Educational Administration: Theory and Practice
2024,30(4), 1102-1109
ISSN:2148-2403
https://kuey.net/

Research Article

A Critical Review Of Machine Learning Approaches To Sentiment Analysis For Stock Market Prediction

Jyotirmoy Roy[1*], Dr Bimal Debnath[2]

[1*]Research Scholar, Dept. of Management, NEHU, Meghalaya, Email: jyotirmoyroy88@gmail.com Ph: 9863082886
[2]Asst. Professor, Dept. of Management, NEHU, Meghalaya, Email: bimal.dn@gmail.com Ph: 9485127950

Citation: Jyotirmoy Roy et al. (2024), A Critical Review Of Machine Learning Approaches To Sentiment Analysis For Stock Market Prediction, Educational Administration: Theory and Practice, 30(4), 1102-1109, Doi: 10.53555/kuey.v30i4.1613

ARTICLE INFO

ABSTRACT

The rapid evolution of machine learning (ML) technologies and their transformative impact on numerous industries has garnered significant interest in their potential for financial market analysis. Given stock markets' volatility and economic significance, understanding and predicting their behaviour is a crucial yet challenging task. This study critically examines the various ML approaches to sentiment analysis for stock market prediction. The primary objective is to synthesise research findings to assess the efficacy of ML models in this domain. While ML models show promise, their accuracy in predicting market movements varies significantly depending on data quality, model complexity, and contextual factors. It also discusses the limitations of current approaches and the need for more robust and adaptable models. The findings suggest that advancements in ML algorithms and data preprocessing techniques could significantly enhance predictive accuracy. This synthesis aims to guide future research towards addressing these gaps and improving sentiment analysis for financial market predictions.

Keywords: *Machine Learning, Methodologies, Challenges, Data Quality, Model Complexity, Contextual factors*

Introduction

In recent years, the intersection of machine learning (ML) and financial market analysis has garnered substantial interest among researchers and practitioners alike (Kearney & Liu, 2014). The allure of applying ML to market data to predict stock returns is undeniable, given the vast potential rewards. One of the most intriguing applications of ML in finance is sentiment analysis, which involves the computational processing of textual data from news articles, social media, financial reports, and other sources to gauge market sentiment and, by extension, predict market movements (Bollen et al., 2011). This study aims to critically examine the current state of ML applications in sentiment analysis for stock market prediction, shedding light on its potential, challenges, and future directions.

The significance of this study lies in the growing reliance on automated systems for financial decision-making. In an era where vast amounts of data are generated daily, ML models offer a sophisticated approach to deciphering market sentiments from textual data, such as news articles, social media, and financial reports (Tetlock, 2007). These models promise enhanced prediction accuracy, potentially leading to more informed investment decisions. However, the complexity and unpredictability of financial markets pose unique challenges to ML models, often resulting in varying degrees of success (Pang & Lee, 2008).

Our study focuses on the methodologies employed in recent studies, their findings, and the factors influencing the performance of these models. By doing so, we aim to provide a comprehensive overview highlighting the advancements in the field and identifying the limitations and areas requiring further investigation. This endeavour is crucial for academic researchers and practitioners in the financial industry, as it can guide the development of more robust and reliable ML models for market prediction.

Following this introduction, the review will proceed as follows: The methodology section will detail the approach for selecting and analysing relevant literature, followed by a comprehensive review of the literature, organised by various preliminary tasks and their results. In the discussion section, the implications of these findings will be elaborated upon, pinpointing the potential reasons for the lack of predictive accuracy. Finally, the review will summarise findings and propose recommendations for future research directions.

2. Methodology

A comprehensive and systematic approach was employed to conduct a critical review that interrogates the efficacy of ML approaches to sentiment analysis for stock market prediction. This methodology section outlines the search strategy, selection criteria, and analytical framework used to examine the relevant literature.

Specific criteria guided our literature review to ensure relevance and quality. We focused on peer-reviewed articles and conference papers published in the last fourteen years, prioritising studies that specifically address the use of machine learning in sentiment analysis for stock market prediction. Additionally, we included significant earlier works to provide historical context. Studies were selected based on their methodological rigour, impact in the field, and the novelty of their approach.

