WICC x Millennium

Social Forum Discourse on Artificial Intelligence and the Effects on Semiconductor Stocks

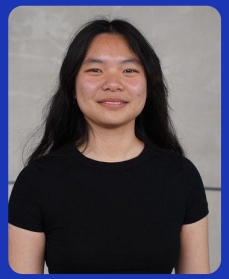
Team 3

April 2025

Table of Contents

Introduction

Meet the (dream) Team!

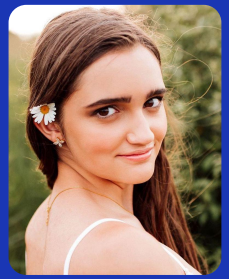


Helen Bian

CS ‘26

Akshika Chawade

CS ‘26



Serena Duncan

CS ‘26

Grace Jin

CS + ECE ‘27

Esha Shah

CS ‘27

Grace Wei

CS ‘26

Natasha Dilamani

Our Mentor

Our Project Timeline

Fall 2024

Team formation

Met with Millennium and learned the scope of the project.

Brainstormed ideas.

Jan. - Feb. 2025

Data Collection

Drafted a preliminary investigation plan and received feedback from our mentor Natasha.

Began collecting stock data and online sentiment data.

Feb. - Mar. 2025

Preliminary Analysis

Continued collecting different types of data.

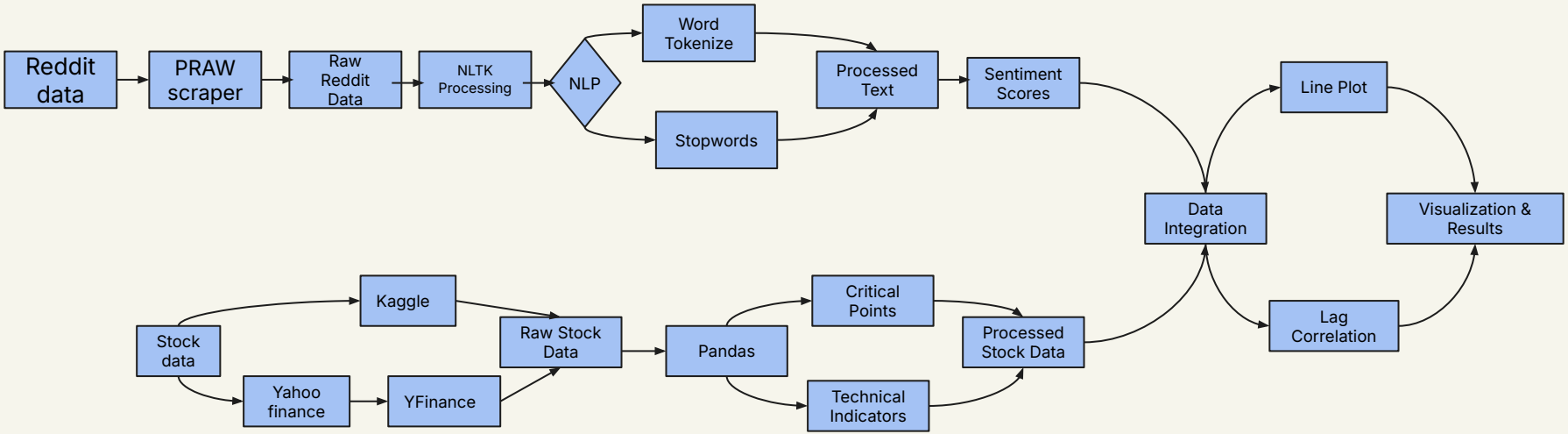
Created programs for sentiment and data analysis.

Apr. 2025

Final Analysis

Collected results from analysis and wrote final conclusions.

Our project architecture



Project Background + Inspiration

- **AI Boom & Market Disruption**
  The rapid growth of AI technologies has reshaped the tech industry, spotlighting companies like NVIDIA, essential to AI hardware and software development.

- **NVIDIA's Rise in Popularity**
  NVIDIA has become a key player in the AI revolution, with its GPUs driving advancements in machine learning, leading to significant stock growth.

- **Major Market Events**
  Notable events, such as the release of **DeepSeek**, caused market disruptions — highlighting the volatility and risks associated with tech stocks.

- **Focus Shift to NVIDIA**
  While the initial goal was to explore sentiment effects on the semiconductor industry, the focus narrowed to NVIDIA due to its prominence in recent market movements.

- **Exploring Sentiment & Stock Correlation**
  Our project investigates the correlation between Reddit sentiment and NVIDIA's stock price, analyzing how online discussions might influence market behavior.

Goals

Sentiment Correlation

Understand the relationship between sentiment and market behavior

Event Detection

Identify sentiment patterns around major market events

Strategy Insight

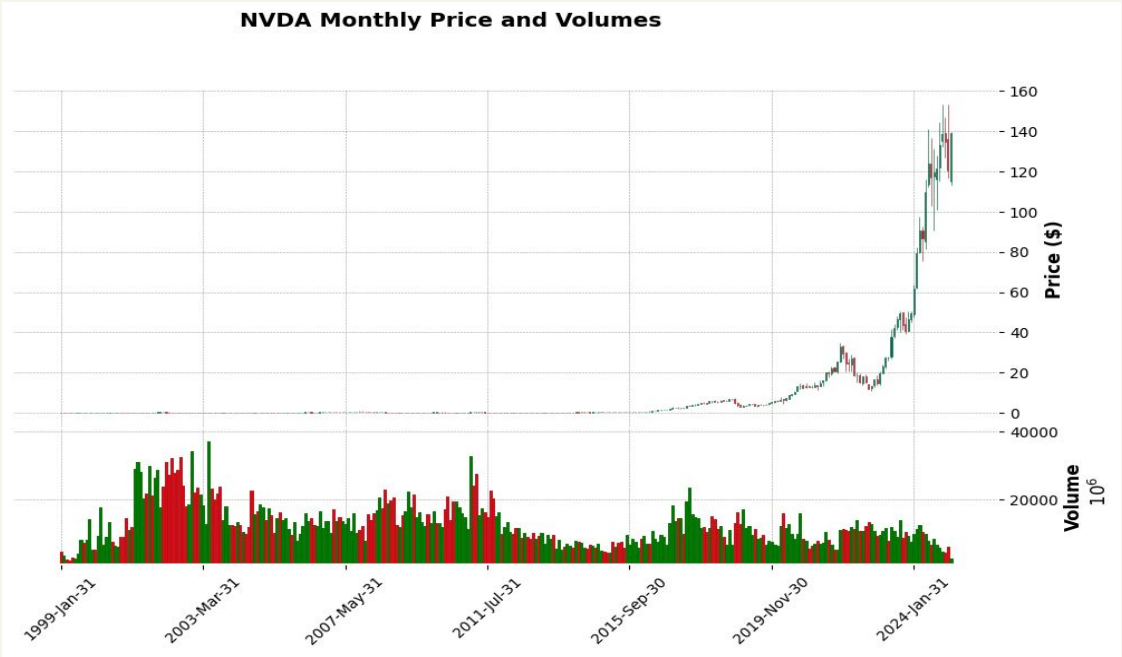
Explore how sentiment trends might inform trading strategies

