**Executive Summary**

**Predicting Stock Returns Using News Sentiment Analysis: A Comparative Study of Machine Learning Models during COVID-19**

Student ID : 720034950

Name : Kota Kobayashi

**Introduction**

In the contemporary financial environment, news sentiment's role in predicting stock market trends has been highlighted. This relationship became even more pronounced amidst the global disruption caused by the COVID-19 pandemic. With the increased dependency on digital news platforms and algorithmic trading during this period, news sentiment analysis has solidified its place as an essential tool in shaping investment decisions.

**Problem Statement & Background:**

Historically, numerous empirical studies have delved into the dynamics between news sentiment and stock returns. However, the COVID-19 pandemic has brought about unique challenges that haven't been previously tackled. Many existing research methodologies, tailored for the pre-pandemic world, now appear outdated. Past research, such as those by Cakra and Trisedya (2015), leaned on regression methodologies, while the current context sees the rise of more advanced models like CatBoost. This discrepancy in methodologies has led to an inconsistent research landscape. Moreover, the narrow scope of many studies, often focusing on specific sectors or broad indices, further exacerbates the fragmentation in the literature. This has left stakeholders, from financial experts to individual investors, facing a multitude of questions regarding model selection, feature importance, and the general applicability of sentiment analysis.

Adding another layer of complexity is the narrow focus of many studies, either targeting specific sectors, like big tech firms, or broad indices like the S&P 500. While a few have ventured into analyzing the NASDAQ, their limited numbers contribute to the existing research void (Lee, 2022). This patchwork of fragmented and sometimes outdated research leaves stakeholders, from financial experts to individual investors, navigating a maze of confusion. They are left grappling with pivotal questions about optimal model selection, feature prioritization, and the broader applicability of sentiment analysis across a varied spectrum of firms.

**Research Objectives:**

To address these challenges and provide actionable insights, this research embarked on answering three pivotal questions:

* *Comparative Analysis:* How do present-day machine learning models perform in leveraging news sentiment for stock return predictions during the pandemic?
* *Feature Significance:* In these models, which variables are pivotal? Does news sentiment hold its ground against other influential factors like volume or the influence of COVID-19?
* *Company-Specific Variations:* How does sentiment analysis fare across different firms and sectors? Are there specific areas where sentiment analysis might be misleading?

**Key Findings:**

* *Model Efficacy & Sectoral Specificity:* XGBoost emerged as the standout performer. However, deeper exploration highlighted performance variations across sectors. Biotechnology firms, like Biogen Inc and Gilead Sciences Inc, exhibited heightened prediction accuracy. Their extended R&D cycles, strict regulatory environments, and continuous innovation resulted in predictable financial trajectories, making them ripe for predictive modeling. This granularity in results underscores the necessity of sector-specific modeling for nuanced predictions.

Table Results of AUC scores

A red line on a white background

Description automatically generated

* *Feature Importance & Sentiment Reliability:* The study emphasized the dominance of traditional stock indicators, especially Volume. Interestingly, sentiment score, despite its relevance, was found less influential than expected. This suggests that investment strategies should not be overly reliant on sentiment analysis alone. A fusion of sentiment scores with traditional stock indicators is paramount for a holistic prediction model.

A group of blue and white bars

Description automatically generated

Figure Feature importance across models

* *Pandemic Data & Predictions:* The importance of granular pandemic data like new COVID-19 cases was accentuated, while binary indicators like pandemic phase were less influential. This finding accentuates the need for detailed, continuous data that reflects dynamic external events, aiding in capturing the intricacies of stock returns during tumultuous periods.
* *Company-Specific Sentiment Variations:* A pronounced correlation between firm size and news sentiment was observed. However, there were anomalies with some firms diverging significantly from general trends, emphasizing the need for a bespoke approach when leveraging sentiment scores in financial strategies.

**Recommendations:**

Drawing from these insights, the following recommendations are proposed:

* **Model Selection for Predictive Analysis:** Financial stakeholders, including analysts and hedge fund managers, should prioritize the XGBoost model. Its superior performance, especially when analyzing biotech firms, makes it a prime candidate. A move towards sector-specific models, like XGBoost tailored for biotechnology, can optimize prediction accuracy.
* **Feature Importance and Sentiment Analysis:** Financial researchers and investors should adopt a balanced stance. Sentiment analysis, while invaluable, should be integrated with traditional stock indicators for a comprehensive prediction model. The study emphasizes the importance of not overly relying on sentiment analysis. A hybrid approach, which melds sentiment scores with traditional stock metrics, is the way forward.
* **Incorporating Pandemic Data:** For those crafting financial models, there's a pronounced need to emphasize granular pandemic data. The study's findings accentuate the importance of including data like new COVID-19 casesover broader indicators. Such granularity offers a dynamic understanding of external influences, especially pivotal during events like the pandemic.
* **Future Research Directions:** The realm of sentiment analysis in stock predictions is vast, holding myriad avenues for exploration. Researchers are poised to delve deeper, either by building on this study's foundation or by charting fresh territories. The pronounced accuracy in biotech firms, for instance, begs for further exploration. Additionally, the consistent success of the XGBoost model could be a launchpad for research aiming to refine its application or merge it with other innovative methodologies.

**Concluding Reflections:**

This study offers a comprehensive insight into the multifaceted relationship between news sentiment and stock returns during a globally disruptive period. It stands as a guiding light for a broad spectrum of stakeholders, providing a roadmap to navigate the intricate terrains of modern financial landscapes. As the world of finance continues to evolve, research endeavors like this will be instrumental in shaping future strategies and investment decisions.

**Ethical Considerations:**

In the realm of computational research, upholding ethical standards is paramount. As this study delves into the intricate interplay of news sentiment and stock returns, it's imperative to ensure that the interpretation and insights derived do not inadvertently propagate misleading or detrimental financial advice. Additionally, leveraging sentiment data from news sources necessitates awareness of potential biases, misrepresentations, or oversimplifications. Throughout this research, a conscientious approach was adopted, anchored firmly in the AREA (Anticipate, Reflect, Engage, Act) framework, ensuring a comprehensive and ethically sound exploration of the topic at hand (EPSRC, n.d.).

**References**

Cakra, Y. E., & Trisedya, B. D. (2015, October). Stock price prediction using linear regression based on sentiment analysis. In *2015 international conference on advanced computer science and information systems (ICACSIS)* (pp. 147-154). IEEE.

Engineering and Physical Sciences Research Council (EPSRC). (n.d.). Research integrity in healthcare technologies. Health Technologies Impact and Translation Toolkit. Retrieved August 27, 2023, from <https://www.ukri.org/councils/epsrc/guidance-for-applicants/what-to-include-in-your-proposal/health-technologies-impact-and-translation-toolkit/research-integrity-in-healthcare-technologies/responsible-research-and-innovation/>

Lee, K. P., & Song, S. (2022). Informational Content of CEO Tweets and Stock Market Predictability. *Available at SSRN 4228651*.