A comprehensive search was conducted across multiple academic databases, including PubMed, IEEE Xplore, and Google Scholar. Keywords such as "machine learning," "sentiment analysis," "stock market prediction," and "financial forecasting" were used in various combinations to identify relevant literature. This was supplemented by a manual search of the reference lists of identified papers to capture any additional pertinent studies.

From each selected study, key information was extracted, including the ML models used, data sources, analysis techniques, and main findings. This data was then systematically analysed to identify common themes, trends, and gaps in the research. Special attention was given to the performance metrics used to evaluate ML models, as these are critical for understanding the effectiveness and practicality of the approaches.

To maintain the rigour of our review, we employed a cross-validation approach, where findings were independently reviewed and verified by multiple team members. This ensured that our interpretations and conclusions were not biased by individual perspectives. Moreover, the review process was iterative, allowing for the incorporation of newly published studies during our analysis, thereby keeping our review up-to-date and comprehensive.

3. Review of Literature

This section meticulously explores a range of scholarly works pertinent to the application of machine learning techniques in sentiment analysis for stock market prediction. By examining a breadth of literature, this section seeks to construct a comprehensive understanding of the current state of machine learning in sentiment analysis within the financial domain.

3.1 Analysis of Model Requirements

➢ Data Preprocessing

The efficacy of ML approaches to sentiment analysis largely depends on the quality of input data, such as news articles, social media, and financial reports. Each data source has unique challenges that require specific preprocessing methods. News articles and financial reports, being more structured, need preprocessing like named entity recognition and coreference resolution for context understanding (Hagenau et al., 2013) and complex feature engineering to quantify event impacts (Schumaker & Chen, 2009). Social media data, especially from platforms like Twitter, is more unstructured and requires noise reduction, text normalisation, and handling of brief content (Bollen et al., 2011). Data quality critically influences ML model performance; poor-quality data can significantly impair the effectiveness of advanced algorithms. Noise reduction, handling missing values, and ensuring representative samples are essential in preprocessing (Liu, 2012). Standard practices like filtering out stop-words and stemming help focus the analysis on relevant words (Agarwal et al., 2011). Therefore, data preprocessing is crucial in determining the success of ML models in sentiment analysis, setting the foundation for building effective models.

➢ Feature Engineering

Feature engineering is vital for the effectiveness of machine learning models in sentiment analysis for stock market prediction. It involves selecting informative features from raw data, which significantly influences predictive capabilities. Traditional techniques like Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) are standard but have limitations, such as high dimensionality and not capturing word order (Ramos, 2003; Salton & Buckley, 1988). Word Embeddings like Word2Vec and Global Vectors for Word Representation (GloVe) provide a more nuanced representation by considering word context (Mikolov et al., 2013; Pennington et al., 2014).

Incorporating domain-specific features, such as market indicators and economic data, alongside textual analysis enhances understanding of market sentiments and prediction accuracy (El-Masry et al., 2002). General sentiment lexicons, like SentiWordNet, are helpful but may not fully capture financial jargon, suggesting creating financial-specific lexicons (Loughran & Mcdonald, 2011). Robust feature selection methods are needed to manage text data's high dimensionality and improve model efficiency (Iguyon & Elisseeff, 2003). Research indicates that combining word embeddings and well-chosen domain-specific features can improve ML models' performance in predicting stock market trends (Bollen et al., 2011). Thus, balancing linguistic nuances with domain-specific knowledge is crucial in feature engineering and selection for sentiment analysis in stock market prediction.

➢ Classifications Algorithms

The use of ML models in sentiment analysis for stock market prediction includes a range of algorithms, from traditional classifiers like Support Vector Machines (SVM) to advanced deep learning methods. SVMs are favoured for their effectiveness in high-dimensional spaces typical of text classification (Hsu et al., 2008). Random Forests, an ensemble learning method, are valued for their ability to reduce overfitting in classification tasks (Breiman, 2001). Ensemble methods like gradient boosting machines enhance predictive performance by combining multiple models to reduce variance and bias (Schapire, 2003).

Comparatively, SVM and Random Forests have been vital in sentiment analysis, especially with well-tuned feature extraction. However, their performance can be limited by the non-linear and complex semantics in financial texts (Huang et al., 2005). Studies like Bao et al. (2017) have shown that while LSTMs excel in capturing financial data's temporal properties, sentiment analysis performance can vary significantly based on the data source.

