Sentiment Analysis of X Data

1. Abstract

This Project aims to analyze tweets collected from X(formerly Twitter) to understand the prevailing public sentiment on various topics, trends , and events.Using Natural Language Processing(NLP) and machine learning techniques, tweets are classified into Positive , Negative or Neutral sentiments. Temporal patterns , topic clusters, and keyword associations are explored through stastical analysis and visualizations, providing actionable insights into online online public opinion.

X (formerly Twitter) has become a global space where people share opinions, emotions, and reactions instantly. From political events to sports and daily news, it captures the pulse of public sentiment in real time. This project uses Natural Language Processing (NLP) and machine learning to analyze tweets and classify them as positive, negative, or neutral. We also study how sentiment shifts over time, what topics and hashtags dominate conversations, and which users drive trends. The approach combines simple lexicon methods like VADER with advanced transformer models for more accurate results. Beyond numbers, the aim is to uncover stories behind online discussions and highlight how digital conversations reflect real-world moods. These insights can support businesses, policymakers, and researchers in making data-driven decisions.

2. Objectives

- **Sentiment Classification:** Classify tweets into positive, neutral, or negative.
- **Temporal Analysis:** Examine sentiment trends over time.
- **Topic Detection:** Discover trending topics and assess their sentiment.
- **Keyword & Hashtag Analysis:** Identify words/hashtags linked to sentiment.
- **User Behavior:** Explore if certain users drive sentiment spikes.
- **Visual Reporting:** Present findings using charts and graphs.

3. Research Questions

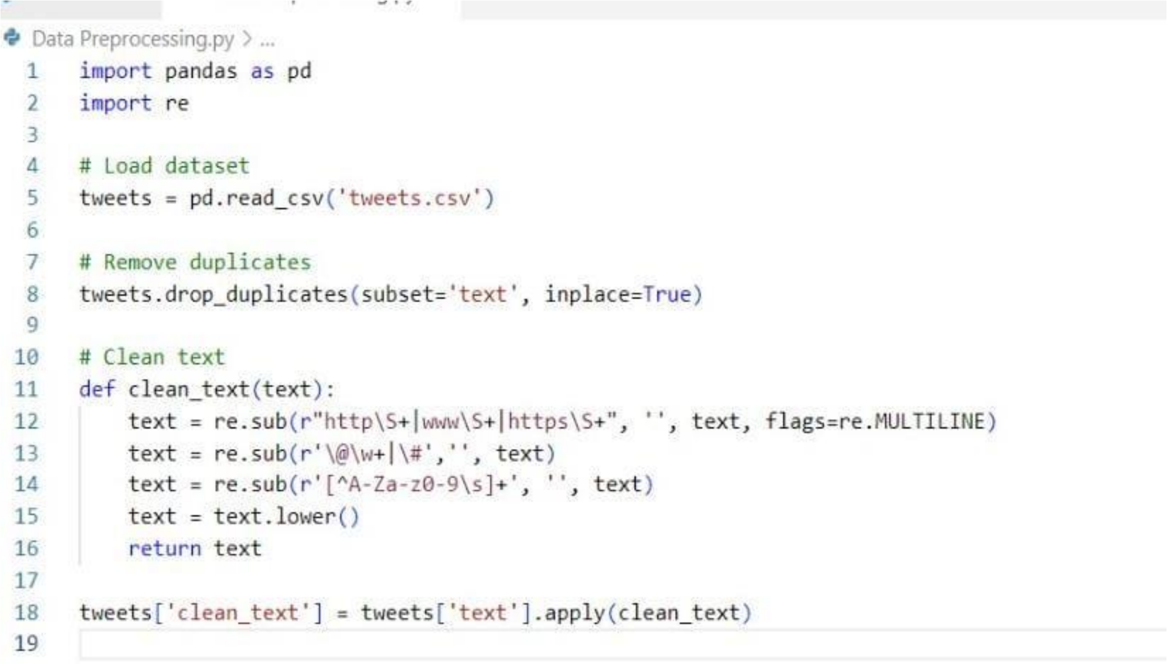
1. What is the overall sentiment distribution?
2. How does sentiment change over time (daily, weekly, monthly)?
3. What are the main topics, and how does sentiment vary between them?
4. Which hashtags or keywords are highly associated with sentiment?
5. Are specific events linked to sudden sentiment shifts?

