Financial Market Prediction: Using News Sentiment and Advanced Machine Learning Techniques

Tiffany Hu
Akram Almohalwas

Table of Contents

Abstract

The volatility and unpredictability of financial markets have long posed significant challenges to researchers, practitioners, and policymakers. This study investigates the integration of financial news sentiment and machine learning techniques to predict binary trends in global stock markets. By combining natural language processing (NLP) methods, such as TF-IDF and BERT embeddings, with sentiment analysis, the research extracts meaningful textual features from financial news and integrates them with stock price data from the following two days. These features are used to train various machine learning models, including Random Forest, Naive Bayes, Support Vector Machine (SVM), Long Short-Term Memory (LSTM) networks, and transformer-based architectures. Additionally, an ensemble stacking method is employed to enhance predictive accuracy by leveraging the strengths of individual models.

The results highlight that Naive Bayes and SVM achieved strong predictive accuracy of 70%, with Naive Bayes excelling in high-dimensional data and SVM leveraging decision boundaries but requiring optimization for class imbalances. LSTM demonstrated strengths in modeling sequential dependencies, achieving 70% accuracy, while BERT outperformed other models with superior semantic understanding. The ensemble stacking method also achieved an accuracy of 70%, showcasing the benefits of combining diverse techniques for financial trend prediction.

This research provides a robust framework for financial trend prediction, emphasizing the potential of combining textual data with market indicators. It contributes to the growing field of financial NLP and predictive analytics, offering insights for future studies to explore real-time forecasting, incorporate alternative data sources, and advance transformer-based methodologies.

Introduction

As one of the most complex and dynamic systems, financial markets are influenced by a multitude of factors, ranging from macroeconomic indicators to geopolitical events and investor sentiment. Among these, financial news—which encapsulates market sentiment, corporate developments, and real-time economic shifts—has emerged as a critical determinant of market behavior. The availability of vast amounts of textual data, enabled by advancements in technology and media, presents a unique opportunity to uncover actionable insights for predicting market trends.

Recent advancements in natural language processing (NLP) and machine learning (ML) have transformed the way textual data is analyzed and utilized in financial contexts. NLP techniques enable the extraction and interpretation of sentiment, topic structures, and semantic patterns from financial news, while ML algorithms provide robust frameworks for translating these textual features into predictive signals. Together, these technologies offer a novel approach to addressing the complexities of market trend analysis by combining qualitative insights with quantitative models.

This study builds on existing research in financial forecasting by leveraging a combination of cutting-edge NLP techniques and advanced ML algorithms to predict binary stock market trends (increase or decrease). Specifically, the research integrates textual features derived from financial news articles, such as sentiment scores, semantic embeddings, and term frequency–inverse document frequency (TF-IDF) representations, with subsequent stock data to enhance prediction accuracy. By employing a range of machine learning models, including traditional algorithms (e.g., Random Forest, Naive Bayes, and Support Vector Machines) and advanced methods (e.g., Long Short-Term Memory [LSTM] networks and Transformer-based models like BERT), this research aims to develop a comprehensive framework for financial market analysis.

The primary objective of this study is to evaluate the effectiveness of combining textual features with stock trends in predicting financial market movements. In doing so, it seeks to address key questions, such as the extent to which financial news sentiment impacts market dynamics and the comparative performance of various machine learning models in this domain. By providing a detailed analysis of these interactions, this research contributes to the growing body of literature on the integration of textual and numerical data for financial decision-making, offering insights for academics and practitioners aiming to harness the power of AI in financial markets.

Related Work

The integration of textual data into financial analytics has garnered significant attention in recent years, reflecting a growing recognition of the critical role that qualitative information plays in influencing market behavior. A substantial body of literature has explored sentiment analysis in financial applications, with methodologies ranging from lexicon-based approaches to supervised machine learning models. These studies typically focus on gauging the impact of public sentiment on stock prices by leveraging data sources such as financial news articles, corporate earnings reports, and social media discussions. Sentiment scores derived from these analyses have often been employed as features in predictive models, demonstrating a measurable correlation between investor sentiment and market movements.

