FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Emil Jacques Nysschens	South Africa	enysschens@gmail.com	
Stephen Osemudiame John-Ebowe	United Kingdom	stephenjohnebowe@gmail.com	
Gift Vavi	South Africa	vavigift@gmail.com	

Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).	
Team member 1	Emil Jacques Nysschens
Team member 2	Stephen Osemudiame John-Ebowe
Team member 3	Gift Vavi

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed. **Note:** You may be required to provide proof of your outreach to non-contributing members upon request.
N/A

Part 1. Assessing Models with Alternative Data

Q1 Data Understanding

i. The paper mainly uses historical stock market data, specifically daily prices including opening price, closing price, high price, low price and trading volume.

From this historical data, various technical indicators are computed to capture trends and patterns. Examples of technical indicators used include moving averages, exponential moving average, relative strength index, moving average convergence divergence, stochastic oscillator and on-balance volume (Mejia et al. 8)

ii. Importance of using such indicators in forecasting stock price trends:

The transformation of raw market data through mathematical methods in technical indicators enables analysts to extract meaningful patterns which reveal market momentum and volatility and trend strength. Technical indicators transform historical price and volume data through mathematical operations to reveal patterns which are concealed within noisy financial time series.

These indicators provide several advantages:

- Moving averages and similar indicators smooth market fluctuations to help investors identify bullish or bearish trends in the market.
- The RSI and Stochastic indicators along with other oscillators enable users to measure market momentum while detecting potential price reversals.
- Bollinger Bands serve as a tool which allows models to determine market risk levels and price variability according to (Murphy, 45).

The application of technical indicators enables machine learning models to extract vital patterns from noisy inputs which leads to better forecasting accuracy. The use of processed features obtained from technical indicators leads to better generalisation and financial interpretation of model predictions compared to using raw prices.

Q2. Security Understanding: IVV (iShares Core S&P 500 ETF)

Fund Overview: The iShares Core S&P 500 ETF (ticker: IVV) is a passively managed exchange-traded fund (ETF) designed to track the S&P 500 Index performance, representing 500 of the largest publicly traded U.S. companies. Managed by BlackRock, IVV provides investors with broad exposure to the U.S. large-cap equity market, covering sectors like technology, healthcare, financials, and consumer discretionary.

Asset Type:

- Type: Equity ETF
- Underlying Index: S&P 500 Index
- Holdings: Approximately 500 large-cap U.S. stocks
- Top Holdings: Apple, Microsoft, Amason, NVIDIA, Alphabet

Historical Price Performance: Since its inception in **May 2000**, IVV has demonstrated strong long-term performance, mirroring the historical growth trajectory of the S&P 500.

- 2000 – 2010: Modest growth with significant volatility due to the Dot-com crash and 2008 Financial Crisis.
- 2010 – 2020: Strong bull market performance, fuelled by technology sector growth.
- 2020 – Present: Recovery from the COVID-19 market crash, followed by volatility from interest rate hikes in 2022-2023. Table 1.0 shows other stats about its history.

Table 1.0: Other Stats About IVV History

Key Historical Statistics	**Price Snapshot**	**Other Notable Features**
- Annualised 10-Year Return (as of 2024): ~11% - Expense Ratio: 0.03% (one of the lowest among ETFs) - Dividend Yield: Approximately 1.4% - Assets Under Management (AUM): Over $300 billion	- All-time High: Around $480 (early 2022) - Recent Price (April 2025): Approximately $450 - 52-week Range: $390–$460	- Liquidity: Highly liquid with narrow bid-ask spreads. - Suitability: Often used by long-term investors seeking exposure to the overall U.S. economy with a low-cost, diversified vehicle. - Risk: Market risk tied to the broader performance of the U.S. economy and large-cap stocks.

Classification versus Regression

The authors selected classification instead of regression because they wanted to predict stock market direction between rise and fall rather than future price values (Mejia et al., 9). The modeling task becomes easier through classification because it transforms into a simple binary decision between rise and fall which traders find more useful. The evaluation becomes simpler with accuracy and F1-score instead of RMSE and the model complexity decreases while preventing overfitting to price fluctuations' noise.

The authors could have defined their classification variable through two different approaches:

Threshold-based Movement: The classification should focus on the magnitude of changes instead of any movement. For example:

- The model should classify the day as up when daily return exceeds +0.5%.
- The model should classify days with returns below -0.5% as down.
- The model should classify days with returns between -0.5% and +0.5% as neutral.

Volatility-adjusted Classes: The classification system should use volatility bands to determine its categories.

- Standard deviation of past returns should determine what constitutes significant market movement.
- The model should classify returns as significant up when they exceed one standard deviation above the mean.
- The model should classify returns as significant down when they fall below one standard deviation below the mean.
- The remaining cases should be classified as no significant change.

