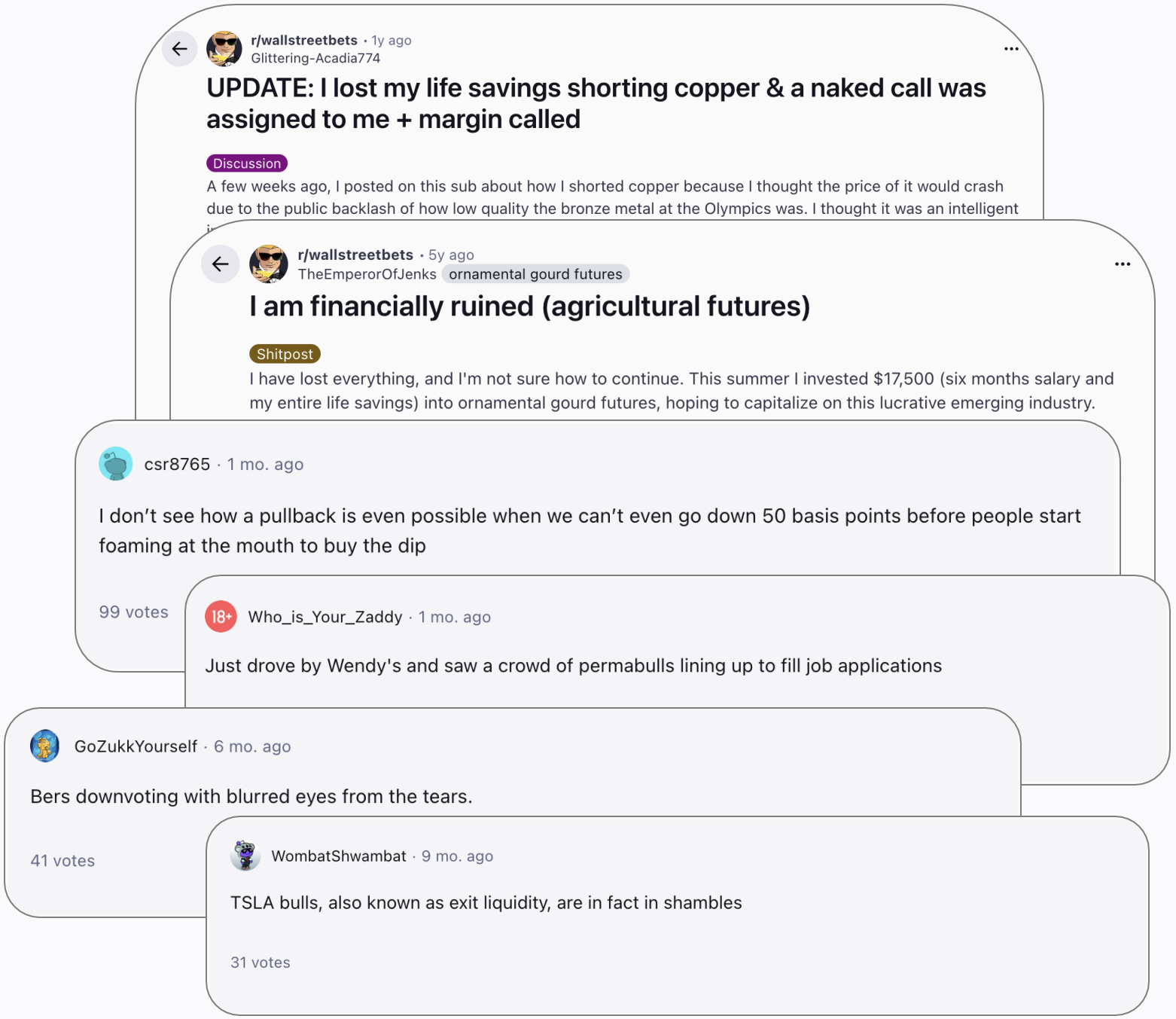
Trading strategies based on r/wallstreetbets sentiment

Case 26: NLP and Sentiment Analysis–Based Trading Strategies

Presented by:
Benjamin Luo



Overview

Section 1 | **Introduction to Natural Language Processing and Sentiment Analysis**

Section 2 | **State of the Field Sentiment Analysis in Finance**

- Trump tweets and the currency market (2025)
- Pump and dump detection (2025)
- Adversarial attacks on financial sentiment with LLMs (2023)
- Financial distress prediction during earnings calls (2023)

Section 3 | **Blueprint for Building a Trading Strategy Based on Sentiment Analysis**

Section 4 | **Trading Strategies Based on r/wallstreetbets Sentiment**

Section 5 | **Concluding Remarks**

Section 1 | Introduction to Natural Language Processing and Sentiment Analysis

Section 2 | State of the Field Sentiment Analysis in Finance

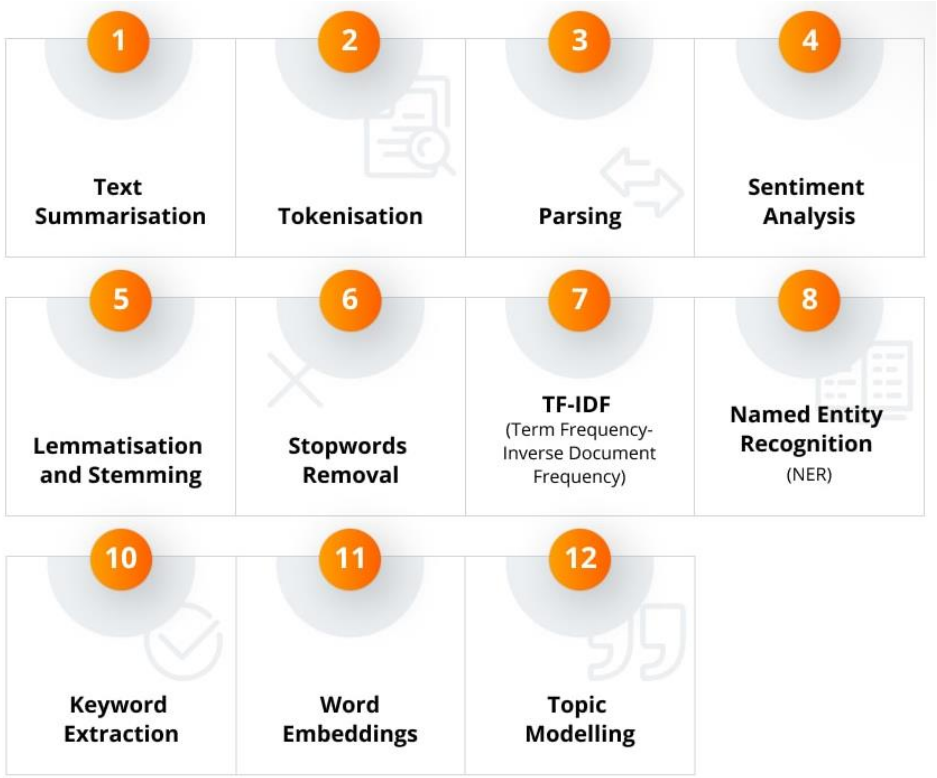
Section 3 | Blueprint for Building a Trading Strategy Based on Sentiment Analysis

Section 4 | Trading via r/wallstreetbets Sentiment

Section 5 | Concluding Remarks

Natural Language Processing (NLP) is a branch of Artificial Intelligence that enables computers to understand, interpret, and generate human language

NLP is used in finance to rapidly process information from **news articles**, **earnings reports**, **SEC filings**, etc.



NLP Applications

Applications in Finance

1. **Sentiment analysis** (Case #26)
2. Financial document summarization (Case #28)
3. Risk assessment and management
4. Chatbots and customer service (Case #27)
5. Regulatory compliance
6. Portfolio management
7. Market research

Future Processing. **Natural Language Processing (NLP): what it is and how it works?** future-processing.com/blog/natural-language-processing-guide/

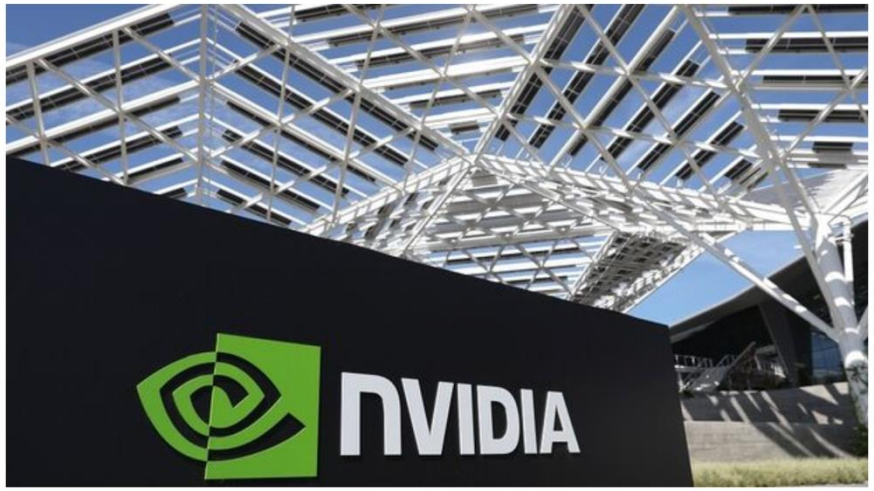
News articles significantly move the stock market

On September 3, 2024, Bloomberg posted a false report that $NVDA was being subpoenaed by the Department of Justice, wiping out 10% of its value (US$300bn)

Technology

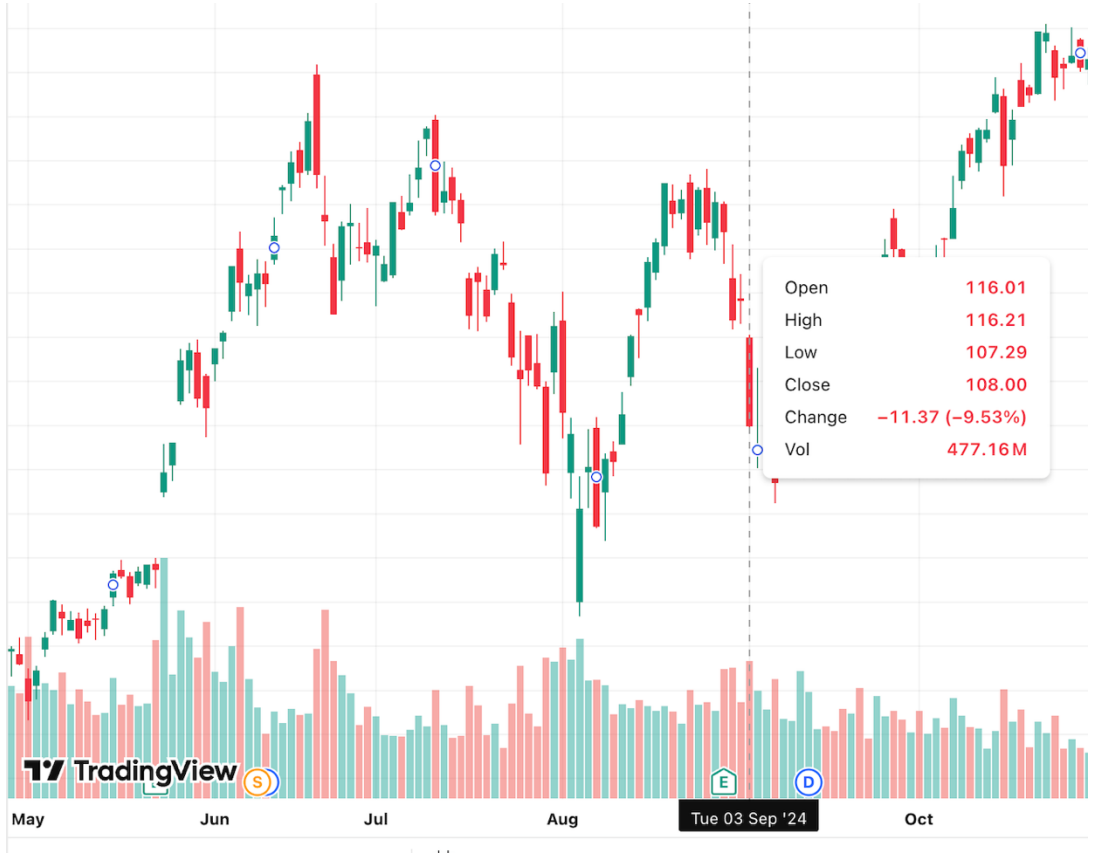
Nvidia Gets DOJ Subpoena in Escalating Antitrust Probe

- Department also sent subpoenas to third parties in case
- Nvidia has built a dominant position in AI computing



Nvidia Subpoenaed in Escalating DOJ Antitrust Probe

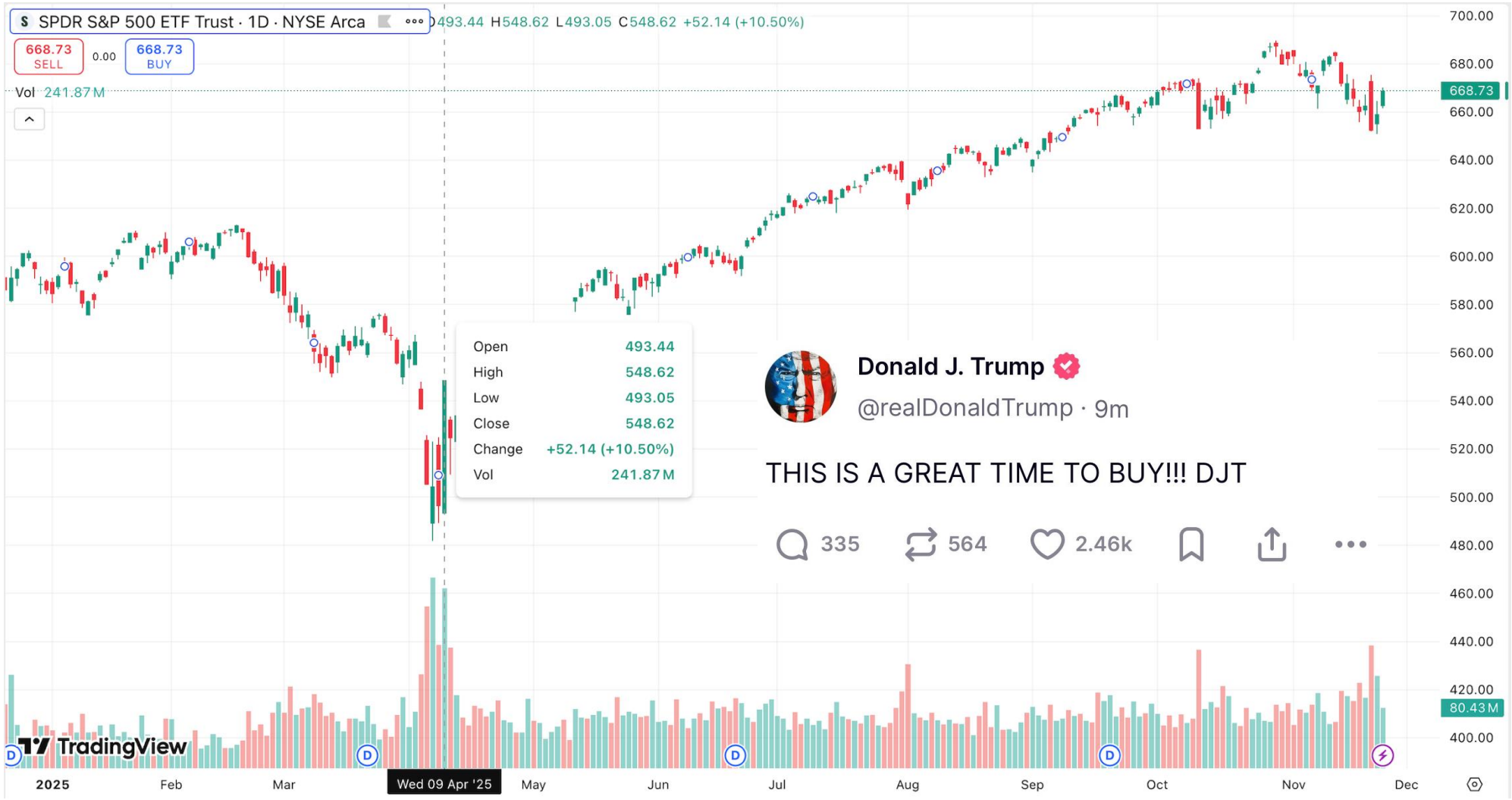
By Ian King and Leah Nylen
September 3, 2024 at 8:24 PM UTC
Updated on September 3, 2024 at 9:34 PM UTC



bloomberg.com/news/articles/2024-09-03/nvidia-gets-doj-subpoena-in-escalating-antitrust-investigation

