Spring
2023

Application of NLP and Neural Network Model in Stock Trading

MATH 5430

JASON JIN, YANG OUYANG, ZHIYUAN FAN, TIANHAO GUAN

Abstract

Exploiting News Sentiment for Short-Term Trading: A Portfolio-Based Approach to Stock Selection

Section 1: Overview of a Sentiment-Driven Trading Approach

Financial markets are influenced by a myriad of factors, including economic indicators, investor sentiment, and news events. Understanding the interplay between these factors is crucial for investors, financial analysts, and policymakers to make informed decisions. In this study, we integrate data from multiple sources to provide a comprehensive analysis of financial market dynamics. The datasets used in this study include historical stock price data, daily returns, market capitalization, and news events. We also use Loughran-McDonald Master Dictionary to conduct sentiment analysis on the news data, further enhancing our understanding of the impact of news events on the market.

Section 2: Gathering Stock Data and Sentiment Scores

We utilized four distinct datasets, in which each dataset contributes unique and interconnected information in our analysis: S&P500.csv, rets.daily.csv, data.csv, and Loughran-McDonald_MasterDictionary_1993-2021.csv (DICT). We will be referring to this master dictionary as DICT in the remaining sections.

S&P500.csv: Contains adjusted closing prices of the S&P 500 index, which accounts for events such as dividends, stock splits, and new stock offerings. This dataset includes a total of 23,947 entries.

rets_daily.csv: Includes daily stock data such as closing prices, returns, ex-dividend returns, market capitalization, and S&P 500 index membership indicators. This dataset has a considerably larger size with 5,386,757 entries and a memory usage of approximately 328.8 MB. For the submission, we included a sample csv file with the latest 40000 observations.

data.csv: Contains stock data, trade dates, news headlines, subjects, and augmented news content. This dataset consists of 444,623 entries and has a memory usage of approximately 50.9 MB.

DICT: A comprehensive financial dictionary used for sentiment analysis. This file contains 86,531 entries and 16 columns.

Section 3: Understanding the Attributes of the Collected Data

In the S&P500_prices.csv file, we have the "Date" and "Adj Close" columns. The adjusted closing price accounts for events like dividends, stock splits, and new stock offerings, providing a more accurate reflection of the index's value.

The rets_daily.csv dataset includes columns like "permno," "prc," "ret," "retx," "mcap," "in_sp," and "in_sp_tm5." These columns provide information about unique stock identifiers, daily closing prices, daily returns, ex-dividend daily returns, market capitalization, and time-varying membership in the S&P 500 index.

The data.csv file contains columns such as "permno," "trade_date," "headline," "subject," and "augmented body." The "headline" column contains the news headline, while the "subject" column refers to the news category or topic. The "augmented body" column provides an

enhanced version of the news content, which may include additional context, explanation, or analysis.

We also utilized the DICT file, which contains columns like "Word," "Negative," "Positive," "Uncertainty," "Litigious," "Strong_Modal," "Weak_Modal," and "Constraining." These columns indicate the sentiment and contextual information associated with each word in the dictionary.

We combined these datasets to create a comprehensive dataset containing all relevant information. We then conducted descriptive and inferential statistical analyses to identify patterns and relationships between stock prices, returns, market capitalization, and news events. Sentiment analysis was carried out using the DICT to understand the impact of news events on the market.

Section 4 : Preparing Data for Analysis and Strategy Implementation

The preprocessing stage is essential for transforming the raw data from various sources into a structured and analyzable format, which facilitates a comprehensive analysis of financial market dynamics. In this study, the preprocessing stage consists of data extraction, transformation, integration, and sentiment analysis.

1. Data Extraction:
Relevant data from the data.csv file, which contains stock data, trade dates, news headlines, subjects, and augmented news content, is extracted using the pandas library. The extraction ensures that only the data within the defined date range (January 1, 2010, to December 31, 2020) is included in the analysis.

2. Data Transformation:
The transformation process involves creating a customized text analyzer that incorporates a stemmer for text preprocessing. The SnowballStemmer is used to reduce words to their root forms, improving the efficiency of sentiment analysis by minimizing redundancy and variations in word forms. The analyzer also removes common English stop words that do not contribute significantly to sentiment analysis.