To summarize, while neural networks offer promising advancements in sentiment analysis for stock market prediction, ensemble methods that integrate multiple algorithms continue to hold a competitive edge.

➢ Model Training and Validation

Training robust machine learning models for sentiment analysis in stock market prediction demands meticulous methodologies for learning from historical data and generalising to new data. This involves standard approaches like K-fold cross-validation to ensure model robustness and strategies to combat overfitting, such as regularisation and dimensionality reduction (Kohavi, 1995; Srivastava et al., 2014). Validating these models goes beyond conventional accuracy metrics, necessitating context-specific considerations like market volatility and trading volume and employing time-series cross-validation for more realistic predictions (Bergmeir et al., 2018). Thus, training and validation require a balanced approach, combining rigorous cross-validation with strategic overfitting prevention to accurately capture the complexities of financial markets.

➢ Evaluation Metrics and Results

Evaluating machine learning models for sentiment analysis in stock market prediction involves various metrics, each offering insights into different facets of model performance. Accuracy, indicating the proportion of correct predictions, is straightforward but can be misleading in imbalanced datasets common in finance (Haibo He & Garcia, 2009). The F1 score balances precision and recall, which is crucial when false positives and negatives have significant costs (Blair, 1979). Root Mean Square Error and Mean Absolute Error are critical for regression models, measuring discrepancies between predicted and actual values in stock prices or returns (Hyndman & Koehler, 2006; Willmott & Matsuura, 2005). Models may perform well in accuracy but show varied results with metrics like F1 score or RMSE (Bao & Datta, 2014). Bollen et al. (2011) highlighted the need to consider market dynamics, which accuracy alone may not fully capture. Statistical significance testing should validate metrics to ensure results reflect the models' actual capabilities, not just chance or overfitting, especially given the volatility and non-stationarity of financial markets (Abdi, 2007; David, 2002). Ultimately, a comprehensive evaluation using various metrics is essential to understand the predictive power of sentiment analysis models in stock market contexts.

3.2 Analysis of Model Inaccuracies

➢ The Role of Data Quality and Quantity

High-quality and extensive data are critical for the effectiveness of machine learning models in sentiment analysis for stock market prediction. The accuracy, reliability, and timeliness of data directly influence the model's predictions, as noted by (Bollen et al., 2011). Data collection and annotation biases, including selection and annotation biases highlighted by Loughran & Mcdonald (2011), can significantly skew model outcomes. Training models on unrepresentative data sets may lead to poor generalisation to different market conditions, as Wiebe et al. (2005) observed. Therefore, ensuring unbiased, comprehensive, and high-quality data is essential for developing accurate and reliable ML models for market prediction.

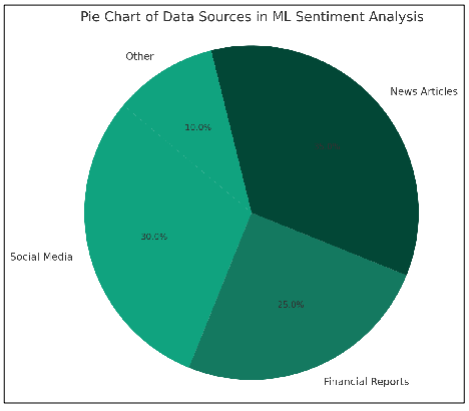

Figure 1: Data Sources in ML Sentiment Anlaysis
Source: Author's Own compilation

➢ The Complexity of Financial Markets

Financial markets are complex and adaptive, involving dynamic, non-linear interactions that impact the development of effective ML models for sentiment analysis and stock market prediction. According to Andersen (2013) markets are not fully efficient or inefficient but evolve with changing external influences, including sentiment. Fama's (1970) Efficient Market Hypothesis suggests that stock prices reflect all available information, posing challenges for predictive accuracy from sentiment analysis. ML models must navigate this complexity and market dynamics, as linear approaches may be insufficient for capturing non-linear relationships (David, 2002). Models must also adapt to rapid market

changes and volatility (Engle, 1982). Despite advancements in ML, capturing the full spectrum of market complexities remains challenging. Continuous evolution and sophistication in ML models are necessary to align with the market's intricate and dynamic nature, as discussed by Kearns (2013). Therefore, while ML models provide valuable insights, their predictive capability is limited by the complex nature of financial markets.