News articles significantly move the stock market

On April 9, 2025, President Trump posted “THIS IS A GREAT TIME TO BUY!!!” 4 hours prior to announcing a 90-day pause on tariffs, driving the S&P up 9.5% (US$4T)



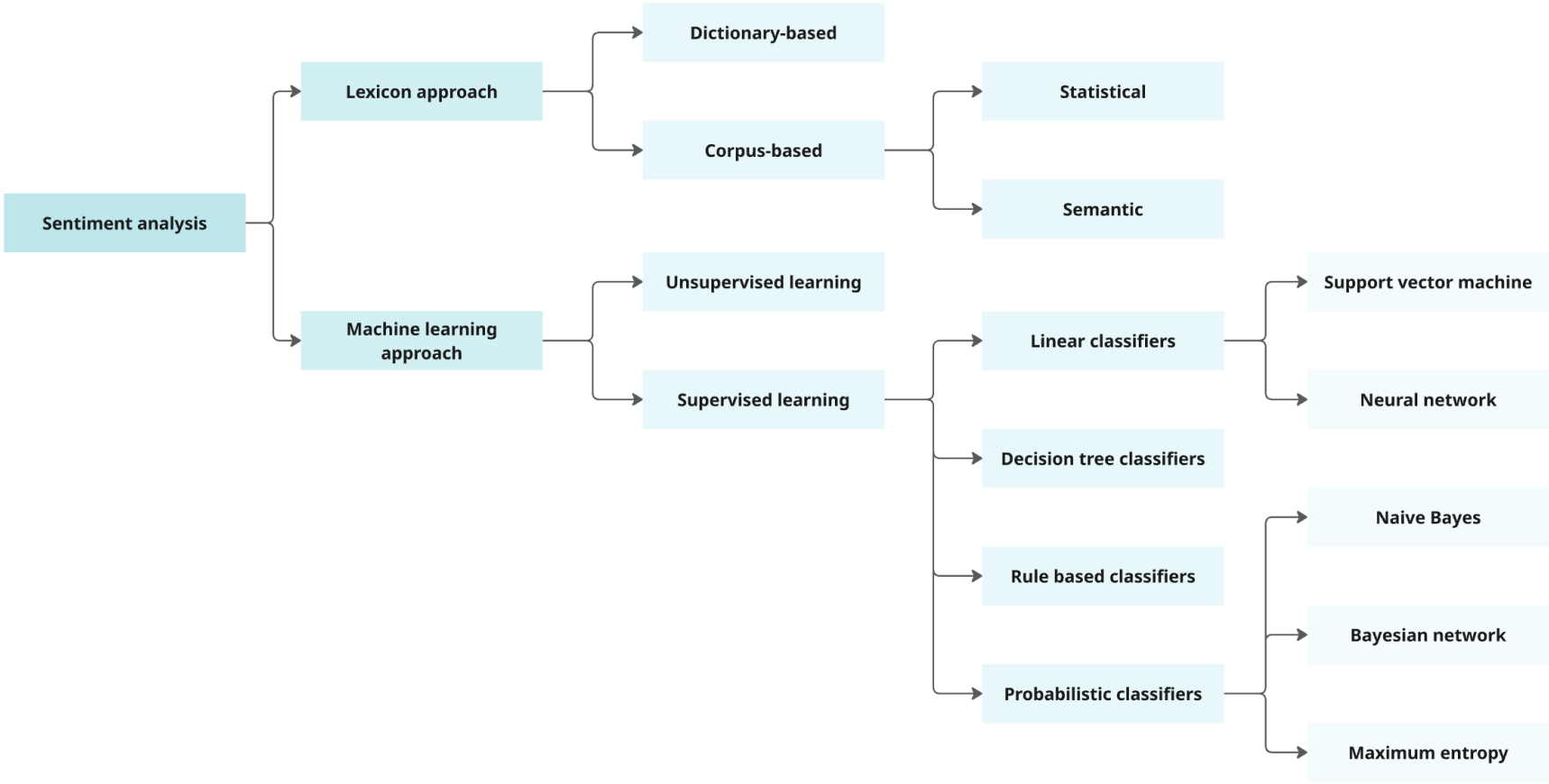
Sentiment analysis is an application of NLP that uses computational methods to determine the emotional tone, opinion, or attitude expressed in text

In finance, sentiment analysis estimates **collective market opinion** (from investors, news, and the public) to forecast price direction

	Predicted sentiment	Actual sentiment
My coffee was **great**	Positive	Positive
My coffee was **awful**	Negative	Negative
My coffee was **not great**	Positive	Negative
My coffee was **not that great**	Positive	Neutral
I did **not** think my coffee was **great**	Positive	Negative
I did **not** expect my coffee to be **this great**	Positive	Positive
I was **disappointed** with the quality of the coffee	Negative	Negative
I **wasn't disappointed** with the coffee quality	Negative	Positive

Sentiment analysis methods are categorized into lexicon and machine learning approaches

Current state of the field research focuses on new context-aware transformer models



Sharma, D., Sabharwal, M., Goyal, V., & Vij, M. (2020). Sentiment analysis techniques for social media data: A review. In A. K. Singh, M. K. Sharma, M. Sabharwal, & M. S. W. P. L. Singh (Eds.), Machine learning and cognitive computing for mobile big data (pp. 75–90). Springer. https://doi.org/10.1007/978-981-15-0029-9_7

Example – Preprocessing

“Next time he says it's a **good time to buy** you’ll go all in only for him to announce a **total embargo of Europe**”

The **pre-processing** pipeline applies several transformations to **standardize the text** and **reduce linguistic variance**

Text Cleaning and Normalization

Transform text to a uniform format to reduce noise

Case | **Special Characters** | **URL, Number, Emoji**

next time he says it's a good time to buy you’ll go all in only for him to announce a total embargo of europe

Tokenization

Transform text into tokens (*usually words*)

next, time, he, says, it, is, a, good, time, to, buy, you, will, go, all, in, only, for, him, to, announce, a, total, embargo, of, europe

Stop Word Removal

Remove frequent words that don't contribute unique semantic or sentiment value

next, time, says, good, time, buy, you, go, all, in, only, announce, total, embargo, europe

Stemming or Lemmatization

Reduces to roots to reduce vocabulary

Stemming | **Lemmatization**

next, time, **say**, good, time, buy, you, go, all, in, only, announce, total, embargo, europe

“next, time, say, good, time, buy, you, go, all, in, only, announce, total, embargo, europe”

Example – Lexicon Dictionary Approach

“Next time he says it's a **good time to buy** you’ll go all in only for him to announce a **total embargo of Europe**”

This approach uses a **curated list** where **words are manually assigned a score** and rules are applied for context

“next, time, say, good, time, buy, you, go, all, in, only, announce, total, embargo, europe”

Score Assignment

Look up the sentiment score for each token

next = 0	time = 0	...	buy = 0.8	good = 0.8	embargo = -0.99

Classification

Classify based on the final score

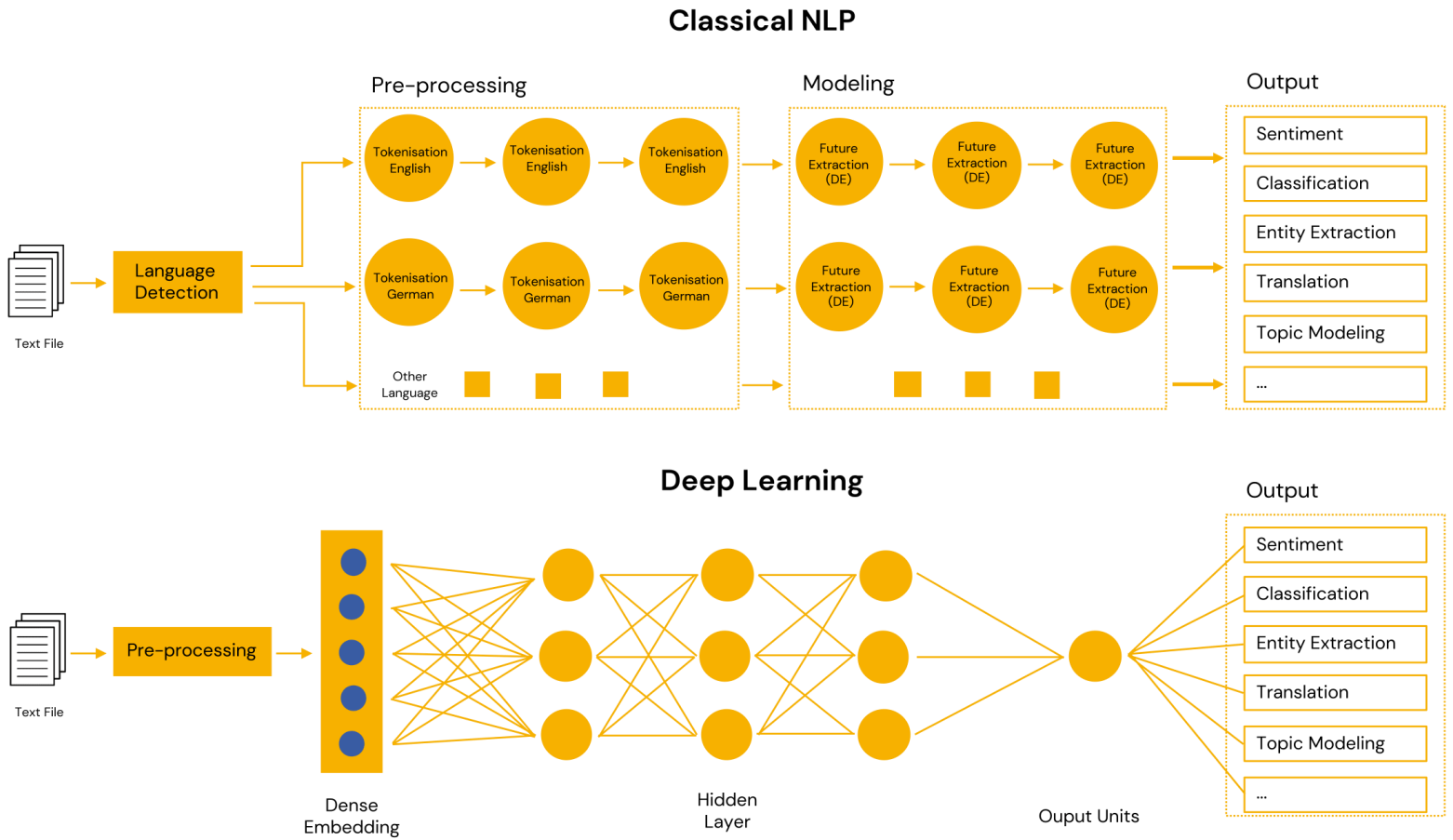
< -1	-1 < 0	0	0 < 1	>1
Very Bad	**Bad**	**Neutral**	**Good**	**Very Good**

0 + 0 + ... + 0.8 + ... + 0.8 + ... - 0.99 = **0.61**

Good

ML-based NLP learns patterns, while classical NLP relies on handcrafted rules and linguistic knowledge

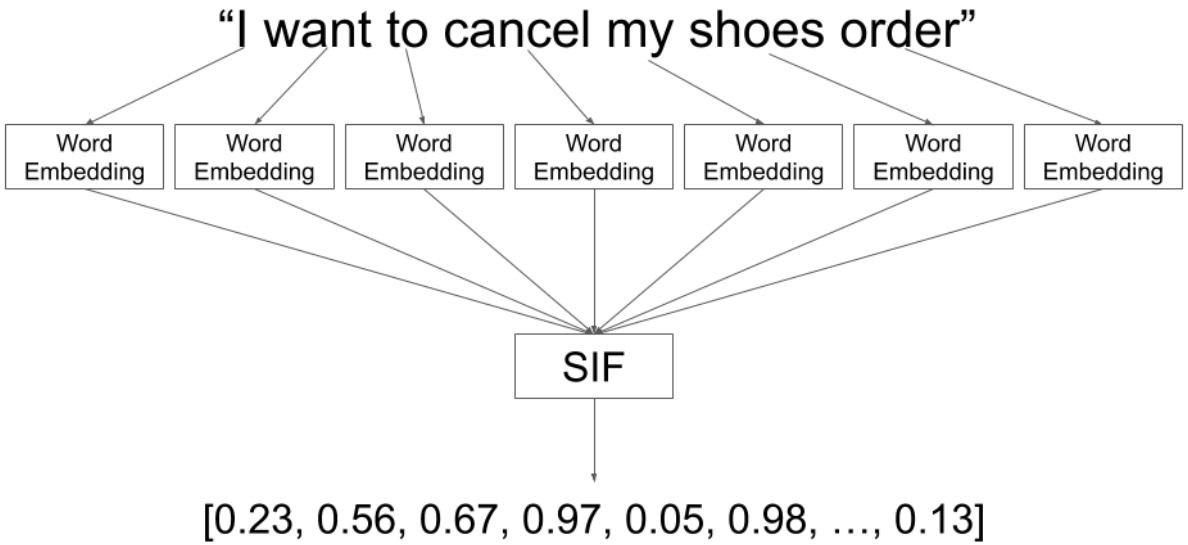
ML-based NLP uses **embeddings** and **neural networks** for automated feature learning on **vast amounts of data**

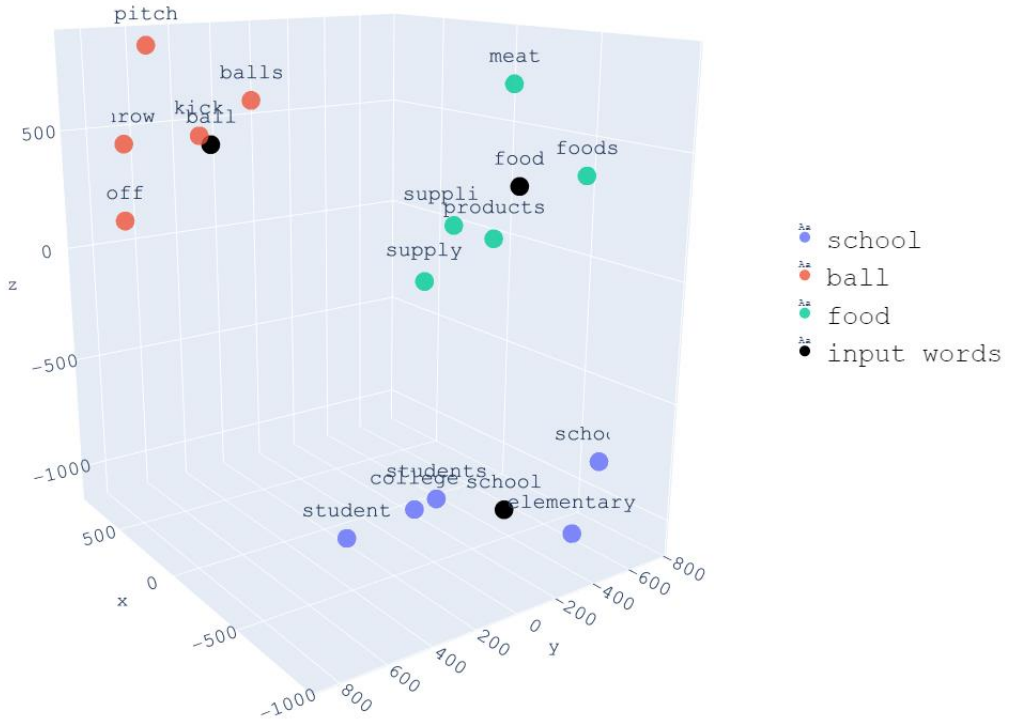


Cherian, T. V., Leelipushpam Paulraj, G. J., Princess, J. B., & Jebadurai, I. J. (2024). A comparative analysis of machine learning and deep learning techniques for aspect-based sentiment analysis. In D. J. Hemanth (Ed.), Computational intelligence methods for sentiment analysis in natural language processing applications (pp. 23–37). Morgan Kaufmann. https://doi.org/10.1016/B978-0-443-22009-8.00006-9