3. Data Integration:
The integrated dataset, created by merging S&P 500 and daily return data, is further enriched by appending sentiment scores derived from the news data. The DICT is utilized for sentiment analysis, to be specific, it was a csv containing a huge amount of words and with professions already counted the frequency of each word in different contexts and their sentiment. Based on that, we created a new sentiment dictionary by assigning positive and negative sentiment scores to stemmed words. The sentiment dictionary is then used to evaluate the sentiment of the augmented news content. The reason why using this new sentiment analysis is because finbert has relatively slow processing speed when dealing with huge text compared with this method. Since we were faced with millions of observations, methods we choose to improve efficiency are very important. Besides, it also provides accurate prediction and analysis since a lot more research is based on that method. For the submission, a sample of the updated dataset is included.

4. Sentiment Analysis:
To minimize computational complexity, duplicate records with the same 'permno' and 'trade_date' values are removed from the dataset, ensuring that each stock and trade date is associated with only one news item. This reduction is performed with the

assumption that the remaining news items sufficiently represent the overall sentiment for each stock and trade date.

Each augmented news item is scored for sentiment using the sentiment dictionary. The sentiment score is calculated as the difference between the number of positive and negative words in the news item, divided by the total number of words in the text. This normalized score effectively captures the overall sentiment of the news item and is assigned to the 'sentiment' column in the dataset.

By carrying out these preprocessing steps, we establish a foundation for our study, ensuring that the data is structured, cleaned, and formatted in a manner that facilitates efficient and accurate analysis of financial market dynamics. The resulting integrated dataset, containing stock prices, returns, market capitalization, news events, and sentiment scores, enables us to perform descriptive and inferential statistical analyses to identify patterns and relationships between these variables.

Section 5 : Sentiment-Based Trading Strategy

This section describes a sentiment-based trading strategy that involves holding the top 10 stocks with the highest positive sentiment scores on a daily basis. The strategy aims to capitalize on the impact of news sentiment on stock prices by investing in stocks with positive news sentiment and liquidating the positions the following day. The methodology and reasoning behind the strategy are detailed below.

1. Creating a Target Portfolio:
The trading strategy focuses on constructing a portfolio of the top 10 stocks with the highest sentiment scores for each trading day. The sentiment scores, calculated during the preprocessing stage, reflect the overall sentiment of the news items associated with each stock. By selecting the top 10 stocks with the highest sentiment scores, the strategy aims to exploit the potential positive impact of favorable news sentiment on stock prices.

2. Trading Frequency:
The strategy involves entering and exiting positions on a daily basis. This short-term trading approach is based on the premise that news sentiment impacts stock prices primarily in the short term, as market participants react to new information. By holding the top 10 positive sentiment stocks for one day, the strategy attempts to capture the short-term price movements driven by news sentiment.

3. Calculating Next-Day Returns:
For each stock in the dataset, the next-day return (t+1) is calculated by shifting the daily return data by one day. This calculation ensures that the strategy's performance can be assessed based on the actual returns generated by the selected stocks on the following trading day.

4. Merging and Cleaning the Data:
The next-day returns are merged with the dataset containing stock data, trade dates, and sentiment scores. The resulting dataset is cleaned to remove any missing or null values, ensuring that the analysis is based on accurate and complete data.

5. Selecting the Top 10 Stocks:
For each trading day, the top 10 stocks with the highest sentiment scores are selected, which we can vary updating the parameter value, and their next-day returns are calculated. This selection process forms the basis of the trading strategy, as it aims to capture the potential positive impact of news

sentiment on stock prices by investing in stocks with the highest positive sentiment scores.

6. Assessing Portfolio Performance:
The average next-day return of the top 10 stocks is calculated for each trading day, providing an estimate of the daily return generated by the sentiment-based trading strategy. The daily returns are then used to assess the overall performance of the portfolio, including cumulative returns and potential growth over the study period.

By implementing this sentiment-based trading strategy, we aim to evaluate the effectiveness of using news sentiment as a predictor of stock price movements. The performance of the portfolio, as assessed by its cumulative returns and growth over time, offers insights into the potential profitability of a trading strategy that relies on news sentiment as a primary investment criterion.

Section 6 : Evaluation

The analysis conducted in this study demonstrates the potential of utilizing news sentiment as a driving factor in a short-term trading strategy. By constructing a portfolio of the top 10 stocks with the highest positive sentiment scores on a daily basis, the strategy aims to capitalize on the impact of news sentiment on stock prices.

