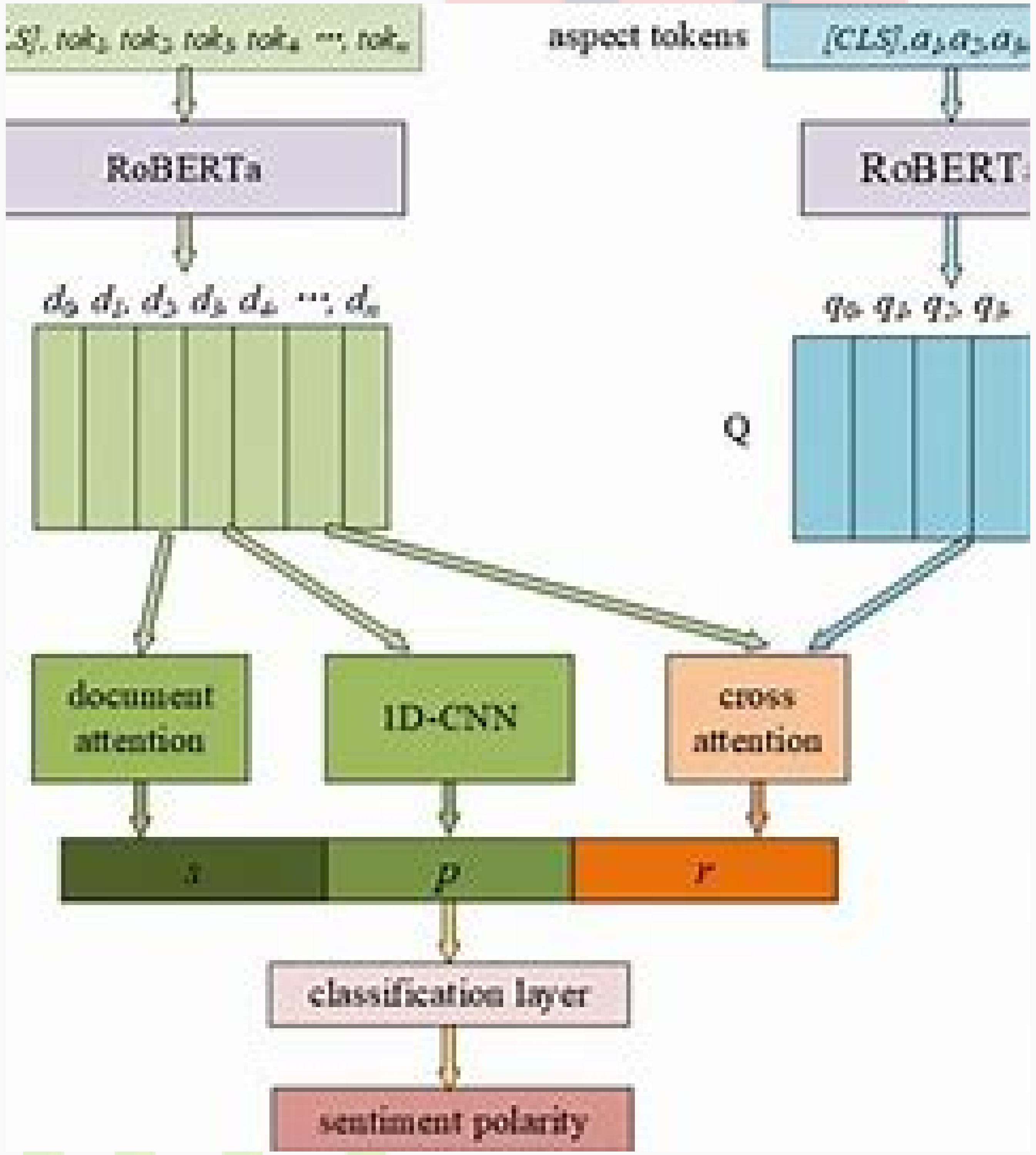


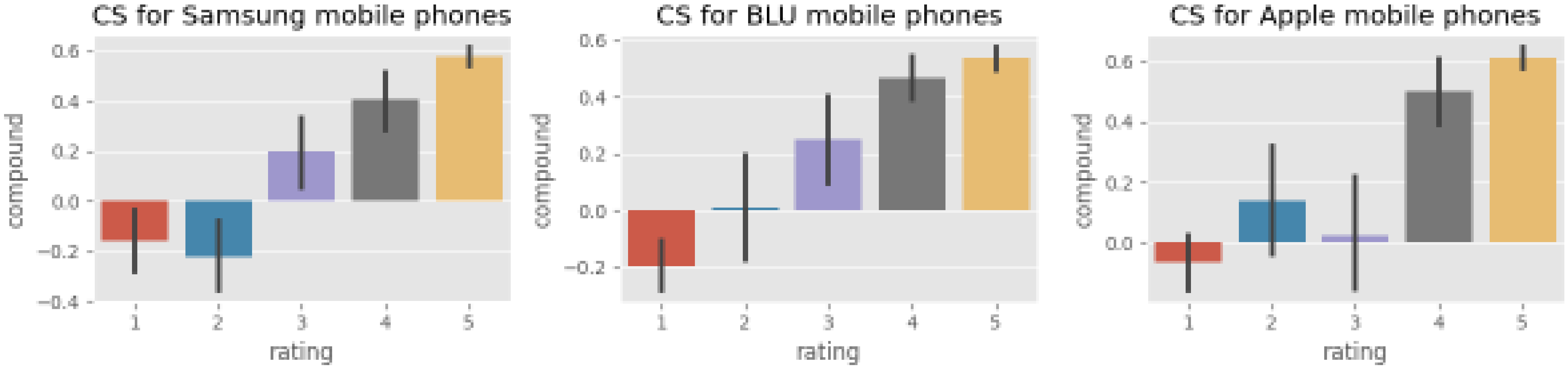