Embeddings are dense, low-dimensional vector representations that capture semantic and syntactic relationships in a numerical space

The embedding step is vital because it converts text into **dense numerical vectors**, which is more effective input for neural networks than **raw, sparse string** representations

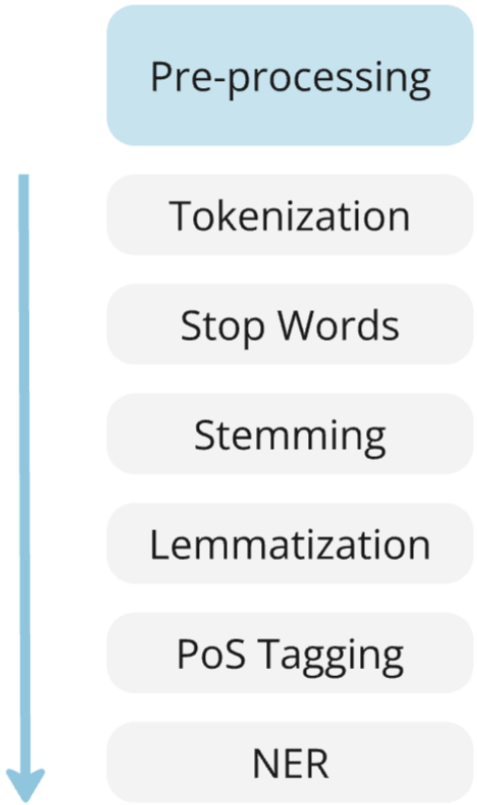




linkedin.com/pulse/word-sentence-embedding-nlp-oybek-eraliev
towardsdatascience.com/visualizing-word-embedding-with-pca-and-t-sne-961a692509f5/

ML Example - Preprocessing

Step 1: Use NLTK and SpaCy to preprocess raw text




```
Tokenization: "my retirement has been delayed by 82 years"
    >> ['my', 'retirement', 'has', 'been', 'delayed', 'by', '82', 'years']


Stop word removal: "Say goodbye to your whole port lil bro"
    >> ['Say', 'goodbye', 'whole', 'port', 'lil', 'bro']


Stemming: "oh my God LMAO bers literally evaporating before our very eyes"
    >> [('bers', 'ber'),
        ('literally', 'liter'),
        ('evaporating', 'evapor'),
        ('before', 'befor'),
        ('very', 'veri'),
        ('eyes', 'eye')]


Lemmetization: "Plunges 14% in sales. Stonk goes up. Fundamental bros explain for me pls"
    >> [('sales', 'sale'), ('goes', 'go')]


Parts-of-speech Tagging: "11 minutes until my life has meaning again"
    >> [('11', 'CD'),
        ('minutes', 'NNS'),
        ('until', 'IN'),
        ('my', 'PRP$'),
        ('life', 'NN'),
        ('has', 'VBZ'),
        ('meaning', 'NN'),
        ('again', 'RB')]

Named Entity Recognition: "I've heard enough. I'm shorting the NASDAQ"
    >> Entity: NASDAQ | Entity Type: ORG (Companies, agencies, institutions, etc.)

```

Step 2: Transform textual data to numerical vectors using 1 of 3 methods

Token Counts →

Vocabulary Map →

Key features (n=16) →

TF-IDF Matrix →

Rows = Sentences

Cols = Features

Values = Importance

Most frequent words →

```
Inputs
"I don't just want ber's money, I want to take their happiness and peace of mind too"
"I'm more upset about bers being happy than I am about losing lots of money"
"Bers in absolute disbelief LMAO"


Bag of Words (Word frequencies)
>> [[0 0 0 1 0 1 0 0 1 1 0 0 1 0 0 0 1 1 0 1 1 1 0 1 1 1 0 2]
    [2 0 1 0 1 0 1 0 0 0 1 0 0 0 1 1 0 1 1 1 0 0 1 0 0 0 1 0]
    [0 1 0 0 0 0 1 1 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0]]
>> {'don': 8, 'just': 12, 'want': 27, 'ber': 5, 'money': 17, 'to': 24, 'take': 21,
    'their': 23, 'happiness': 9, 'and': 3, 'peace': 20, 'of': 19, 'mind': 16,
    'too': 25, 'more': 18, 'upset': 26, 'bers': 6, 'being': 4, 'happy': 10,
    'losing': 14, 'lots': 15, 'in': 11, 'absolute': 1, 'disbelief': 7, 'lmao': 13}


 Term-Frequency Inverse Document Frequency (Word frequencies across documents)
>> ['happy' 'just' 'lmao' 'losing' 'lots' 'mind' 'money' 'peace' 'upset' 'want']
>> "Shape: (3,16)"
>> [[0.         0.30746099 0.         0.         0.30746099 0.30746099
     0.         0.30746099 0.         0.         0.         0.30746099
     0.23383201 0.30746099 0.         0.61492198]
    [0.         0.         0.3349067  0.         0.         0.
     0.44036207 0.         0.         0.44036207 0.44036207 0.
     0.3349067  0.         0.44036207 0.        ]
    [0.52863461 0.         0.40204024 0.52863461 0.         0.
     0.         0.         0.52863461 0.         0.         0.
     0.         0.         0.         0.        ]]


Word Embedding (Word2Vec) (Words to dense vector representations)
>> ['bers', 'I', 'about', 'of', 'want', 'LMAO', 'disbelief', 'absolute', 'money']
>> "Embedding for 'bers': [ 0.000233    0.00511642  0.00902829 -0.00931126]"
```

Step 3: Train a Naïve Bayes model using labeled data then predict unseen data

```
1  # Supervised ML (Naive Bayes)
2  data = ['BYND made me believe in magic. Cause I never seen money disappear like that',
3          'Bought $BYND on the giga dump 💀 the money was gone in seconds',
4          'BYND TO THE MOON']
5  sentiment = [0, 0, 1]
6
7  vect = CountVectorizer().fit(data).transform(data)
8  model = MultinomialNB(alpha = 0.1)
9  model.fit(vect, data['sentiment'])
10
11 preds = model.predict(vect.transform(['y BYND dump?',
12                                       'CALLS ON BYND']))
13
14 >> [0, 1]
```

Labels

Model training

Unseen data

Predictions

NLP sentiment analysis accelerates financial forecasting but struggles with market-specific jargon and human nuance

Category	Advantages	Limitations
Speed & Scale	**Real-Time Processing** of massive, unstructured data volumes (news, social).	Difficulty distinguishing genuine sentiment from **Data Noise** (bots, manipulation).
Insight	**Early Signal Detection** for market shifts and impending events.	Struggles with **Domain Ambiguity** (financial jargon and context-specific terms).
Analysis	**Scalability** for monitoring global markets and thousands of assets.	Poorly handles complex human language like **Negation and Sarcasm**.
Accuracy	Enhances **Due Diligence** and **fundamental analysis**.	Vulnerable to **Bias in Training Data** and fails at long-term **Context Dependency**.

Section 1 | Introduction to Natural Language Processing and Sentiment Analysis

Section 2 | State of the Field

Sentiment Analysis in Finance

- Trump tweets and the currency market (2025)
- Pump and dump detection (2025)
- Adversarial attacks on financial sentiment with LLMs (2023)
- Financial distress prediction during earnings calls (2023)

Section 3 | Blueprint for Building a Trading Strategy Based on Sentiment Analysis

Section 4 | Trading via r/wallstreetbets Sentiment

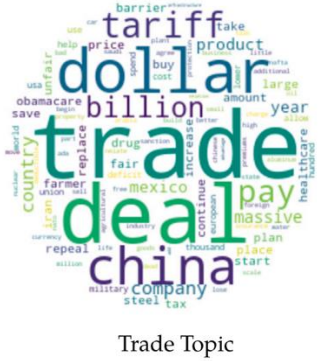
Section 5 | Concluding Remarks

Trump's Tweets containing macroeconomic and trade content significantly appreciate the U.S. dollar and reduce foreign exchange (FX) volatility

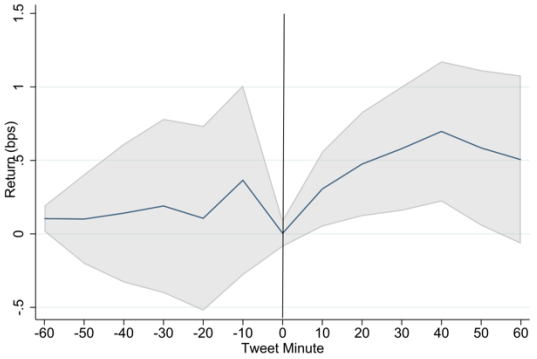
The study uses textual analysis to identify the subset of tweets that contain information on macroeconomic policy then measure their systematic impact on FX market outcomes like USD returns and volatility

Methodology

1. **Tweet filtering**: Use textual analysis (Dictionary and Topic Modeling) to isolate macroeconomic and trade policy Tweets
2. **Impact**: Track high-frequency data for 14 currency pairs to measure changes in **USD Returns, Volatility, and Trading Volume** around the exact minute a tweet was posted
3. **Causality**: Employ robust panel regressions and event studies to confirm that the Tweets uniquely caused **USD appreciation** and a **decline in market volatility**, independent of scheduled economic news

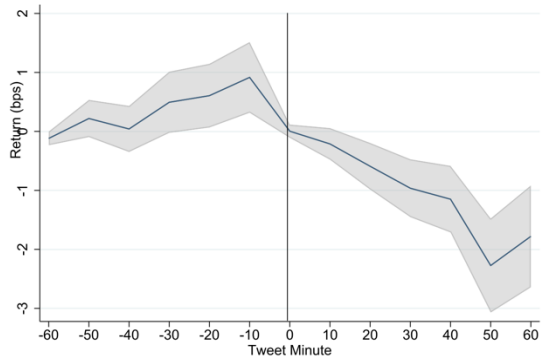


Trade Topic



Positive tweets v Normalized return of 14 currencies

Macroeconomics Topic

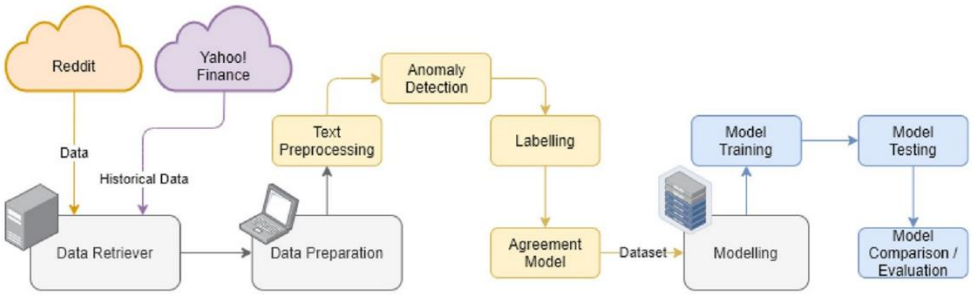


Negative tweets v Normalized return of 14 currencies

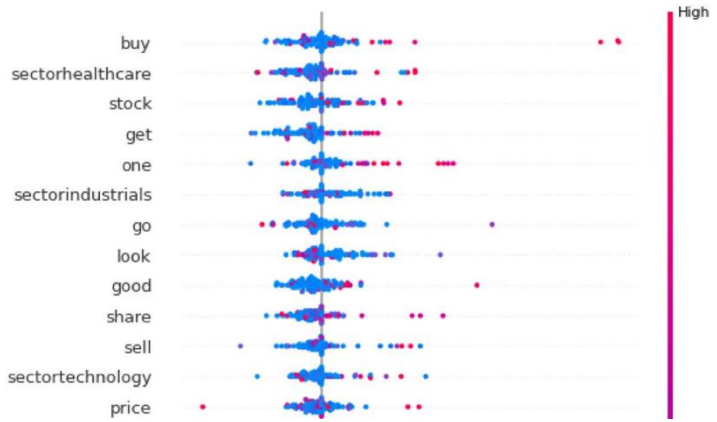
Filippou, I., Gozluklu, A. E., Nguyen, M. T., & Viswanath-Natraj, G. (2025). Signal in the noise: Trump tweets and the currency market. Journal of International Money and Finance, 156, 103343.

Combining NLP to analyze manipulative language on social media and anomaly detection on stock price/volume patterns can effectively detect Pump & Dumps

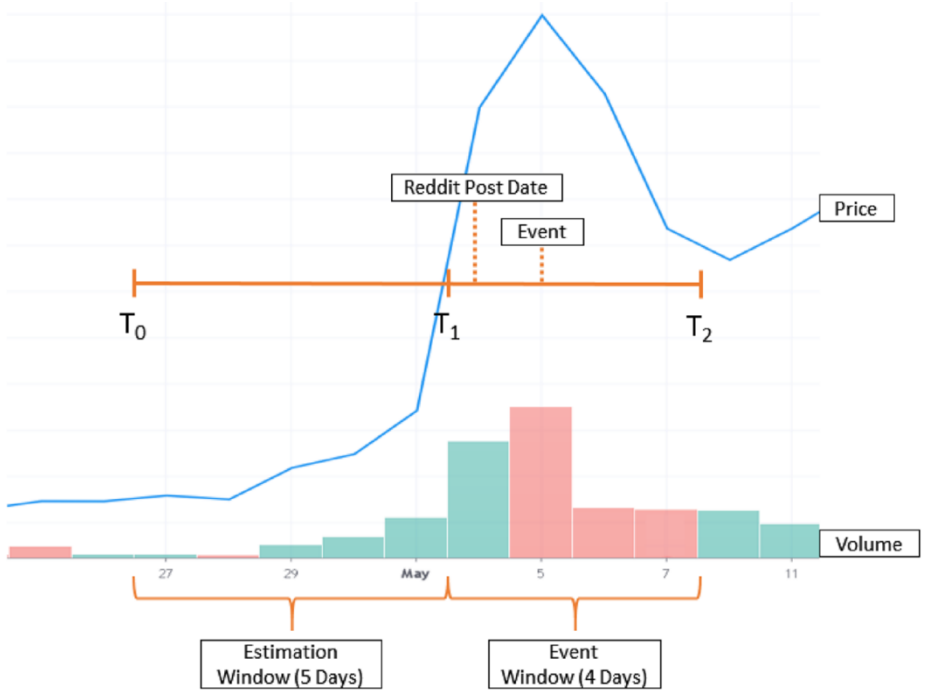
Predictive models analyzing language from online social media forums successfully detected Pump & Dump stock manipulations with an accuracy of 85% and an F1-score of 62%



Experiment Workflow



Feature Importance



Time Window for Data Collection and Analysis

Nam, D., & Skillicorn, D. B. (2025). Detecting Pump & Dump stock market manipulation from online forums. Digital Finance, 7(1), 1–20. https://doi.org/10.1007/s42521-024-00121-4

Adversarial attacks on financial sentiment with LLMs (2023)