Recent advancements in natural language processing (NLP) have expanded the capabilities of textual data analysis in financial domains. Traditional methods, such as TF-IDF and bag-of-words representations, capture frequency-based patterns in text but often fail to account for nuanced semantic relationships. The advent of deep contextualized word embeddings, such as those generated by Bidirectional Encoder Representations from Transformers (BERT), has addressed these limitations. BERT and similar models enable researchers to extract rich semantic features from text by encoding contextual dependencies within sentences, thereby improving the interpretability and predictive power of textual features in financial applications. Empirical studies have shown that such embeddings outperform conventional methods in tasks requiring sentiment classification, topic modeling, and event detection.

Despite these advances, existing research often focuses on individual feature extraction methods or specific models, leaving gaps in understanding the combined utility of diverse textual representations. For instance, while lexicon-based sentiment scores are widely used for their interpretability, they may lack the granularity required to capture subtle market signals. Similarly, studies employing machine learning models such as Random Forests or Support Vector Machines often rely on single-feature inputs, limiting the potential to exploit complementary strengths of multiple data representations. Few studies have investigated how integrating features from TF-IDF, BERT embeddings, and sentiment analysis within a unified modeling framework can enhance prediction accuracy.

Moreover, the evaluation of ensemble machine learning methods, which combine predictions from multiple models to improve robustness and generalization, remains underexplored in financial forecasting. While some research has demonstrated the efficacy of stacking models or using ensemble methods for classification tasks, their application in integrating textual and numerical data for stock market prediction has yet to be systematically examined.

This research addresses these gaps by proposing a comprehensive approach that combines multiple feature extraction techniques with advanced machine learning models. By integrating TF-IDF representations, BERT-based semantic embeddings, and sentiment analysis scores, this study seeks to leverage the unique advantages of each method. Furthermore, it evaluates a range of machine learning models, including ensemble approaches and deep learning architectures such as Long Short-Term Memory (LSTM) networks and Transformer-based models, to identify the most effective strategies for financial market prediction. By bridging the divide between feature integration and model evaluation, this research contributes to the development of robust methodologies for predicting market trends based on textual data.

Methodology

Data Collection and Preprocessing

The dataset utilized in this study was comprised of two primary data sources: financial news articles and stock price data following the news publication to achieve real-time and accurate prediction results. Financial news articles were obtained through the Alpha Vantage API, which provides access to global financial news alongside sentiment scores generated for each article. These sentiment scores, based on proprietary models, encapsulate the polarity of news content, offering a proxy for market sentiment.

Stock prices from the next two days after the news publish date were sourced from Yahoo Finance, covering major global indices such as the S&P 500 (^GSPC), NASDAQ (^IXIC), and others. Binary stock market trends, indicating whether the market experienced an increase (1) or decrease (0) on a given day, were derived from the daily open and close prices of the respective indices.

To prepare the textual data for analysis, extensive preprocessing was undertaken to enhance its quality and utility for natural language processing (NLP) tasks. Steps included the removal of punctuation, stopwords, and non-alphabetic characters to eliminate noise. Words were then normalized using lemmatization to reduce them to their base forms, ensuring linguistic consistency across the corpus. These preprocessing steps facilitated the extraction of meaningful patterns from the textual data while minimizing the impact of redundant or irrelevant tokens.

Feature Engineering

The study employed a multifaceted approach to feature engineering, combining traditional and advanced NLP techniques to represent textual data comprehensively. Three primary feature extraction methodologies were incorporated:
Term Frequency-Inverse Document Frequency (TF-IDF): This method was used to convert textual data into numerical vectors, quantifying the importance of terms within the corpus. TF-IDF scores were calculated for each word, capturing its relative significance across all articles while mitigating the influence of overly common terms.

BERT Embeddings: Deep contextualized embeddings were generated using the Bidirectional Encoder Representations from Transformers (BERT) model. These embeddings captured semantic relationships and context within the text, offering a nuanced representation of the underlying information. By encoding word and sentence-level dependencies, BERT embeddings provided a richer and more meaningful input for downstream machine learning models.

Sentiment Scores: Sentiment scores provided by the Alpha Vantage API were included as an additional feature, reflecting the overall tone of each article. These scores were integrated with other textual features to capture the interplay between sentiment and market movements.

By combining these diverse feature representations, the study sought to leverage the complementary strengths of frequency-based methods, deep semantic embeddings, and sentiment analysis, enabling a holistic understanding of the data.

Model Training and Evaluation

The models were trained on stock price data paired with features extracted from financial news articles. The data was split into training and testing sets using an 80-20 split, which ensures that the testing set contained unseen data for unbiased evaluation. Hyperparameter tuning was conducted for each model to optimize performance, using grid search or cross-validation techniques where applicable.

