Multi-Modal Stock Information Retrival and Prediction

Anonymous Author(s)*

ACM Reference Format:
Anonymous Author(s). 2024. Multi-Modal Stock Information Retrival and Prediction. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROBLEM STATEMENT

Financial markets comprise individual investors, institutional investors, and algorithms. Along with this diversity in participants, there is also a wide range in their objectives—from long-term wealth gain to intra-day trading, as well as hedging and dividend-based profits and risk leveraging. Together, these elements combine into a vast pool of information that includes news, price data, predictions, earnings calls, expert opinions, and public sentiment. No single type of investor can analyze all this information simultaneously, and ultimately, they categorize it into **Qualitative Data** and **Quantitative data**, with a challenging gap in understanding between the two. While **Qualitative data** can be interpreted through financial sentiment analysis, **Quantitative data** is analyzed using price-action based mathematical models. Our aim is to integrate both approaches to effectively retrieve information in all modalities for a user.

Financial sentiment analysis specifically poses significant challenges due to the specialized terminology and the scarcity of labeled data specific to this sector. Standard models fall short due to the niche language employed in financial contexts. We propose that pre-trained language models are a viable solution, as they require fewer labeled instances and can be adapted further with domain-specific training. In this study, we introduce ProMod (proposed model), a language model derived from BERT, designed specifically for NLP tasks within the financial sector. Our findings demonstrate enhancements across all key metrics when compared to the existing top-performing results for two financial sentiment analysis datasets. Notably, ProMod achieves superior performance even with a smaller training dataset and by fine-tuning only portions of the model, surpassing other advanced machine learning approaches.

2 MOTIVATION

Market prices in an open economy encapsulate all known information about the assets being traded. When fresh information emerges, market participants modify their positions, causing prices to realign, thus making it exceedingly difficult to consistently outperform the market. Nevertheless, the definition of "new information" is subject to change as advancements in information retrieval technologies develop. Early adopters of these technologies may experience a temporary competitive advantage.

Analyzing financial texts, such as news articles, analyst reports, and official corporate communications, is a crucial source of new information. The daily production of vast amounts of such content makes manual analysis and extraction of actionable insights unfeasible for any one organization. As a result, the automated analysis of sentiment or polarity in financial texts using natural language processing (NLP) methods has risen in popularity over the recent years.

In financial markets, the analysis of price information—such as open, high, low, close, and volume data—is crucial for understanding market dynamics and forecasting future trends. This data encapsulates the market's behavior within specific time frames, providing insights into investor sentiment, market liquidity, and potential price movements. Accurate models of price information help investors make informed decisions, allowing for strategies that capitalize on market inefficiencies or anticipated price changes.

3 LITERATURE REVIEW

Several studies have explored sentiment analysis in financial news. These include research papers such as "Sentiment Analysis of Financial News Articles" by Li et al. (2014) and "Predicting Stock Prices with Financial News Articles" by Zhang et al. (2018). Additionally, there are numerous commercial sentiment analysis tools available in the market. In the realm of integrating market price data with sentiment analysis for stock market prediction, a notable recent study is "Combined deep learning classifiers for stock market prediction: integrating stock price and news sentiments" (2023) by B L, S., and B R, S., published in Kybernetes. This research introduces a unique sentiment analysis-based prediction framework that considers both stock data and news sentiment data. Key elements include extracting features from stock data using technical indicators (MACD, RSI, MA) and processing news data for sentiment analysis. The study employs advanced techniques like deep neural networks and self-improved whale optimization algorithms for sentiment classification and optimized deep belief networks for stock prediction. The model's performance, superior to traditional classifiers in certain metrics, highlights the potential of integrating sentiment analysis with market data for stock market predictions

4 NOVELTY

We try to use the potential of NLP transfer learning techniques, which offer a robust solution to the aforementioned limitations. The principal idea of these models is that training them on extensive text collections and subsequently using these pre-learned weights to initialize downstream tasks can lead to superior outcomes. The range of initialization can extend from a mere word embedding layer to the entirety of the model. In theory, this method addresses the problem of limited labeled datasets, as language models leverage the task of next-word prediction to capture semantic meanings without

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

needing labeled inputs. The fine-tuning process then focuses solely on applying this learned semantic knowledge to label prediction.

A key feature of transfer learning approaches is their ability to adapt language models further by pre-training them on specific, unlabeled text from the relevant domain. This adaptation enables the models to grasp the semantic nuances within texts that are characteristic of a particular field, which often vary widely from those found in general language corpora. This method is particularly advantageous for specialized sectors such as finance, where distinct terminologies and expressions prevail.

