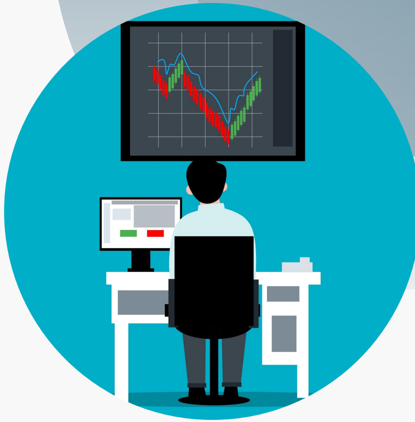
Subreddit's Sentiment Impact On Stock Performance



Kevin Shi, Lixiao Yang, Chris Fluta, Vandan Patel

Table of contents

01

Introduction and Motivation

Intro & Motivation

- Hedge funds were heavily shorting GameStop (GME), betting on its decline
    - Decline in Physical Retail
    - Competition (sony, microsoft, nintendo)
    - Declined Profits due to Covid -19
    -
- January 22 2021 users of r/wallstreetbets initiated a short squeeze on GameStop, pushing their stock prices up significantly

- Rose from $17.25 to a pre market value of $500 per share

- Melvin Capital required a $2.75 billion bailout from other hedge funds, including Citadel, to stay afloat

Research Question: Does the sentiment towards a brand on their subreddit affect the brand's stock performance

02

Methodology

Methodology

Objective: Investigate whether the sentiment towards brand on their subreddit affects the stock performance of GameStop, Tesla, and Nvidia.

Importance: Understanding the influence of online communities on stock prices can help investors and analysts make more informed decisions.

Data Collection

- Reddit Data: Use Reddit API (praw) to collect posts from r/GameStop, r/Tesla, and r/Nvidia
- Stock Data: Obtain historical stock prices from Yahoo Finance (yfinance) for GameStop, Tesla, and Nvidia

Data Preprocessing

- Reddit Data Cleaning: Remove URLs, mentions, special characters, and irrelevant posts
- Remove useless fields and those not needed for sentiment analysis and drop any null values

Methodology…

<u>**Sentiment Analysis**</u>

- VADER Sentiment Analysis: Use NLTK's VADER to analyze the sentiment of each Reddit post
- BERT (Bidirectional Encoder Representations from Transformers) models for sentiment Analysis
  analyzes the relationships between words in a sentence in both directions
  - FinBERT
    - Pre-trained NLP model to analyze sentiment of financial text

  - BERTweet
    - Designed for processing and understanding tweets and trained on 850 million English tweets

  - RoBERTa (Robustly Optimized BERT Approach)
    - Optimized variant of BERT that improves performance by using more data and longer training times *(trained on 160GB text data)*

Data Extraction

Why we extracted the data we extracted?

- Attempted to extract the top 10,000 posts from the past year
- Gamestop
  - Extracted 990 posts
  - 2023-04-05 to 2024-04-03
- Tesla
  - Extracted 534 posts
  - 2009-12-19 to 2023-11-22
- Nvidia
  - Extracted 988 posts
  - 2023-04-06 to 2024-04-03

```python
def extract_posts(_reddit, _subreddit_name, _time_filter):
    """
    Function to extract raw top posts from ‘r/____’ from the past time filter and save to a file
    :param _reddit: an authenticated Reddit instance
    :param _subreddit_name: a string, name of the subreddit
    :param _time_filter: a string, "year" or "all"
    :return: _posts_df, dataframe of raw subreddit posts
    """
def create_relevant_df(df):
    """
    Function to create a df with relevant columns and merge 'selftext' and 'title' columns into one column
    :param df: dataframe of raw subreddit posts
    :return: cleaned_df, a dataframe with relevant columns and a ‘text’ column
    """
    # Extract the required columns from the original dataframe
    selected_columns = df[['score', 'created_utc', 'title', 'selftext']]

    # Combine 'title' and 'selftext' into a new 'text' column
    selected_columns['text'] = selected_columns['title'].fillna('') + ' ' + selected_columns['selftext'].fillna('')

    # Create a new dataframe with the 'score', 'date', and 'text' columns
    cleaned_df = selected_columns[['score', 'created_utc', 'text']]

    return cleaned_df
def clean_text(text):
    """
    Function to clean 'text' column
    :param text: strings in 'text' column of the df
    :return: text, cleaned strings for ‘text’ column
    """
    text = re.sub( pattern: r'<[^<]+?>', repl: ' ', text)  # remove HTML tags
    text = text.lower()  # Convert text to lowercase
    text = re.sub( pattern: r'http\S+|www\S+|https\S+', repl: '', text, flags=re.MULTILINE)  # Remove URLs
    text = re.sub( pattern: r'[^a-z\s]', repl: '', text)  # Remove special characters, numbers, and punctuation
    text = re.sub( pattern: r'\s+', repl: ' ', text).strip()  # Remove extra spaces

    return text
```