➢ Misalignment of Sentiment with Market Indicators

Market sentiment, as gauged from sources like news or social media, does not always align with actual market performance. Discrepancies often occur, with sentiment analysis sometimes indicating trends opposite to actual market movements (Tetlock, 2007). Factors like 'noise traders', who act on irrelevant data, can cause sentiment analyses to be out of sync with market indicators (DeLong et al., 1987). Timing issues further complicate matters; sentiment analysis often lags behind the market's current state due to data processing delays Engle and Patton (2001) and markets may react to news faster than sentiment analysis can keep up (Joulin et al., 2008). Despite being ideal, real-time sentiment analysis poses challenges in aligning accurately with trading activities and requires sophisticated infrastructure (Kraus & Feuerriegel, 2017). Therefore, while sentiment analysis provides insights, its alignment with market indicators is complex and affected by various factors, including timing lags and market dynamics.

In summary, the application of ML in sentiment analysis for stock market prediction is a multifaceted field that requires careful consideration of data quality, model choice, feature engineering, and validation approaches. While promising, the complexity of financial markets and the dynamic nature of market sentiment pose significant challenges to the predictive accuracy of these models.

4. Results and Discussion

This section of the critical review will interpret the findings, considering the broader implications, practical significance, and potential directions for future research based on the analysis of ML techniques in sentiment analysis for stock market prediction.

➢ Common Challenges and Pitfalls

Applying ML in sentiment analysis for stock market prediction presents significant challenges affecting model development and outcome interpretation. Natural language processing (NLP) models often struggle with the nuances of human language, such as irony and sarcasm, which are critical in understanding sentiment in financial contexts (González-Ibáñez et al., 2011). Contextual understanding in economic texts is also challenging, as words may have different meanings in different situations (Li, 2018).

A major issue in ML models is overfitting, where models trained on noisy financial data can perform poorly on new data due to learning irrelevant patterns (Hawkins, 2004). This issue is compounded in sentiment analysis by the changing nature of market-related sentiment (Li et al., 2014). Overfitting can lead to incorrect assumptions about the relationship between sentiment and market movements, and a phenomenon known as data snooping can result in models that appear effective by chance (Sullivan et al., 1999).

These challenges highlight the need for sophisticated NLP techniques and robust validation strategies to ensure the predictive reliability of ML models in financial sentiment analysis.

➢ Impact of Market Dynamics

Market dynamics, including volatility and unforeseen events like economic crises or political turmoil, significantly impact the effectiveness of sentiment analysis models in stock market prediction. Volatility, reflecting the variability of trading prices, can both mirror and intensify market sentiment, often complicating predictive models' ability to accurately translate sentiment changes into market movements (Pérez-Rodríguez, Torra, & Andrada-Félix, 2019). Similarly, unpredictable 'black swan' events can abruptly alter market sentiment, leading to the quick invalidation of model predictions if not accounted for (Baker & Wurgler, 2007). The adaptability of models, especially those employing advanced techniques like reinforcement learning or deep learning, is crucial in handling financial markets' non-linear and dynamic nature (Kercheval & Zhang, 2015). Models like the Autoregressive Conditional Heteroskedasticity (ARCH) are designed to adjust to such market volatility (Engle, 2000). Therefore, the success of sentiment analysis in predicting stock returns heavily depends on the models' ability to effectively respond to and incorporate rapid market changes and unexpected events

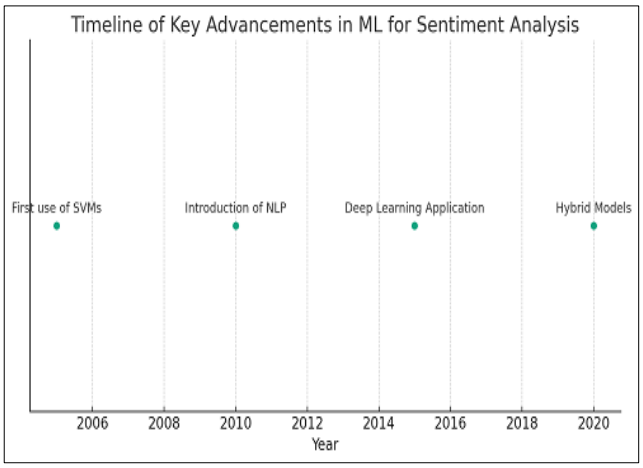

Figure 2: Key Advancements in ML for Sentiment Analysis
Source: Author's own compilation