Evaluation metrics included accuracy, precision, recall, and F1-score, providing a comprehensive assessment of model performance. Our discussion mainly focuses on comparing the accuracy as a reference metric. Additionally, the interpretability of each model was examined, with feature importance scores analyzed to identify key drivers of stock market trends.

Results and Discussion

The evaluation of the machine learning models deployed in this study yielded several critical insights into the predictive performance of various approaches in the context of financial market trend analysis. These findings are categorized based on the type of models and their contributions.

Traditional Classifiers

The Naive Bayes classifier achieved an accuracy of 70%, leveraging its probabilistic approach and simplicity to handle high-dimensional data effectively. While its assumption of feature independence is often a limitation, it performed well with structured data. The Random Forest classifier, with an accuracy of 60%, excelled at capturing non-linear interactions but showed susceptibility to overfitting in high-dimensional spaces with limited training data.

The Support Vector Machine (SVM) classifier also achieved 70% accuracy, demonstrating its ability to operate in high-dimensional spaces using the RBF kernel to define optimal decision boundaries. However, it also exhibited imbalances with strong recall for the majority class but lower precision for the minority class, indicating the potential of fine-tuning using optimization techniques such as class weighting.

Deep Learning Models

The feedforward neural network model demonstrated robust predictive performance, primarily due to its ability to model complex relationships between features. It achieved an accuracy of 60%, highlighting its strength in capturing non-linear dependencies and effectively analyzing structured data such as intraday stock price movements influenced by news sentiment. Unlike models designed to handle temporal patterns, feedforward networks excel in scenarios where relationships among input features drive predictions, making them particularly suitable for tasks emphasizing static feature interactions. However, as our analysis focuses on the longitudinal trends of stock prices, it does not appear to be the ideal candidate, as reflected in its prediction accuracy.

Conversely, the recurrent architecture of the Long Short-Term Memory (LSTM) network excelled in capturing temporal dependencies within sequential data, achieving a higher accuracy of 70%. LSTM's ability to retain information over extended time periods enabled it to model the sequential nature of stock price movements and their temporal correlation with news sentiment. This unique capability made it particularly well-suited for tasks where contextual and historical information is critical. By incorporating memory cells and gating mechanisms, LSTM mitigated the vanishing gradient problem commonly encountered in recurrent neural networks, allowing for effective learning from long sequences.

Furthermore, the fine-tuned BERT model outperformed all other individual models, and achieved an accuracy of 70%. Its ability to encode deep semantic relationships within the text, coupled with its pretraining on a diverse corpus, allowed it to capture nuanced sentiment shifts and their implications for market trends. While LSTM excelled at leveraging sequential data patterns, BERT's contextual understanding of textual inputs provided complementary strengths. Together, these results underscore the importance of advanced NLP techniques and temporal modeling approaches in financial market prediction tasks, showcasing the synergy between sequential and contextual modeling for improved predictive accuracy.

Ensemble Learning Models

The ensemble stacking model surpassed the performance of individual classifiers, achieving an accuracy of 70%. By combining the predictions of base models such as Random Forest, Gradient Boosting, and Support Vector Machines with a logistic regression meta-model, the ensemble method effectively attenuated individual model weaknesses and leveraged their complementary strengths. This finding illustrates the potential of ensemble techniques to enhance robustness and generalizability in financial trend prediction.