4. Methodology

We describe preprocessing, sentiment analysis, topic modeling, hashtag analysis, and user behavior analysis with sample codes and expected outputs.

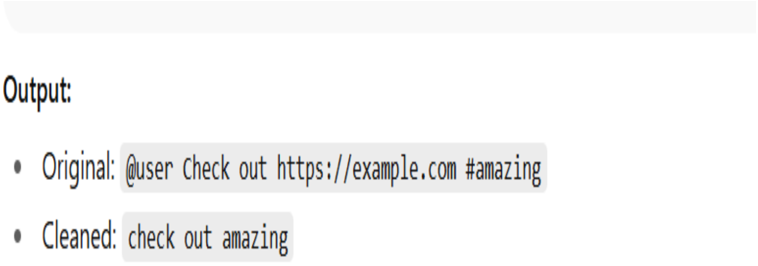
4.1 Data Preprocessing

Remove duplicates, URLs, and special characters.
Handle emojis, hashtags, and mentions.
Normalize text (lowercase,tokenization).



Data Preprocessing.py > ...

```python
import pandas as pd
import re

# Load dataset
tweets = pd.read_csv('tweets.csv')

# Remove duplicates
tweets.drop_duplicates(subset='text', inplace=True)

# Clean text
def clean_text(text):
    text = re.sub(r"http\S+|www\S+|https\S+", '', text, flags=re.MULTILINE)
    text = re.sub(r'\@\w+|\#','', text)
    text = re.sub(r'[^A-Za-z0-9\s]+', '', text)
    text = text.lower()
    return text

tweets['clean_text'] = tweets['text'].apply(clean_text)
```