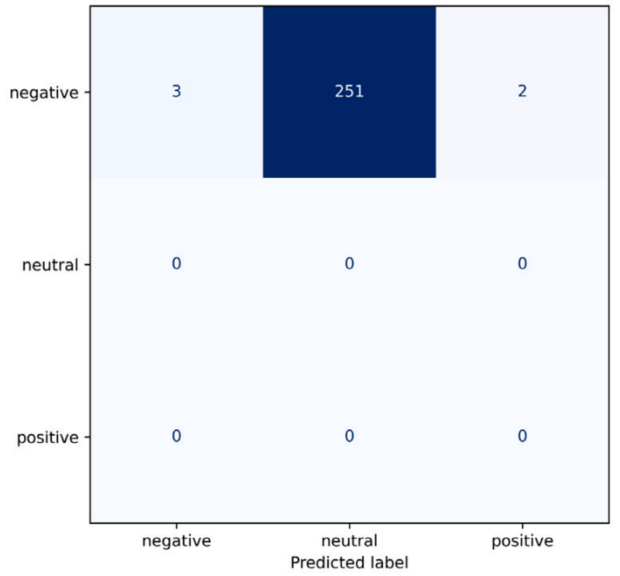
LLMs can effectively manipulate financial sentiment to deceive traditional keyword-based analysis methods, while context-aware models demonstrate greater robustness

This study focused on taking sentences originally classified as negative and altering them using GPT-3 so that sentiment models would classify them as neutral or positive

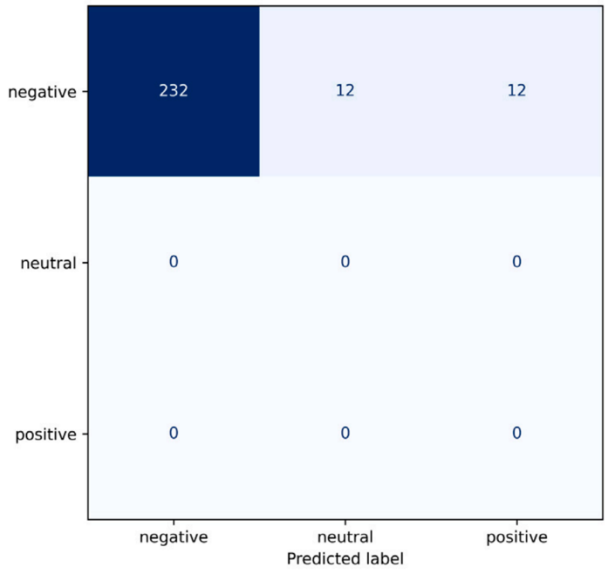
Methodology

1. **Dataset:** Identify a sample of sentences with a confirmed negative sentiment
2. **Adversarial Attack (GPT-3):** Replace negative words with non-negative synonyms
3. **Evaluation:** Classify the attacked sentences using the **Keyword-based approach** and **FinBERT**
4. **Comparison:** Measure the percentage of attacked sentences that the classifiers incorrectly label as neutral or positive

(A) Attacked KW-Sentiment



(B) Attacked FB-Sentiment



Leippold, M. (2023). Sentiment spin: Attacking financial sentiment with GPT-3. Finance Research Letters, 55, 103957

Financial distress prediction during earnings calls (2023)

Integrating speech emotions and text sentiment from earnings calls with financial data significantly improves the accuracy of dynamic financial distress prediction

This paper synthesizes multimodal (text, audio, and financial) methods for dynamic financial-distress forecasting, highlighting the added value of sentiment/emotion features and LSTM-based modeling

LSTM Model for Financial Stress Prediction

1. **Emotional Speech:** A **pre-trained CNN model** classified audio recordings from earnings calls into 8 emotional states using 180 spectral features
2. **Text Sentiment:** The domain-specific **FinBERT** model and dictionary-based text emotional features were used for sentiment scoring
3. **Financial Indicators: 20 financial features** (e.g., liquidity, profitability) and the Altman's Z-score from quarterly financial statements were included

Top sentiment and emotional features on model output

Rank	Financial distress	Safe position
1	Anticipation (text)	Joy (text)
2	Trust (text)	FinBERT sentiment (text)
3	Surprise (text)	Sadness (text)
4	Disgust (text)	Fear (text)
5	Fear (speech)	Happy (speech)

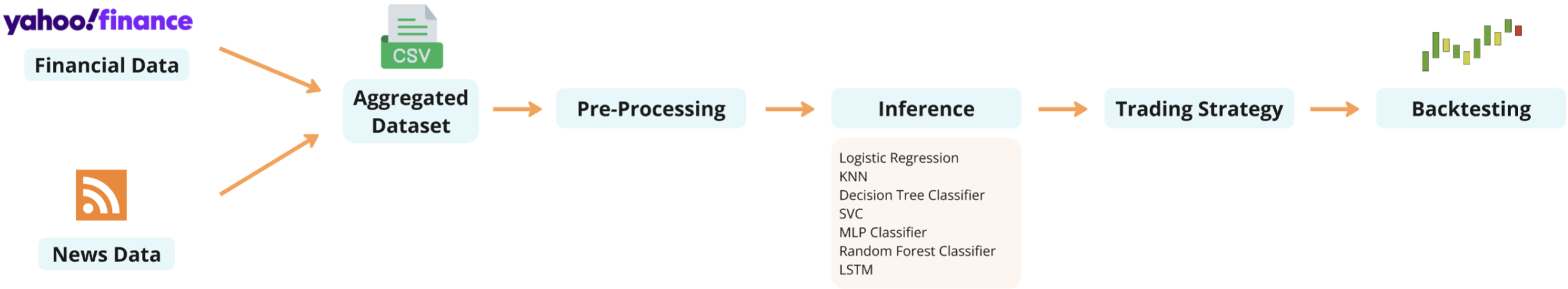
Neural Computing and Applications. (2023). Financial distress prediction: A multimodal dynamic approach. 35, 21463–21477. https://doi.org/10.1007/s00521-023-08470-8

Trump tweets and the currency market (2025)	Tweets and news can move markets
Pump and dump detection (2025)	Be wary of the financial ‘news’ you see online!
Adversarial attacks on financial sentiment with LLMs (2023)	News articles might use weird words to **trick NLP algorithms**
Financial distress prediction during earnings calls (2023)	Sentiment analysis is a **multi-modal problem** (video, audio, text)

Section 3 | Blueprint for Building a Trading Strategy Based on Sentiment Analysis

Objective

Our goal is to use NLP to extract information from news headlines, assign a sentiment, and then build a trading strategy



Finance Dataset

Yahoo Finance provides historical price data for a given ticker

APPL, MSFT, AMZN, GOOG, WMT, JPM, TSLA, NFLX, and ADBE data was retrieved from 2010 to 2018

```python
# 1. Initialize an empty list to store all the individual DataFrames
all_data_frames = []

for ticker in tickers:
    ticker_yf = yf.Ticker(ticker)

    # Fetch the history for the current ticker
    data_temp = ticker_yf.history(start=start, end=end)

    # Add the ticker identifier column
    data_temp['ticker'] = ticker

    # 2. Append the new DataFrame to the list
    all_data_frames.append(data_temp)

# 3. Concatenate all DataFrames in the list into one final DataFrame
df_ticker_return = pd.concat(all_data_frames)
```

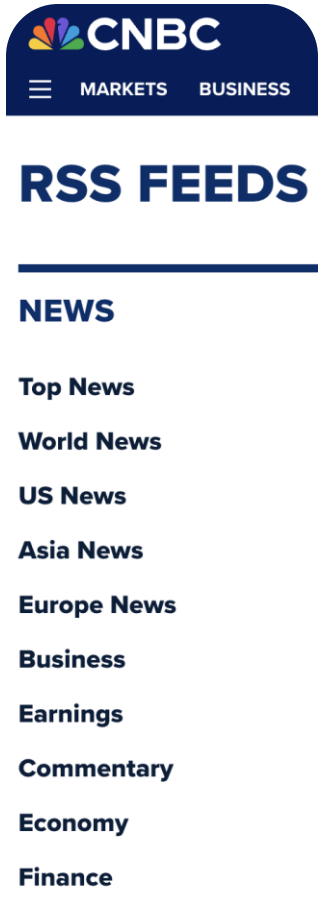
Data Query

Date	Open	High	Low	Close	Volume	Dividends	Stock Splits	ticker
2010-01-04 00:00:00-05:00	6.40	6.43	6.37	6.42	493729600	0	0	AAPL
2010-01-05 00:00:00-05:00	6.44	6.47	6.40	6.43	601904800	0	0	AAPL
2010-01-06 00:00:00-05:00	6.43	6.45	6.32	6.33	552160000	0	0	AAPL
2010-01-07 00:00:00-05:00	6.35	6.36	6.27	6.32	477131200	0	0	AAPL
2010-01-08 00:00:00-05:00	6.31	6.36	6.27	6.36	447610800	0	0	AAPL
2010-01-11 00:00:00-05:00	6.38	6.39	6.25	6.30	462229600	0	0	AAPL
2010-01-12 00:00:00-05:00	6.27	6.29	6.19	6.23	594459600	0	0	AAPL
2010-01-13 00:00:00-05:00	6.23	6.33	6.12	6.32	605892000	0	0	AAPL
2010-01-14 00:00:00-05:00	6.30	6.31	6.27	6.28	432894000	0	0	AAPL
2010-01-15 00:00:00-05:00	6.33	6.35	6.17	6.18	594067600	0	0	AAPL
2010-01-19 00:00:00-05:00	6.25	6.45	6.22	6.45	730007600	0	0	AAPL
2010-01-20 00:00:00-05:00	6.45	6.46	6.28	6.35	612152800	0	0	AAPL
2010-01-21 00:00:00-05:00	6.36	6.40	6.21	6.24	608154400	0	0	AAPL
2010-01-22 00:00:00-05:00	6.20	6.22	5.91	5.93	881767600	0	0	AAPL

Daily Price Data

The dataset consists of 9470 news headlines spanning 30 tickers from 2011-2018

Data was sourced from News RSS feeds, then RegEx was used to extract ticker information



CNBC News RSS Feed

```
<item>
<link>https://www.cnbc.com/2025/11/27/green-light-away-from-ai-trade-two-etf-ceos-see-a-key-market-shift.html</link>
<guid isPermaLink="false">108232454</guid>
<metadata:type>cnbcnewsstory</metadata:type>
<metadata:id>108232454</metadata:id>
<metadata:sponsored>false</metadata:sponsored>
<title>'Green light' away from AI trade: Two ETF executives see a key market shift underway </title>
<description>
<![CDATA[ Markets may have entered a new cycle. Here's why. ]]>
</description>
<pubDate>Thu, 27 Nov 2025 21:00:01 GMT</pubDate>
</item>

<item>
<link>https://www.cnbc.com/2025/11/27/sec-investigates-jefferies-over-first-brands-collapse-report-says.html</link>
<guid isPermaLink="false">108233065</guid>
<metadata:type>cnbcnewsstory</metadata:type>
<metadata:id>108233065</metadata:id>
<metadata:sponsored>false</metadata:sponsored>
<title>SEC investigates Jefferies over First Brands collapse, report says</title>
<description>
<![CDATA[ The FT said the SEC is looking into whether Jefferies gave investors enough information on their exposure to the failed auto business. ]]>
</description>
<pubDate>Thu, 27 Nov 2025 18:02:34 GMT</pubDate>
</item>

<item>
<link>https://www.cnbc.com/2025/11/26/stocks-making-the-biggest-moves-midday-dell-arwr-urbn.html</link>
<guid isPermaLink="false">108232531</guid>
<metadata:type>cnbcnewsstory</metadata:type>
<metadata:id>108232531</metadata:id>
<metadata:sponsored>false</metadata:sponsored>
<title>Stocks making the biggest moves midday: Dell, Arrowhead Pharmaceuticals, Urban Outfitters & more</title>
<description>
<![CDATA[ These are the stocks posting the largest moves midday. ]]>
</description>
<pubDate>Wed, 26 Nov 2025 17:06:02 GMT</pubDate>
</item>
```

RSS Data (XML)

datetime	headline	ticker
9/9/2019 8:18	$JPM $BAC:Trump°¶s Trade War Starts t	JPM
9/9/2019 7:05	We also explained the importance of $GD	XOM
9/9/2019 6:55	$COST $HD Worth keeping an eye on aga	HD
9/9/2019 6:49	We own some Verizon for the dividend va	VZ
9/9/2019 5:34	$NRG $HD $LOW $AAPL $GS:Chances of	HD
9/9/2019 10:01	Coca-Cola, Merck and United Health go e	MRK
9/8/2019 9:21	Goldman Sachs Group $GS Downgraded	GS
9/8/2019 9:19	$MSFT Trading News - High Win Rate Trig	MSFT
9/8/2019 8:31	Merck & Co., Inc. $MRK Earns Buy Rating f	MRK
9/8/2019 8:03	Merck & Co., Inc. $MRK Earns Buy Rating f	MRK
9/8/2019 7:31	$XOM Entering equal legs area $70.58 - $	XOM
9/8/2019 6:10	Morgan Stanley Raises Walmart $WMT Pr	WMT
9/8/2019 3:39	IBM $IBM Upgraded to °ßHold°® at ValuEn	IBM
9/8/2019 22:50	Coca-Cola Consolidated $COKE Cut to H	KO
9/8/2019 22:17	$CSCO - Cisco Is A Buy Following Post-E	CSCO
9/8/2019 21:10	$PG - Procter & Gamble: Factors Underlyi	PG
9/8/2019 21:04	$MSFT $SPX:Microsoft CEO Satya Nadell	MSFT
9/7/2019 9:14	so is $CAT a bellwether for the global ecc	CAT
9/7/2019 9:10	Intel Co. $INTC CFO Acquires $249,430.6	INTC
9/7/2019 7:25	Richard Carapaz misses Tour of Britain aft	V
9/7/2019 7:19	Home Depot $HD Upgraded to °ßBuy°® by	HD
9/7/2019 7:13	$CSCO $T:2 Cheap Dividend Stocks You	CSCO
9/7/2019 6:50	On 8/26 $WMT triggered for me by going	WMT
9/7/2019 5:31	JPMorgan Chase & Co. $JPM Rating Incre	JPM
9/7/2019 5:25	Timothy Patrick Flynn Purchases 2,000 Sh	UNH

Conversion to CSV and Ticker Identification

Aggregated Data

Price and news data are merged based on ticker and date

"eventRet" measures the price change spanning from the day before the news release to the day following it

ticker	headline	date	eventRet	Close
AMZN	Whole Foods (WFMI) -5.2% following a downgrade	2011-05-02	0.02	10.06
NFLX	Netflix (NFLX +1.1%) shares post early gains after C	2011-05-02	-0.01	3.39
MSFT	The likely winners in Microsoft's (MSFT -1.4%) Sky	2011-05-10	-0.02	19.70
MSFT	Microsoft (MSFT -1.2%) and Skype signed their dea	2011-05-10	-0.02	19.70
MSFT		2011-05-10	-0.02	19.70
AMZN	Amazon.com (AMZN -1.7%) shares slip as commen	2011-05-12	-0.01	10.30
GOOG	It's been some time coming, but Google (GOOG -1	2011-05-16	-0.01	12.82
MSFT	Accusing underwriters of digging out their late 19	2011-05-19	0.00	19.09
MSFT	If you bought LinkedIn (LNKD, now legging higher	2011-05-19	0.00	19.09
MSFT	From Jens Heycke, the top 5 things you could buy	2011-05-19	0.00	19.09
MSFT	"OK. This is almost hilarious. P/E Ratio for $LNKD ri	2011-05-19	0.00	19.09
MSFT	LinkedIn (LNKD) is off to an astonishing start on its	2011-05-19	0.00	19.09
MSFT		2011-05-20	-0.02	18.92
AMZN	Amazon (AMZN -1.2%) moves nicely off its low for	2011-05-23	-0.03	9.81
JPM	The investigation by New York AG Eric Schneiderr	2011-05-23	-0.04	28.65
MSFT	LinkedIn (LNKD) shares could fall by 50%, suggest	2011-05-23	-0.02	18.67
MSFT		2011-05-24	-0.01	18.65
NFLX	Today's strength in Netflix (NFLX +4.8%) is attribut	2011-05-25	0.06	3.71
TSLA	Tesla Motors (TSLA +3.6%) announces a follow-on	2011-05-25	0.10	1.93