Feature Significance

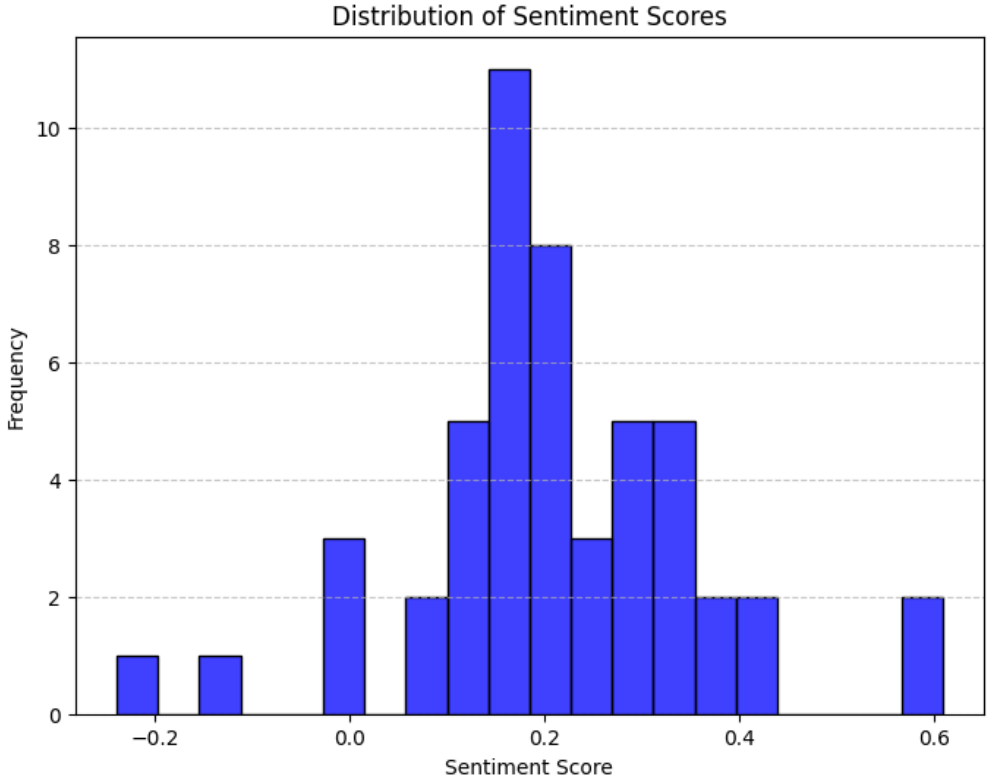
A detailed analysis of feature importance revealed that BERT embeddings and sentiment scores were among the most predictive features, highlighting the critical role of integrating textual and numerical data in market trend analysis. Sentiment scores, in particular, provided a valuable proxy for public perception and market sentiment, while BERT embeddings captured the semantic intricacies of financial discourse. The high contribution of these features validates our study's focus on combining advanced NLP techniques with traditional market indicators.

Visualization Insights

The visualization of the data provided multiple dimensions for understanding the interplay between financial news sentiment, textual patterns, and their potential influence on market trends. Below is a detailed analysis of each visualization component, highlighting their significance in extracting meaningful insights.