Q3 Section 2: Data

2.1 Data collection

2.2 Technical indicator calculation

2.3 Data processing

2.4 Feature selection

2.5 Classification variable definition

Section 3: Methodology

3.1 Neural network model

3.2 Hyperparameter optimisation with Genetic Algorithm

3.3 Feature selection using LASSO

3.4 Training and testing procedure

Table 1.1: Dividing Descriptive statistics from Models

Aspect	**Descriptive Statistics**	**Predictive Models**
Purpose	Understand the structure, distribution, and relationships in data	Predict future outcomes or classify new data
Example	Pearson Correlation measures the linear relationship between two variables	LASSO Regression selects important predictors and builds a model by shrinking coefficients toward zero for irrelevant features
Process	Summarises data	Learns from data
Outputs	Correlations, means, variances, graphs	Predictions, selected features, coefficients
Dependency on Labels	No	Yes

Optimisation Process of Technical Indicators

The authors applied a two-step optimisation strategy to maximise predictive ability of technical indicators:

Genetic algorithm optimisation

- A Genetic Algorithm is first used to optimise the parameter settings of each technical indicator, such as choosing the most effective time windows for moving averages or the RSI (Mejia et al., 6).
- The GA evolves populations over generations, selecting parameter combinations that yield the highest predictive performance (Mejia et al., 6).

LASSO feature selection

- After parameter optimisation, the authors use LASSO regression to automatically select the best subset of indicators by penalising less useful ones, thus reducing dimensionality (Mejia et al., 7).

How the Authors Improve Predictive Power

The combination of optimised parameters and feature selection enhances predictive power significantly:

- Customisation of Indicators: Fine-tuning standard technical indicators to fit the specific dynamics of the emerging markets improves signal quality.
- Noise Reduction: LASSO shrinks noisy or redundant inputs, reducing overfitting and focusing the neural network on truly informative features.
- Model Generalisation: By reducing the number of irrelevant features, the model becomes more robust and performs better on unseen data.

Why It Is Important to Optimise Indicators for Neural Networks

Optimising technical indicators is crucial because:

- Neural networks are sensitive to irrelevant features, and too many noisy inputs can lead to overfitting (Goodfellow et al., 97).
- Better feature quality means faster training and higher prediction accuracy, improving both efficiency and effectiveness (Hastie et al., 217).
- Dimensionality reduction lowers computational costs, enabling faster and more stable learning even on large financial datasets (Mejia et al., 7).

Q4. Investigating the Features

a. In the paper by Sagaceta Mejia et al., the researchers constructed several technical analysis features that were calculated using the Python Pandas TA (Technical Analysis) library.

b. When deploying a neural network it is necessary to configure a layer of inputs. These inputs form an array and, in the context of artificial neural networks are referred to as 'features'. (Hardesty 2017) These features can be values of a sample of external data or the inputs of other neurons. The difference between features and methods are that features are 'raw data', by virtue of the fact that it forms the input layer in MLP (Multi-layer Perceptron).

   A method in programming on the other hand is a function that transforms inputs. A method typically takes a number of arguments (inputs) and the method or function will produce a particular output through the transformation of the input (argument) (Telles 2002).

   Lastly, a model is a mathematical or programmatic representation of a real-world situation in which it supports decision making. (Winston 2004)

c. The features drawn from the Pandas-TA (Technical Analysis) library are divided into the following indicator categories (Johnson 2021):

1.	Candles	6.	Statistics
2.	Cycles	7.	Trend
3.	Momentum	8.	Utility
4.	Overlap	9.	Volatility
5.	Performance	10.	Volume

d. The researchers employed two techniques in the optimization of their MLP model, in line with their stated objective of formulating an emerging market ETF trend predictor:

   i. Early stopping to prevent overfitting

   ii. Using a subset of 5 features to train the model. Deep learning neural networks are notoriously resource hungry and this step was introduced to ensure the efficient use of computer processing resources.

   The researchers summarise the results of these two optimization techniques in Table 6 on page 9. From the table, it is clear that the training time is significantly reduced when compared to training the dataset on all the identified features and, it is demonstrated, that the predictive accuracy of the limited feature set is improved, most likely due to the increase in epochs (training iterations).