Merged Dataset

spaCy provides an all-in-one NLP preprocessing function via en_core_web_lg

Each headline is converted to an embedding (numeric vector) of length 300

Pre-trained embeddings

(Headlines, Vector Dimensions)

Vector of the 1st headline

```
1  import en_core_web_lg
2
3  nlp = en_core_web_lg.load()
4
5  all_vectors = np.array([np.array([token.vector for token in nlp(s)])
6                          .mean(axis=0)*np.ones((300)) for s in sentiments_data['headline']])
7
8  all_vectors.shape
9
10 >> (9470, 300)
11
12
13 all_vectors[0]
14
15 >> array([ 2.47068703e-04,  1.47522390e-01, -9.74819958e-02, -1.23402007e-01,
16            2.00748015e-02, -1.31392732e-01, -6.77573532e-02, -1.31395116e-01,
17            1.56753976e-02,  1.95713353e+00, -1.32420510e-01,  1.92656741e-01,
18            ...])
```

Pre-processing Code

Inference is run on various ML models then evaluated against labeled data

Labeled data is provided by Kaggle, where each headline has a positive (1) or negative (0) label

```
models = []
models.append(('LR', LogisticRegression()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('SVM', SVC()))
models.append(('NN', MLPClassifier()))
models.append(('RF', RandomForestClassifier()))


for name, model in models:

    kfold = KFold(n_splits=num_folds, random_state=seed, shuffle=True)
    cv_results = cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)

    res = model.fit(X_train, Y_train)

    train_result = accuracy_score(res.predict(X_train), Y_train)
    train_results.append(train_result)

    test_result = accuracy_score(res.predict(X_test), Y_test)
    test_results.append(test_result)

    name
    >> LR

    cv_results.mean(), cv_results.std()
    >> 0.878714, 0.009366

    train_result, test_result
    >> 0.896063 0.875748
```

ML Models

K-Folds Eval

Model Training

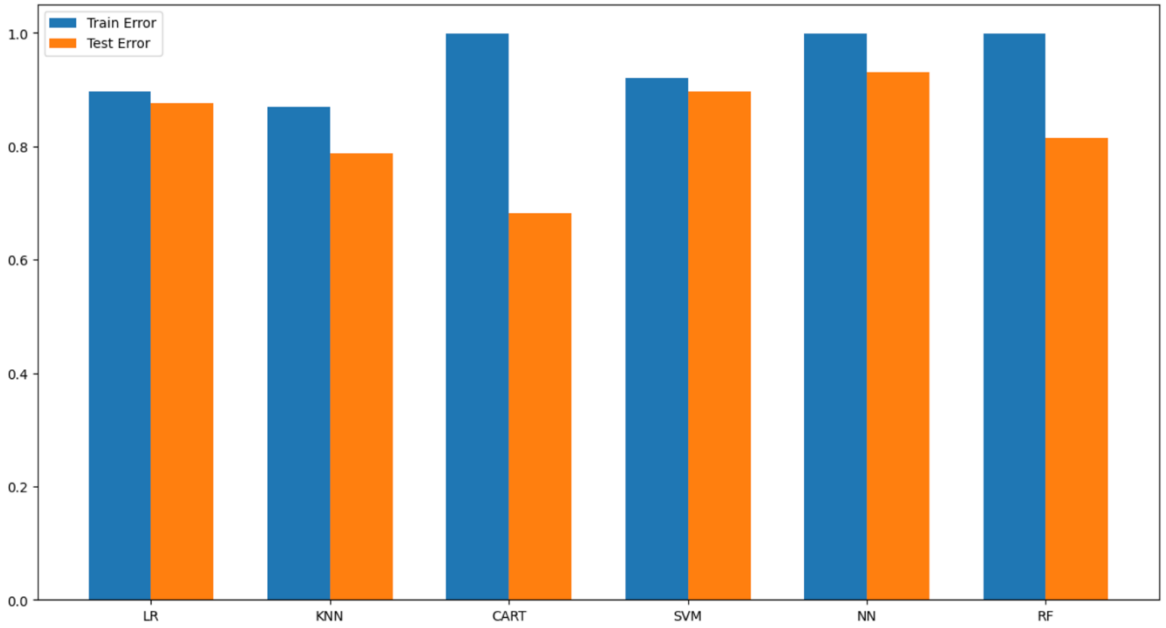
Training Results

Test Results

K-Fold Results

Eval Results

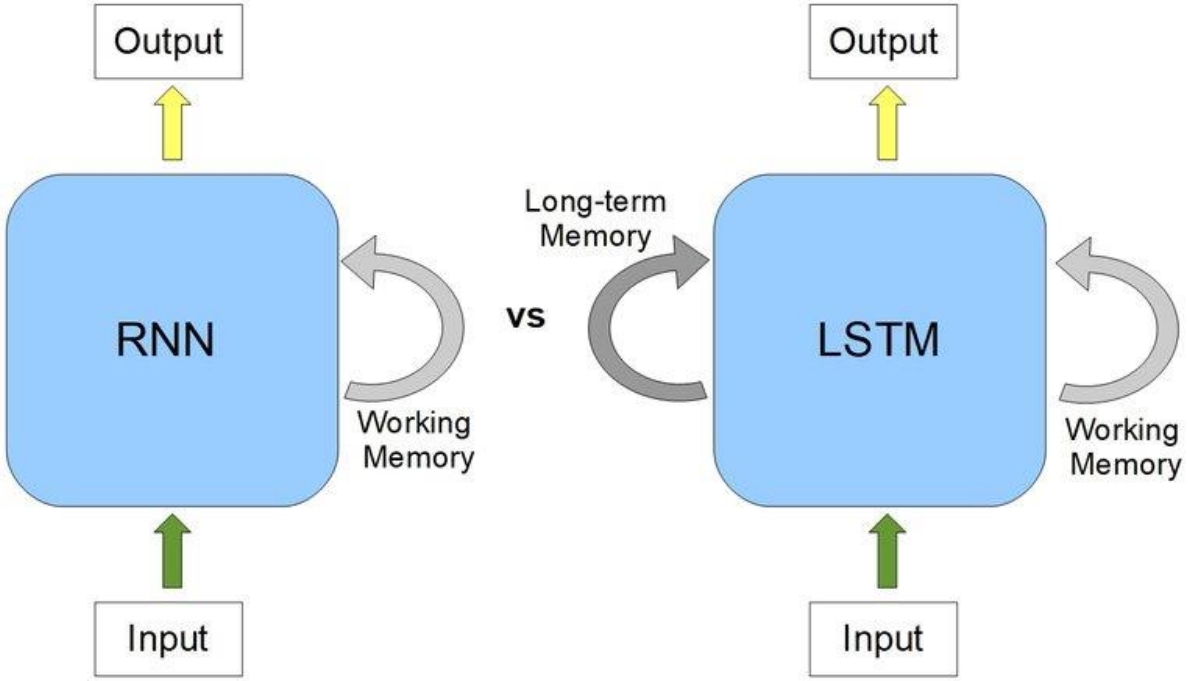
Training Code



Model Comparison

LSTMs are variants of recurrent neural networks that incorporate long-term memory to remember past data

LSTMs selectively remember important information and forget irrelevant details across long sequences of text, effectively capturing complex dependencies

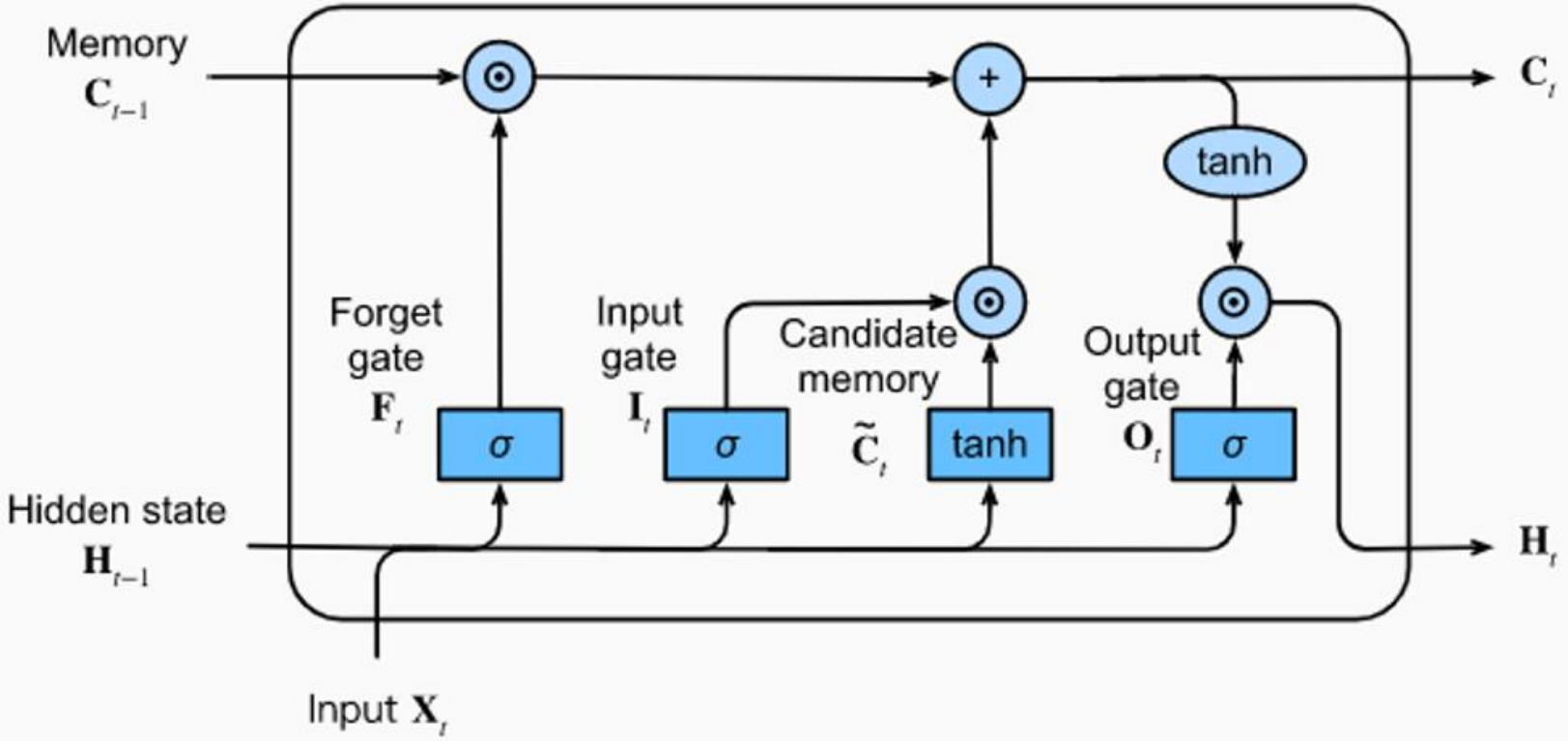


Recurrent Neural Network vs Long Short-Term Memory Model

PhenomNet: Bridging Phenotype-Genotype Gap: A CNN-LSTM Based Automatic Plant Root Anatomization System - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/RNN-v-s-LSTM-a-RNNs-use-their-internal-state-memory-to-process-sequences-of-inputs_fig1_341131167 [accessed 29 Nov 2025]

“BREAKING: TSLA Put Options Now Classified as Treason Under the Sedition Act of 1918; All Put Holders to Be Immediately Deported to El Salvador”

The **Forget Gate** disregards the serious context “treason” and “sedition” because the **Input Gate** recognizes the absurdity of the subsequent context “immediately deported to El Salvador”



Ottavio Calzone (2022). An Intuitive Explanation of LSTM. https://medium.com/@ottaviocalzone/an-intuitive-explanation-of-lstm-a035eb6ab42c

Results from the LSTM outperform classical ML methods according to test accuracy (97.7%)

```
1  def create_model(input_length=50):
2      model = Sequential()
3
4      model.add(Embedding(input_dim=20000, output_dim=300, input_length=input_length))
5      model.add(LSTM(units=100, dropout=0.2, recurrent_dropout=0.2, kernel=GlorotUniform))
6      model.add(Dense(1, activation='sigmoid'))
7
8      model.build(input_shape=(None, input_length))
9      model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
10
11     return model
12
13
14 model_LSTM = KerasClassifier(build_fn=create_model, input_length=50, epochs=3, verbose=1)
15 model_LSTM.fit(X_train_LSTM, Y_train_LSTM)
16
17 accuracy_score(model_LSTM.predict(X_train_LSTM), Y_train_LSTM)
18 >> 0.9987931814753357
19
20 accuracy_score(model_LSTM.predict(X_test_LSTM), Y_test_LSTM)
21 >> 0.9771207321365716
```

Architecture

Embedding Layer

LSTM Layer

Hidden Layer

Model Training

LSTM Code

Enter the trade if daily sentiment increased by 0.5 and the price is above the 15-day simple moving average (SMA)

This is a trend-following momentum-based strategy with a technical filter and sentiment catalyst

Trading Strategy Components

Backtesting was performed on multiple stock combinations across various time periods

The graph summarizes the backtesting results over time, including entries, exits, sentiment, and P/L

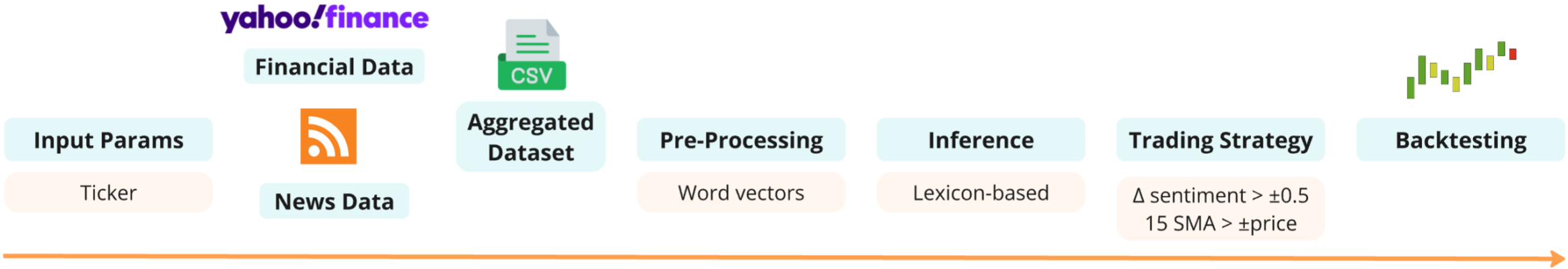
Cumulative P/L

Trade P/L

Price Graph

Sentiment

Backtesting Results



Data Limitations

- News articles provide **delayed sentiment**
- Single news source
- Small **dataset** (n=9470)
- Focused on **non-volatile stocks** (large cap)
- Did not **benchmark** against SP500

Model Limitations

- Simple **technical indicators** (15SMA)
- Simple **sentiment model** (Lexicon-based rather than ML-based)
- **Binary sentiments** (0 or 1)

Section 1 | Introduction to Natural Language Processing and Sentiment Analysis

Section 2 | State of the Field Sentiment Analysis in Finance

Section 3 | Blueprint for Building a Trading Strategy Based on Sentiment Analysis