➢ Efficacy of Machine Learning Techniques

Machine learning (ML) techniques have shown varying success in sentiment analysis for stock market prediction. These techniques, particularly natural language processing (NLP), have been crucial in analysing

large datasets to understand market sentiment, with studies like Liu, Hsaio, and Miao (2019) demonstrating NLP's effectiveness in extracting sentiment from financial news and social media. However, the predictive performance of ML techniques, including support vector machines (SVM) and neural networks, often varies, sometimes limited to short time horizons and specific market conditions (Patel et al., 2015). Ensemble methods, which combine multiple algorithms, have shown promise in outperforming individual models in some cases (Tsai & Hsiao, 2010). The effectiveness of ML in sentiment analysis largely depends on data quality, model appropriateness, and the ability to handle complex, non-linear relationships. Challenges like overfitting and underfitting, or lack of clear sentiment signals related to stock prices, can lead to failures. ML's potential in stock prediction through sentiment analysis is significant but depends on multiple factors, including the model's ability to accurately capture and interpret market sentiment's intricate and dynamic nature.

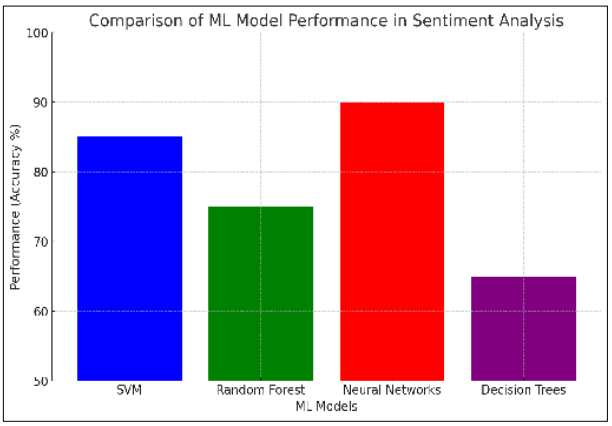

Figure 3: Comparison of ML Model Performance in Sentiment Analysis
Source: Author's Own compilation

➢ **Comparison with Existing Theories and Models**

The Efficient Market Hypothesis (EMH) posits that stock prices reflect all available information, potentially limiting the effectiveness of sentiment analysis by ML models. However, sentiment analysis can provide advantages, especially in short-term trading, by capturing emotional responses not yet reflected in stock prices, challenging the strong form of EMH but aligning with its semi-strong or weak forms. Sentiment analysis might reveal emotional reactions and investor sentiment not immediately apparent in price movements, suggesting markets are not always fully efficient.

Studies have shown that ML models can exploit market inefficiencies, indicating that markets do not always incorporate all available information, contrary to EMH. This leads to broader theoretical implications, where the EMH may not fully account for the complexities of human emotion and decision-making. Behavioral Finance Theory, which considers psychological factors in financial analysis, and the Noise Trader Theory, highlighting the impact of irrational traders on market inefficiencies, provide more comprehensive frameworks for understanding sentiment analysis's implications.

In summary, ML in sentiment analysis both complements and challenges traditional financial theories like the EMH, suggesting a more nuanced understanding of market dynamics is needed. This includes recognising the role of investor psychology and sentiment in market movements, aligning with the concept that market predictability involves both rational and irrational behaviours. This discussion suggests a need for expanded theoretical models that integrate insights from behavioural finance, acknowledging the complexity and nuances of market dynamics and efficiency.

➢ **Implications for ML Techniques**

The critical review reveals that developing and refining ML techniques for sentiment analysis in stock market prediction involves balancing model complexity with interpretability and adapting to market dynamics. Advanced ML models, like deep learning, are adept at identifying complex patterns but often lack transparency, highlighting the need for a balance between sophistication and understandability. The models should also incorporate temporal market dynamics, such as momentum and information relevance over time, and adapt to the ever-changing stock market through techniques like online learning.

Improvements could include hybrid approaches that combine ML with traditional forecasting models, sophisticated feature engineering to capture human sentiment and behaviour better, and enhancing model robustness to prevent overfitting. Investing in Explainable AI (XAI) could help make complex models more transparent and trustworthy.