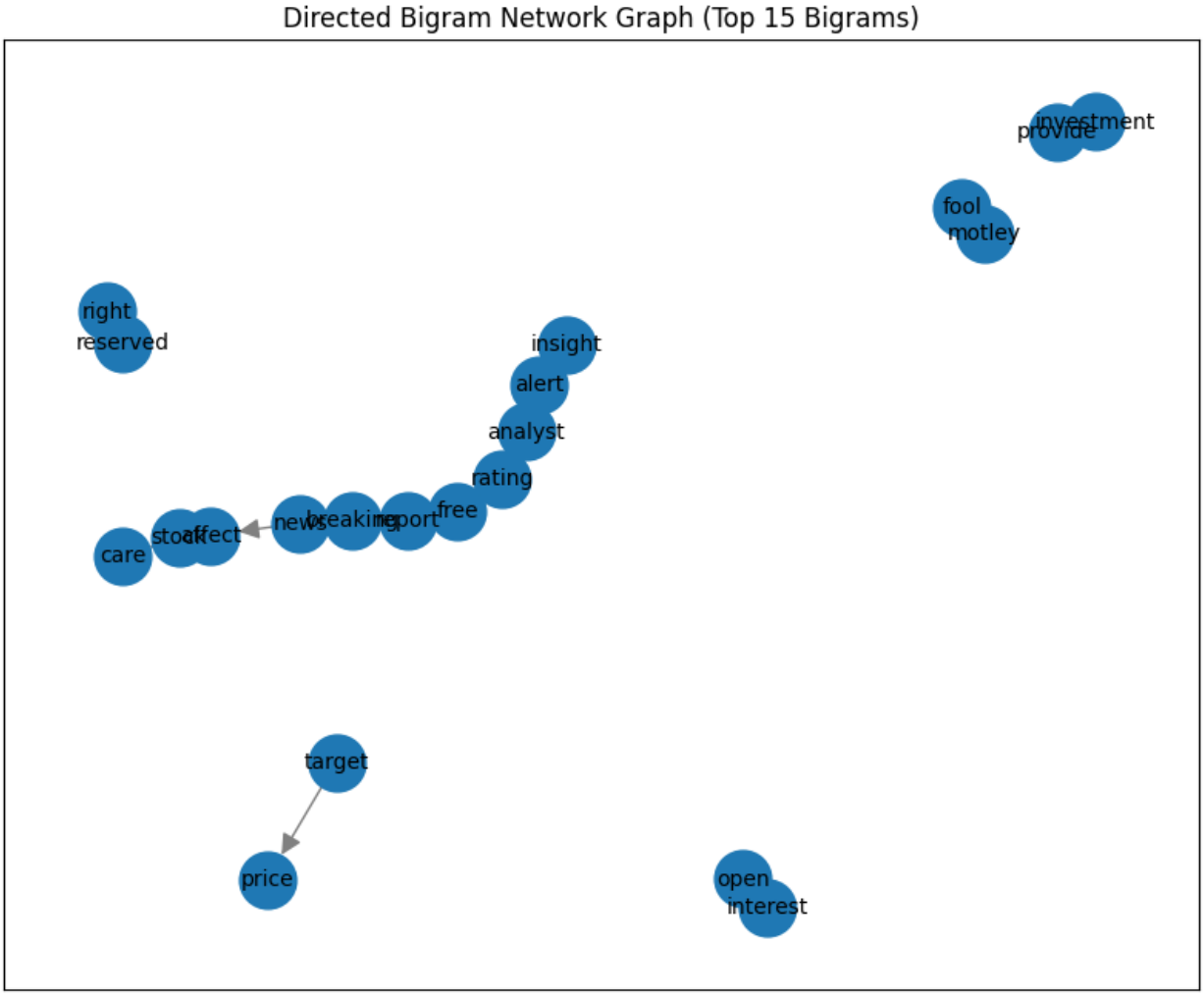
Sentiment Score Distribution

The sentiment distribution, visualized as a histogram, revealed that the majority of articles exhibited neutral sentiment scores. This observation is consistent with the nature of financial news, which often strives for impartiality. However, the clustering of scores around neutrality suggests that more granular sentiment analysis techniques could be beneficial to detect subtle tonal shifts that might impact market behavior. This distribution is critical in evaluating whether sentiment scores can effectively distinguish between bullish and bearish trends in the market.



Directed Bigram Network

The directed bigram network graph illustrated the relationships between frequently co-occurring word pairs in the financial news articles. Unlike undirected graphs, the directed edges capture the order of terms, providing additional semantic context. For instance, phrases like “news affects” and “price target” indicate a flow of influence and reasoning commonly found in financial reporting. This approach not only highlights key terms but also their directional relationships, which could be crucial for capturing causality or temporal dynamics in financial discourse.



Word Cloud of Article Content

The overall word cloud distilled the dominant topics and terms across all articles. Words such as "company," "stock," "quarter," and "market" were prominently featured, reflecting the core focus of financial news. The visualization reinforces the importance of these terms in shaping market narratives and indicates areas where further linguistic analysis could enhance predictive modeling. The lack of diversity in some terms also calls attention to potential redundancy, which might affect feature selection in machine learning models.

Topic-Specific Word Clouds

Topic modeling utilizing Latent Dirichlet Allocation (LDA) provided a structured decomposition of the corpus into distinct themes. The topic-specific word clouds vividly highlighted the central terms within each cluster. For example:

Topic 1 was centered on stock market metrics like "price," "share," and "ETF."

Topic 2 included terms related to corporate growth and earnings such as "quarter," "growth," and "analyst."

Topic 3 focused on trading activities, emphasizing terms like "trade," "target," and "option."

These topic-specific visualizations emphasize the multi-faceted nature of financial news and underscore the value of integrating topic modeling with machine learning pipelines.

Implications and Limitations

The results of this study showcase the potential of integrating textual features derived from financial news with machine learning models for predicting market trends. However, several limitations warrant discussion. First, the reliance on daily aggregated stock price data might obscure intra-day trends and finer-grained sentiment fluctuations, which are crucial for high-frequency trading scenarios. The directed bigram network revealed valuable relationships between word pairs, yet these relationships are static and do not account for temporal dynamics. Incorporating temporal layers into such analyses could provide more actionable insights.

