Stock Movement Prediction based on Topics of News

Ruxi Zheng
Center for Data Science
New York University
rz2778@nyu.edu

Kexin Zhang
Center for Data Science
New York University
kz1267@nyu.edu

Zhuojian Wei
Center for Data Science
New York University
zw2219@nyu.edu

Qihang Tang
Center for Data Science
New York University
qt2087@nyu.edu

Abstract

The ability to detect the stock trends is very important for investment decisions. However, stock prices can be affected by various information, which makes it hard to predict. News is one of the resources that can provide information for the market. To examine if the information embedded in News can help improve stock price forecasting, we conducted topic modeling and predictive modeling. As a result, we found that stock price predictions for 9 out of 11 industry sectors elicit improvement by adding News topics as information.

1 Introduction

Financial markets are complex and hard to predict since they are highly influenced by large amounts of information. News is a kind of resource that can provide much information. Incorporating News data to identify stock trends could help traders know markets better and make informed investment decisions. In our project, we tend to find out if news topics could help with stock prediction and if so, we would delve into the relationship between news and stock prediction.

To approach this, we extracted New York Times articles and stock market data for 11 industry sectors and performed topic modeling on the articles using Latent Dirichlet Allocation (LDA) and BERT-based Topic Modeling (BERTopic), and predictive modeling on the stock prices using Long Short-Term Memory (LSTM) neural network, experimenting different hyperparameters, embedding models, model structures, etc. More specifically, topic models outputted the topic distributions to add into the LSTM models and we examined if the LSTM performance can be improved by adding topics as predictors based on their Root Mean Square Errors (RMSE).

Based on the results, there are decreases in RMSE for predictions of 9 industry sectors, especially for Industrials, Health Care, and Communication Services sectors. Therefore, we conclude that topics have effectiveness in improving stock price forecasting by providing more information to the predictive models.

2 Related Work

There have been some works done to explore the relationship between financial news and stock price. In Stock Price Prediction Using News Sentiment Analysis[1], it finds a strong relationship between financial new articles and stock prices and evaluates ARIMA, RNN, and Facebook Prophet models in the task. The results shows that RNN works best in the context of stock price.

3 Approach

Our project consists of two steps to reach our goal: topic modeling and predictive modeling. We test if the LSTM model combined with topic scores would perform better in 11 industries' stock prediction than the LSTM baseline which only historical prices are used as predictors. For topic modeling, the primary input is a corpus of word embedding of News articles, which is from the preprocessed News data by applying text cleaning and text representation techniques and undergoes topic modeling to discern underlying themes or topics present within the News. The method employed for this purpose is LDA and BERTopic. The output from this phase is topic distributions for each article and keywords for each topic. Later the topic distributions are aggregated to daily level. In the subsequent phase, predictive modeling, we choose LSTM to predict stock price. For each of sector, we have an LSTM model combined with these daily topic distributions, aligned with the same transaction date, and an LSTM baseline model to evaluate whether the integration of topic scores would help to predict stock prices. The features used for the training of LSTM baseline are $close\_price_{t-i}$ for $i \in [1, 7]$, $i \in \mathbb{N}$, which is the close price for each sector at transaction date t - i. The output is $close\_price_t$.

4 Experiments

4.1 Data

The first type of data we use is **New York Times articles (NYT)** from 2017 to 2023 (May), comprising 350,598 articles in total with their post time and content. We mainly use title, headline, and leading paragraph as the text content we feed into the later topic models and remove the articles that have less than 3 sentences. In terms of text preprocessing, a rigorous process is conducted such as removing URL and noise patterns, stop words, and lemmatization. We further split the data into 2017-2018 articles for topic model hyperparameter fine-tuning, 2019-2022 articles for final topic model training, and 2023 articles as the temporal / holdout set.

The second type of data we used is the daily market performance of the US stock market. As News articles are often not targeted to any specific firms, we analyze the performances of industries instead of individual firms in the US stock market. We use the industrial indices from **S&P Sectors** to represent our targeted industries [2], and we use **yfinance** library to extract the daily closing price of the industries indices from Yahoo Finance. The industries covered in our project are Energy (XLE), Materials (XLB), Industrials (XLI), Consumer Discretionary (XLY), Consumer Staples (XLP), Heath Care (XLV), Financials (XLF), Information Technology (XLK), Communication Services (XLC), Utilities (XLU), and Real Estate (XLRE).