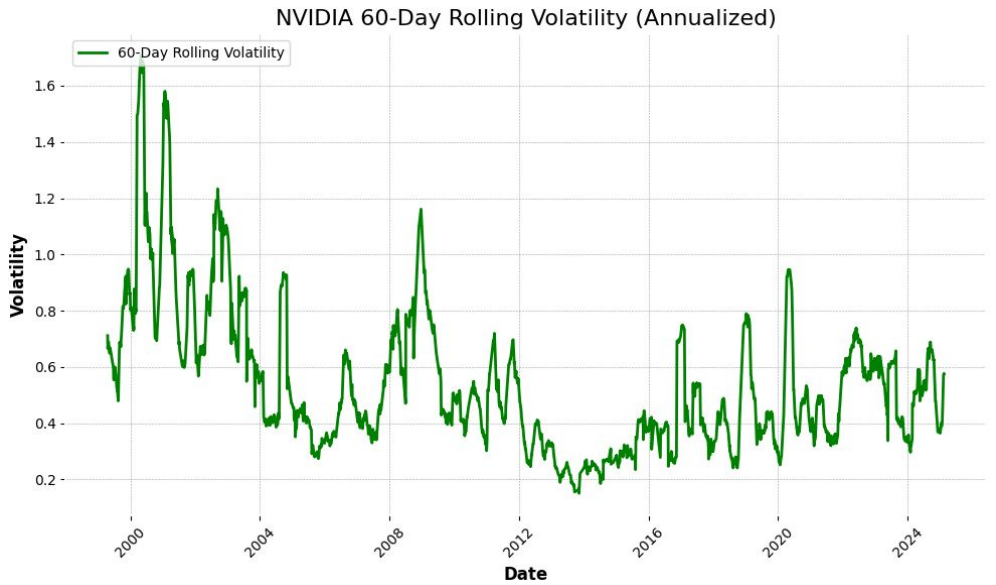
A Look Into NVIDIA:

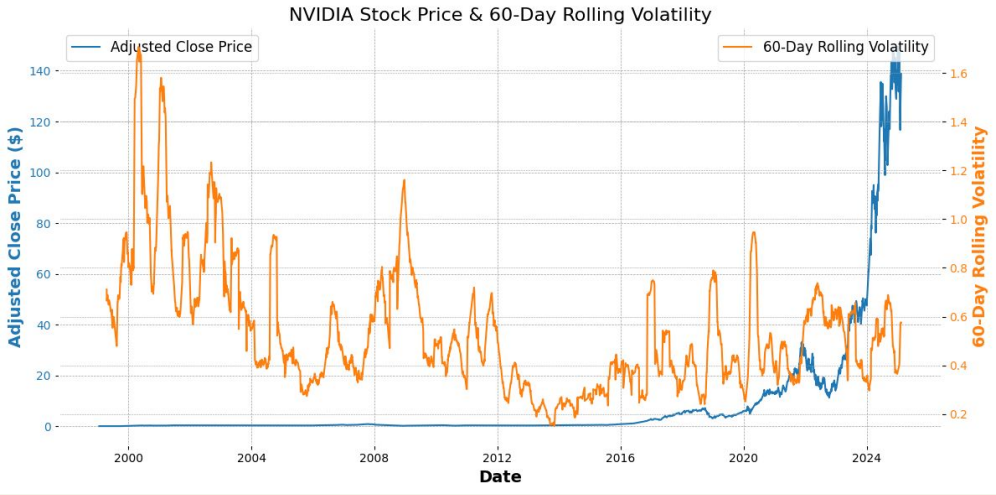
An exploratory data analysis

Observations and Statistics

- The stock has experienced explosive recent growth-dramatic upward momentum, particularly from 2020 to 2024, with the most dramatic rise occurring in the 2023-2024 period.
- Price range: From trading below $20 for many years, the stock has increased to well over $140 by 2024 (approximately 7-8x increase).
- As the price has increased, the stock has shown greater price volatility, with larger daily and weekly price swings.







Observations

- Volatility peaks in the early 2000s, but this could be due to the very low market price of the stocks during the time
  - Also coincides with the “Dot-Com Bubble burst” era of the American economy
- Spikes after 2008, 2020
- General downward trend
  - Opposite of upward growth in market price

Narrowing Our Vision: 2020-2025

- AI Boom and the Rise of NVIDIA's Market Influence (2020s)
- More recent and relevant data to understand market behavior
- NVDA stock started dramatically increasing around 2020, which attracted more attention from investors and the public eye, which led for more commentary that we could analyze.
- This time frame encompasses multiple market cycles, including the COVID-19 market crash (March 2020), recovery, which results in diverse data points
- 2020-2025 marks the transition period when NVIDIA shifted from being primarily viewed as a gaming/crypto mining company to becoming the dominant AI computing platform, fundamentally changing investor perceptions.