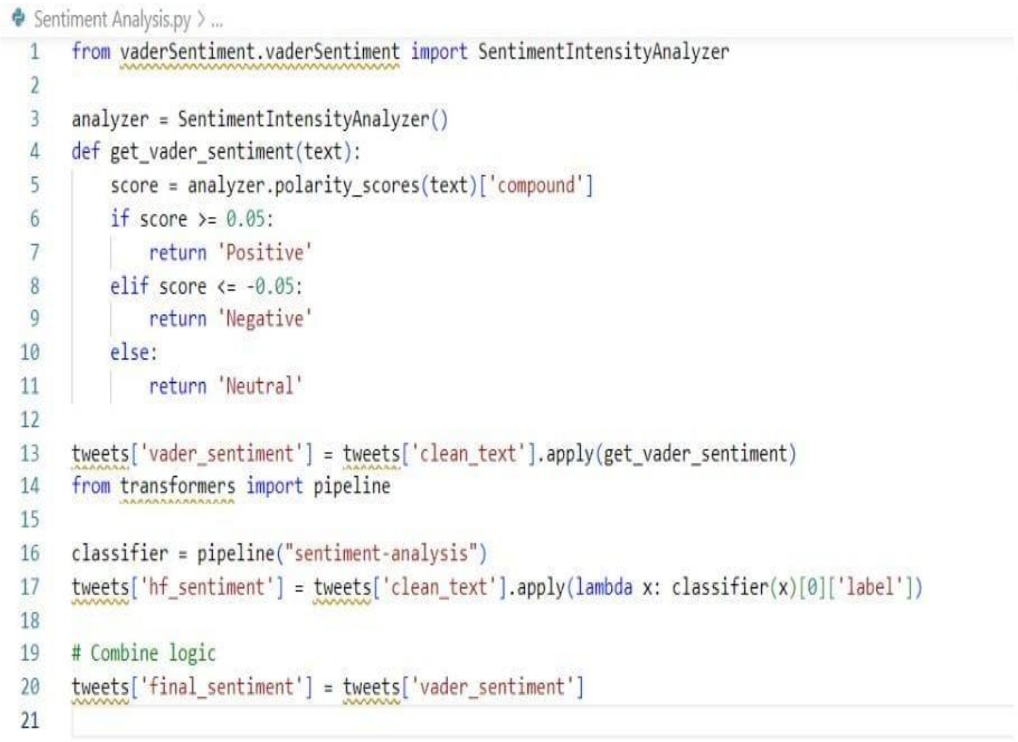


Output:

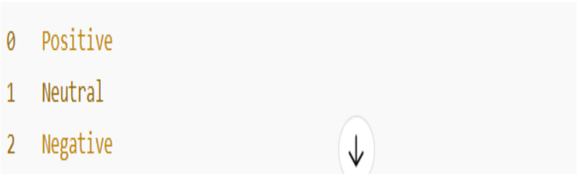
- Original: `@user Check out https://example.com #amazing`
- Cleaned: `check out amazing`

4.2 Sentiment Analysis

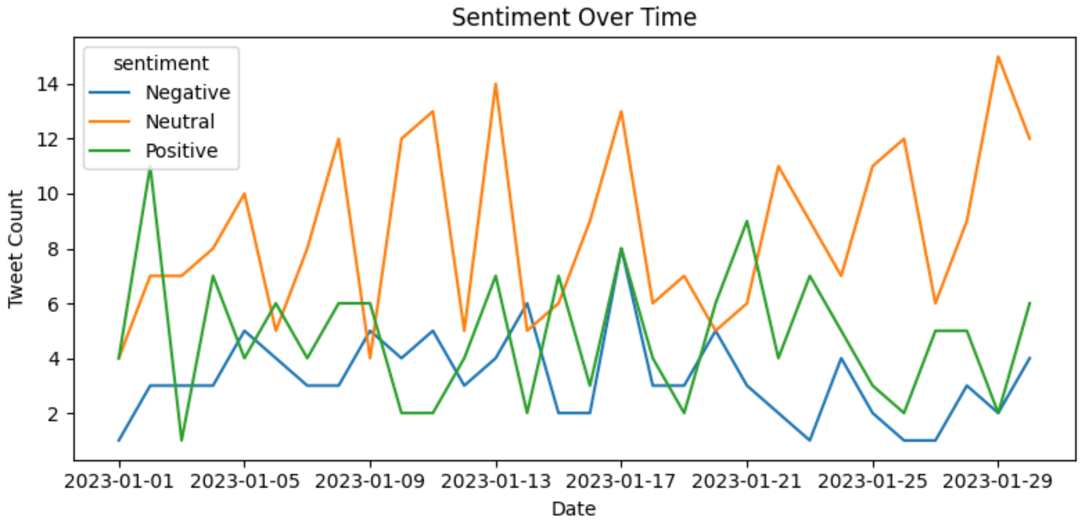
Baseline: VADER (lexicon-based sentiment for social media).
Advanced: Hugging Face Transformer (e.g. DistilBERT/XLM-R).