**Fig. Flow Chart**
(Visual Representation of VADER)

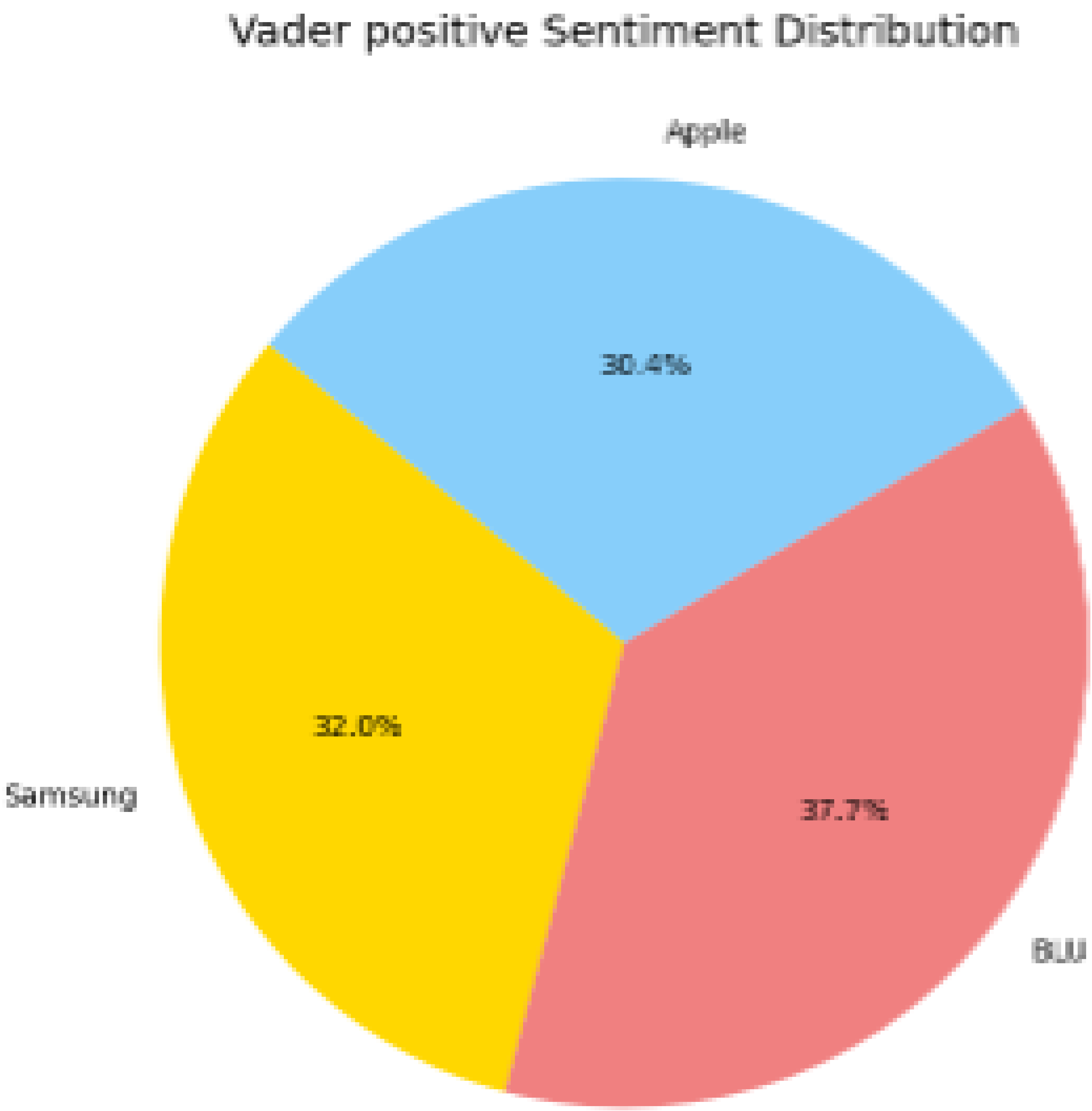
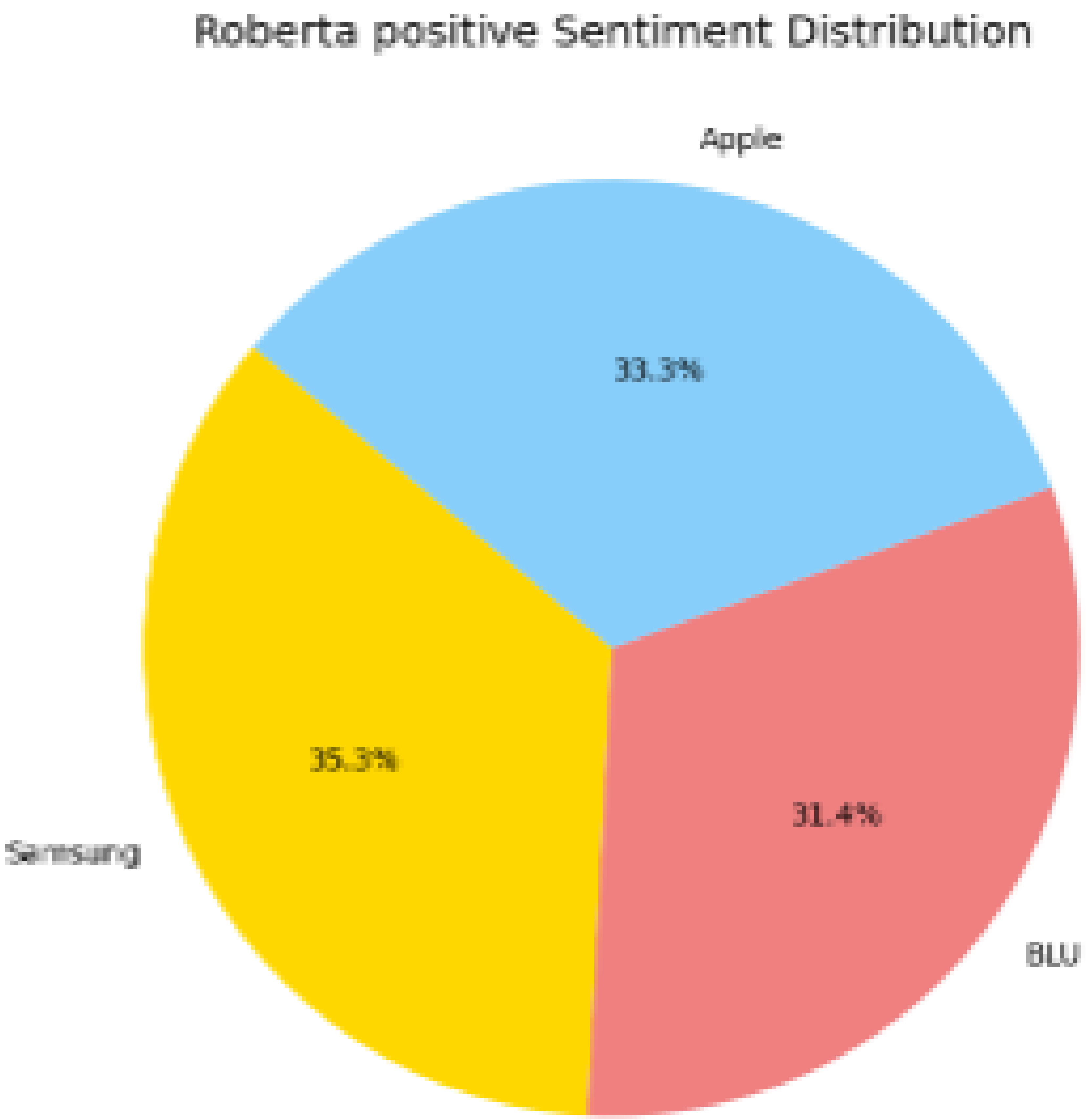
# roBERTa