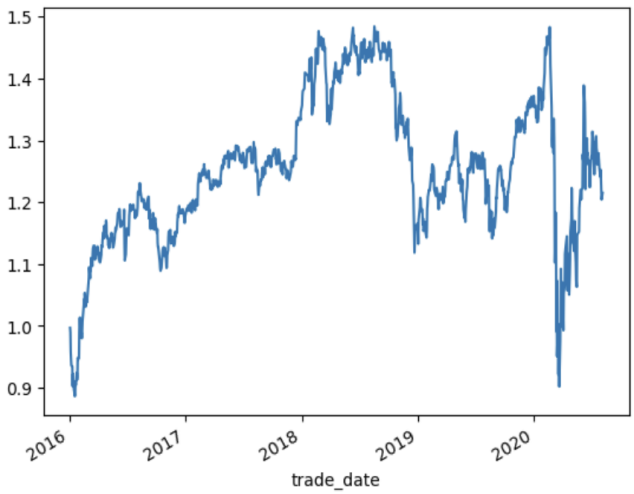
The results indicate that the sentiment-based trading strategy yields promising returns, with approximately 50% profit over a four-year period (**Graph i**). This performance highlights the effectiveness of using news sentiment as an investment criterion, particularly in short-term trading scenarios where investors seek to exploit short-term price movements driven by news events.

However, it is essential to consider certain limitations and caveats while interpreting these results. The study's findings are based on historical data and may not necessarily be indicative of future performance. Additionally, the study does not account for transaction costs, taxes, or other practical constraints that may influence the actual returns generated by the trading strategy.

Despite these limitations, the study provides valuable insights into the potential benefits of incorporating news sentiment into investment decisions. The results suggest that news sentiment can serve as a useful indicator for identifying stocks with potential short-term price appreciation. As such, investors and traders seeking to enhance their decision-making processes may consider incorporating sentiment analysis into their investment strategies.

The sentiment-based trading strategy presented in this study demonstrates the potential of news sentiment as a predictor of stock price movements. Although further research is needed to validate these findings and address the limitations mentioned above, the study's outcomes offer valuable insights into the practical applications of sentiment analysis in the financial markets.

Graph i. Return on Portfolio over time



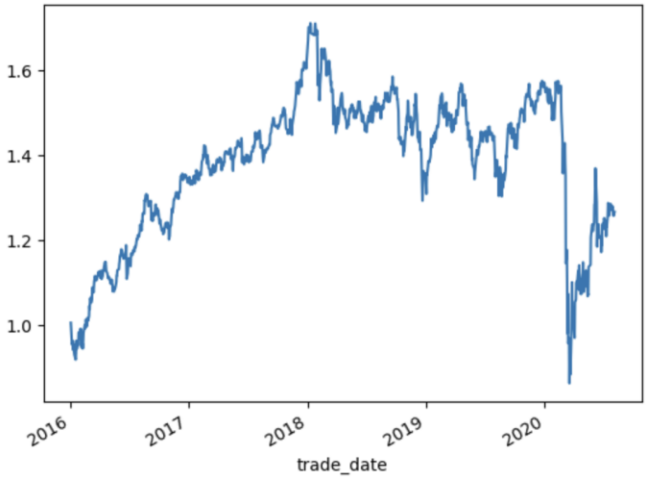
Section 7: Using LSTM

We examined the application of Long Short-Term Memory (LSTM) networks for sentiment analysis in our short-term trading strategy. LSTMs, a specialized type of recurrent neural network, excel at handling sequential data, such as text sentiment analysis. We used an LSTM model to predict sentiment scores for financial news articles and built a daily portfolio of the top 10 stocks with the highest positive sentiment scores.

The LSTM model featured a vocabulary size of 10,000 words, a sequence length of 128, and an embedding dimension of 64. We trained the model on 80% of the data and validated it on the remaining 20%. To minimize training time and computational demands, the LSTM model was trained for only 3 epochs with a batch size of 64.

The LSTM-based trading strategy showed promising results, outperforming the brute strategy discussed in Section 6 in terms of short-term returns—experiencing more than 60% growth in two years and approximately 50% in four years (**Graph ii**). These findings indicate that LSTM networks can effectively capture the temporal dependencies in financial news articles and predict short-term stock price movements based on news sentiment.

Graph ii. Return on Portfolio over time



It's important to note that these results come from a simplified LSTM model. Further improvements in the model's performance may be achieved by tuning hyperparameters, employing more complex architectures, increasing the bag of words or dimensions, or even using transformers such as BERT or FinBERT. However, due to our large size datasets and the limitation of time, we were unable to implement hyperparameter tuning as one single run takes hours.

Section 8: Conclusion

This group project demonstrated the potential of sentiment-based trading strategies using both the brute mathematical approach and LSTM networks. Both methods yielded promising results, showcasing the value of incorporating sentiment analysis into investment decision-making processes.