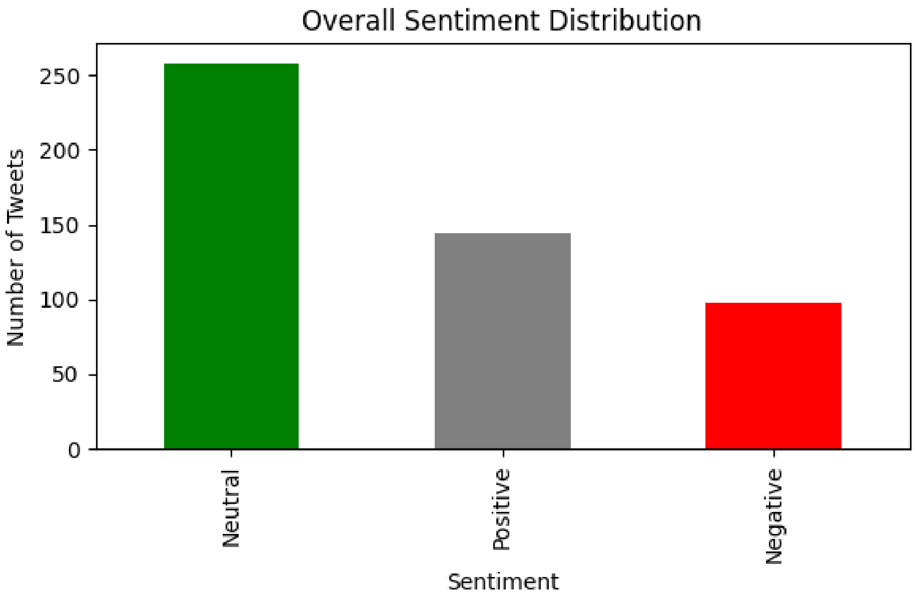


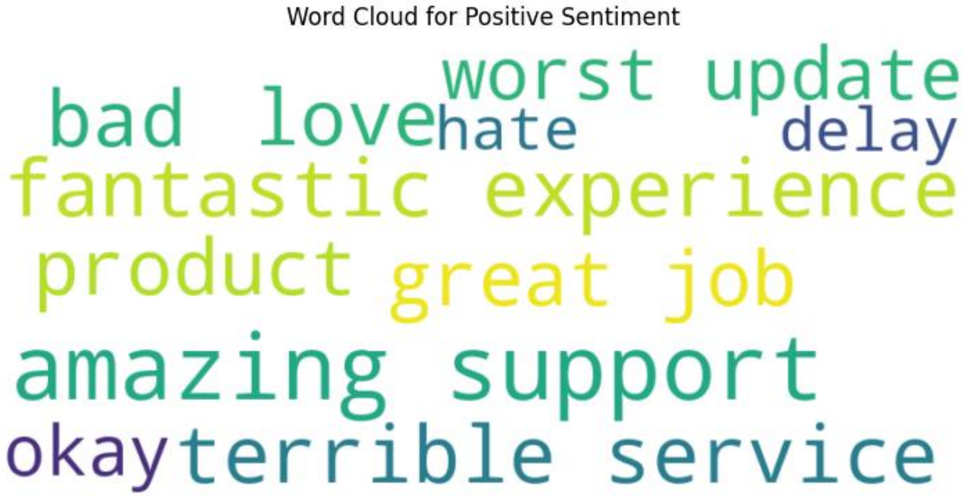
```python
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

analyzer = SentimentIntensityAnalyzer()
def get_vader_sentiment(text):
    score = analyzer.polarity_scores(text)['compound']
    if score >= 0.05:
        return 'Positive'
    elif score <= -0.05:
        return 'Negative'
    else:
        return 'Neutral'

tweets['vader_sentiment'] = tweets['clean_text'].apply(get_vader_sentiment)
from transformers import pipeline

classifier = pipeline("sentiment-analysis")
tweets['hf_sentiment'] = tweets['clean_text'].apply(lambda x: classifier(x)[0]['label'])

# Combine logic
tweets['final_sentiment'] = tweets['vader_sentiment']