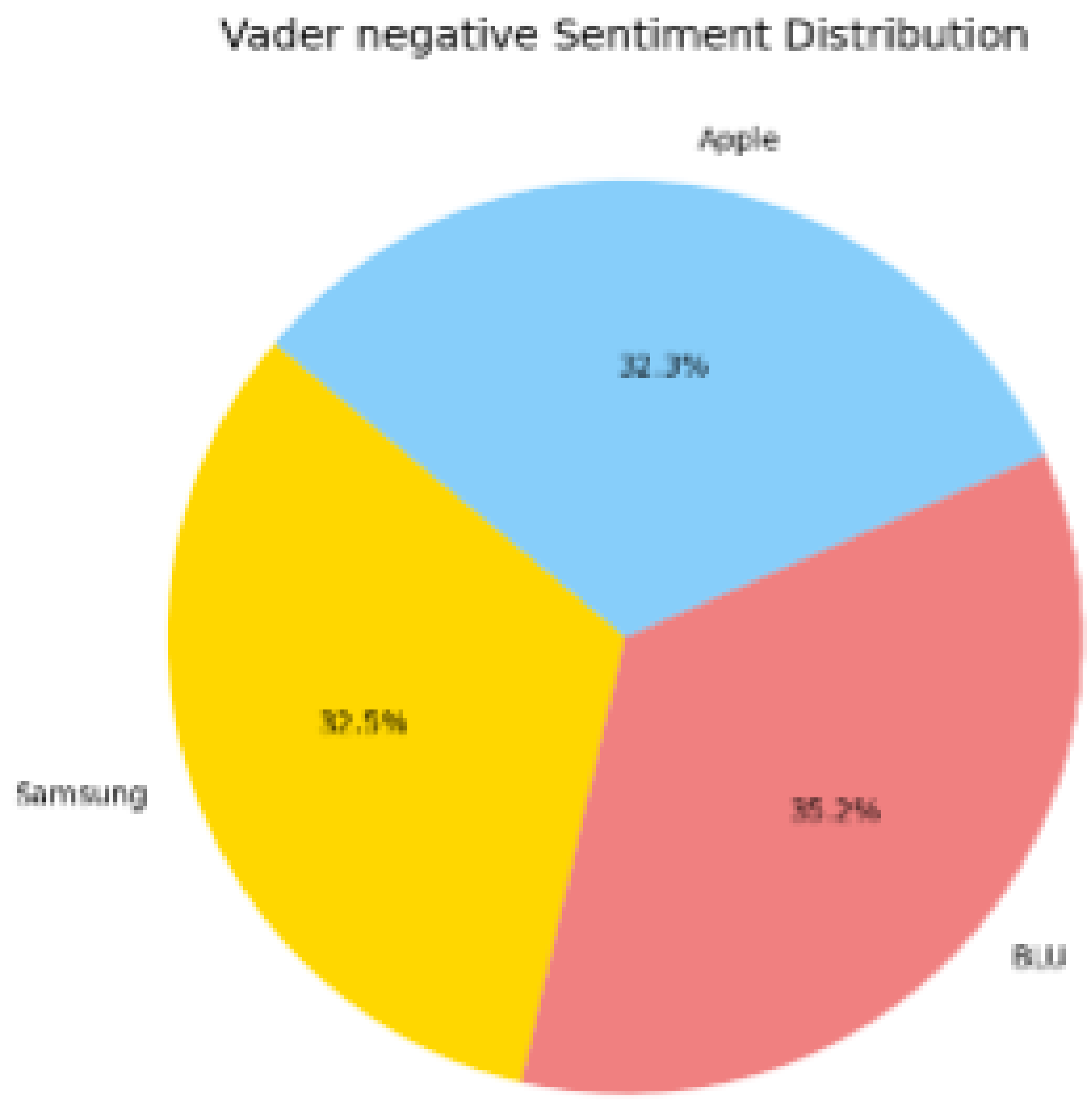
- Robustly Optimized BERT Pretraining Approach.
- It is a transformer-based model, which means that it uses a neural network to learn the relationships between words in a text.
- RoBERTa is more accurate than VADER, but it is also more complex and requires more training data.
- It is also slower than VADER, which makes it less ideal for real-time applications.