4.2 Evaluation method

For **topic modeling** including LDA and BERTopic, we used the coherence score and human judgement as an evaluation metric. The method of evaluation discussed here can be characterized as measuring the level of relevance and connection among words within a topic, focusing on their interpretability. The objective of the topic coherence metrics used in this project is to evaluate the quality of the topics in a manner that reflects a human perspective [3].

For **LSTM**, since the target is continuous data, we used RMSE as the evaluation metric for both the LSTM baseline and the LSTM model combined with topic scores from BERT to check alignment of predicted and actual stock prices. We compared the RMSE of the baseline model and the model combined with topic scores to check if the topic score would improve the performance of LSTM stock prediction model.

4.3 Experimental details

For **LDA**, we fine-tuned the best number of topics and learning decay that result in the highest log likelihood using the 2017 and 2018 NYT articles by Grid Search. As a result, topic number of 5 and learning decay of 0.5 were selected to be the hyperparameters of the final LDA model and the model was trained on the 2019-2022 NYT articles.

For **BERTopic**, three parts can be modified or fine-tuned: embedding models, UMAP parameters for dimension reduction, and HDBSCAN parameters for hierarchical clustering. For embedding

models, we tried the proposed best pre-trained sentence transformer (all-MiniLM-L6-v2) listed on the GitHub page of sbert, and the highest-rank free model that provides local access (which is the ember-v1 model of the 7th rank) of the MTEB leaderboard on Huggingface [4]. We experimented with various parameter combinations for UMAP and HDBSCAN to optimize the BERTopic model. For UMAP, we evaluated three different sets of parameters: 'n_neighbors', 'n_components', and 'min_dist'. Similarly, for HDBSCAN, we explored two combinations involving 'min_cluster_size' and 'min_samples'. Our selection criterion for the optimal combination was based on achieving the highest coherence score. The most effective combination we identified was as follows: For UMAP, we used 'n_neighbors' = 10, 'n_components' = 2, and 'min_dist' = 0.01. For HDBSCAN, the chosen parameters were 'min_cluster_size' = 100 and 'min_samples' = 100. In terms of embedding models, the two choices yielded similar coherence score.

For **LSTM**, we tried epochs equals 30, 50, 100 with the number of neurons in 100, 200, 400 and the number of hidden layers in 2, 4. We found that for LSTM with epochs equals 50, the number of neurons equals 200 and the number of hidden layers equals 4 performed the best with l2 regularizer equals 1e-6. For learning rate, we used the default learning rate 0.001 for Adam optimizer.