Market Sentiment

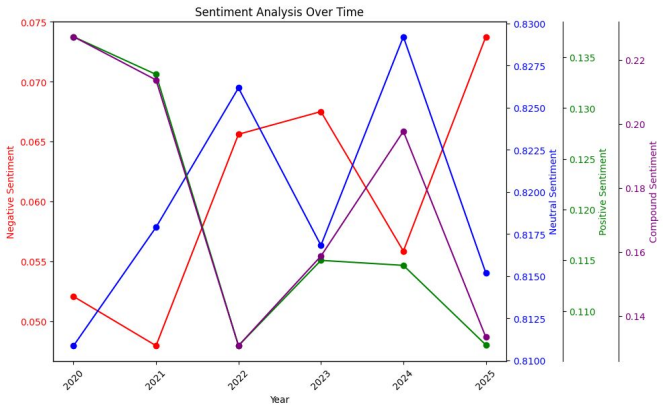
Data Source: Reddit

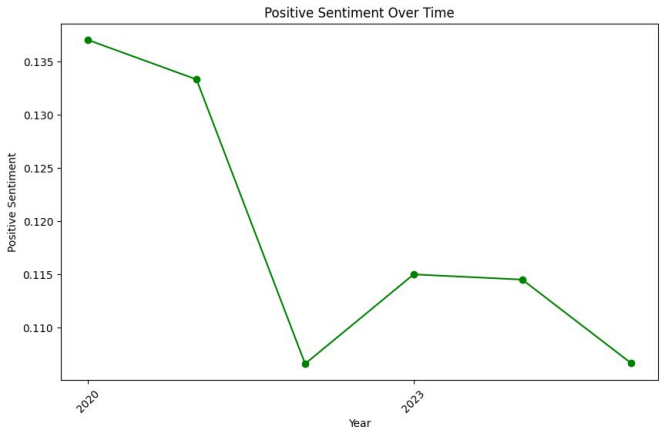
- Data collected from Reddit posts across subreddits r/stocks and r/investing from the years 2020-2025
- We scraped Reddit with PRAW (Python Reddit API Wrapper) with target keyword filtering (“nvidia”,”nvda”) and post metadata
- We chose Reddit for its less-filtered sentiment that isn’t available in traditional financial news sources

```python
import praw
import pandas as pd
import datetime
import time
from tqdm import tqdm

def scrape_subreddit(reddit, subreddit_name, keywords, start_date, end_date, limit=500):


    start_timestamp = int(start_date.timestamp())
    end_timestamp = int(end_date.timestamp())

    subreddit = reddit.subreddit(subreddit_name)
    posts = []

    query = " OR ".join(keywords)
    print(f"Searching for: {query}")

    search_count = 0


    for post in tqdm(subreddit.search(query, sort='new', time_filter="all", limit=None),
                     desc="Searching posts", unit="post"):
        search_count += 1


        if len(posts) >= limit:
            break


        if start_timestamp <= post.created_utc <= end_timestamp:
            title_lower = post.title.lower()
            selftext_lower = post.selftext.lower() if hasattr(post, 'selftext') else ""
            matched_keywords = []
            for keyword in keywords:
                keyword_lower = keyword.lower()
                if keyword_lower in title_lower or keyword_lower in selftext_lower:
                    matched_keywords.append(keyword)

            if matched_keywords:
                posts.append({
                    'post_id': post.id,
                    'title': post.title,
                    'author': str(post.author) if post.author else "[deleted]",
                    'created_utc': post.created_utc,
                    'date': datetime.datetime.fromtimestamp(post.created_utc).strftime('%Y-%m-%d %H:%M:%S'),
                    'score': post.score,
                    'num_comments': post.num_comments,
                    'permalink': f"https://www.reddit.com{post.permalink}",
                    'selftext': post.selftext if hasattr(post, 'selftext') else "",
                    'matched_keywords': ", ".join(matched_keywords)
                })
                print(f"Found matching post from {datetime.datetime.fromtimestamp(post.created_utc).strftime('%Y-%m-%d')}")
        if search_count % 10 == 0:  # Every 10 posts
            time.sleep(1)  # sleep to avoid hitting rate limits

    print(f"Collected {len(posts)} posts matching the criteria")
```