Data Preprocessing

- Standard text cleaning
- Merge stock and sentiment data
- Adjust merged stock and sentiment data

```python
import yfinance as yf


def merge_with_historical_stock_data(stock_ticker, sentiment_df):
    """
    Function to merge sentiment data with stock data
    :param stock_ticker: a string (stock abbreviation)
    :param sentiment_df: sentiment_df dataframe
    :return: merged_data, a dataframe with the merged sentiment data with stock data
    """
```

```python
def adjust_sentiment_scores_merged_data(merged_data):
    """
    Function to adjust sentiment scores in merged data by taking the average of each models' scores on each day
    :param merged_data: a dataframe with the merged sentiment data with stock data
    :return: merged_data_aggregated, a dataframe with the merged sentiment data with stock data with adjusted sentiment scores
    """
```

	index	Date	Open	High	Low	Close	Adj Close	Volume	score	text	FinBERT_sentiment_score	BERTweet_sentiment_score	RoBERTa_sentiment_score	vader_sentiment_score	FinBERT_sentiment_score_adj
0	0	2023-04-05	22.469999	22.469999	21.23	22.07	22.07	3638100	182.0	really warehouse i mean sure its happened befo...	0.906191	0.931234	0.717845	-0.4438	0.943087
1	0	2023-04-05	22.469999	22.469999	21.23	22.07	22.07	3638100	139.0	this	0.979983	-0.797670	-0.347687	0.0000	0.943087
2	1	2023-04-06	22.000000	22.670000	21.77	22.40	22.40	2506900	62.0	anyone else win a golden ticket and never get ...	0.999028	0.000000	-0.292112	0.8821	0.999028

03

Experimental Results

BERT-Based Models Structure

- BERT-Based Models
  - Advanced NLP models based on neural networks using Transformer architecture
  - Fine-tuned for sub-tasks and fields
  - Known for achieving state-of-the-art performance
- Comparison of Three BERT-based Models

Model	**FinBERT**	**BERTweet**	**RoBERTa**
Purpose	Financial sentiment analysis	General purpose	General purpose
Training Data	Financial texts	Tweets	General domain text
Parameters	110M	134M	125M
Fine-tuned Tasks	Sentiment analysis in finance	Sentiment analysis in social media text	Various NLP tasks (classification, QA, etc.)

Model Sentiment Scores (Range -1 to +1)

Sentiment Scores Pre-Adjustment

GameStop:

	FinBERT_sentiment_score	BERTweet_sentiment_score	RoBERTa_sentiment_score	vader_sentiment_score
positive	604	236	375	340
negative	57	256	286	248
zero	0	169	0	73

Tesla:

	FinBERT_sentiment_score	BERTweet_sentiment_score	RoBERTa_sentiment_score	vader_sentiment_score
positive	343	14	29	141
negative	18	333	332	48
zero	0	14	0	172

Nvidia:

	FinBERT_sentiment_score	BERTweet_sentiment_score	RoBERTa_sentiment_score	vader_sentiment_score
positive	521	51	78	392
negative	161	510	604	96
zero	0	121	0	194

Sentiment Scores Post-Adjustment

GameStop:

	FinBERT_sentiment_score_adj	BERTweet_sentiment_score_adj	RoBERTa_sentiment_score_adj	vader_sentiment_score_adj
positive	226	122	174	133
negative	12	102	64	101
zero	0	14	0	4

Tesla:

	FinBERT_sentiment_score_adj	BERTweet_sentiment_score_adj	RoBERTa_sentiment_score_adj	vader_sentiment_score_adj
positive	301	13	26	126
negative	16	294	291	45
zero	0	10	0	146

Nvidia:

	FinBERT_sentiment_score_adj	BERTweet_sentiment_score_adj	RoBERTa_sentiment_score_adj	vader_sentiment_score_adj
positive	192	15	17	186
negative	46	210	221	30
zero	0	13	0	22