4.4 Results

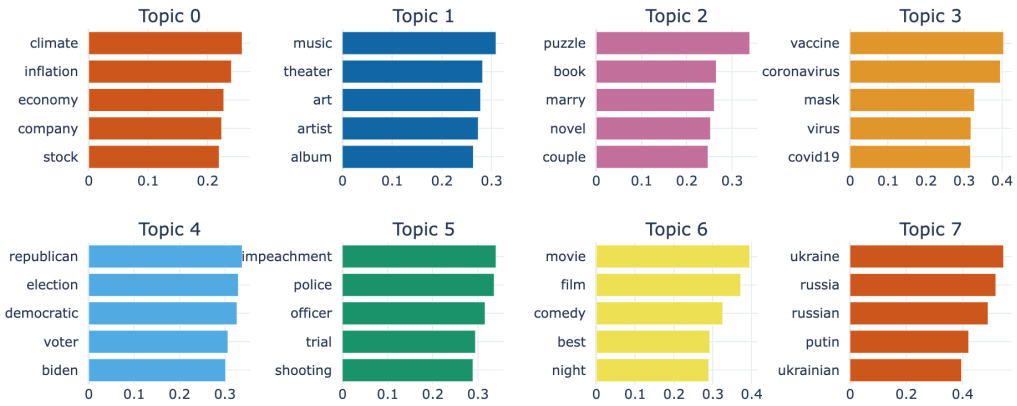


Figure 1: Topic Word Scores

Based on the evaluation methods we used, we found that BERTopic model (coherence score: 0.79) performs better than LDA (coherence score: 0.43). Figure 1aids in the interpretation of the **BERTopic** model's output. This figure effectively illustrates the 10 identified topics, highlighting the cohesiveness of the words within each topic. The output of BERTopic generates topics and associated probability scores for each article. We then aggregate this data to a daily level, creating a consolidated view where each date is characterized by a predominant topic and its probability. These aggregated topics' probabilities are subsequently integrated as new features into our LSTM prediction model.

Table 1 shows the result of **LSTM** comparing the baseline model and the model combined with topic scores. There's a decrease in RMSE for models combined with topic scores from the BERTopic model with all-MiniLM-L6-v2 embedding in Materials, Industrials, Consumer Discretionary, Consumer Staples, Health Care, Financials, Information Technology, Utilities, and Real Estate sectors. Also, Figure 2 in Appendix A shows the line graphs of LSTM predictions for the 11 industry sectors, where the grey lines, blue lines, and red lines represent the actual stock prices, prediction from LSTM baseline, and prediction from LSTM topic, respectively. To better understand our proposed method and results, the GitHub Repository of our source code can be found here.

5 Analysis

For the topics generated, the number of topics generated is larger than the number of industries, suggesting an industry may be related to multiple topics. Based on the top keywords of each topic, while none of the topics can be manually classified as a full representation of a specific industry, the diversity of topics suggests that the NYT articles have a wide coverage of topics. While the coherence

Sectors	LSTM Baseline	LSTM with Topics	Percentage of Changes
Energy	3.06	7.98	160.78%↑
Materials	2.16	2.09	3.24%↓
Industrials	3.66	1.97	46.17%↓
Consumer Discretionary	4.96	4.24	14.52%↓
Consumer Staples	1.38	0.98	28.99%↓
Health Care	3.58	2.11	41.06%↓
Financials	1.02	1.00	1.96%↓
Information Technology	4.98	3.61	27.51%↓
Communication Services	1.83	2.59	41.53%↑
Utilities	1.68	1.36	19.05%↓
Real Estate	1.00	0.98	20.00%↓

Table 1: LSTM in different sectors with RMSEs after 2023(2 decimal places)

score is not an absolute standard, a score of 0.79 confirms that the topics found by the model are highly relevant and consistent.

For the LSTM models, adding topic scores can decrease the prediction error in nine out of the eleven industries, with the largest decrease of error in the Industrials sector (46% decrease of RMSE). The results suggest that for the nine industries with improvement, our topics found contain highly correlated information that contributes to the forecasts of these industries. However, for the two industries with no improvement, the error can be more than doubled, suggesting that our topics found not only are unable to cover relevant information about the two industries but also serve as noisy input that blurred the model's forecast.

In general, adding topic scores has improved our baseline model, as the topics of NYT articles found have a wide coverage. However, it may be either that the News articles we used lack information on Energy and Communication Services or that the topics about energy and communication services extracted are not as coherent and informative as the other topics, which leads to the decrease in performance on the predictions of these two industries.

6 Conclusion

In summary, our project shows that topics of financial news articles can be high-level summaries of the information of articles that correlates with the movement of stock prices. By using distribution of topics for the articles (topic scores), we can decrease the prediction error in most of the industries covered in this project.

However, our findings include some limitations. As our temporal testing period only covered 5 months, whether the topic distribution can remain stable for a longer period needs to be tested. The BERTopic model needs to be further fine-tuned and the news article source may need to be expanded to yield topics that can cover all industries of interest. Also, our baseline LSTM model only used prices of the past 7 days: the prediction power of topics in a more complicated model that has more diverse input features needs to be further evaluated. The frequency of updating the BERTopic model also needs to be tested if topic modeling is to be used in a long-term operating algorithm.