Traditional models of price analysis often utilize non-sequential approaches, treating each data point as independent of others. However, this method overlooks the inherent sequential nature of financial data, where each price point can be significantly influenced by preceding events. This sequential dependency is pivotal for accurately predicting future market behaviors, as patterns tend to develop over time, reflecting the cumulative effects of trading behaviors and external factors on price movements.

Given the limitations of non-sequential models in capturing these temporal dependencies, there is a strong motivation to employ models like Long Short-Term Memory (LSTM) networks. LSTMs are designed to process data sequences by maintaining a memory of past data points, which enables them to learn and recognize the temporal patterns that are critical in financial time series. This capability makes them highly effective for financial applications where past price actions are predictors of future activities, such as in trend following or mean reversion strategies.

The use of LSTMs in analyzing financial price data is motivated by their ability to integrate and learn from the complete historical sequence of price movements. This allows for a deeper understanding of market conditions, aiding in the prediction of future price behavior based on established patterns. By leveraging LSTMs, financial analysts can better interpret complex market dynamics and enhance the predictive accuracy of their models, leading to more strategic decision-making in trading and investment.

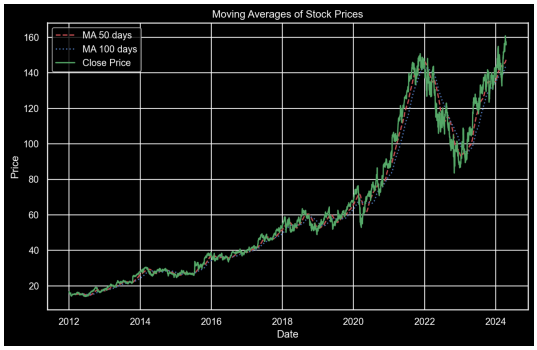


Figure 1: 50 vs 100 Moving Average

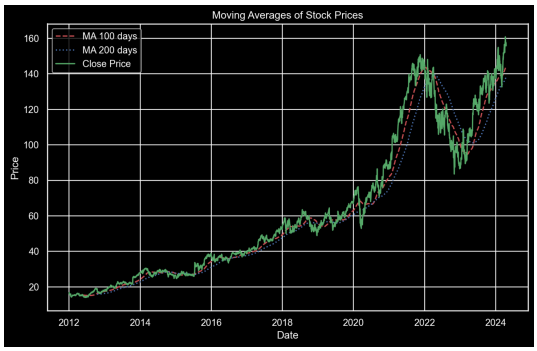


Figure 2: 100 vs 200 Moving Average

5 METHODOLOGY

Two different models were trained for tasks of predicting the market sentiment and for predicting the price trend for better understanding of the market.

5.1 Price Trend Prediction

Moving Average of the particular stock is considered for this task, which is a technical indicator that investors and traders use to determine the trend direction of securities. It is calculated by adding up all the data points during a specific period and dividing the sum by the number of time periods. Moving averages of 50, 100 and 200 days for a given stock is calculated 1 2.

Many models were trained, including Linear Regression, Decision Tree, Random Forest, MLP Regressor, and SVM, but the **Long Short-Term Memory (LSTM)** model was chosen over the others for stock market prediction due to its performance and for the following reasons:

1. **Capability to Capture Long-Term Dependencies:** This feature is critical for understanding stock market trends. Sequence training optimizes LSTM to retain historical information, allowing it to detect patterns that influence future stock prices.

2. **Complexity and Non-Linearity:** Although other models, such as Decision Trees and Random Forests, can handle nonlinear relationships, LSTM effectively captures the intricate dynamics and nonlinearities found in stock market data through sequence training.

3. **Hyper parameter Tuning:** All models underwent hyper parameter tuning, but the LSTM's parameters were specifically optimized for stock market prediction using sequence training. This procedure ensures that the model effectively uses its architectural features to capture the complexities of stock market data.

5.2 Price Sentiment Prediction

Our methodology for financial sentiment analysis begins with the BERT architecture, a transformative model based on a multi-layered transformer mechanism. This architecture excels at understanding context within text by using attention mechanisms that evaluate the relationship and importance of words in sentences.

BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. This approach allows the model to capture a more nuanced understanding of language context and flow than previous models. BERT is structured as a multi-layer bidirectional Transformer encoder. This architecture leverages an attention mechanism that learns contextual relations between words

(or sub-words) in a text. The innovation of BERT lies in its ability to fine-tune on a variety of downstream tasks with minimal task-specific adjustments, significantly improving performance on numerous NLP tasks. We enhance BERT's capabilities through a process known as sequence classification fine-tuning. Here, BERT is adapted to classify sentiments by training on labeled financial datasets, allowing it to recognize and predict sentiments accurately.