Q5. Optimization

a) Cross-validation is a method used in machine learning whereby a dataset is divided into a pre-specified number of blocks which are then used to train and test the outcome of several predictive algorithms. (StatQuest 2018)

b) The number of blocks used in cross-validation, is referred to as the number of folds. For instance, a common number of folds is 10 and this is what the researchers used as well. This means, the data was divided into 10 equally sized blocks and the number of k is 10, meaning 10-fold cross validation was used to test and train the model.

c) Distance measures are methods to quantify the distance between data points in one (or more dimensions). The Jaccard distance is calculated as the ratio of the intersection and union of a combination set. ($\frac{A \cap B}{A \cup B}$)

d) The Jaccard distance is equal to 1 minus the Jaccard similarity (Rajaraman 2014). As such, it is a measure of *dissimilarity*. Other measures encountered in lessons previously are the Euclidean and Manhattan distance measures. In Euclidean distance, an application of Pythagorean geometry is encountered whereby two points in the same space or dimension are set equal to the hypotenuse of a right-angled triangle. For the distance between two points this will be:

$$\text{Euclidean distance} = \sqrt{(\mathrm{x}_2 - x_1)^2 + (y_2 - \mathrm{y}_1)^2}$$

Closely related to this is the 'Manhattan' distance measure which relies on absolute values (and not squaring) to calculate distance. For comparison, this is how one would calculate the Manhattan distance between two points:

$$\text{Manhattan distance} = |\mathrm{x}_2 - \mathrm{x}_1| + |\mathrm{y}_2 - \mathrm{y}_1|$$

e. The researchers indicate that an optimal solution would be one where the MLP model assists investors in timing their entry and exit points into emerging market ETFs as well as serving as an indicator of risk to ensure that investors attain maximum profit for minimum loss.

Due to the limitations of the model they do warn however that investors would be best advised to also implement the model into a broader investment framework but provide no details on how such a framework should be constructed.

Step 1

The researchers set out to find an optimal solution (processor efficient using a subset of total features) to the problem of predicting an emerging market ETF's performance.

From the results published, the researchers conclude that the performance of an emerging market ETF is driven by quantitative factors whereas that of a developed market ETF is driven by more qualitative factors. The MLP method utilised in the study is a highly quantitative method and therefore more suited to the analysis of an emerging market ETF.

Step 2

As illustrated in Table 1.2, the accuracy of the model increases with the number of features up to 5 before steeply decreasing with 6 features. The results of the study therefore indicate that a subset of 5 features is the optimal number of features to employ in the MLP model described.

The most useful features in predicting the performance of each ETF fund differs by fund (presumably due to sectoral focus of each fund) are as follows:

Table 1.2: Specific Features Useful from the Study

Archer's on balance volume	AOBV_LR_2	
Bollinger band percent	BBP_5_2.0	
Balance of power	BOP	Momentum
Correlation trend indicator	CTI_12	
Decreasing	DEC_1	Cycle
Even better sinewave	EBSW_40_10	
Increasing	INC_1	
	J_9_3	
	K_9_3	
	ZS_30	
	STOCHk_14_3_3	
Williams % R	WILLR_14	
	TTM_TRND_6	
Stochastic relative strength index	STOCHRSIk_14_14_3_3	

Part 2. Evaluating One Particular Type of Alternative Data (User Guide: Social Media)

1. Sources of Data

One major source of social media data is public social media platforms such as Facebook, X (formerly Twitter), Reddit, Instagram, LinkedIn, and TikTok. These platforms provide a rich stream of user-generated content, including posts, comments, shares, and various engagement metrics. They serve as direct channels for capturing users' opinions, interactions, and the relationships between individuals and organizations. Another valuable source of social media comes from blogs and forums, including websites like Seeking Alpha, Reddit (in specialized threads), and niche industry-specific forums. These platforms are useful for gathering unfiltered insights, market sentiment, and in-depth discussions about companies, financial markets, and investment products.

Financial news aggregators and data providers also contribute significantly to social media data collection. Services like Bloomberg, Yahoo Finance, and Refinitiv often integrate social media sentiment analysis into their platforms. Additionally, specialized providers such as Social Listening, Brandwatch, and Meltwater offer dedicated tools and APIs for collecting, filtering, and analyzing social media data in real time.

2. Types of Data

a. **Text Data:** Text data is any information represented in written or digital textual form, consisting of characters, words, sentences, or symbols (Zhai and Massung). These include posts, comments, messages, and articles contain valuable information about opinions, sentiment, and discussions. Natural language processing (NLP) techniques are crucial for extracting meaning from text data.

b. **Sentiment Data:** Sentiment data refers to textual or digital data that captures subjective opinions, emotions, or attitudes (e.g., positive, negative, neutral) expressed by individuals, often analyzed through techniques like sentiment analysis to derive actionable insights (Xu et al. 1-16; Yue et al. 617-663). Sentiment analysis is widely used to gauge market sentiment and predict stock price movements.

c. **Social Networking Data:** This include information about user connections, followers, and interactions. This reveals influential users, communities, and information diffusion patterns.

d. **Engagement Metrics Data:** Metrics such as likes, shares, retweets, and comments indicate user engagement and popularity. These metrics can be used to measure the reach and impact of information (Perreault and Mosconi 3568-3577).

e. **Image and Video Data:** Image and video data refer to digital representations of visual information involving images composed of pixel grids capturing static visuals and videos which are sequences of such frames depicting motion (Guan 609-631). Platforms like Instagram and TikTok contain visual data that can be analyzed to understand brand perception, consumer behavior, and trends.