References

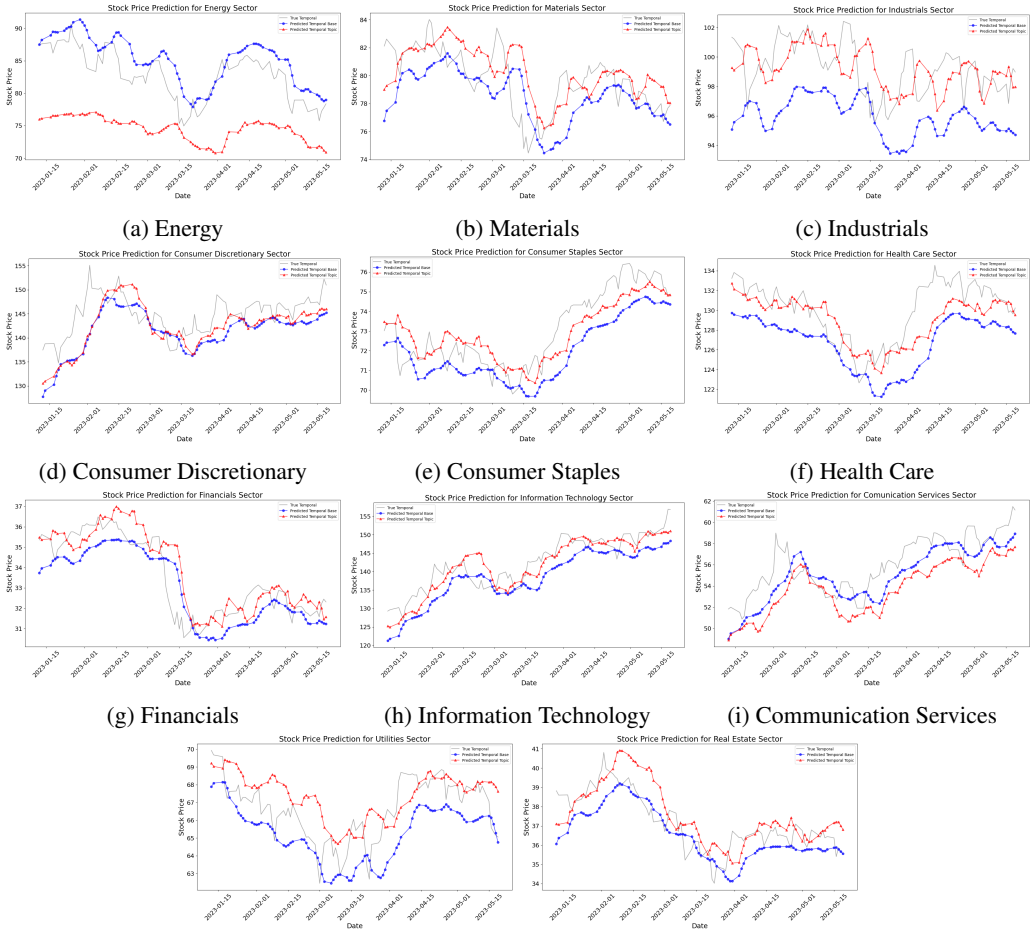
[1] Saloni Mohan, Sahitya Mullapudi, Sudheer Sammeta, Parag Vijayvergia, and David C. Anastasiu. Stock price prediction using news sentiment analysis. In *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 205–208, 2019.

[2] Spdr exchange traded funds (etfs) - all sectors. `https://www.sectorspdrs.com/allsectors`. Accessed: 2023-11-01.

[3] Konstantina Andronikou. Topic modeling with bertopic. *THEANALYTICSLABS*, 2022.

[4] Massive text embedding benchmark (mteb) leaderboard. `https://huggingface.co/spaces/mteb/leaderboard`. Accessed: 2023-11-14.

A Visualizations of LSTM Predictions



(a) Energy (b) Materials (c) Industrials

(d) Consumer Discretionary (e) Consumer Staples (f) Health Care

(g) Financials (h) Information Technology (i) Communication Services

(j) Utilities (k) Real Estate

Figure 2: Stock Price Prediction for the 11 Industry Sectors