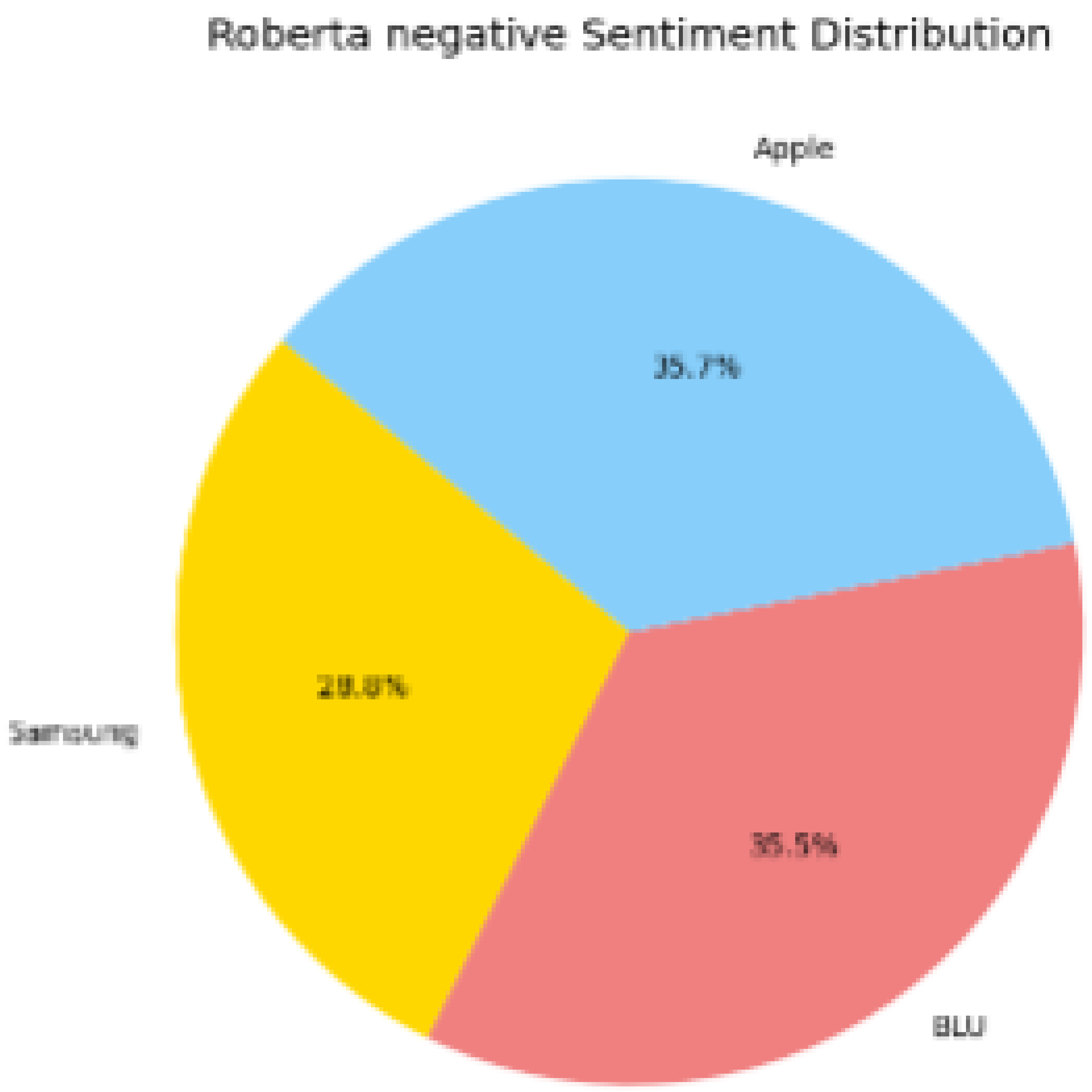


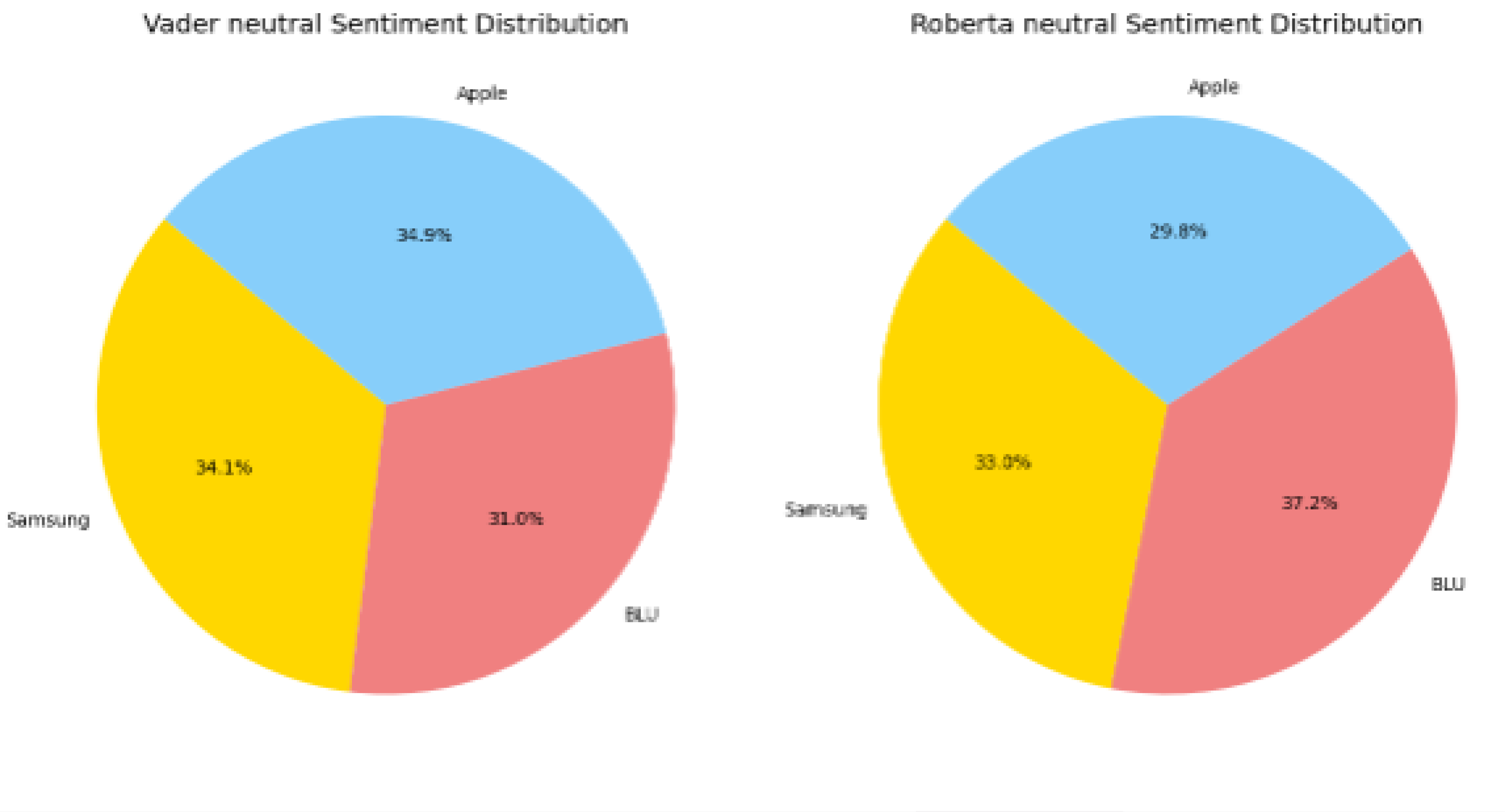

# RESULTS AND INFERENCES


CS for Samsung mobile phones
CS for BLU mobile phones
CS for Apple mobile phones
compound
rating
0.6
0.4
0.2
0.0
−0.2
−0.4
1
2
3
4
5