3. Quality of Data

One major factor on the quality of social media data is authenticity and reliability. It can be difficult to verify whether user accounts are genuine and whether the content they post is accurate (Ismail and Latif 254-261). The presence of fake accounts and the spread of misinformation can significantly distort any analysis conducted using such data. Another important aspect of data quality is representativeness, social media users do not necessarily reflect the broader population, which means that insights drawn from such data might be biased or limited in scope (Kaschesky et al. 2003-2012). This can affect the validity of conclusions made in studies or business decisions.

The volume and velocity of social media data also presents additional challenges to the quality of data collected from social media. Posts are generated in massive quantities and at high speed, which can quickly become overwhelming (Manovich 1-17). This requires efficient data management and processing techniques to ensure meaningful insights can be extracted in a timely manner. Furthermore, social media data is often filled with noise and bias (Morstatter and Liu 1-13). Irrelevant information, spam, and strongly opinionated or polarizing content are common. As such, filtering and cleaning are critical preprocessing steps to isolate valuable, objective information from the clutter. Lastly, API limitations and changes imposed by social media platforms can complicate data collection thus affecting quality (Lomborg and Bechmann 256-265). Rate limits, evolving terms of service, and frequent updates to API structures often disrupt access and require continual adjustments to data-gathering tools and workflows.

4. Ethical Issues

One major ethical concern in using social media data is privacy. Collecting and analyzing user-generated content without explicit consent can violate individuals' privacy rights (Custers et al. 268-295). To address this, data should be anonymized and aggregated to reduce the risk of exposing personal information. Another important ethical issue is bias and fairness. Social media platforms often reflect societal biases, and algorithms trained on such data can inadvertently perpetuate or even amplify these inequalities (Morstatter and Liu 1-13; Saxena et al. 1-45). Developers must therefore prioritize fairness and accountability throughout the model development process.

Data security is also an ethical issue, protecting sensitive user data from breaches and unauthorized access is not optional but a necessity. Implementing strong security protocols and complying with relevant data protection laws helps safeguard this information (Herath et al. 155-179; Shukla et al. 41-59). Additionally, ethical issues arise from the potential for market manipulation. Leveraging social media data to spread false information or influence investor sentiment is not only unethical but illegal (Selvakumar et al. 225-250). This underscores the need for clear regulatory oversight and strict adherence to ethical standards. Finally, informed consent remains a foundational principle for ethics. Researchers and practitioners should be transparent about their data collection and analysis practices, and, wherever possible, seek informed consent from users whose data is being utilized (Williams et al. 27-52).

5. Python Code to Import and Structure Data

The Python implementation in Figure 1.0 offers a foundation for collecting, processing, and analyzing social media data. By applying these techniques and building upon the existing research literature, analysts can derive meaningful insights from the vast amount of social media content generated daily.

```
!pip install praw
import praw
import pandas as pd
from datetime import datetime

# Reddit API credentials
reddit = praw.Reddit(client_id='p7UMF3D4H3VFZ-Xe2E3FrA',
                     client_secret='jQQR4I7IDTOo3hg7CA5ty-oGVvmFIw',
                     user_agent='DataBot by u/Deep_Engineering_773')

# Define subreddits and limits
subreddits = ['investing', 'stocks', 'wallstreetbets']
post_limit = 500
comment_limit = 50

# Collect posts and comments
all_posts = []

for subreddit_name in subreddits:
    subreddit = reddit.subreddit(subreddit_name)

    # Get top posts
    for post in subreddit.top(time_filter='month', limit=post_limit):
        post_data = {
            'id': post.id,
            'title': post.title,
            'text': post.selftext,
            'created_utc': datetime.fromtimestamp(post.created_utc),
            'score': post.score,
            'upvote_ratio': post.upvote_ratio,
            'num_comments': post.num_comments,
            'author': str(post.author),
            'subreddit': subreddit_name,
```

Fig. 1.0: Importing and Structuring the Reddit Data into Useful Data Structures Using Python