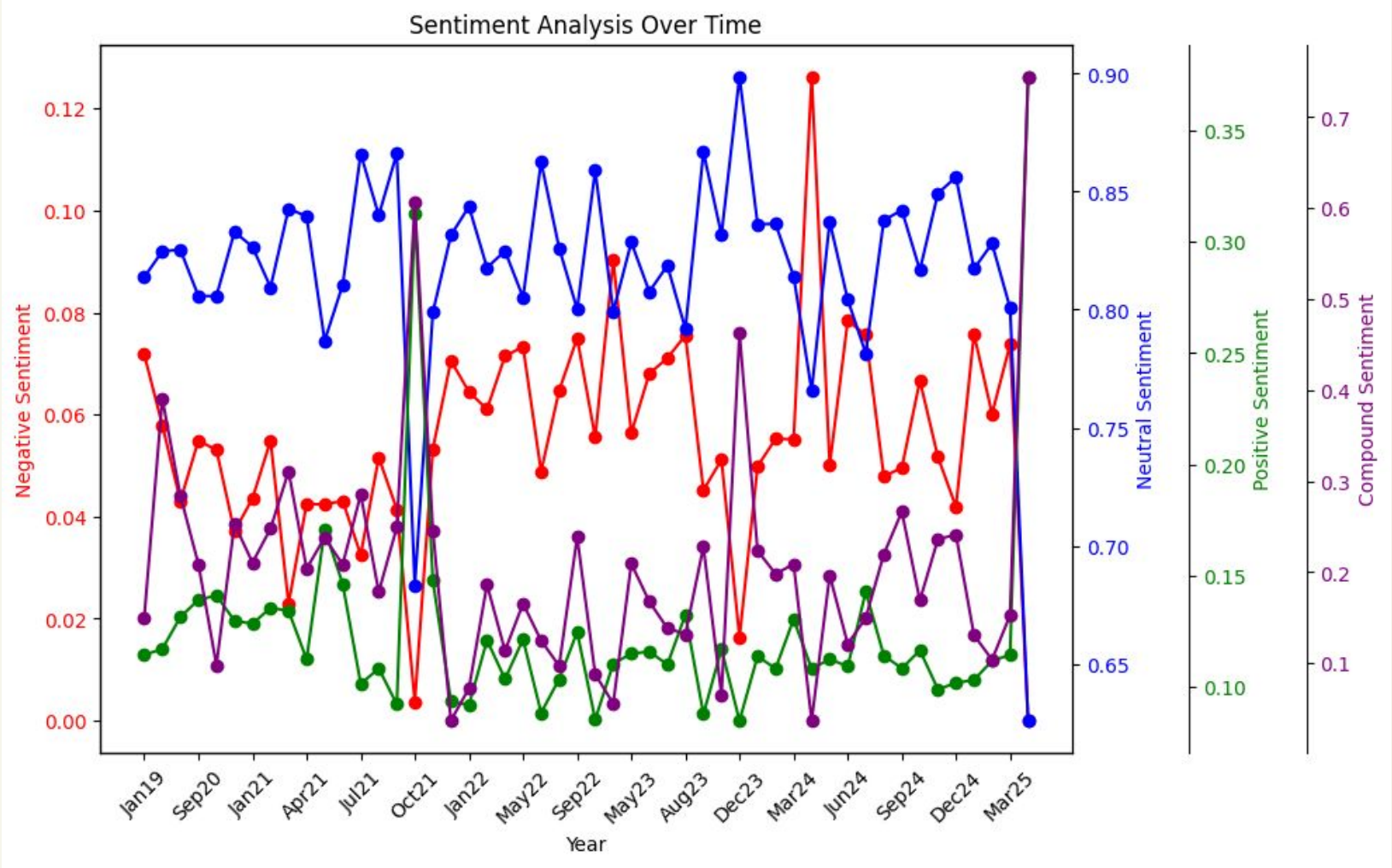
Critical Part in Data Collection Script

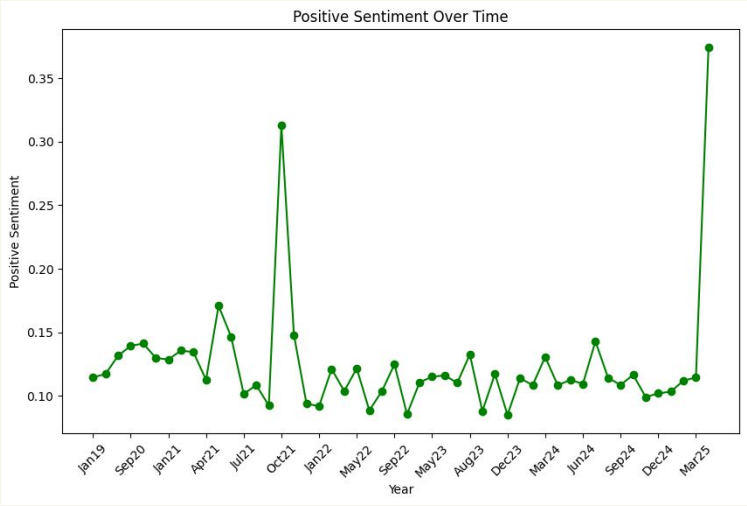
Sentiment Over Time

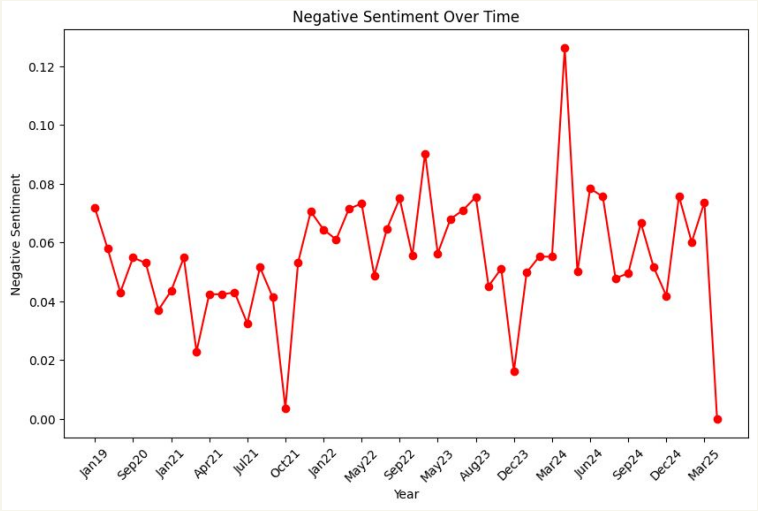


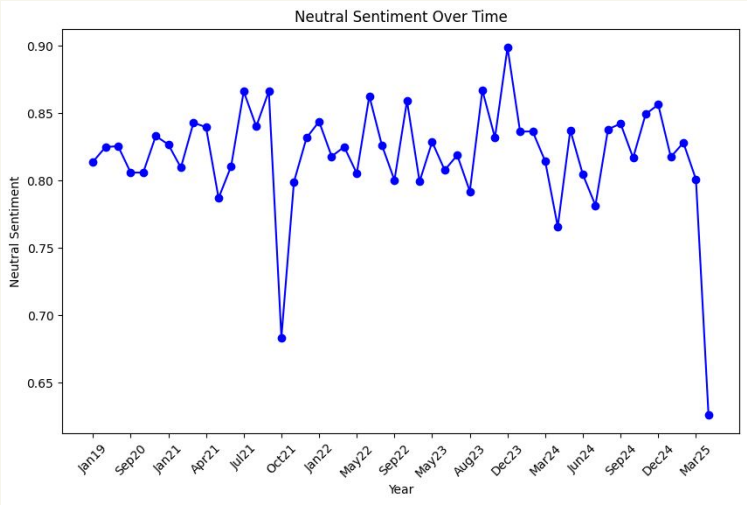


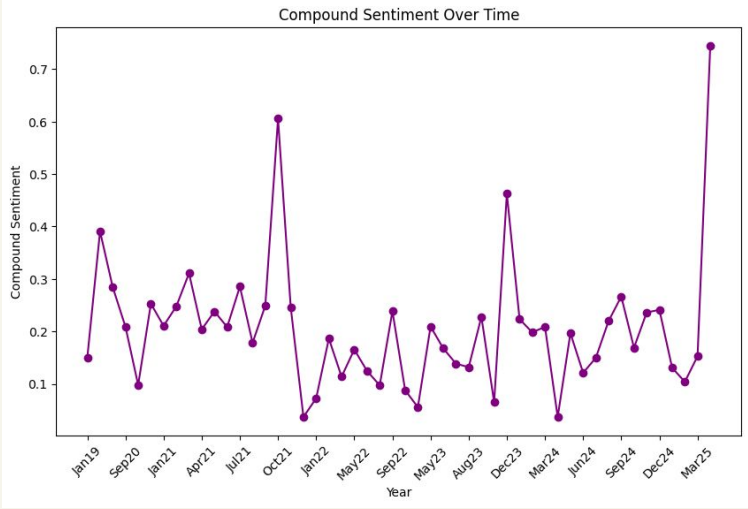
- Used NLTK to do the following (2021-2025):
  - Find average sentiment across years
  - Find top 15 words across years
  - Find average sentiment per word
- Why NLTK?
  - Easy-to-use sentiment analysis tool
  - Built-in tools for:
    - Stopword removal
    - Tokenization
    - Sentiment scoring
  - *Limitation*: Basic sentiment analysis - might miss context or sarcasm
- What did we find?
  - Positive sentiment around NVIDIA highest in 2020
    - Drops **significantly** from 2020-2022
  - Slowly recovers through 2023-2024 but never reached 2020 levels again
  - Drops again in 2025