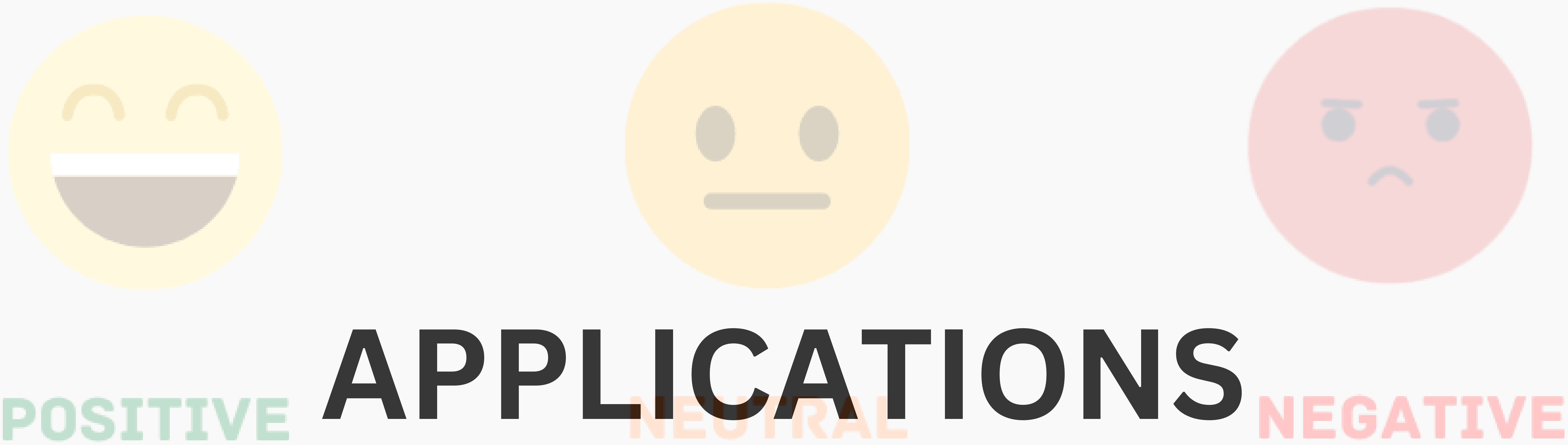

Vader positive Sentiment Distribution
Apple
30.4%
Samsung
32.0%
37.7%
BLU

Roberta positive Sentiment Distribution
Apple
33.3%
Samsung
35.3%
31.4%
BLU


Vader negative Sentiment Distribution
Apple
32.3%
Samsung
32.5%
BLU
35.2%


Roberta negative Sentiment Distribution
Apple
35.7%
Samsung
28.8%
BLU
35.5%


Vader neutral Sentiment Distribution
Apple
34.9%
Samsung
34.1%
31.0%
BLU
Roberta neutral Sentiment Distribution
Apple
29.8%
Samsung
33.0%
37.2%
BLU

# APPLICATIONS



1. **Product Improvement:** Identify areas of concern or satisfaction to improve product quality and features.
2. **Reputation Management:** Monitor and manage a brand's online reputation by addressing customer feedback promptly.
3. **Market Research:** Gain insights into market trends, consumer preferences, and competitive intelligence.
4. **Customer Service:** Respond to negative reviews and provide solutions to enhance customer satisfaction.
5. **Recommendation Systems:** Enhance product recommendations based on user sentiment and preferences.

# LIMITATIONS

1. **Context Understanding:** Sentiment analysis may struggle with sarcasm, irony, or nuanced language, leading to misinterpretations.
2. **Data Bias:** Biased or unrepresentative data can lead to skewed results, especially in the case of fake reviews.
3. **Multilingual Challenges:** Handling sentiment analysis for reviews in multiple languages can be complex.

# POTENTIAL FUTURE

1. **Emotion Detection:** Moving beyond positive/negative sentiment to identify specific emotions in reviews.
2. **Cross-Lingual Analysis:** Improved handling of sentiment in reviews written in various languages.
3. **Ethical Considerations:** Addressing bias, fairness, and privacy concerns in sentiment analysis.
4. **Real-Time Analysis:** Providing businesses with immediate insights for quick decision-making.

# THANK YOU