BERT-Based Experimental Results

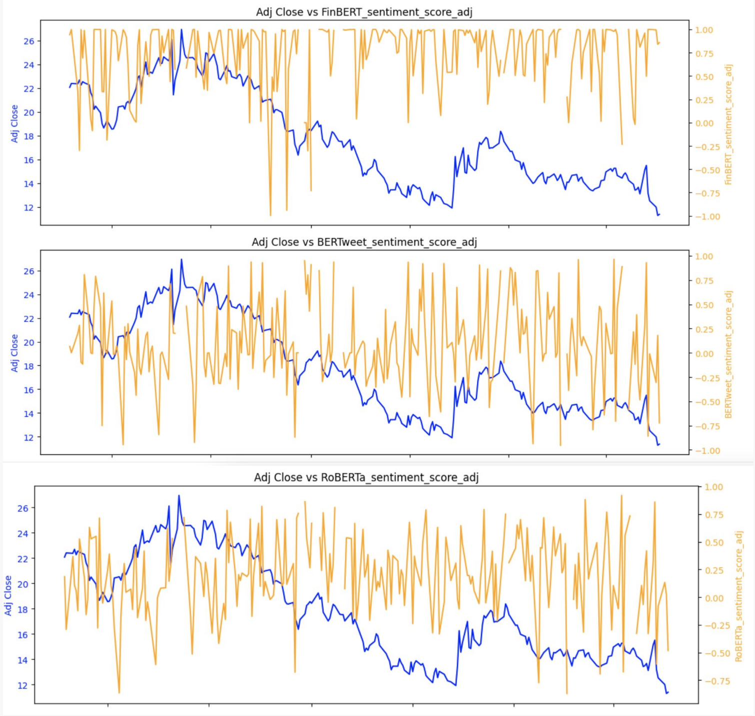
GameStop results across three models

- **We wanted to see the relationship between the sentiment scores and the adjusted daily close price of the stock**

- ***The visual comparison of adjusted close stock price vs results of the FinBERT model***
  (On the Top Right)

- ***The visual comparison of adjusted close stock price vs the BERTTweet model***
  (On the Middle)

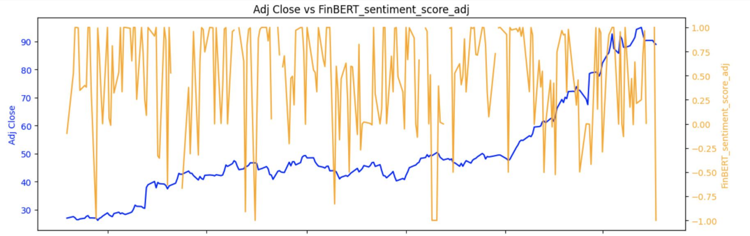
- ***The visual comparison of adjusted close stock price vs results of the RoBERTa model***
  (On the Bottom Right)

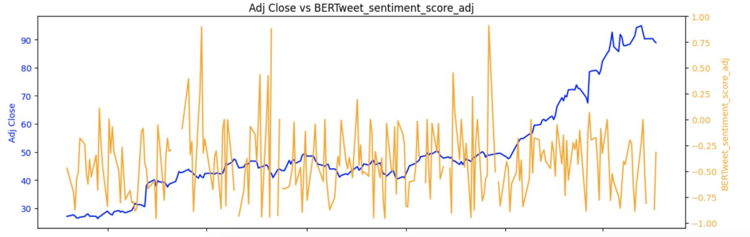


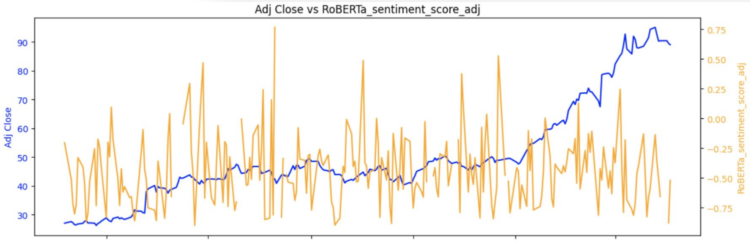
BERT-Based Experimental Results

Nvidia results across three models:

- ***The visual comparison of adjusted close stock price vs results of the FinBERT model***
  (On the Top Right)
- ***The visual comparison of adjusted close stock price vs the BERTTweet model***
  (On the Middle)
- ***The visual comparison of adjusted close stock price vs results of the RoBERTa model***
  (On the Bottom Right)







BERT-Based Experimental Results

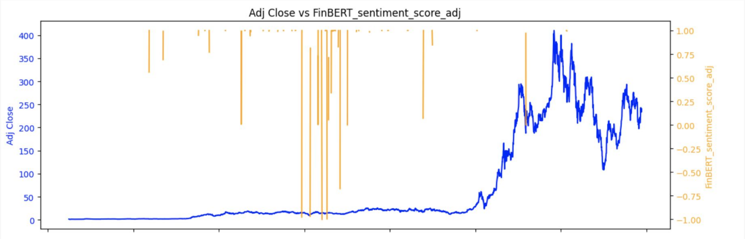
Tesla results across three models:

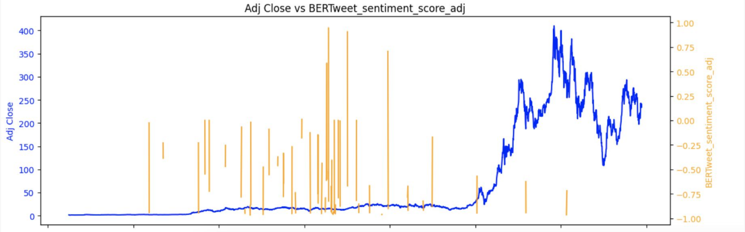
- ***All time top post data for Tesla due to insufficient data for past year***

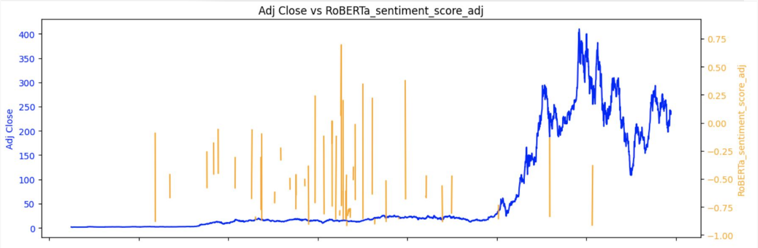
- ***The visual comparison of adjusted close stock price vs results of the FinBERT model***
  (On the Top Right)

- ***The visual comparison of adjusted close stock price vs the BERTTweet model***
  (On the Middle)

- ***The visual comparison of adjusted close stock price vs results of the RoBERTa model***
  (On the Bottom Right)







VADER Structure

- Valence Aware Dictionary and sEntiment Reasoner
    - Tuned for Social Media
    - Excels at short text sentiment

- VADER scores text by looking at individual token sentiments as well as the compound score

