Context

With advancements in artificial intelligence technology, companies across diverse industries are pioneering innovative approaches to their analyses. Particularly within the financial sector, firms are integrating fundamental financial indicators with alternative data sources. When alluding to alternative data, we refer to non-traditional datasets, including web-scraped news articles/microblogs, satellite imagery, credit card transactions, social media data, and more. In our investigation, our focus lies on User Generated Content (UGC) generated by authors within Virtual Investing Communities. By merging this alternative data with conventional financial metrics, we aim to glean insights into the potential impact of market sentiment on company financial indicators.

**To achieve this, we outline the following steps**:

1. Collect and aggregate alternative data from social media, financial media, and other relevant sources.
2. Employ text scraping techniques and conduct sentiment analysis using the OpenAI package. Subsequently, construct a data frame that can be merged with our traditional financial metrics data frame.
3. Merge these datasets and conduct analysis to address our research questions.

Research Questions

1. **Can sentiment analysis of User Generated Content (UGC) provide enhanced predictive capabilities for identifying the directional movement of stock returns?**

a. Can this predictive power be observed at varying temporal frequencies, such as daily, weekly, monthly, and/or quarterly intervals?

b. How do these predictive outcomes differ when stratifying data by individual firm, industry, market, and/or clusters based on market capitalization?

c. Can actionable trading strategies be derived from the predictive insights obtained from sentiment analysis of UGC?

d. What is the duration of any observed momentum resulting from the sentiment analysis, if present?

e. How does the sentiment polarity (positive, negative, neutral) of UGC mentions interact with the volume of mentions to influence the returns of our group of interest (individual firms, industry level, etc.)?

1. **What are the prevalent characteristics of companies that attract a higher volume of attention from Virtual Investing Communities?**

a. Does the market capitalization (size of the company) influence the level of attention received from the specified Virtual Investing Community?

b. Which industries tend to garner more attention from Virtual Investing Communities compared to others?

1. **How does the volume of mentions in User Generated Content (UGC) for each company influence the trading volume of the respective company's stock shares?**

a. Can the predictive power of UGC mention volume regarding stock trading volume be observed across various temporal resolutions, including daily, weekly, monthly, and/or quarterly intervals?

b. To what extent do these predictive outcomes differ when segmenting the data by individual firm, industry, market, and/or clusters based on market capitalization?

c. What is the duration of any observed momentum resulting from the sentiment analysis of UGC mention volume regarding stock trading volume, if indeed present?

d. How does the sentiment polarity (positive, negative, neutral) of UGC mentions interact with the volume of mentions to influence the trading volume of stock shares? Can we derive a potential trading strategy?

References

**Algorithmic Trading and the Market for Liquidity** <https://faculty.haas.berkeley.edu/hender/ATMonitor.pdf>

**Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement.** <file:///C:/Users/joel3/Downloads/stockmicroblogging_ICIS_2011.pdf>

**News versus Sentiment: Predicting Stock Returns from News Stories** <https://www.federalreserve.gov/econresdata/feds/2016/files/2016048pap.pdf>

**Stock Prediction Using Twitter Sentiment Analysis** <https://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf>

**The Relationship Between Twitter Sentiment and Stock Performance: A Decision Tree Approach** <https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/2add0fa5-1ea6-4b52-af8f-77bad6ebb791/content>

**A sentiment analysis approach to the prediction of market volatility** <https://www.frontiersin.org/articles/10.3389/frai.2022.836809/full>

Potential Datasets

**YouTube API:** [**https://developers.google.com/youtube/v3/getting-started**](https://developers.google.com/youtube/v3/getting-started)

**News API:** [**https://newsapi.org/**](https://newsapi.org/)

**Bing Search API:** [**https://www.microsoft.com/en-us/bing/apis/bing-news-search-api**](https://www.microsoft.com/en-us/bing/apis/bing-news-search-api)

**Dow Jones Developer Platform API’s:** [**https://developer.dowjones.com/site/docs/factiva\_feeds/analytics/xml\_reference/index.gsp#**](https://developer.dowjones.com/site/docs/factiva_feeds/analytics/xml_reference/index.gsp)

**StockTwits API:** [**https://rapidapi.com/stocktwits/api/stocktwits/**](https://rapidapi.com/stocktwits/api/stocktwits/)

**Stock Market and Financial News API:** [**https://eodhd.com/financial-apis/stock-market-financial-news-api/**](https://eodhd.com/financial-apis/stock-market-financial-news-api/)

**News Data:** [**https://ir.nist.gov/reuters/**](https://ir.nist.gov/reuters/)

username: trc2

password: KT9LWkUo