The training process involves an initial unsupervised phase, where ProMod, our financial-specific BERT model, is trained on a large corpus of financial texts without requiring labeled data. This step helps the model grasp the nuanced language of finance.

Subsequently, supervised training refines ProMod's abilities, focusing on specific sentiment analysis tasks within the financial domain. During this phase, ProMod is fine-tuned with labeled financial data, significantly improving its performance in sentiment classification.

Finally, to adapt to the evolving financial language, we continuously expand ProMod's vocabulary. New tokens that emerge in financial communications are added, ensuring the model remains current with industry jargon and effective in processing and understanding new and specific financial terms. This comprehensive approach enables ProMod to deliver superior performance in financial sentiment analysis tasks.

6 DATABASE AND CODE

Data was collected from Yahoo Finance of about 10-12 years using the library 'yfinance'. Initially, basic preprocessing steps were performed on the data like handling missing values, scaling features etc. Then a sequential LSTM model is built with multiple layers and dropout regularization to prevent over fitting. Lastly, the training is conducted over 50 epochs with a batch size of 32, showing decreasing loss over epochs, which indicates learning progression of the model.

The ProMod model undergoes pre-training using diverse financial datasets, followed by refinement and evaluation across three specific financial sentiment classification tasks:

Financial Phrase Bank: Comprises 4,840 sentences selected from financial news and meticulously labeled by 16 experts in financial markets. AnalystTone dataset: Contains 10,000 randomly selected phrases from analyst reports in the Investext database, with annotations for positive, negative, and neutral sentiments. Earning Calls: Utilized for further training and validation. SEC Forms: Also included to enrich the model's learning and adaptability to real-world financial documents. These datasets are instrumental in enhancing ProMod's capability to accurately analyze and classify financial sentiments.

A simple webpage application using streamlit is then constructed in which we can give a stock symbol as an input and it yields the stock data about that particular company from some date to present day offering real-time data, and also the report for sentiment analysis and the other associated charts.

7 EVALUATION

The plot shows the predicted price and actual price. We can infer from this plot that the model predicts well 3. [!h] For the sentiment classification model we have the classification metrics, the model

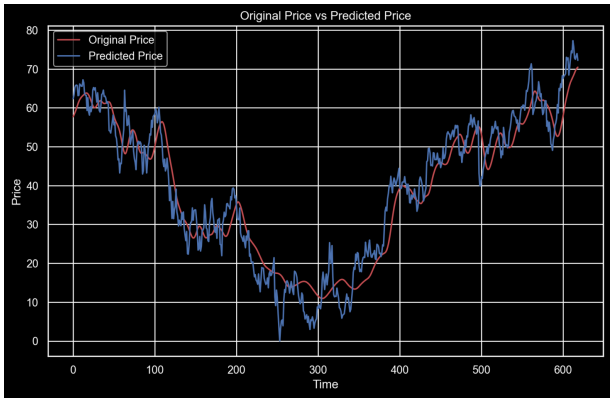
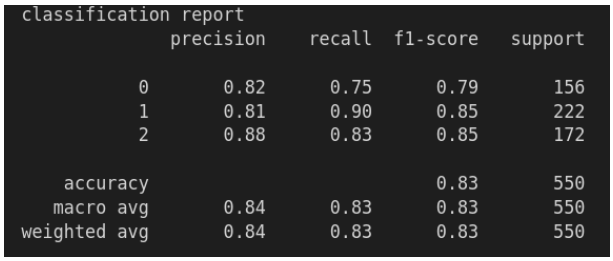


Figure 3: Original vs Predicted Price

Dataset	BERT cased	ProMod cased	ProMod+vocab
PhraseBank	0.64	0.74	0.83
AnalystTone	0.77	0.74	0.81

Table 1: Performance comparison of BERT and ProMod on various datasets



classification report

	precision	recall	f1-score	support
0	0.82	0.75	0.79	156
1	0.81	0.90	0.85	222
2	0.88	0.83	0.85	172
accuracy			0.83	550
macro avg	0.84	0.83	0.83	550
weighted avg	0.84	0.83	0.83	550

Figure 4: Classification results on sentiment testing

outputs a total accuracy of 0.83 with a weighted F1 of 0.83. This is better than all current model performing models that we have found.

Lastly a better decision can be made by the users by combining the two outputs which predicted the trend of the price and the current sentiment of the market.