6. Exploratory Data Analysis (EDA) of Sample Data

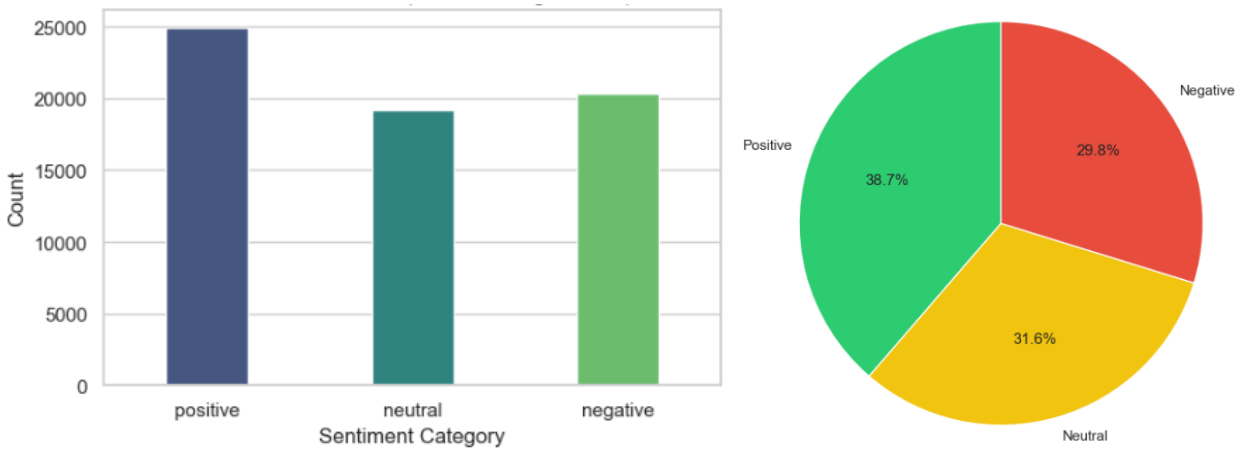

Fig. 1.1: Sentiment Distribution Analysis of the Reddit Data

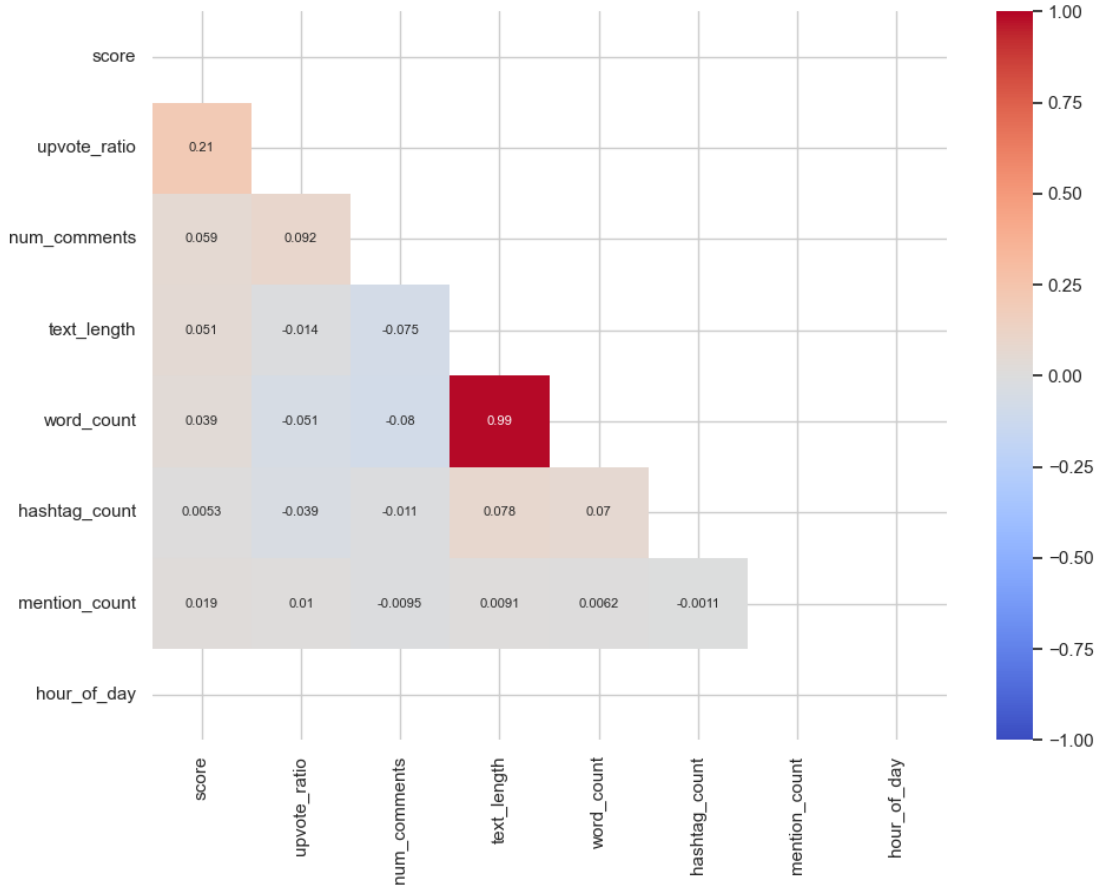

Fig. 1.2: Correlation Analysis of the Reddit Data

Figure 1.1 illustrates the sentiment analysis of the Reddit data having 38.7% positive, 29.8% negative, and 31.6% neutral sentiments while Figure 1.2 shows the correlation matrix heatmap displaying relationships