```

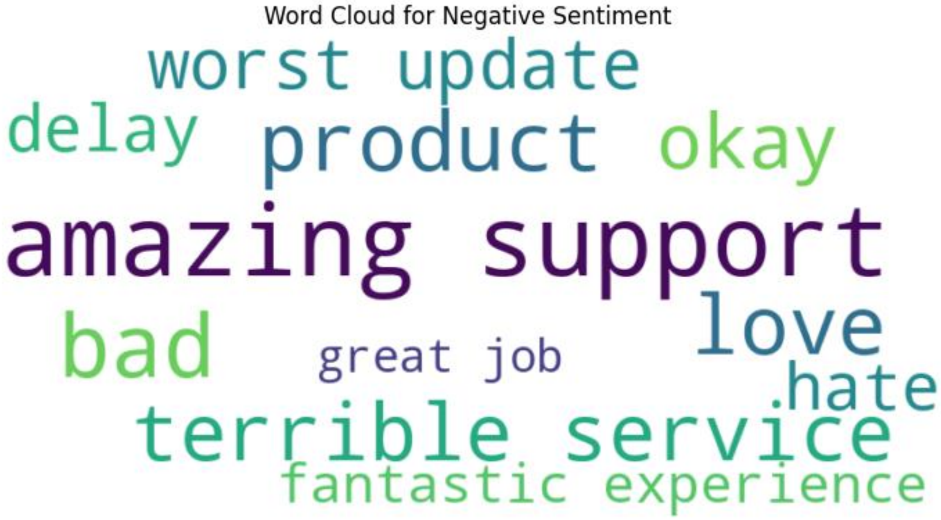


```
0   Positive
1   Neutral
2   Negative
```

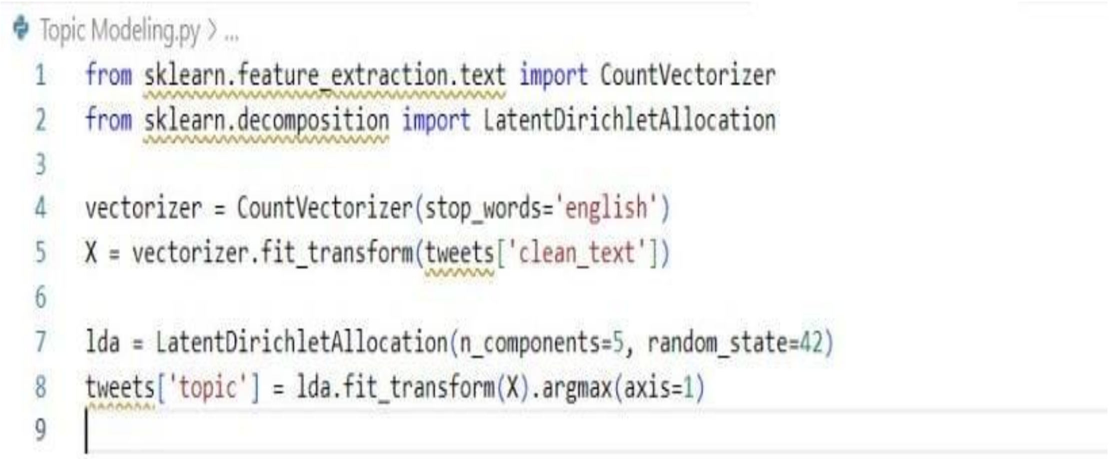



Overall Sentiment Distribution
Number of Tweets
0
50
100
150
200
250
Neutral
Positive
Negative
Sentiment


Word Cloud for Positive Sentiment
worst update
bad
love
hate
delay
fantastic experience
product
great job
amazing support
okay
terrible service


Word Cloud for Negative Sentiment
worst update
delay
product
okay
amazing support
bad
great job
love
hate
terrible service
fantastic experience