- Default VADER vs AFINN
    - Default VADER is more similar to BERT

NLTK

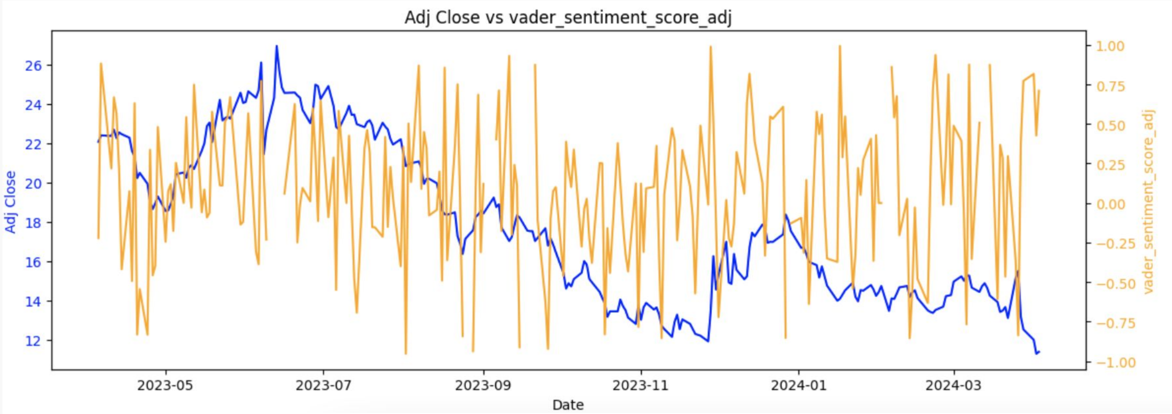
VADER Results

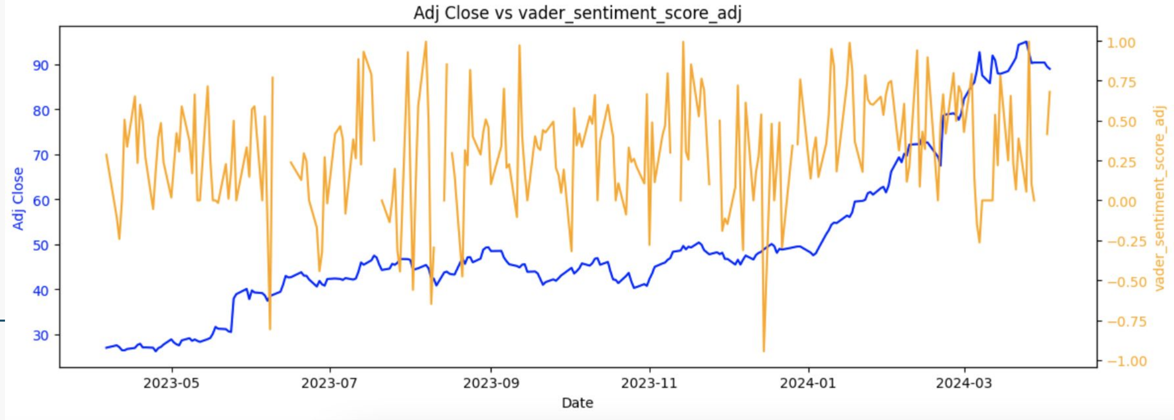
Gamestop Results

- ***The visual comparison of adjusted close stock price vs the NLTK VADER model** (Top Right)*

Nvidia Results

- ***The visual comparison of adjusted close stock price vs the NLTK VADER model** (Bottom Right)*

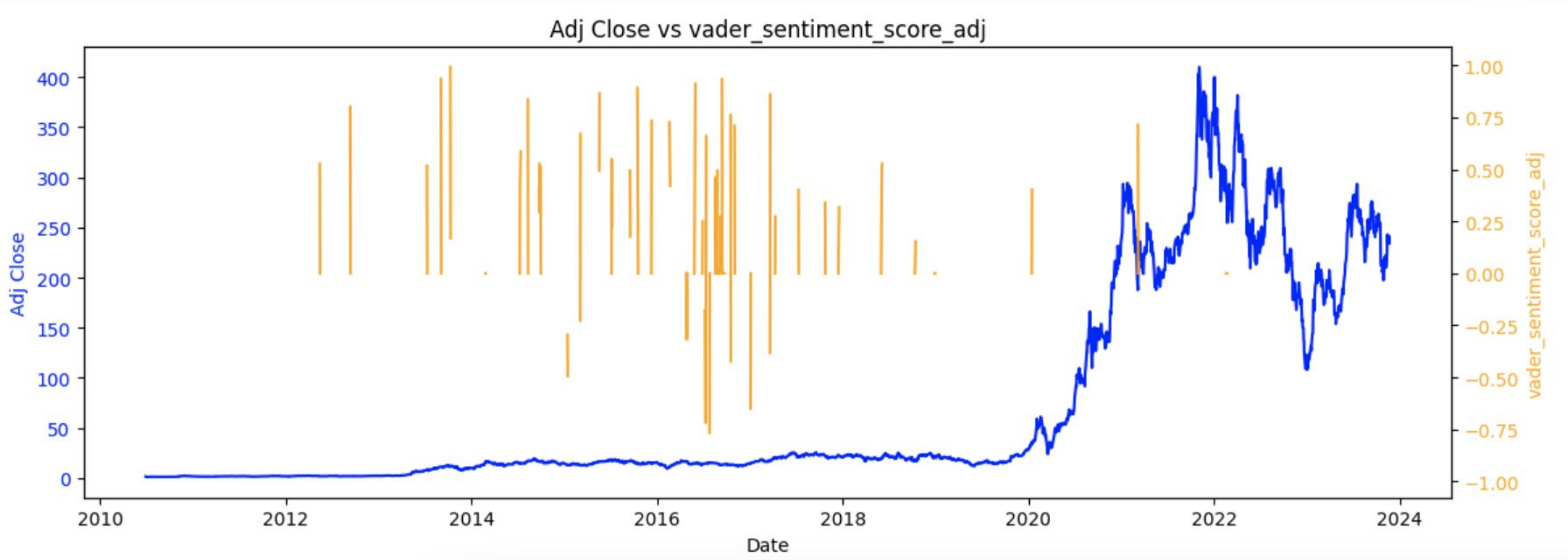




VADER Results

Tesla Results

- ***The visual comparison of adjusted close stock price vs the NLTK VADER model***
  (On the Bottom)



04

Conclusion

Conclusion

- One lesson learned is that with data extractions using Reddit APIs, you cannot pull top posts for two years or a specified time frame but only for "all", "day", "hour", "month", "week", or "year"

  If we were to do it differently we would find a way to get data for each day from the past 2 years

- In the future, if we had more funds, time, and computational power, we could potentially develop and train our own model

- Though findings did not directly show a relationship between online sentiment and stock price, we still believe it has an impact - we would expand to other sources like twitter, stock-specific subreddits, etc.

- Recent Events:

  Roaring Kitty posted a screenshot on Reddit late Sunday - paid $175 million building a position in game stock -> Stock rose nearly 75% at market open

Thank You