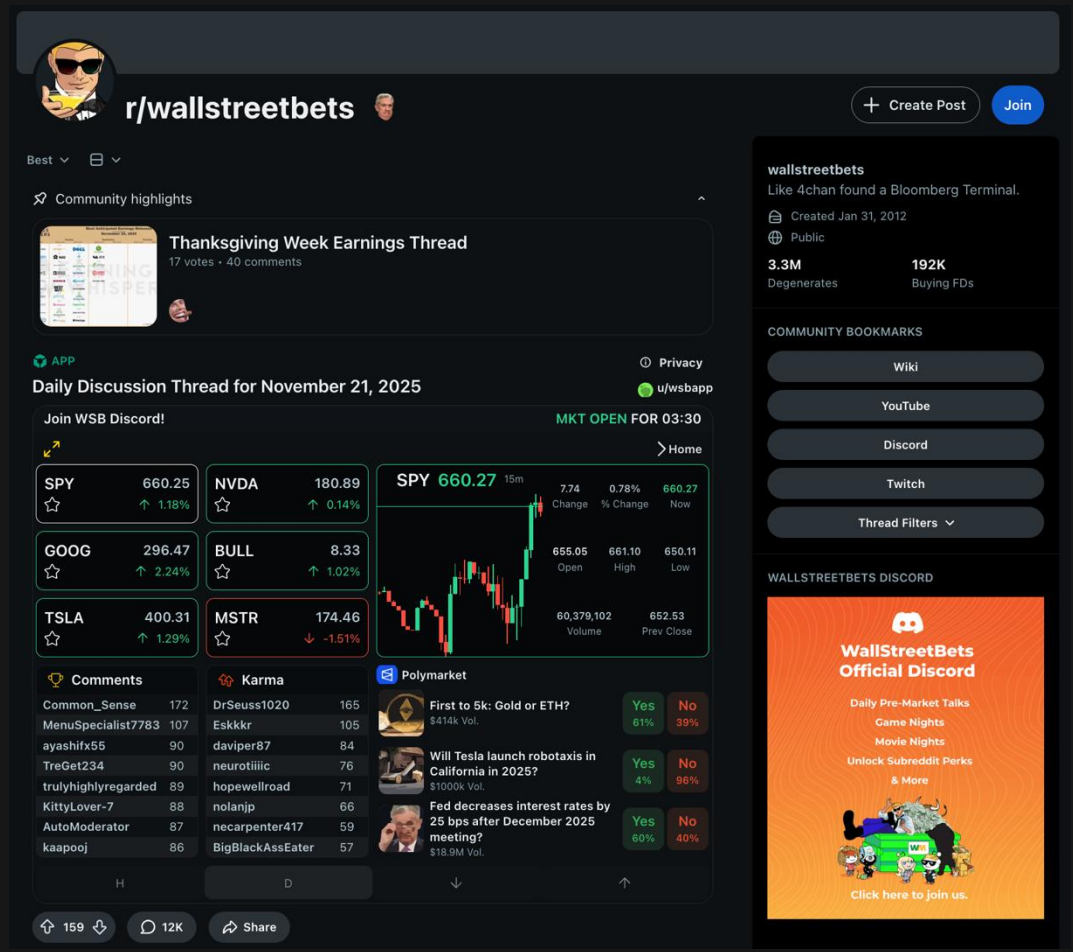
Section 4 | **Trading via r/wallstreetbets Sentiment**

Section 5 | Concluding Remarks

What is r/wallstreetbets?

r/wallstreetbets is a subreddit where retail investors discuss high-risk, highly speculative stock and option trading

The r/wallstreetbets **daily discussion thread** can provide **real-time** market news and **sentiment signals**



Limitations

- **Bots and trolls** create noise and false signals
- **Linguistic complexity** (typos, slang, sarcasm, ambiguity)
- **Low predictive power** (**reactive** sentiment, short-term focus)

reddit.com/r/wallstreetbets/

The terminology deviates from standard financial language, requiring models to be carefully trained to correctly interpret the true intent

Phrase	Meaning
Ber	Bear; someone who wants the price to go down
Bol	Bull; someone who wants the price to go up
Diamond hands	Holding onto a position for dear life
Paper hands	Closing a position early on
To the moon 🚀	Extreme optimism
Hedgies	Hedge funds
Bagholder	An investor left holding the “bag” after a price crash
Wendy’s	The traditional, unofficial employment location for WSB traders who have lost all their money
Regard	A euphemism for “retard”; a bad trader
Exit liquidity	Bagholders who are used for liquidity by large investors
0dte	0 days-to-expiry options

BearyChristmas223 • 1mo ago

Things visible from space:

- Great Wall of China
- Giza pyramids
- The bags of BYND shareholders

Main-Economist67 • 19d ago

I refuse to take any accountability for my trades. All of my losses are due to bad luck, all of my gains are due to skill.

bullrfuk • 19d ago

WSB whenever there is the slightest disturbance: "WW3 is coming"

WSB when the disturbance ends: "it was so obvious"

The BEAR after earnings!

GemmyBoy999 • 28d ago

I'm somewhat of a wall street philanthropist myself

Visual_Enthusiasm_73 • 17d ago

Quick! Everybody take out loans and lines of credits and let's pump this back up!

jsie-iaiqhsi816278 • 3mo ago

☝it will be

👈frankly 👉

the greatest👌

🖐economic crash🖐

👌in the

☝history of

👈 amaerica👉

They will say. Mr 🍊 no one can crash it like you 👈👉

hv876 • 7d ago

Top 1% Commenter

They called you an investor

---Right--Tackle--- • 18d ago

Top 1% Commenter

At this rate I'm gonna have to return my groceries

Sweg_OG • 18d ago

too scared to even open the app anymore

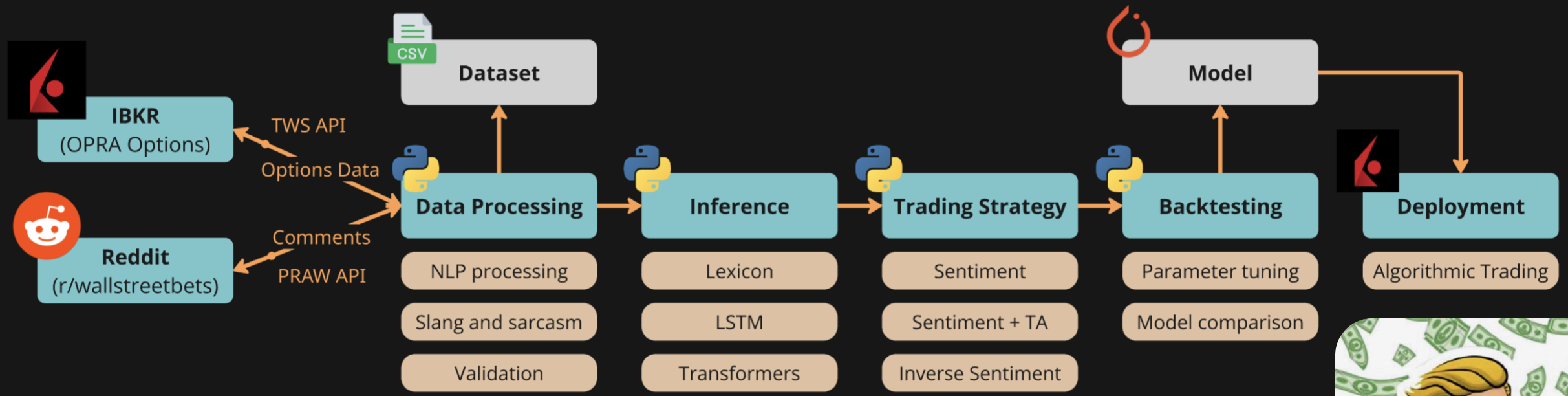
KittyLover-7 • 6mo ago

Top 1% Commenter

My greed is the greed they warned about in the bible

Develop a trading algorithm for 0DTE SPY options using r/wsb sentiment

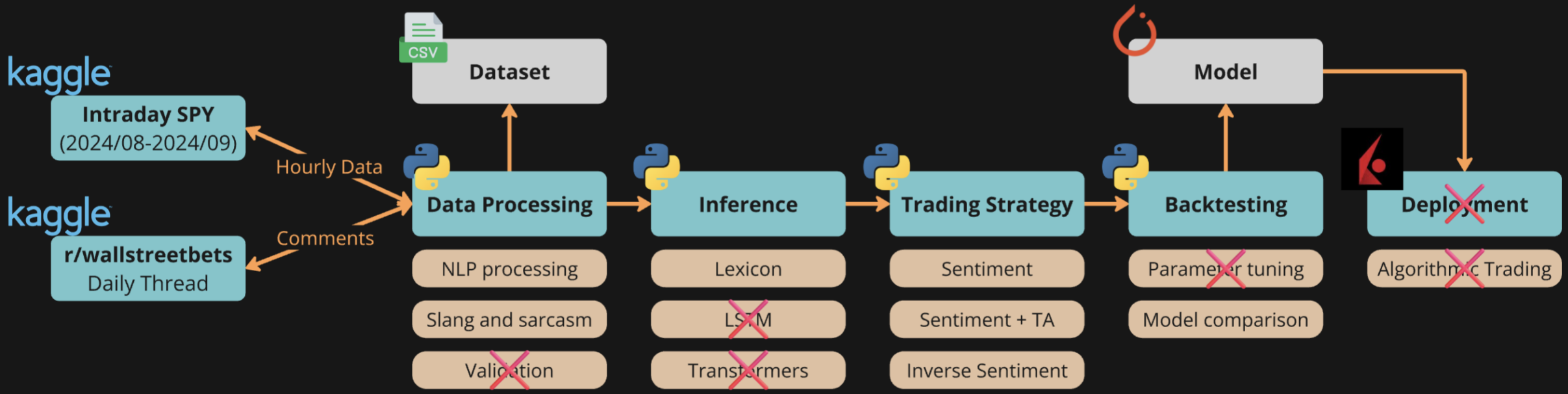
IBKR and Reddit APIs provide real-time data for algorithmic trading



The new plan after I realized I had no clue what I was doing

Acquiring high quality data is difficult!

Good data needs $$$ (i.e. r/wallstreetbets batch data, labeled sentiment data)



What are 0DTE SPY options?

Zero Days To Expiration (0DTE) SPY options are contracts tied to the SP500 ETF (SPY) that expire on the same day they are traded

0DTE options offer the potential **for extremely high, leveraged returns** on small price movements due to their low premium cost and rapid directional sensitivity

$SPY 684C Nov 28 (0DTE)

High = 0.76

Open = 0.35

Low 0.01

$SPY ETF Nov 28

High = 683.66

Low = 680.50

Disclaimer

DO NOT TRADE NAKED 0DTE CALLS, (especially on margin) YOU WILL GET EVAPORATED

THIS IS ~~FINANCIAL~~ FRIENDLY ADVICE!!!

Disclaimer

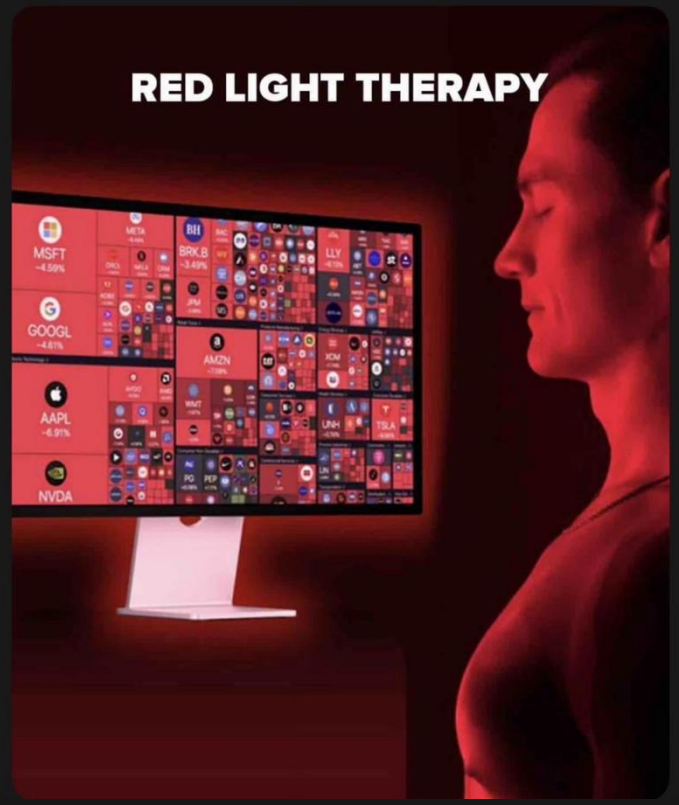
This could be you.

Trade with **caution**!!

Skincare first ladies & gentlemen 💪

Laughing through the pain 🩸🩸🤦🏻‍♀️🤦🏻‍♀️

$VFV $TSLA $MSTR $ASML $GOOGL $NVDA $PLTR



Kaggle contains intraday $SPY ETF and options data

I couldn't find any intraday $SPY options data so I fell back to using ETF data with higher lot sizes

quote_unixtime	dte	underlying_last	c_delta	c_gamma	c_theta	c_iv	c_volume	c_bid	c_ask	strike
1630526400	0	451.85	1	0	-0.00053		726	4.65	4.79	447
1630526400	0	451.85	0.95299	0.04262	-0.04563	0.11477	2407	3.67	4.1	448
1630526400	0	451.85	1	0	-0.0003		2955	2.4	3.07	449
1630526400	0	451.85	1	0	0		5519	1.68	1.79	450
1630526400	0	451.85	1	0	-0.00057		22637	0.71	0.78	451
1630526400	0	451.85	0.29441	1.16447	-0.05499	0.01498	134982	0.05	0.06	452
1630526400	0	451.85	0.05114	0.14827	-0.01562	0.03466	209154	0.01	0.02	453
1630526400	0	451.85	0.01411	0.0371	-0.00529	0.04878	71909	0	0.01	454
1630526400	0	451.85	0.01091	0.02065	-0.00513	0.06738	24042	0	0.01	455
1630526400	0	451.85	0.00889	0.01345	-0.00538	0.08576	6068	0	0.01	456
1630699200	0	453.08	0.93179	0.07373	-0.03555	0.15611	2525	3.88	4.33	449
1630699200	0	453.08	0.96046	0.11036	-0.00477	0.11438	12496	3	3.15	450
1630699200	0	453.08	0.94354	0.1873	-0.00512	0.08583	20724	1.96	2.19	451
1630699200	0	453.08	1	0	0		114804	1	1.1	452
1630699200	0	453.08	0.51334	1.54214	-0.06519	0.02152	234622	0.12	0.15	453
1630699200	0	453.08	0.02369	0.09933	-0.00517	0.02836	199570	0	0.01	454
1630699200	0	453.08	0.01436	0.03771	-0.00537	0.04791	78033	0	0.01	455
1630699200	0	453.08	0.01071	0.02091	-0.00505	0.06717	26995	0	0.01	456
1630699200	0	453.08	0.00914	0.01351	-0.00488	0.08536	8904	0	0.01	457

EOD $SPY Options
(2010 - 2023)
n = 4.2m

timestamp	open	high	low	close	volume
2024-08-19 4:00	559.54	559.54	553.98	554.15	523
2024-08-19 4:01	554.15	554.22	554.05	554.05	982
2024-08-19 4:02	554.07	554.09	553.95	553.95	1718
2024-08-19 4:03	553.95	553.98	553.68	553.68	1151
2024-08-19 4:04	553.68	553.79	553.65	553.77	5232
2024-08-19 4:05	553.7	553.74	553.3	553.3	1374
2024-08-19 4:06	553.29	553.36	553.19	553.31	114
2024-08-19 4:07	553.27	553.33	553.12	553.15	4268
2024-08-19 4:08	553.17	553.36	553.16	553.32	155
2024-08-19 4:09	553.25	553.5	553.25	553.5	1321
2024-08-19 4:10	553.5	553.61	553.49	553.58	1057
2024-08-19 4:11	553.64	553.66	553.5	553.61	1529
2024-08-19 4:12	553.61	553.66	553.56	553.66	213
2024-08-19 4:13	553.64	553.67	553.57	553.67	594
2024-08-19 4:14	553.63	553.63	553.5	553.5	29
2024-08-19 4:15	553.59	553.76	553.55	553.76	579
2024-08-19 4:16	553.71	553.92	553.71	553.82	133
2024-08-19 4:17	553.84	553.84	553.84	553.84	55
2024-08-19 4:18	553.79	553.88	553.71	553.71	1050