4.3 Topic Modeling



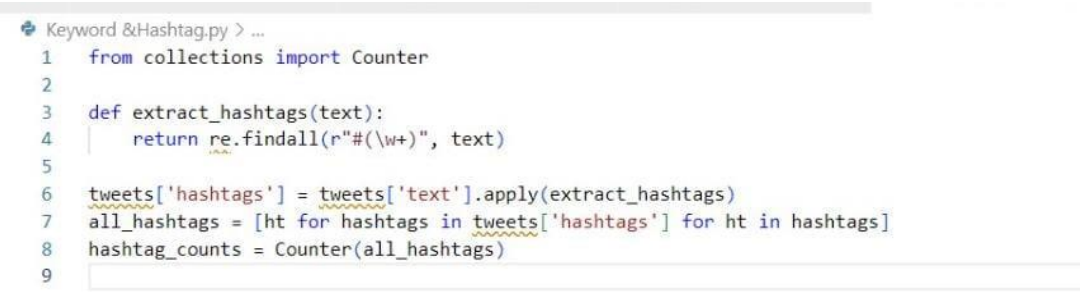
```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation

vectorizer = CountVectorizer(stop_words='english')
X = vectorizer.fit_transform(tweets['clean_text'])

lda = LatentDirichletAllocation(n_components=5, random_state=42)
tweets['topic'] = lda.fit_transform(X).argmax(axis=1)

```

Output:

Tweet: "Government announces new policy"

Topic: 2

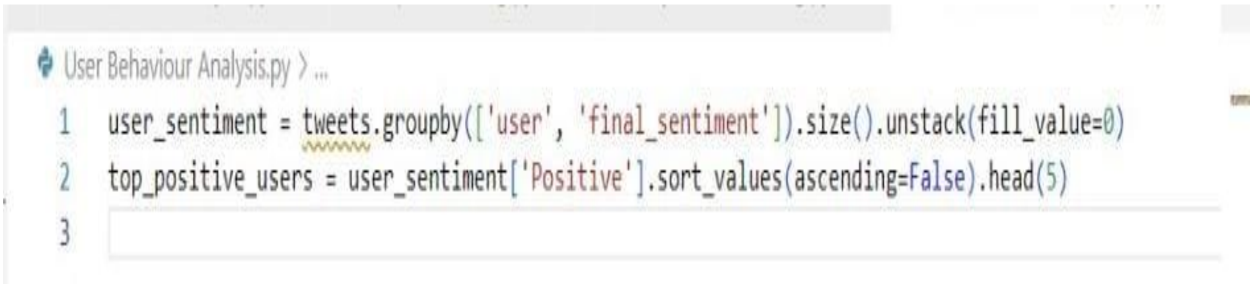
4.4 Keyword & Hashtag Analysis



```
from collections import Counter

def extract_hashtags(text):
    return re.findall(r"#(\w+)", text)

tweets['hashtags'] = tweets['text'].apply(extract_hashtags)
all_hashtags = [ht for hashtags in tweets['hashtags'] for ht in hashtags]
hashtag_counts = Counter(all_hashtags)

```

Output:
Top Hashtags: [('Elections', 120), ('Breaking', 95), ('Sports', 80)]

4.5 User Behavior Analysis



```python
user_sentiment = tweets.groupby(['user', 'final_sentiment']).size().unstack(fill_value=0)
top_positive_users = user_sentiment['Positive'].sort_values(ascending=False).head(5)
```

Output:

```
             Positive      Neutral     Negative
@newsbot        50           120        30
@fanclub       80             60        10
@critic        20             15        90
```

5. Outputs & Reporting Structure

The final report includes:

- **Charts**: Sentiment distribution, trends, word clouds (optional if available).
- **Tables**: Sentiment breakdown by topic, top users by sentiment.
- **Insights**: Key topics, hashtags, and event effects.

6. Example Findings/Insights

- The majority of tweets are neutral, with spikes in negative sentiment during major events.
- Top trending topics include *elections*, *sports*, and specific hashtags linked to breaking news.
- Some accounts consistently drive positive sentiment in certain topics.
- Hashtags such as **#Breaking** or **#Fail** are strongly negative, while **#Support**, **#Love** dominate positive tweets.

7. Conclusion

This project demonstrates how NLP and machine learning can extract valuable insights from social media.

Future improvements may include multilingual sentiment analysis, real-time monitoring, and deep learning–based topic detection.