between various features of posts on Reddit. The strongest correlation (i.e. 0.99) exists between text length and word count, which is expected as they measure similar attributes. Other notable correlations include a weak positive relationship (i.e. 0.21) between score and upvote ratio, suggesting higher-scored posts tend to have better upvote ratios. Most other correlations are quite weak indicating minimal relationships between features like mention count, hashtag count, and number of comments.

7. Short Literature Search

Research (Sun et al. 1-32) has examined the impact of social media on various aspects of corporate finance, such as IPO performance, mergers and acquisitions, and investor relations. In general, social media as alternative data has gained significant attention in academic research. The following studies highlight key developments in this area:

Sentiment Analysis and Market Prediction: A pioneering study (Bollen et al. 1-8) found that Twitter mood states could predict changes in the Dow Jones Industrial Average with 86.7% accuracy. Further study (Bartov et al. 25-57) demonstrated that aggregate opinion from Twitter can help predict earnings surprises and abnormal stock returns. A different study (Chen et al. 1367-1403) examined user-generated content on Seeking Alpha and found that the sentiment of articles and comments predicted stock returns and earnings surprises.

Information Dissemination and Market Efficiency: A study (Sprenger et al. 926-957) found that tweet sentiment relates to abnormal returns and message volume correlates with trading volume. Subsequently, another study (Cookson & Niessner 173-228) used StockTwits data to demonstrate how investor disagreement stems from different interpretations of information rather than information asymmetry.

Topic Modeling and Event Detection: A study (Dimpfl & Jank 172-192) analyzed how internet search queries can serve as proxies for investor attention and help predict market volatility. Another research (Mazboudi & Khalil 115-124) found that social media activity reduces information asymmetry around Merger & Acquisition announcements.

Methodological Advances: A study (Azar and Lo 1-22) developed novel approaches to extract actionable signals from Twitter data for trading strategies while another study (Renault 25-40) Used StockTwits messages to create intraday sentiment indicators that predict intraday stock returns. Further to this, a different study (Li and van Rees 50-69) investigated the incremental information value of social media over traditional news sources.

References

Azar, Pablo, and AW Lo. "The Wisdom of Twitter Crowds: Predicting Stock Market Reactions to FOMC Meetings via Twitter Feeds." *MIT Open Access Articles*, 2016, pp. 1–22, https://dspace.mit.edu/handle/1721.1/109079

Bartov, Eli, et al. "Can Twitter Help Predict Firm-Level Earnings and Stock Returns?" *The Accounting Review*, vol. 93, no. 3, May 2018, pp. 25–57, https://publications.aaahq.org/accounting-review/article-abstract/93/3/25/4062

Bollen, Johan, et al. "Twitter Mood Predicts the Stock Market." *Journal of Computational Science*, vol. 2, no. 1, Mar. 2011, pp. 1–8, https://www.sciencedirect.com/science/article/pii/S187775031100007X?casa_token=2IP62wEMhbQAAAAA:1d7yx3wDRQ_JeeVBQRuqGu-JLss5X1oq1WsWx3AfSm3etIC8NuAqmuCURkqcE_KPVjWiiZJA_Mo

Chen, Hailiang, et al. "Wisdom of Crowds: The Value of Stock Opinions Transmitted through Social Media." *Review of Financial Studies*, vol. 27, no. 5, Feb. 2014, pp. 1367–1403, https://academic.oup.com/rfs/article/27/5/1367/1581938

Cookson, J. Anthony, and Marina Niessner. "Why Don't We Agree? Evidence from a Social Network of Investors." *The Journal of Finance*, 2020, pp. 173–228, https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12852?casa_token=zwy8GDyyVTQAAAAA:zqrVkGnYU34p5DoBAVhdVJpdNWZpylvXe2fY9vFoBvlQEvR0JnyajiVIu2gOWilXpE-p362ZXs9pFrNK