The review on applying ML techniques to sentiment analysis in stock market prediction identifies both challenges, such as capturing complex market dynamics and ensuring high data quality, and opportunities, like the potential of hybrid models and reinforcement learning. It calls for ongoing innovation in ML, advocating a comprehensive approach that merges ML with various analytical methods, human insights, and disciplines like economic theory and behavioral science. The review also emphasizes the need for advanced data analysis techniques, particularly for unstructured data, and highlights the importance of addressing data biases.

Currently, sentiment analysis in stock market prediction, intersecting finance, computer science, and data science, shows promise but faces challenges in accuracy, robustness, and interpretability, preventing its mainstream adoption. Key steps for practical application include developing extensive, unbiased datasets, models capable of processing diverse data types, and fostering collaboration between financial experts and ML practitioners to create interpretable and relevant models.

The field is on the brink of significant progress, with overcoming these limitations central to its advancement. Integrating emerging ML techniques, improving data quality, and adopting interdisciplinary approaches are crucial for realizing the full potential of sentiment analysis in this domain.

5. Conclusion

Our review critically examined the use of machine learning in sentiment analysis for stock market prediction, uncovering both promising potentials and notable challenges. We found that while ML models offer sophisticated tools for market analysis, their effectiveness is significantly influenced by data quality, model complexity, and external market factors.

Looking forward, we encourage further research in several key areas. These include the development of more adaptable ML models that can respond to rapid market changes, exploring cross-lingual and cross-cultural applications of sentiment analysis, and integrating sentiment data with other financial indicators for a more comprehensive approach to market prediction. Additionally, research is needed focusing on the ethical implications of using ML in financial decision-making.

The continued advancement in ML and sentiment analysis holds significant potential for transforming stock market prediction. Improved accuracy and adaptability of these models could lead to more informed and effective investment strategies, potentially changing the landscape of financial analysis and decision-making.

In conclusion, integrating machine learning in sentiment analysis for stock market prediction is a rapidly evolving field with substantial opportunities and challenges. As this technology continues to develop, researchers and practitioners must work collaboratively to harness its potential responsibly, ensuring that it serves as a tool for enhanced, ethical financial decision-making.

References

1. Abdi, H. (2007). The Bonferonni and Šidák corrections for multiple comparisons. In *Encyclopedia of Measurement and Statistics*. Sage Publications, Inc. https://methods.sagepub.com/reference/encyclopedia-of-measurement-and-statistics
2. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of twitter data. In Proceedings of the Workshop on Language in Social Media (LSM 2011). *Proceedings of the Workshop on Language in Social Media (LSM 2011)*.
3. Andersen, E. S. (2013). The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective. *Evolutionary Economics: Post-Schumpeterian Contributions*, 1–238. https://doi.org/10.4324/9781315072012
4. Baker, M., & Wurgler, J. (2007). Investor Sentiment in the Stock Market. *Journal of Economic Perspectives*, *21*(2), 129–151. https://doi.org/10.1257/jep.21.2.129
5. Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLoS ONE*, *12*(7). https://doi.org/10.1371/journal.pone.0180944
6. Bao, Y., & Datta, A. (2014). Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures. *Management Science*, *60*(6), 1371–1391. https://doi.org/10.1287/mnsc.2014.1930
7. Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics and Data Analysis*, *120*. https://doi.org/10.1016/j.csda.2017.11.003
8. Blair, D. C. (1979). Information Retrieval. *Journal of the American Society for Information Science*, *30*(6), 374–375. https://doi.org/10.1002/asi.4630300621
9. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007
10. Breiman, L. (2001). Random forests. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2001)*, *12343 LNCS*.
11. David, R. C. (2002). Analysis of Financial Time Series. *Journal of Financial Research*, *25*(3), 445–446. https://doi.org/10.1111/1475-6803.00029
12. DeLong, J. B., Shleifer, A., Summers, L., & Waldmann, R. (1987). The Economic Consequences of Noise Traders. In *Russell The Journal Of The Bertrand Russell Archives*. https://doi.org/10.3386/w2395
13. El-Masry, A. M., Ghaly, M. F., Khalafallah, M. A., & El-Fayed, Y. A. (2002). Deep Learning for Event-Driven Stock Prediction. *Journal of Scientific and Industrial Research*, *61*(9).
14. Engle, R. F. (2000). Dynamic Conditional Correlation - A Simple Class of Multivariate GARCH Models. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.236998
15. Engle, R. F., & Patton, A. J. (2001). What good is a volatility model? *Quantitative Finance*, *1*(2), 237–245. https://doi.org/10.1088/1469-7688/1/2/305
16. Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383. https://doi.org/10.2307/2325486
17. González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in Twitter: A closer look. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational*