Sentiment Analysis Over Time
Negative Sentiment
Neutral Sentiment
Positive Sentiment
Compound Sentiment
Year
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.65
0.70
0.75
0.80
0.85
0.90
0.10
0.15
0.20
0.25
0.30
0.35
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Jan19
Sep20
Jan21
Apr21
Jul21
Oct21
Jan22
May22
Sep22
May23
Aug23
Dec23
Mar24
Jun24
Sep24
Dec24
Mar25


Positive Sentiment Over Time
Positive Sentiment
0.35
0.30
0.25
0.20
0.15
0.10
Jan19
Sep20
Jan21
Apr21
Jul21
Oct21
Jan22
May22
Sep22
May23
Aug23
Dec23
Mar24
Jun24
Sep24
Dec24
Mar25
Year

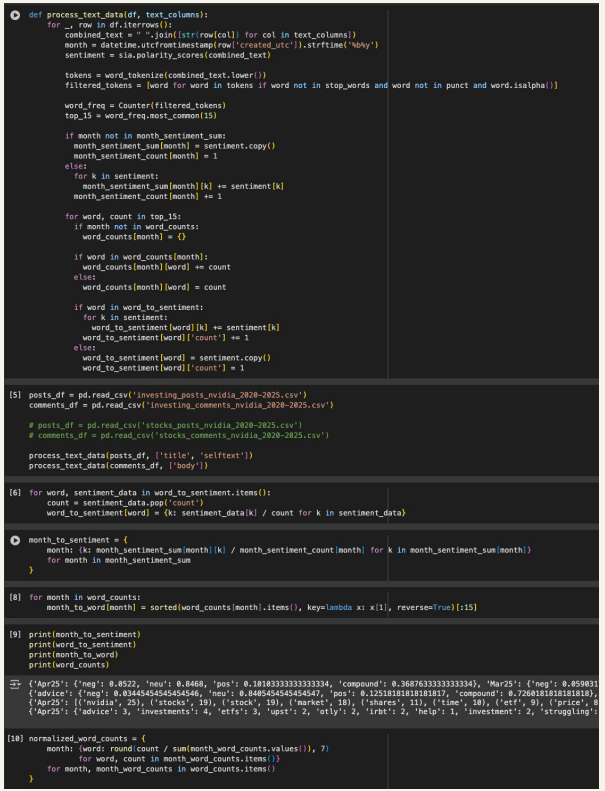

Negative Sentiment Over Time
Negative Sentiment
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Jan19
Sep20
Jan21
Apr21
Jul21
Oct21
Jan22
May22
Sep22
May23
Aug23
Dec23
Mar24
Jun24
Sep24
Dec24
Mar25
Year


Neutral Sentiment Over Time
Neutral Sentiment
0.90
0.85
0.80
0.75
0.70
0.65
Jan19
Sep20
Jan21
Apr21
Jul21
Oct21
Jan22
May22
Sep22
May23
Aug23
Dec23
Mar24
Jun24
Sep24
Dec24
Mar25
Year


Compound Sentiment Over Time
Compound Sentiment
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Jan19
Sep20
Jan21
Apr21
Jul21
Oct21
Jan22
May22
Sep22
May23
Aug23
Dec23
Mar24
Jun24
Sep24
Dec24
Mar25
Year

Prevalence of AI

- What we looked at:
  - Top 15 most common words in r/stock or r/investing reddit posts & comments about NVIDIA (2021 - 2025)
- Key trends we saw:
  - **Early years (2021-2022)** → Conversations dominated by:
    - *Stock splits*, *shares*, *crypto*, *gaming*, *mining*
  - **Starting in 2023** → Clear rise of *AI* mentions
    - "*AI*" enters top words in 2023
    - Becomes **2nd most common** word by 2024
    - **Most common** word by 2025
- Other Emerging Keywords:
  - *Deepseek* (AI-related) appears in 2025
  - Consistent growth of *NVIDIA* mentions over time (likely tied to AI leadership)



```
def process_text_data(df, text_columns):
    for _, row in df.iterrows():
        combined_text = " ".join([str(row[col]) for col in text_columns])
        month = datetime.utcfromtimestamp(row['created_utc']).strftime('%b%y')
        sentiment = sia.polarity_scores(combined_text)

        tokens = word_tokenize(combined_text.lower())
        filtered_tokens = [word for word in tokens if word not in stop_words and word not in punct and word.isalpha()]

        word_freq = Counter(filtered_tokens)
        top_15 = word_freq.most_common(15)

        if month not in month_sentiment_sum:
          month_sentiment_sum[month] = sentiment.copy()
          month_sentiment_count[month] = 1
        else:
          for k in sentiment:
            month_sentiment_sum[month][k] += sentiment[k]
          month_sentiment_count[month] += 1

        for word, count in top_15:
          if month not in word_counts:
            word_counts[month] = {}

          if word in word_counts[month]:
            word_counts[month][word] += count
          else:
            word_counts[month][word] = count

          if word in word_to_sentiment:
            for k in sentiment:
              word_to_sentiment[word][k] += sentiment[k]
            word_to_sentiment[word]['count'] += 1
          else:
            word_to_sentiment[word] = sentiment.copy()
            word_to_sentiment[word]['count'] = 1
```

```
[5] posts_df = pd.read_csv('investing_posts_nvidia_2020-2025.csv')
    comments_df = pd.read_csv('investing_comments_nvidia_2020-2025.csv')

    # posts_df = pd.read_csv('stocks_posts_nvidia_2020-2025.csv')
    # comments_df = pd.read_csv('stocks_comments_nvidia_2020-2025.csv')

    process_text_data(posts_df, ['title', 'selftext'])
    process_text_data(comments_df, ['body'])
```

```
[6] for word, sentiment_data in word_to_sentiment.items():
        count = sentiment_data.pop('count')
        word_to_sentiment[word] = {k: sentiment_data[k] / count for k in sentiment_data}
```

```
month_to_sentiment = {
    month: {k: month_sentiment_sum[month][k] / month_sentiment_count[month] for k in month_sentiment_sum[month]}
    for month in month_sentiment_sum
}
```