Custers, Bart, et al. "Privacy Expectations of Social Media Users: The Role of Informed Consent in Privacy Policies." *Policy & Internet*, vol. 6, no. 3, Sept. 2014, pp. 268–295, https://onlinelibrary.wiley.com/doi/abs/10.1002/1944-2866.POI366?casa_token=sTBDvbFb3LUAAAAA:wCWdm3sJv5k2UovpegmkTy4sl0U_NKjSuZRs049IjtuF7XHZX7vanByFxvSmCEjT830C3XB1fwAQkGfH

Dimpfl, Thomas, and Stephan Jank. "Can Internet Search Queries Help to Predict Stock Market Volatility?" *European Financial Management*, vol. 22, no. 2, 2016, pp. 171–192, https://onlinelibrary.wiley.com/doi/abs/10.1111/eufm.12058?casa_token=BTxcAwQl6EcAAAAA:Nu01-6yC2XYt7S4mxLQbw-MkRhceiQ-gF--lRIfXfGQmg1q0JRcv_IdS6JKxYaOxrVbn_OuCVnXwtRjI

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT Press, 2016.

Guan, Ling. *Multimedia Image and Video Processing*. CRC Press, 2017, https://books.google.com/books?hl=en&lr=&id=eTfNBQAAQBAJ&oi=fnd&pg=PP1&dq

Hardesty, Larry. Explained: Neural Networks. 14 April 2017. https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer, 2009.

Herath, H. M. S. S., et al. "Data Protection Challenges in the Processing of Sensitive Data." *Data Protection*, Springer Nature Switzerland, 2024, pp. 155–179, https://link.springer.com/chapter/10.1007/978-3-031-76473-8_8

Ismail, Shahrinaz, and Roslina Abdul Latif. "Authenticity Issues of Social Media: Credibility, Quality and Reality." *Ir.unikl.edu.my*, 2013, pp. 254–261, http://ir.unikl.edu.my/jspui/handle/123456789/1504

Johnson, Kevin. *Pandas – Technical Analysis*, 2021. https://github.com/twopirllc/pandas-ta

Kaschesky, Michael, et al. "Bringing Representativeness into Social Media Monitoring and Analysis." *46th Hawaii International Conference on System Sciences*, Jan. 2013, pp. 2003–2012, https://ieeexplore.ieee.org/abstract/document/6480083/?casa_token=BAf5BIDJHa4AAAAA:YkYgXWTWz7Z8BzHG8ztdgc0Q8LxkuMtyZxDvbxxTSVT7HXoxBhcR9MMWMVO_ZaFDROlBFgJYFnE

Li, Ting, et al. "More than Just Noise? Examining the Information Content of Stock Microblogs on Financial Markets." *Journal of Information Technology*, vol. 33, no. 1, Mar. 2018, pp. 50–69, https://journals.sagepub.com/doi/abs/10.1057/s41265-016-0034-2?casa_token=5UFCpWQNYxkAAAAA:87yy3nyH-Xkn8sBfa6f92kde7YGtB0KE-DVW3nIjflCZbj7XZ9KicKM6pOVFc845T6S_ZWMb-S2RTw

Lomborg, Stine, and Anja Bechmann. "Using APIs for Data Collection on Social Media." *The Information Society*, vol. 30, no. 4, July 2014, pp. 256–265, https://www.tandfonline.com/doi/abs/10.1080/01972243.2014.915276?casa_token=b8CKWEusdAgAAAAA:XdLE9wtlDlsb8erzMw6mli2kWYa0s_P3XXuYMYnTEntHSfR653dtqirt_686fKYNt_SNuoDC4_t-MA

Manovich, L. *Trending: The Promises and the Challenges of Big Social Data*. 2011, pp. 1–17, https://www.academia.edu/download/35983875/64-article-2011_trending.pdf

Mazboudi, Mohamad, and Samer Khalil. "The Attenuation Effect of Social Media: Evidence from Acquisitions by Large Firms." *Journal of Financial Stability*, vol. 28, 2017, pp. 115–124, https://www.sciencedirect.com/science/article/pii/S1572308916301991?casa_token=jqnA2rg1eBIAAAAA:_f-yosxXvxDTr_WQSQM7xHA9PhSRp-m1-1IvX6cnq_zMfLdumdya8iMLEjlPQobQeXO3x9mJrGQ

Morstatter, Fred, and Huan Liu. "Discovering, Assessing, and Mitigating Data Bias in Social Media." *Online Social Networks and Media*, vol. 1, June 2017, pp. 1–13, https://www.sciencedirect.com/science/article/pii/S2468696416300040?casa_token=QeCzjoIkNJsAAAAA:PwKkood-LSh6QGhdnqS4G-mUlzEnN2Cb1S6X_SZTLjjgLXB3ThTKwoRwm_U6rMNPpqbPRXHpvsw

Murphy, John J. Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications. New York Institute of Finance, 1999.