Intraday $SPY ETF
(2024/08– 2024/09)
n = 5.7k

SPY Options 2010-2023 EOD. https://www.kaggle.com/datasets/benjaminbtang/spy-options-2010-2023-eod?resource=download
SPY (SPDR S&P 500 ETF Trust) Intraday Data. https://www.kaggle.com/datasets/yoannjaisenlollmun/s-and-p-500-intraday-data-aug-19-2024-to-sep-17-2024

Kaggle hosts r/wallstreetbets datasets

The dataset was created by running the **Python Reddit API Wrapper (PRAW)** daily

author	datetime	score	text	tag
scott_jr	2023-06-16 20:36	1	Watch til 1 10	Meme
merakibret	2023-06-16 20:24	8	Entered an Iron Condor on ADBE yesterday at 455p 460p 500c and 505	Gain
VisualMod	2023-06-16 20:24	1	User Report Total Submissions 1 First Seen In WSB 1 we	Gain
VisualMod	2023-06-16 20:24	2	That was a very wise move	Gain
DreamcatcherEgg	2023-06-16 20:35	2	All you have to do is repeat this same winning trade 10 or 12x in a row to g	Gain
rebelo55	2023-06-16 20:38	2	You re doing it wrong Remember Buy high sell low img emote t5 2th5	Gain
AutoModerator	2023-06-16 20:24	1	Hey if you haven t already please reply to this comment with your positi	Gain
merakibret	2023-06-16 20:37	1	Honestly I ve been furiously learning how to trade options for a few year	Gain
lliorca336	2023-06-16 20:23	4	I don t care what you say YOOHOOS ARE DELISH	Weekend Discussion
ghostofwhiskey	2023-06-16 20:31	8	Juneteenth might be the dumbest thing I ve ever heard	Weekend Discussion
lotus_bubo	2023-06-16 20:34	3	For months the proportion of options to share volume keeps widening a	Weekend Discussion
BunnyGoHops	2023-06-16 20:04	8	Happy Juneteenth all my black stock brokers	Weekend Discussion
Nerdcubing	2023-06-16 20:24	8	I just lost 3 weeks worth of profit in 3 technically 2 days Feeling ficking I	Weekend Discussion
HI_IM_MR_MEESEE	2023-06-16 20:00	7	enis	Weekend Discussion
ban_evader3	2023-06-16 20:01	5	Okay I need to 3x everything by next week	Weekend Discussion
Blind--Squirrel	2023-06-16 20:11	5	LMAO Bers can t get even one proper red day stairs down elevator up	Weekend Discussion
Neither_Meat8226	2023-06-16 20:15	5	I just need SPY to open at 442 on Tuesday	Weekend Discussion
Express-Campaign	2023-06-16 20:00	5	img emote t5 2th52 4267	Weekend Discussion
Rare-ish_Bird	2023-06-16 20:01	4	25	Weekend Discussion

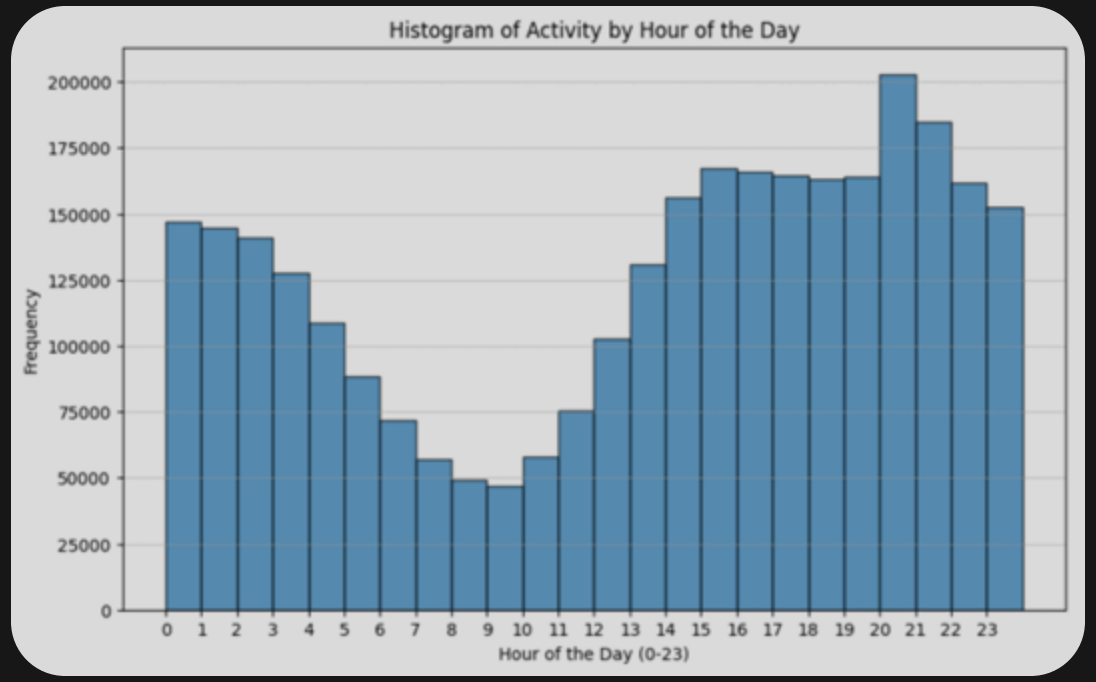
r/wallstreetbets Data
(2023 - 2025)
n = 415k (daily discussion)

415k datapoints / 2 years = 568 comments per day?

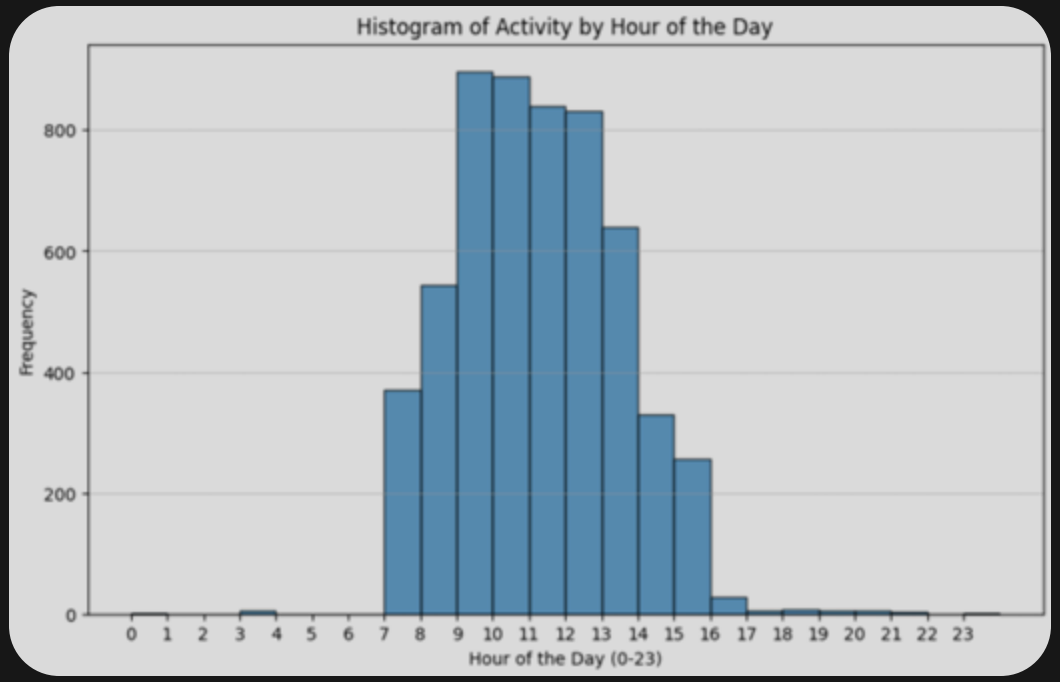
r/WallStreetBets Posts & Comments. https://www.kaggle.com/datasets/curiel/rwallstreetbets-posts-and-comments

There is lots of missing data within trading hours

I disaggregated comments based on the hour they were posted, then compared against a manually extracted dataset



Kaggle dataset



Manually extracted dataset
(25/11/28 daily discussion)

Pre-processing mainly targeted slang, stop words, emojis, and images

I took a gentler approach as the inference is limited to lexicon-based approaches which benefit from having more context

datetime	score	text	cleaned_text	entities
2024-08-30 10:36	9	New York has the gall to call itself the city that never sle	new york gall call city never sleeps nyse closes damn ca	CALL, CITY, DAMN, DAY, EVERY, GALL, NEVER, NEW, N
2024-08-30 10:05	8	I m scared lost 50 of my port yesterday img emote t	scared lost port yesterday	PORT
2024-08-30 10:12	7	Holiday ahead Don t do revenge trades today Loss is l	holiday ahead revenge trades today loss loss	DON, LOSS
2024-08-30 10:27	7	Whats that shampoo called again Head and shoulders	whats shampoo called head shoulders	AGAIN, AND, THAT, WHATS
2024-08-30 10:39	7	Can all the bulls join me in praying the away for today s	bulls join praying away today open	BULLS, FOR, JOIN, OPEN
2024-08-30 10:43	7	Seen too many eerie coincidences to not believe this su	seen many eerie coincidences believe sub heavily mon	EERIE, FOR, MANY, NOT, SEEN, SUB, THIS, TOO
2024-08-30 10:17	6	If you get naked your calls Print	naked calls print	CALLS, GET, NAKED, PRINT, YOU, YOUR
2024-08-30 10:25	6	Ban Bet Lost u ircphoenix made a bet that NVDA woul	ban bet lost u ircphoenix made bet nvda go within days	AND, BAN, BET, DAYS, DID, FOR, HTTP, JOIN, MADE, N
2024-08-30 10:51	6	Too traumatized from yesterday Is this a fake pump i	traumatized yesterday fake pump	FAKE, FROM, THIS, TOO
2024-08-30 10:00	10	Reminder to self just need 20 gain and not be greedy t	reminder self need gain greedy today	AND, GAIN, JUST, NEED, NOT, SELF
2024-08-30 10:25	5	Market just doesn t hit the same without a 3 trillion earn	market hit without trillion earnings report	DOESN, HIT, JUST
2024-08-30 10:28	5	The number of people commenting here is about to cra	number people commenting crater due lost nvda bank	ABOUT, DUE, FOR, HERE, NVDA
2024-08-30 10:30	5	Oh my god is it finally bear season My puts have been	oh god finally bear season puts getting cooked year	BEAR, BEEN, GOD, PUTS, YEAR
2024-08-30 10:53	4	As much as I d love a drop for a better buy in It ain t hap	much love drop better buy happening	AIN, DROP, FOR, LOVE, MUCH
2024-08-30 10:14	9	got back from orlando and im greeted by a bunch of ho	got back orlando greeted bunch homeless people hate	AND, BACK, BUNCH, FROM, GOT, HATE, NYC, ORLAN
2024-08-30 10:06	8	NVDA overnight crew is underappreciated	nvda overnight crew underappreciated	CREW, NVDA
2024-08-30 10:08	4	Ban Bet Lost u imsocuteimsopretty made a bet that R	ban bet lost u imsocuteimsopretty made bet rklb go wi	AND, BAN, BET, DID, FOR, HTTP, JOIN, MADE, NOT, RK
2024-08-30 10:21	4	good morning pre market regards	good morning pre market regards	GOOD, PRE
2024-08-30 10:33	4	I can buy 1 share of NVDA or I can pay to be with a hool	buy share nvda pay hooker mins hmmm	FOR, HMMM, MINS, NVDA, PAY, SHARE, WITH

Cleaned dataset and entities

Sentiment was estimated by various models, including pre-trained and lexicon-based

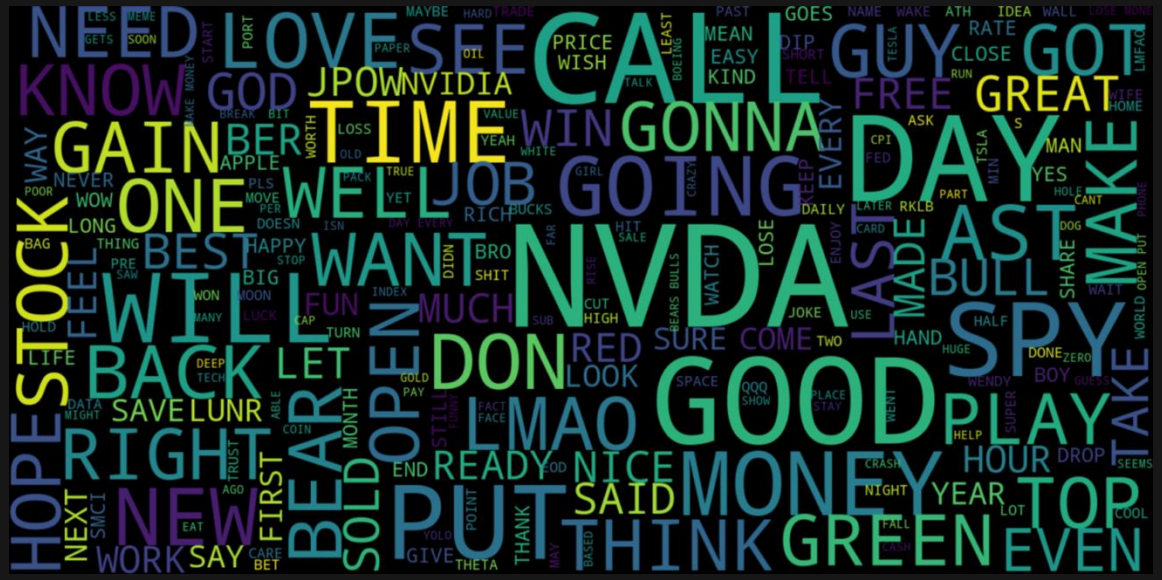
Supervised methods were unavailable due to a **lack of labeled data**

text	score_vader	score_lexicon	score_financial_lex	score_finbert	ensemble_sentiment	nltk_sentiment
New York has the gall to call itself the city that never	-0.4019	0	0.5	-0.0304499	0.016912529	-1
I m scared lost 50 of my port yesterday img emot	-0.6369	0	0	-0.1800176	-0.204229405	-1
Holiday ahead Don t do revenge trades today Loss	-0.6486	-0.5	-0.9	-0.5146028	-0.640800709	-1
Whats that shampoo called again Head and shoulde	0	0	0	-0.0215936	-0.005398397	0
Can all the bulls join me in praying the away for toda	0.5719	0.5	0.6	0.0492978	0.430299451	1
Seen too many eerie coincidences to not believe this	-0.3612	0	0	0.05637364	-0.07620659	-1
If you get naked your calls Print	0	1	0.5	-0.1187983	0.345300435	0
Ban Bet Lost u ircphoenix made a bet that NVDA w	-0.836	0	0	-0.020517	-0.214129257	-1
Too traumatized from yesterday Is this a fake pump	-0.7003	0	0	-0.4449662	-0.286316547	-1
Reminder to self just need 20 gain and not be greed	0.2732	0	0	0.15173679	0.106234197	1
Market just doesn t hit the same without a 3 trillion e	0	0	0	-0.8043476	-0.201086912	0
The number of people commenting here is about to	-0.25	0	0	-0.1490292	-0.099757306	-1
Oh my god is it finally bear season My puts have be	0.2732	-1	-0.55	0.17173053	-0.276267367	1
As much as I d love a drop for a better buy in It ain t	0.7184	0	0	0.05399267	0.193098168	1
got back from orlando and im greeted by a bunch of	-0.3818	0	0	-0.2034807	-0.146320178	-1
NVDA overnight crew is underappreciated	0	0	0	-0.9416267	-0.235406684	0
Ban Bet Lost u imsocuteimsopretty made a bet tha	-0.836	0	0	-0.0247155	-0.215178875	-1
good morning pre market regards	0.4404	0	0	0.13886596	0.14481649	1
I can buy 1 share of NVDA or I can pay to be with a h	0.2023	0.5	0.4	0.03761633	0.284979084	1