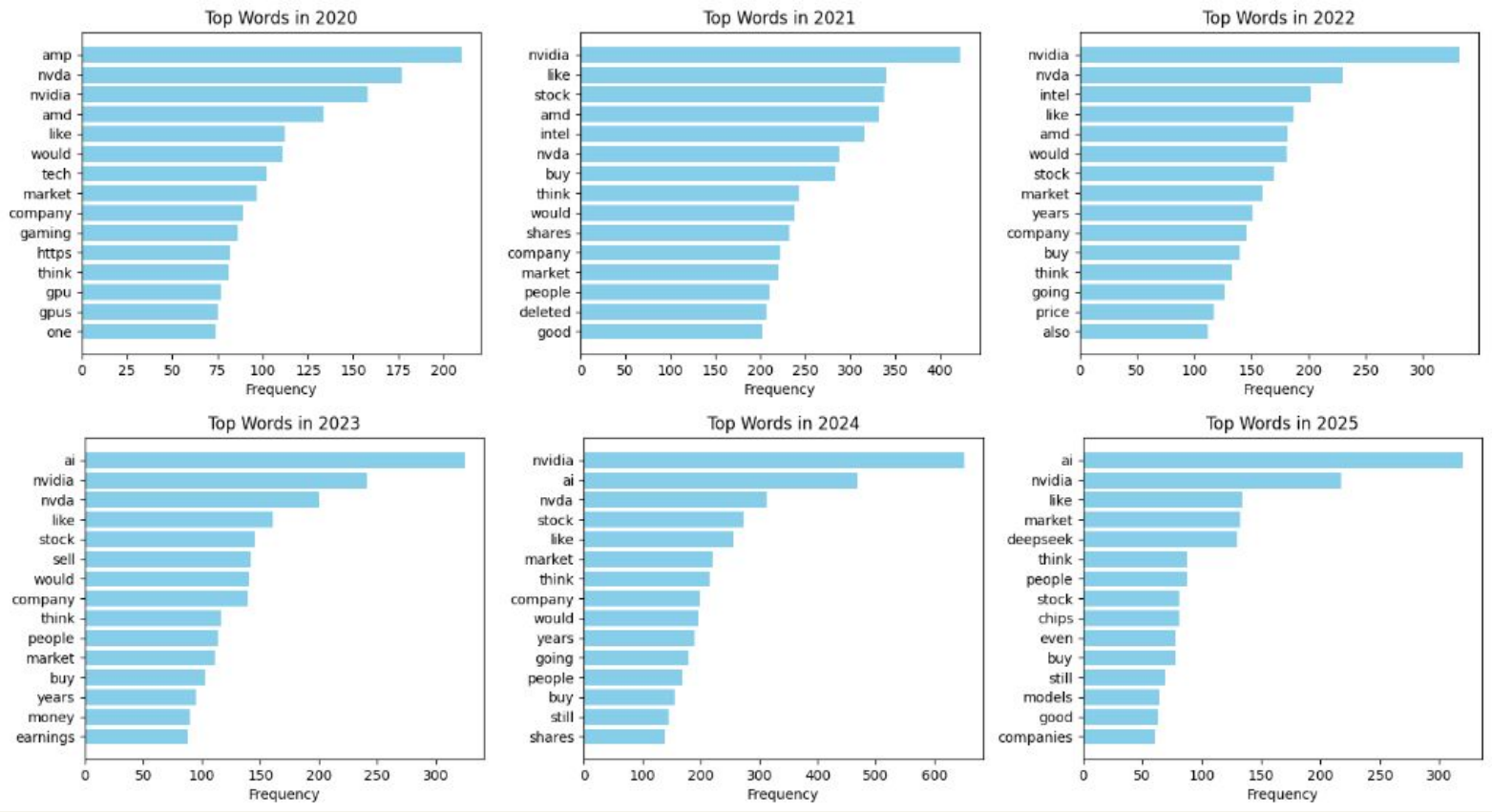
```
[8] for month in word_counts:
        month_to_word[month] = sorted(word_counts[month].items(), key=lambda x: x[1], reverse=True)[:15]
```

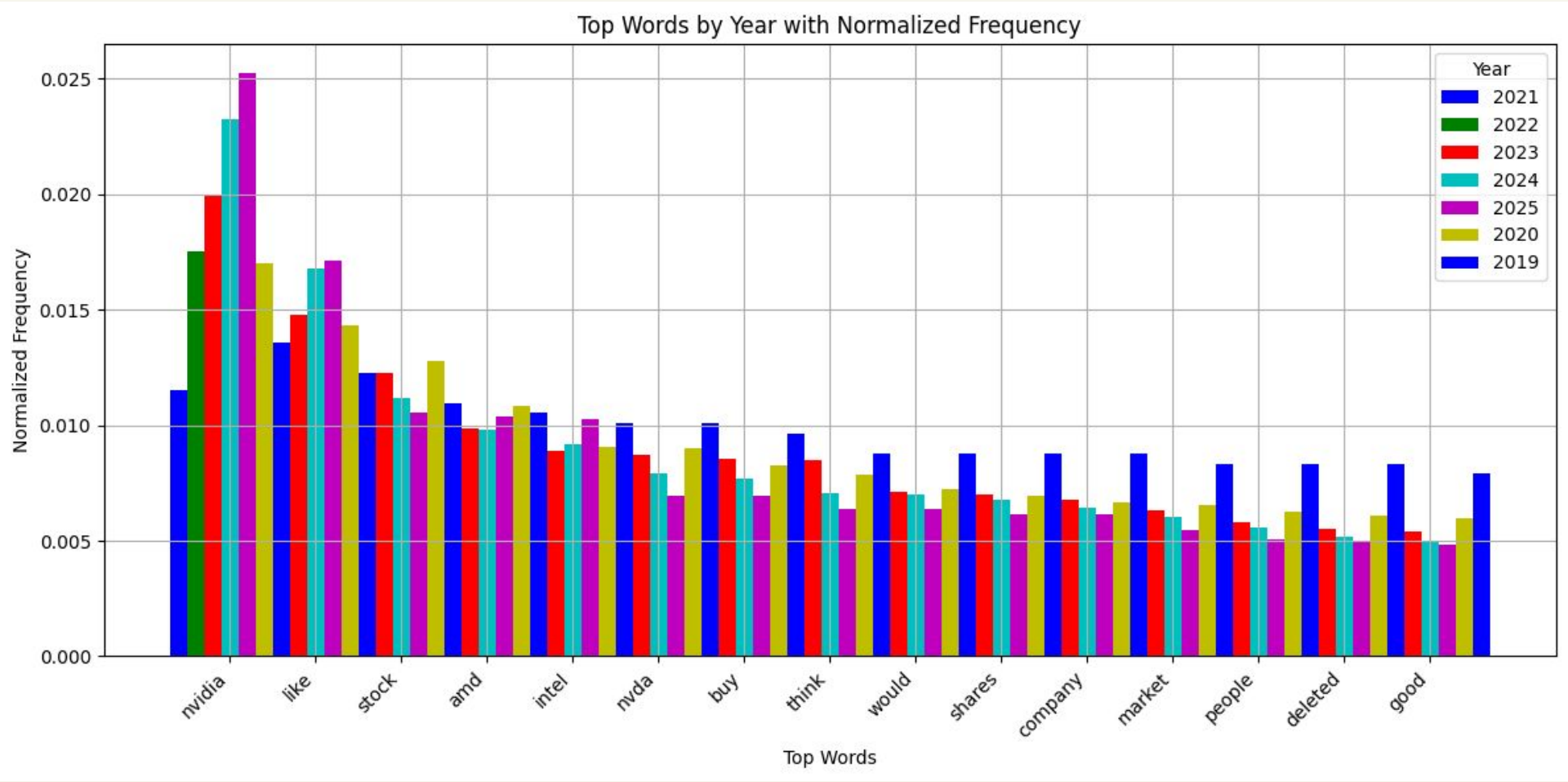
```
[9] print(month_to_sentiment)
    print(word_to_sentiment)
    print(month_to_word)
    print(word_counts)
```

```
{'Apr25': {'neg': 0.0522, 'neu': 0.8468, 'pos': 0.10103333333333334, 'compound': 0.3687633333333334}, 'Mar25': {'neg': 0.059031
{'advice': {'neg': 0.03445454545454546, 'neu': 0.8485454545454547, 'pos': 0.12518181818181817, 'compound': 0.7268181818181818},
{'Apr25': [('nvidia', 25), ('stocks', 19), ('stock', 19), ('market', 18), ('shares', 11), ('time', 10), ('etf', 9), ('price', 8
{'Apr25': {'advice': 3, 'investments': 4, 'etfs': 3, 'upst': 2, 'otly': 2, 'irbt': 2, 'help': 1, 'investment': 2, 'struggling':
```