Perreault, Marie-Catherine, and Elaine Mosconi. "Social Media Engagement: Content Strategy and Metrics Research Opportunities." *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018, pp. 3568–3577, https://scholarspace.manoa.hawaii.edu/handle/10125/50339

Rajaraman and Ullman. *Mining of Massive Datasets*. Cambridge University Press, 2014. http://infolab.stanford.edu/~ullman/mmds/book.pdf

Renault, Thomas. "Intraday Online Investor Sentiment and Return Patterns in the U.S. Stock Market." *Journal of Banking & Finance*, vol. 84, Nov. 2017, pp. 25–40, https://www.sciencedirect.com/science/article/pii/S0378426617301589?casa_token=kgeb3FMzpokAAAAA:WcEm2_6WAp-ikphMWezMpnmpYTMMQveIaZFxiBpLsncj0JpEgllNaCCM4p-BwY5OeC67J3imIKM

Sagaceta Mejia, Gilberto, Ivan Marsura, Sergio Romero-Torres, and Edgar Sucar. "An Intelligent Approach for Predicting Stock Market Movements in Emerging Markets Using Optimised Technical Indicators and Neural Networks." The Open Economics Journal, vol. 15, no. 1, 2022, https://www.degruyter.com/document/doi/10.1515/econ-2022-0073/html.

Saxena, Akrati, et al. "FairSNA: Algorithmic Fairness in Social Network Analysis." *ACM Computing Surveys*, Association for Computing Machinery, Mar. 2024, pp. 1–45, https://dl.acm.org/doi/abs/10.1145/3653711?casa_token=htF60EQVuwAAAAAA:AuvIenYOl1-X2YpyfGFBk6adzEIwUwAxPcmLAFJI_6qGAhumd8lHzGNmU7c9vS8jY0ImtiYp8xit2A&casa_token=ZtKA1UkZgm0AAAAA:61sH3zHbgnM3gSkjGfE_S9ICSwGW_QnuUuyr6v7WA2jEW29exQgND-YTg8M-PAbb32Uw4TeoySERug

Selvakumar, P., et al. "Social Media Influence on Market Sentiment." *Advances in Business Strategy and Competitive Advantage*, IGI Global, Dec. 2024, pp. 225–250, https://www.igi-global.com/chapter/social-media-influence-on-market-sentiment/365116

Shukla, Samiksha, et al. "Data Security." *Data Ethics and Challenges*, 2022, pp. 41–59, https://link.springer.com/chapter/10.1007/978-981-19-0752-4_3

Sprenger, Timm O., et al. "Tweets and Trades: The Information Content of Stock Microblogs." *European Financial Management*, vol. 20, no. 5, May 2014, pp. 926–957, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-036X.2013.12007.x?casa_token=6o5S6ca_EvUAAAAA:tUjw6FHh-G4ars1fgtRXZo7SK6j9IwAs7LEI3TDnRt4PZWEvHfe4AL3H9AFjA21rE_P0zgOaHOvqKEu2

StatQuest with Josh Starmer. "Machine Learning Fundamentals: Cross Validation" Youtube, 24 April 2018. https://www.youtube.com/watch?v=fSytzGwwBVw

Sun, Yunchuan, et al. "A Survey on Alternative Data in Finance and Business: Emerging Applications and Theory Analysis." *Financial Innovation*, 2024, pp. 1–32, https://link.springer.com/article/10.1186/s40854-024-00652-0

Telles, Matt. *C# Black Book*. Coriolis, 2002.

Williams, Matthew L., et al. "Users' Views of Ethics in Social Media Research: Informed Consent, Anonymity, and Harm." *The Ethics of Online Research*, vol. 2, Dec. 2017, pp. 27–52, https://www.emerald.com/insight/content/doi/10.1108/S2398-601820180000002002/full/html

Winston, Wayne. *Operations Research Applications and Algorithms Fourth Edition*. Brooks/Cole Cengage Learning. 2004.

Xu, Qianwen Ariel, et al. "A Systematic Review of Social Media-Based Sentiment Analysis: Emerging Trends and Challenges." *Decision Analytics Journal*, vol. 3, June 2022, pp. 1–16, https://www.sciencedirect.com/science/article/pii/S2772662222000273

Yue, Lin, et al. "A Survey of Sentiment Analysis in Social Media." *Knowledge and Information Systems*, vol. 60, no. 2, July 2019, pp. 617–663, https://link.springer.com/article/10.1007/s10115-018-1236-4

Zhai, Chengxiang, and Sean Massung. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. New York, Ny, Usa Association for Computing Machinery and

Morgan &↣ Claypool, 2016,

https://books.google.com/books?hl=en&lr=&id=WoKkDAAAQBAJ&oi=fnd&pg=PR15&dq