Sentiment signals from various methods

Word clouds visually represent frequent words in comments with positive or negative sentiment

Comments were disaggregated on sentiment (+1, 0, or -1), key words were identified, then word clouds were created based on word frequency



Prevalent words with positive sentiment

Prevalent words with negative sentiment

Methodology from: WallStreetBets Reddit Posts Analysis. https://www.kaggle.com/code/gpreda/wallstreetbets-reddit-posts-analysis#Data-visualization

Consolidated Dataset

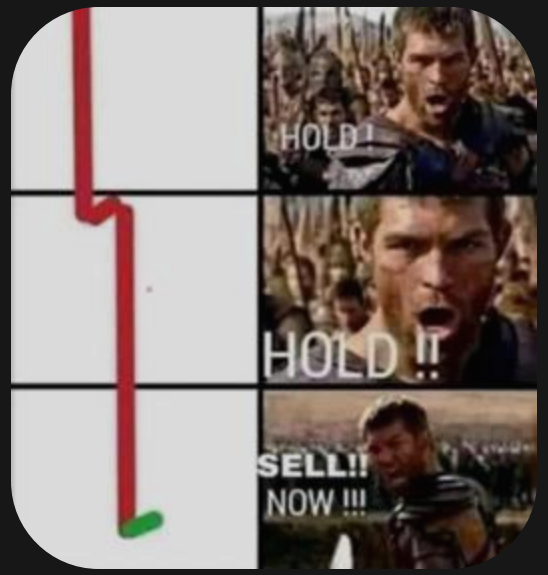
The finance and r/wsb datasets were merged on 60-minute time intervals

The sparsity of the r/wsb dataset (6033 comments over 22 days) necessitated longer time intervals

timestamp	open	high	low	close	volume	ensemble_sentiment	nltk_sentiment	score_finbert	post_volume
2024-08-21 9:00	559.77	560.45	559.35	560.35	3011844	0	0	0	0
2024-08-21 10:00	560.37	562.11	558.58	559.28	6542591	-0.02923887	0	-0.1302242	16
2024-08-21 11:00	559.225	560.33	558.55	558.89	3924618	0.105224272	0.14285714	0.0634399	28
2024-08-21 12:00	558.89	560.31	554.73	560.13	3158057	0.004626789	-0.02439024	-0.0605554	41
2024-08-21 13:00	560.14	560.54	559.88	560.3999	1821014	0.031020412	0	-0.0238197	73
2024-08-21 14:00	560.43	561.71	560.18	560.75	3998324	-0.068828206	-0.0990099	-0.1249868	101
2024-08-21 15:00	560.74	561.05	559.22	560.7	11372890	-0.01613846	0.03448276	-0.1091561	29
2024-08-21 16:00	560.7	560.83	560.58	560.83	1874328	0.07027151	0.5	0.060886	2
2024-08-22 9:00	562.56	563.18	561.63	562.3599	4142169	-0.110831971	-0.33333333	0.1324166	3
2024-08-22 10:00	562.38	562.475	560.16	560.53	4718841	-0.001223657	-0.06666667	-0.075528	15
2024-08-22 11:00	560.53	561.2499	557.06	557.7401	5498801	0.020867507	0.09615385	-0.099528	52
2024-08-22 12:00	557.73	560.07	557.261	557.58	4992877	0.026354441	0.09090909	-0.0629848	77
2024-08-22 13:00	557.5894	558.42	555.24	555.52	5579650	0.043531714	0.046875	-0.000757	64
2024-08-22 14:00	555.53	557.3	555.265	556	4125473	-0.008577098	-0.1147541	-0.0538898	61
2024-08-22 15:00	556.03	557.18	554.98	556.21	14459445	0.026160496	-0.06122449	-0.0103839	49
2024-08-22 16:00	556.2	556.35	555.74	556.18	3756251	0	0	0	0
2024-08-23 9:00	559.53	559.83	558.83	559.78	3683878	-0.019153367	-0.375	-0.065676	8
2024-08-23 10:00	559.77	563.09	558.35	560.53	12183113	0.055728339	0.10344828	0.0057616	29
2024-08-23 11:00	560.5	562.5	559.52	560	4892434	0.047403105	0.13953488	0.0155256	43

Merged dataset

Baseline Trading Strategies



HODL (Hold On for Dear Life)

- **Buy and hold**

DCA (Dollar Cost Averaging)

- **Incremental buying** (24hrs)

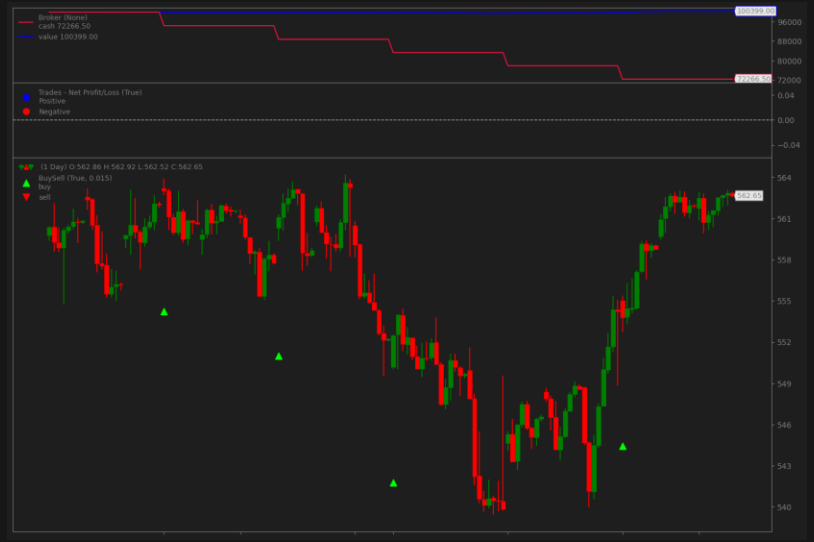
TA (Technical Analysis)

- **Buy** if price crosses 10 EMA and RSI is oversold (<30)
- **Sell** if price crosses 10 EMA and RSI is overbought (>70)

Baseline Trading Strategies

HODL (Hold On for Dear Life)

- **P/L**: 0.41%
- **Sharpe**: -3.34
- **Max Drawdown**: 4.2%



DCA (Dollar Cost Averaging)

- **P/L**: 0.40%
- **Sharpe**: -20.28
- **Max Drawdown**: 0.58%

TA (Technical Analysis)

- **P/L**: 0.23%
- **Sharpe**: -9.25
- **Max Drawdown**: 1.29%

Sentiment Trading Strategies

The FinBERT sentiment was normalized from -1 to 1 and used as the signal

Sentiment-only

- **Buy**: sentiment > 0.2
- **Sell**: sentiment < -0.2

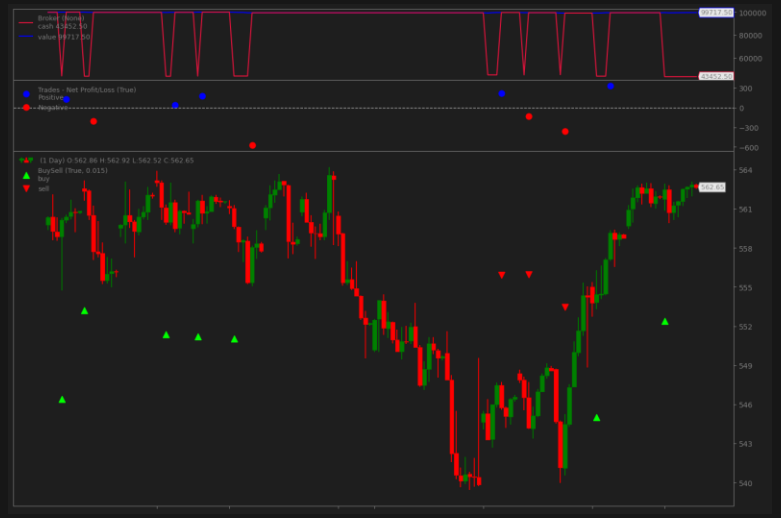
Sentiment + TA

- **Buy**: price > EMA and sentiment > 0
- **Sell**: RSI > 70 or sentiment < -0.1

Inverse Sentiment

- **Buy**: sentiment < -0.7
- **Sell**: sentiment > 0.7

Sentiment Trading Strategies



Sentiment-only

- **P/L**: -0.28%
- **Sharpe**: -14.83
- **Max Drawdown**: 0.84%

Sentiment + TA

- **P/L**: -0.23%
- **Sharpe**: -22.71
- **Max Drawdown**: 0.71%

Inverse r/wsb

- **P/L**: -0.15%
- **Sharpe**: -6.96
- **Max Drawdown**: 2.35%

Model Comparison – Kaggle Datasets (V-shaped $SPY)

1

HODL (Hold On for Dear Life)

- **P/L**: 0.41%
- **Sharpe**: -3.34
- **Max Drawdown**: 4.2%

2

DCA (Dollar Cost Averaging)

- **P/L**: 0.40%
- **Sharpe**: -20.28
- **Max Drawdown**: 0.58%

3

TA (Technical Analysis)

- **P/L**: 0.23%
- **Sharpe**: -9.25
- **Max Drawdown**: 1.29%

6

Sentiment-only

- **P/L**: -0.28%
- **Sharpe**: -14.83
- **Max Drawdown**: 0.84%

5

Sentiment + TA

- **P/L**: -0.23%
- **Sharpe**: -22.71
- **Max Drawdown**: 0.71%

4

Inverse r/wsb

- **P/L**: -0.15%
- **Sharpe**: -6.96
- **Max Drawdown**: 2.35%

Datasets can be manually created using PRAW and yfinance

This dataset spans 2025 Nov 24-26 with **29186 comments** (vs 6033) and **1170 time intervals** (vs 144)

```python
import pandas as pd

submission_id = "1p5f9mw"
submission_url = "https://www.reddit.com/r/wallstreetbets/.../"

def extract_all_comments(submission_id):

    submission = reddit.submission(id=submission_id)
    submission.comments.replace_more(limit=None)

    all_comments = submission.comments.list()

    for comment in all_comments:

        comment_dict = {
            "id": comment.id,
            "author": author_name,
            "score": comment.score,
            "created_utc": comment.created_utc,
            "created_datetime": timestamp_str,
            "body": comment.body.replace('\n', ' ')
        }
        comments_data.append(comment_dict)

    df = pd.DataFrame(comments_data)
```

Code to get r/wsb data

```python
def fetch_intraday_data(ticker, start, end, interval):
      data = yf.download(tickers=ticker, start=start, end=end)

      is_after_start = (market_hour > 9) | ((market_hour == 9) & (market_minute >= 30))
      is_before_end = (market_hour < 16) | ((market_hour == 16) & (market_minute == 0))

      data_filtered = data[is_after_start & is_before_end].copy()

      return data_filtered



spy_data = fetch_intraday_data(TICKER, START_DATE, END_DATE)
```

Code to get intraday $SPY data

Model Comparison – Curated Dataset (Bull trend)

1

HODL (Hold On for Dear Life)

- **P/L**: 2.49%
- **Sharpe**: -5.69
- **Max Drawdown**: 0.72%

6

DCA (Dollar Cost Averaging)

- **P/L**: 0.09%
- **Sharpe**: -96.52
- **Max Drawdown**: 0.03%

5

TA (Technical Analysis)

- **P/L**: 0.97%
- **Sharpe**: -12.36
- **Max Drawdown**: 0.41%

2

Sentiment-only

- **P/L**: 1.37%
- **Sharpe**: -8.23
- **Max Drawdown**: 0.23%

6

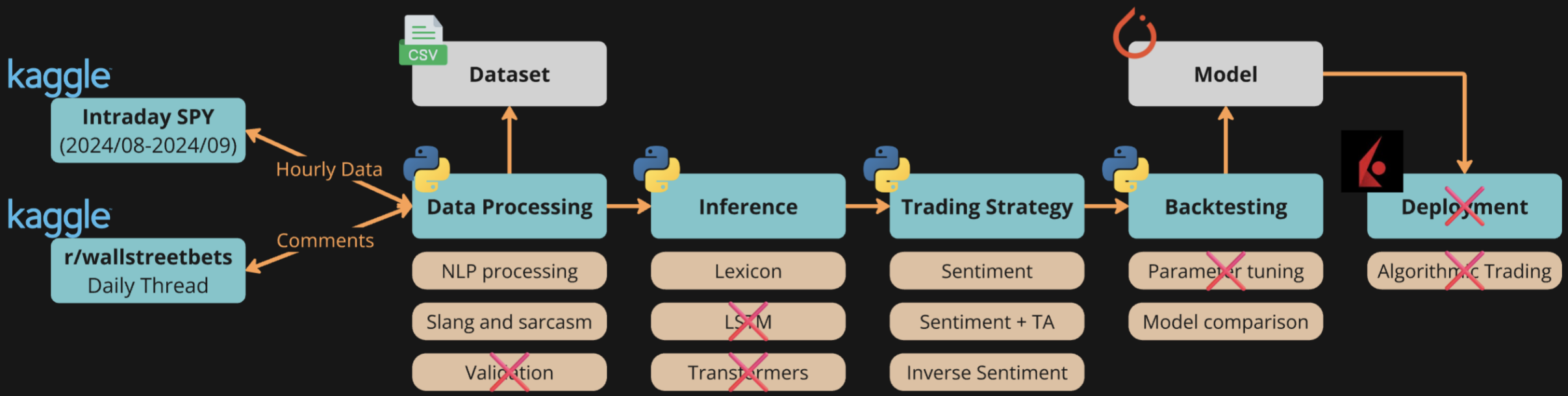
Sentiment + TA

- **P/L**: 0.88%
- **Sharpe**: -8.96
- **Max Drawdown**: 0.24%

4

Inverse r/wsb

- **P/L**: 1.09%
- **Sharpe**: -9.91
- **Max Drawdown**: 0.48%



Limitations

- **Garbage in = garbage out**: The r/wsb dataset was too sparse to generate diverse signals. The data itself is very noisy
- **Unlabeled data**: Supervised sentiment analysis methods were unavailable

Conclusions

- **r/wsb is not reliable** for sentiment-based trading due to noise, misinformation, and its responsive nature

Section 1 | Introduction to Natural Language Processing and Sentiment Analysis

Section 2 | State of the Field Sentiment Analysis in Finance

Section 3 | Blueprint for Building a Trading Strategy Based on Sentiment Analysis

Section 4 | Trading via r/wallstreetbets Sentiment

Section 5 | **Concluding Remarks**

Key Takeaways

Section 1 | **Introduction to Natural Language Processing and Sentiment Analysis**

- Sentiment analysis: Split into lexicon and ML-based approaches
- Preprocessing: Cleaning → Tokenization → Stop word removal → Stemming or lemmatization
- ML-based sentiment analysis: Includes feature representation (words to vectors)

Section 2 | **State of the Field Sentiment Analysis in Finance**

- News and social media can move markets but be skeptical of what you see online
- Sentiment analysis can be multi-modal (video, audio, text)

Section 3 | **Blueprint for Building a Trading Strategy Based on Sentiment Analysis**

- Developed ticker-specific strategies based on news articles from RSS feeds and Yahoo Finance
- LSTM models outperform classical ML methods for sentiment signal prediction

Section 4 | **Trading via r/wallstreetbets Sentiment**

- Social media-based sentiment strategies are susceptible to noise, misinformation, and lag