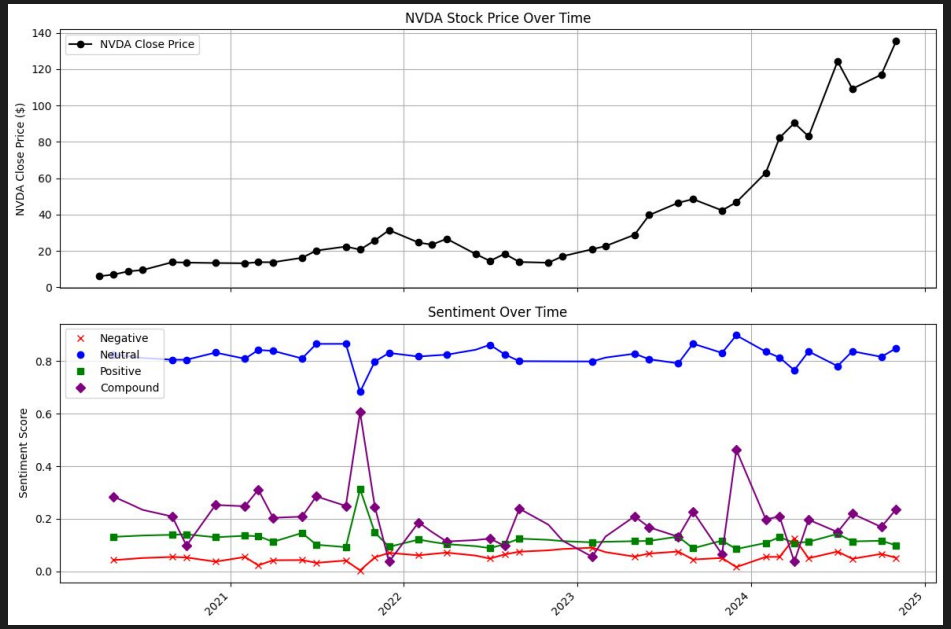
```
[10] normalized_word_counts = {
        month: {word: round(count / sum(month_word_counts.values()), 7)
                for word, count in month_word_counts.items()}
        for month, month_word_counts in word_counts.items()
    }
```

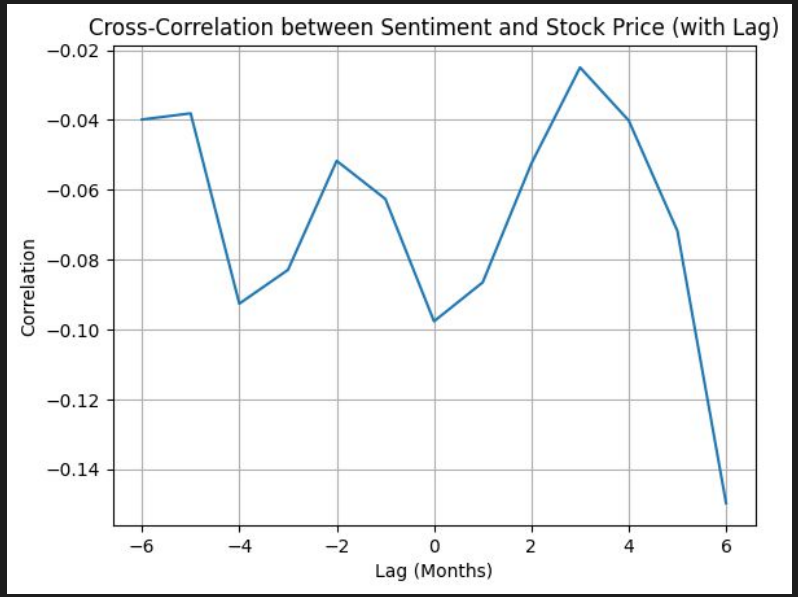
Critical Part in Sentiment Analysis Script