Linguistics: Human Language Technologies, *2*, 581–586.
18. Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, *55*(3), 685–697. https://doi.org/10.1016/j.dss.2013.02.006
19. Haibo He, & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239
20. Hawkins, D. M. (2004). The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1–12. https://doi.org/10.1021/ci0342472
21. Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2008). A Practical Guide to Support Vector Classification. *BJU International*, *101*(1), 1396–1400. http://www.csie.ntu.edu.tw/%7B~%7Dcjlin/papers/guide/guide.pdf
22. Huang, W., Nakamori, Y., & Wang, S. Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers and Operations Research*, *32*(10), 2513–2522. https://doi.org/10.1016/j.cor.2004.03.016
23. Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001
24. Iguyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. In *Journal of Machine Learning Research* (Vol. 3, pp. 1157–1182).
25. Joulin, A., Lefevre, A., Grunberg, D., & Bouchaud, J.-P. (2008). *Stock price jumps: news and volume play a minor role*. https://doi.org/https://doi.org/10.48550/arXiv.0803.1769
26. Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. In *International Review of Financial Analysis* (Vol. 33, pp. 171–185). https://doi.org/10.1016/j.irfa.2014.02.006
27. Kearns, M. (2013). Machine Learning for Market Microstructure and High Frequency Trading. In *High Frequency Trading - New Realities for Traders, Markets and Regulators* (pp. 1–21).
28. Kercheval, A. N., & Zhang, Y. (2015). Modelling high-frequency limit order book dynamics with support vector machines. *Quantitative Finance*, *15*(8), 1315–1329. https://doi.org/10.1080/14697688.2015.1032546
29. Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *IJCAI International Joint Conference on Artificial Intelligence*, *2*, 1137–1143.
30. Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, *104*, 38–48. https://doi.org/10.1016/j.dss.2017.10.001
31. Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., & Chen, Y. (2014). The effect of news and public mood on stock movements. *Information Sciences*, *278*, 826–840. https://doi.org/10.1016/j.ins.2014.03.096
32. Liu, B. (2012). Sentiment Analysis and Opinion Mining Mining. *Synthesis Lectures on Human Language Technologies*, *5*(1).
33. Loughran, T., & Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, *66*(1). https://doi.org/10.1111/j.1540-6261.2010.01625.x
34. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.
35. Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135. http://www.nowpublishers.com/article/Details/INR-001
36. Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications*, *42*(4), 2162–2172. https://doi.org/10.1016/j.eswa.2014.10.031
37. Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162
38. Ramos, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries. *Proceedings of the First Instructional Conference on Machine Learning*, *242*, 29–48.
39. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0
40. Schapire, R. E. (2003). The Boosting Approach to Machine Learning: An Overview. *Nonlinear Estimation and Classification*, *171*(9), 149–171. https://doi.org/10.1007/978-0-387-21579-2_9
41. Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news. *ACM Transactions on Information Systems*, *27*(2), 1–19. https://doi.org/10.1145/1462198.1462204
42. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.
43. Sullivan, R., Timmermann, A., & White, H. (1999). Data-Snooping, Technical Trading Rule Performance, and the Bootstrap. *The Journal of Finance*, *54*(5), 1647–1691. https://doi.org/10.1111/0022-1082.00163
44. Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, *62*(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x
45. Tsai, C.-F., & Hsiao, Y.-C. (2010). Combining multiple feature selection methods for stock prediction:

Union, intersection, and multi-intersection approaches. *Decision Support Systems*, *50*(1), 258–269. https://doi.org/10.1016/j.dss.2010.08.028

46. Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. In *Language Resources and Evaluation* (Vol. 39, Issues 2–3, pp. 165–210). https://doi.org/10.1007/s10579-005-7880-9
47. Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82. https://doi.org/10.3354/cr030079