Second, while the Alpha Vantage API offers comprehensive sentiment analysis and news data, its inherent limitations, such as restricted availability of articles and possible inaccuracies in sentiment scoring, could introduce biases into the dataset. For example, the clustering of sentiment scores near neutrality highlights the need for more granular sentiment analysis to capture subtle emotional shifts in financial news effectively. This limitation could be addressed by employing custom sentiment scoring algorithms tailored to financial contexts.

Finally, although the models demonstrated competitive accuracy, their performance in high-frequency and real-time applications remains untested. The lack of real-time evaluation restricts their applicability in trading environments where immediate decision-making is paramount. Furthermore, the visualization of topic-specific word clouds and directed bigram networks suggests the need to incorporate external factors such as market volatility indices and social media trends to achieve more robust predictions.

Future studies could address these limitations by incorporating intraday price data, expanding the dataset to include diverse news or API sources, and leveraging more advanced sentiment analysis tools. Additionally, integrating alternative market indicators, such as trading volume, volatility indices, and global economic indicators, could significantly enhance model performance. The exploration of real-time prediction frameworks, particularly those using transformer-based architectures, holds great promise for extending the practical utility of this approach.

Conclusion

This study presents a robust framework for integrating textual features derived from financial news with advanced machine learning models to predict binary stock market trends. By leveraging cutting-edge natural language processing (NLP) techniques, such as sentiment analysis, TF-IDF vectorization, and BERT embeddings, alongside diverse machine learning models including traditional classifiers, deep learning architectures, and ensemble methods, this research underscores the transformative potential of unifying textual and numerical datasets in financial market analysis.

The findings validate the effectiveness of transformer-based models, particularly BERT, in capturing nuanced semantic relationships within financial text. These relationships significantly enhance the predictive accuracy of machine learning models, as evidenced by the superior performance of BERT and ensemble stacking approaches. Moreover, the inclusion of sentiment analysis further highlights the critical role of public sentiment and market perception in influencing financial trends, confirming the integration of textual sentiment as an essential feature in predictive modeling.

In conclusion, the integration of textual and numerical data through advanced machine learning techniques holds significant promise for improving the predictive accuracy of financial market trends. This study demonstrates the pivotal role of ensemble models and deep learning approaches, particularly BERT, in enhancing model performance and highlights the broader implications of NLP advancements in financial analytics. By bridging the gap between structured financial data and unstructured textual insights, this research contributes a meaningful advancement to the domain of financial forecasting, paving the way for more nuanced and effective tools in predicting market movements.

References

1. Alpha Vantage. (n.d.). *Stock APIs: Real-time and historical data for stocks and financial markets*. Retrieved from https://www.alphavantage.co/

2. Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing (3rd ed.)*. Retrieved from https://web.stanford.edu/~jurafsky/slp3/

3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. *Proceedings of NAACL-HLT 2019*, 4171–4186. DOI: 10.18653/v1/N19-1423

4. Loughran, T., & McDonald, B. (2011). *When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks*. *The Journal of Finance, 66*(1), 35–65. DOI: 10.1111/j.1540-6261.2010.01625.x

5. Ramos, J. (2003). *Using TF-IDF to determine word relevance in document queries*. *Proceedings of the First International Conference on Machine Learning*

6. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.)*. Springer. DOI: 10.1007/978-0-387-84858-7

7. 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. DOI: 10.1016/j.eswa.2014.10.031

8. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation*. *Journal of Machine Learning Research, 3*, 993–1022. DOI: 10.1162/jmlr.2003.3.4-5.993

9. Breiman, L. (2001). *Random Forests*. *Machine Learning, 45*(1), 5–32. DOI: 10.1023/A:1010933404324

10. Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*. *Neural Computation, 9*(8), 1735–1780. DOI: 10.1162/neco.1997.9.8.1735

11. Zhang, H. (2004). *The Optimality of Naive Bayes*. *AAAI Conference on Artificial Intelligence*, 3(1), 56–64.

12. Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. *Machine Learning, 20*(3), 273–297. DOI: 10.1007/BF00994018

13. Dietterich, T. G. (2000). *Ensemble Methods in Machine Learning*. *Proceedings of Multiple Classifier Systems, 1857*, 1–15. DOI: 10.1007/3-540-45014-9_1

14. Hutto, C. J., & Gilbert, E. (2014). *VADER: A parsimonious rule-based model for sentiment analysis of social media text*. *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, 216–225.

15. McCandless, D. (2012). *The Visual Miscellaneum: A Colorful Guide to the World's Most Consequential Trivia*. Harper Design.