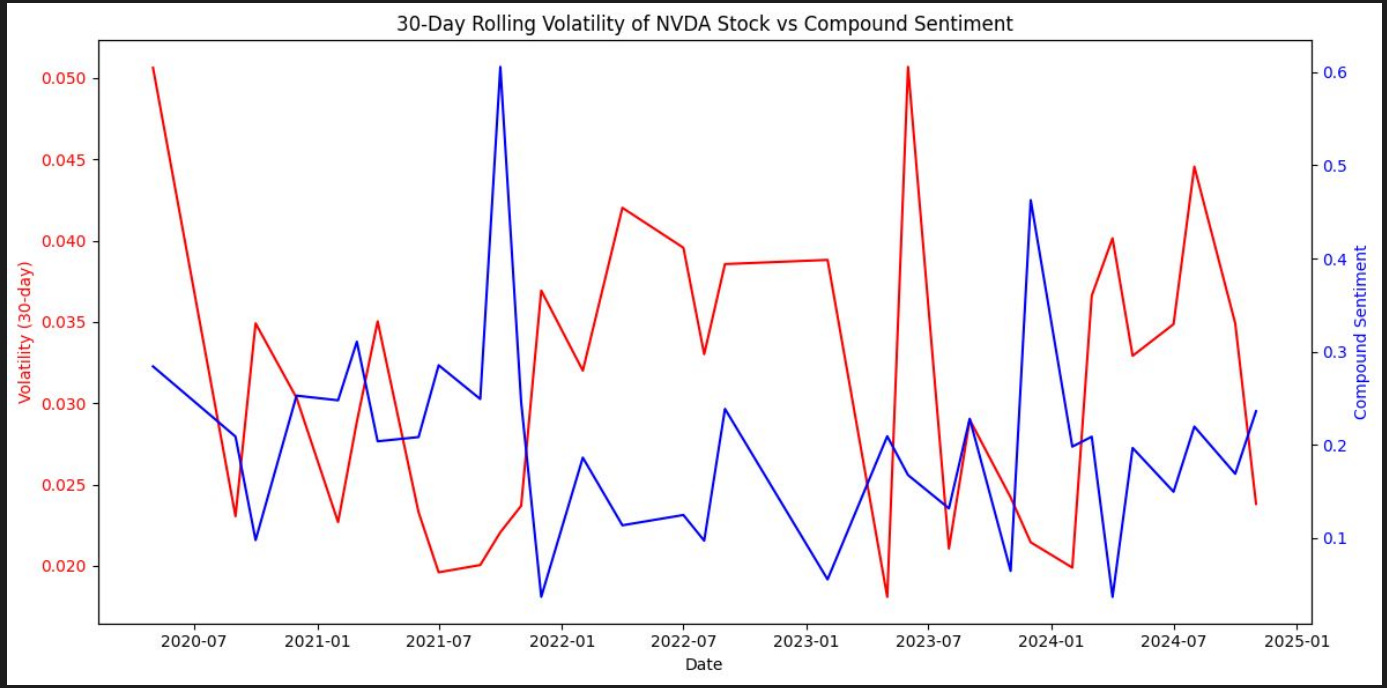

Top Words in 2020
amp
nvda
nvidia
amd
like
would
tech
market
company
gaming
https
think
gpu
gpus
one
0 25 50 75 100 125 150 175 200
Frequency
Top Words in 2021
nvidia
like
stock
amd
intel
nvda
buy
think
would
shares
company
market
people
deleted
good
0 50 100 150 200 250 300 350 400
Frequency
Top Words in 2022
nvidia
nvda
intel
like
amd
would
stock
market
years
company
buy
think
going
price
also
0 50 100 150 200 250 300
Frequency
Top Words in 2023
ai
nvidia
nvda
like
stock
sell
would
company
think
people
market
buy
years
money
earnings
0 50 100 150 200 250 300
Frequency
Top Words in 2024
nvidia
ai
nvda
stock
like
market
think
company
would
years
going
people
buy
still
shares
0 100 200 300 400 500 600
Frequency
Top Words in 2025
ai
nvidia
like
market
deepseek
think
people
stock
chips
even
buy
still
models
good
companies
0 50 100 150 200 250 300
Frequency

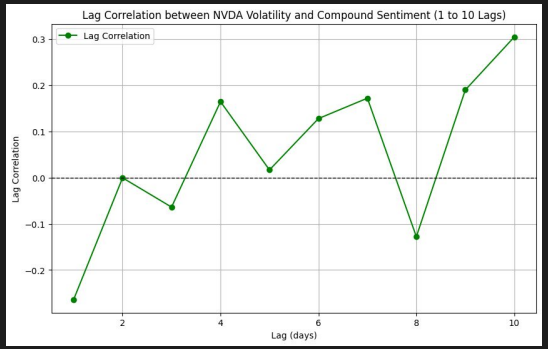

Top Words by Year with Normalized Frequency
Normalized Frequency
0.025
0.020
0.015
0.010
0.005
0.000
Year
2021
2022
2023
2024
2025
2020
2019
nvidia
like
stock
amd
intel
nvda
buy
think
would
shares
company
market
people
deleted
good
Top Words

Combined Analysis









Summarizing Findings

Sentiment does not significantly affect stock growth

The correlation between sentiment (both positive and negative) and NVDA stock price growth is minimal. Although fluctuations in sentiment were observed, they did not show a clear, direct impact on the stock's short-term or long-term performance.

Sentiment and volatility correlation is weak

While there are periods of heightened sentiment, especially with terms like "AI," the sentiment does not show a strong correlation with stock price volatility. The stock price volatility does not seem to increase during periods of high sentiment, suggesting that sentiment alone is not a key driver of price fluctuations for NVDA.

Limitations

- Difficult to collect comprehensive and balanced data set (e.g. samples across years)
- Issues collecting data with API, and having to pivot at the last minute → difficult to collect many samples
- Uses sentiment data from a single platform, which may not fully capture broader market sentiment
- Difficult distinguishing between social media sarcasm and genuine analysis without more advanced NLP techniques

Beyond Our Project

To improve our current project, we would:

- Diversify our data sources beyond Reddit (e.g. Twitter, YouTube comments, paywalled stock forums)
- Use fine-tuned transformer models (e.g., BERT, RoBERTa) for more nuanced sentiment classification
- Examine the effects on other companies to gain a better understanding of the semiconductor and GPU manufacturing market, as well as the tech industry beyond computer hardware

Final Takeaways

- Learned a lot of new skills — got a lot of exposure to new technologies
  - Got experience web scraping, using external APIs, NLP and text processing techniques
- Learned better data wrangling techniques
- Learned about the stock market and visualizations and metrics for analyzing stock market events

Thank you!